Università degli Studi di Modena e Reggio Emilia

Dottorato di ricerca
in Information and Communication Technologies

Ciclo XXXIII

Technologies for the Monitoring of Rail Tracks and Railway Traffic Safety Systems

Candidato: Davide Bombarda

Relatore (Tutor): Prof. Giorgio Matteo Vitetta
Correlatore (Co-tutor): Ing. Giovanni Ferrante, Alstom

Coordinatore del Corso di Dottorato: Prof.ssa Sonia Bergamaschi

Anno Accademico 2019 – 2020

Ringraziamenti

Giunto alla conclusione del mio periodo di dottorato, sono molte le persone che mi sento di ringraziare per il sostegno e l'aiuto che mi hanno dato.

La mia famiglia e i miei amici, che mi hanno sostenuto anche nei momenti meno facili che sono capitati, non da ultimo una pandemia globale.

Il mio tutor accademico, Prof. Giorgio Matteo Vitetta, che mi ha saputo guidare nel percorso che mi ha portato sin qui, dandomi i migliori consigli e offrendomi la sua conoscenza. L'impegno profuso, a dispetto delle avversità, merita più di un ringraziamento.

Anche il personale di Alstom merita un ringraziamento, a partire dal co-tutor, Ing. Giovanni Ferrante e l'Ing. Pier Alessandro Aisa che mi hanno seguito durante questi tre anni, per poi proseguire con l'Ing. Pasquale Capasso e l'Ing. Giuseppe Casulli che, in questi anni, mi hanno affiancato per un periodo nel portare avanti alcune attività di studio e ricerca, alcune delle quali presentate in questo manoscritto. Ringrazio tutti quanti per il tempo che mi hanno saputo dedicare.

Un sincero ringraziamento va anche al personale, docenti, tecnici e ricercatori del centro ARCES dell'Università di Bologna, che mi hanno accolto presso i loro laboratori specializzati nel campo del monitoraggio dell'integrità strutturale. La loro conoscenza e la dotazione dei loro laboratori sono stati fondamentali per introdurmi al mondo delle onde ultrasoniche guidate e per svolgere le mie attività di ricerca, purtroppo interrottesi a causa della pandemia. Una menzione particolare va al Dott. Salvatore Lottante, con il quale ho condiviso il percorso di introduzione alle onde ultrasonche guidate.

Abstract

In recent years, railway engineers are required to support higher levels of traffic at increased speeds over the railway infrastructure; this is pushed by the demand for faster and more frequent passenger trains or for heavier freight trains. These results must be achieved together with an improvement of the safety and robustness of the railway network, at competitive costs.

In this thesis, various aspects related to the introduction of new technologies in the systems guaranteeing a safe operation of the railway network are investigated. One of the main topics addressed in this work is related to the introduction of ultrasonic guided waves for the monitoring of the integrity of rails. Another topic concerns development of an instrument improving the installation and maintenance of track circuits, a fundamental equipment for safe operation of railway traffic.

Rail tracks undergo massive stresses that can affect their structural integrity with the possible result of a rail breakage; this may cause derailments and, consequently, a possible loss of rolling stock material and lives. Therefore, the activities of track maintenance and inspection are of paramount importance. In recent years, the use of various technologies for the monitoring of rails and the detection of their defects has been investigated but substantial research efforts are still required to achieve more reliable results while increasing the scanning speed. It is expected that, in the near future, an important role in track maintenance and inspection will be played by the ultrasonic guided wave technology. In this manuscript, the two main technical approaches to the use of ultrasonic guided waves to monitor rail integrity available in literature are analyzed and compared. Then, the approach usually followed to analyze the propagation of ultrasonic guided waves into solid waveguides is described. This provides a baseline to the development of innovative methods and systems able to detect early-stage defects in rails, and may lead to the development of a commercial product.

The track circuit is used to detect the presence of a train in a given section of track. A carrier is injected at one end of the section and its presence at the other end is checked. A train is detected when its axis short-circuits the rails, thus preventing the injected current from reaching the receiver. Track circuits also allow to perform train-track communication, and to infer the presence of a rail breakage that interrupts the electrical continuity of the circuit. This system, performing safety critical operations, needs a careful and precise calibration during installation and maintenance, which should be accomplished under normal track operation. This calls for the use of a portable and low-cost albeit accurate device to quantify the real performance of the system. To carry out this task, Alstom has begun to develop a portable device, called track circuit contactless signal analyzer (TCCSA). Its signal acquisition hardware and software can also be employed to extract specific signal features; these can be fed into a machine learning software able to classify the detected signal and thus assessing the health status of the track circuit components or of the rails. In this thesis, it is shown how the development of this system has been carried out, with the implementation of new features and functionalities.

Keywords: SHM, UGW, rail, track, maintenance.

Abstract (italiano)

Tecnologie per il monitoraggio dei binari ferroviari e per i sistemi di sicurezza della circolazione ferroviaria

Negli ultimi anni, agli ingegneri ferroviari è stato richiesto di supportare l'aumento del traffico ferroviario e l'incremento delle velocità di percorrenza sull'infrastruttura ferroviaria, sulla spinta della domanda per treni passeggeri più frequenti e veloci, e di treni merci più pesanti. Questi risultati devono essere conseguiti insieme ad un miglioramento delle prestazioni in termini di sicurezza e robustezza della rete ferroviaria, a costi competitivi.

In questa tesi sono investigati vari aspetti relativi all'introduzione di nuove tecnologie nei sistemi che garantiscono la sicurezza della gestione del traffico ferroviario. Uno dei principali argomenti affrontati è l'introduzione delle onde ultrasoniche guidate nel controllo dell'integrità delle rotaie. Un secondo punto di interesse è relativo allo sviluppo di uno strumento che migliora le procedure di installazione e manutenzione dei circuiti di binario, strumenti essenziali per assicurare la sicurezza della circolazione ferroviaria.

Le rotaie sono sottoposte a forti stress, che possono influire negativamente sulla loro integrità strutturale, con il possibile risultato di una rottura della rotaia; ciò potrebbe causare il deragliamento di un treno e, quindi, la perdita di materiale rotabile e vite umane. Pertanto, le attività di manutenzione ed ispezione del binario sono di importanza capitale. Negli ultimi anni, è stato investigato l'uso di varie tecnologie per il monitoraggio dei binari e per l'individuazione dei difetti; tuttavia, sono ancora richiesti notevoli sforzi per migliorare l'affidabilità dei risultati ed incrementare la velocità con cui i controlli vengono eseguiti. Ci si aspetta che, in un prossimo futuro, un ruolo importante nella manutenzione e nell'ispezione dei binari sarà giocato dalla tecnologia delle onde ultrasoniche guidate. In questo scritto sono analizzati e confrontati i due principali approcci disponibili in letteratura sull'uso delle onde ultrasoniche guidate nel monitoraggio dell'integrità delle rotaie. Di seguito, è descritto l'approccio seguito per analizzare la propagazione delle onde ultrasoniche guidate all'interno di guide d'onda costituite da materiali solidi. Ciò fornisce una base per lo sviluppo di metodi e tecniche innovative per l'individuazione preliminare di difetti nelle rotaie, e può allo sviluppo di un prodotto commerciale.

Il circuito di binario è utilizzato per determinare la presenza di un treno in una data sezione di binario. Una portante viene immessa ad una estremità della sezione e si valuta la sua presenza all'altra estremità. Quando un treno corto-circuita le rotaie interrompe il circuito ed impedisce alla corrente immessa di raggiungere il ricevitore; viene così segnalata la presenza di un treno nella sezione. I circuiti di binario consentono anche la comunicazione terra-treno e l'individuazione di rotture della rotaia che interrompano la continuità elettrica del circuito. Questi sistemi, eseguendo operazioni critiche dal punto di vista della sicurezza, richiedono una calibrazione precisa durante l'installazione e la manutenzione, che dovrebbe essere compiuta mantenendo il binario in servizio. Per fare questo, l'azienda Alstom Ferroviaria S.p.A. ha iniziato a sviluppare un dispositivo portatile, chiamato *track circuit contactless signal analyzer* (TCCSA). Il suo hardware ed il suo software per l'acquisizione del segnale possono anche essere utilizzati per estrarre caratteristiche specifiche del segnale; da queste caratteristiche si possono estrarre, mediante tecniche di apprendimento automatico, informazioni sullo stato di salute del binario e delle componenti del circuito di binario. In questa tesi viene descritto come è stato portato avanti lo sviluppo del sistema, con l'implementazione di nuove caratteristiche e funzionalità.

Parole chiave: SHM, UGW, rotaie, binari, manutenzione.

Contents

Chapter 1

Summary

1.1 Context

In the last decades, government authorities (and, in particular, the European Union, EU) are encouraging a *modal shift* towards rail transport. The position of rail transport is competitive, in terms of performance and efficiency, respect to all the other means of transport (road, air, internal and maritime navigation), as established by the *European union agency for railways* (ERA). Rail is, in fact, one of the most safer and ecologic modes ways to travel and to move goods at medium-long distances [1].

The modal share level of railway transport in the EU from 2012 to 2018 has remained quite stable (around 6.7% for passenger and 11.3% for freight). It is important to say, though, that the level of railway traffic has increased: the increment at EU level kept in line with the general increase in traffic level among all the means of transport. The volume of rail passenger traffic from 2009 onward seen a stable increase of 1.6% per year, while in 2017 and 2018[1] the freight level traffic show an interesting increase, which may change the overall trend of the last decade [1]. In the case in which the modal shift policies should succeed[2], the volume of traffic on the railway network it is expect to grow further [4].

From the point of view of railway infrastructure managers, the push towards the modal shift is translated by the different stakeholders into different inputs. Government authorities, railway safety agencies, railway transport companies and the public have usually contrasting requests; the role of infrastructure managers is to summarize all the expressed *needs* and *constraints* in a compromise capable to support the expected increment of traffic. Government authorities encourage a competition system among the various transport companies, both in the market (the case of high speed rail or freight transport) or for the market (the case of local public transport). Government authorities also urge towards an increasingly cost-effective management of the railway infrastructure. The public (comprising both travelers and companies which make use of the rail transport to ship their goods) ask to governments and transport companies competitive costs of travel, with increasing levels of punctuality. In addition, they demand faster long distance passenger trains, frequent local public transport, heavier and longer freight trains. Moreover, public and government authorities expect and pretend ever growing levels of safety, and drive the railway safety agencies to emanate stricter regulations. These contrasting requirements and constraints yet today ask for an accurate planning of railway timetables and a careful infrastructure maintenance. With traffic increment, the performance levels in terms of safety and reliability of railway network must rise. There is a problem, though. To make room for the increased traffic, the planned maintenance intervals need to be shorter; moreover, it will became harder to reach the targets for safety and reliability because of the stricter regulations and the costs reduction demanded. This without taking into account that the more intense infrastructure exploitation causes wear and tear, increasing the need for maintenance operations.

To keep up with the expected increasing rate of traffic, the railway infrastructure managers are accelerating the existing process of migration from cyclic and corrective maintenance towards the concept of *predictive* maintenance. The *diagnostic* stays at the basis of such a change in maintenance paradigm. Yet today diagnostic services are useful to establish the the health status of infrastructural assets, and this is worthwhile to guarantee the operational safety, prevent faults and optimize maintenance intervention. Diagnostic can be applied to the different sub-systems composing the railway infrastructure: railway superstructure, energy, telecommunications, safety systems, and rolling stock material. Different types of diagnostic exist, divided into two categories: fixed diagnostic and mobile diagnostic. Fixed diagnostic systems are spread over the railway network and are used

[1] The last two years for which data are available at the moment of writing.

[2] And in the case that the effects of the actual Covid-19 pandemic does not affect permanently the expected rate of increase.

to determine the health status of safety systems (interlocking[3] and level crossings), of the track, or to ensure that vehicles passing in a given point are not experimenting a danger condition. Mobile diagnostic systems are useful to determine the health status of the railway infrastructure as a whole, by means of specialized diagnostic vehicles properly equipped [2].

One of the purposes of railway infrastructure manager supplier companies (like Alstom) is to develop and make available ever more accurate and well-performing diagnostic systems and services, fixed or mobile.

1.2 Contents of the Manuscript

The PhD activity summarized in this manuscript regarded the study of diagnostic systems employed to determine the health status of two of the fundamental elements in the railway infrastructure: the track and the safety systems. More specifically, the study has focused on the systems to inspect (or monitor) the integrity of rails by means of ultrasonic guided waves, and on the systems to acquire and analyze the operating parameters of track circuits for installation, maintenance and monitoring purposes.

1.2.1 Rail Integrity Inspection

The track is used to transmit train wheel forces to the track bed and to guide vehicles. Being laid in open-air, it is exposed to harsh environmental conditions like corrosion and thermal stresses, whose effects are complex and variable [14]. The weight of trains is discharged on a very small surface of the rails [5], also named *contact patch* [20]: therefore, the rail is subjected to intense bending and shear stresses, plastic deformation and wear, leading to degradation of its structural integrity with time [3]. Moreover, the rail can also be subjected to mechanical and thermal abuse during installation and track maintenance operations [14]; rails might also contain internal fabrication defects slipped from quality control. All these factors could lead to a rail breakage.

A broken rail has harmful consequences on the safety of railway operation: it is one of the major causes of derailment[4]. According to the European railway agency (ERA) safety overview for 2017 [4], in the period between 2011 and 2015 broken rails are the main form of precursors to accidents (inconvenients that, under other circumstances, could have led to an accident)[5]. In the specified period, in the whole European Union (EU-28), in mean 4,445 broken rails per year were found over a total of 11,222 precursors to accidents.

Since a derailment causes damages to the railway infrastructure and to the rolling stock, in some cases spread of dangerous substances, and in the most severe cases, injuries and death, it is necessary to limit its probability of occurrence. Preventing rail breakage is one of the ways to reduce the risk of derailment [4]. Even the rail breakage in itself is an event harmful for the regularity of railway operation: it is a fault that must be solved before the traffic on the interested track can be reopened, provoking delays.

To prevent rails from breaking (and other equivalently deleterious accidents) the activities of track inspection and maintenance are of paramount importance. By regularly (or continuously) inspecting the track health status it is possible to detect in advance (and kept monitored) critical situations, such as defects in rails [8]. This allows to plan maintenance activities when these have the lower impact possible on the commercial exploitation of the track[6].

To perform track inspection and monitoring, different techniques have been studied and adopted in the years [8], [23]. Since current solutions are still sub-optimal respect to railway infrastructure companies requests, the research is yet active nowadays. In the light of the expected increase in traffic discussed previously, it is even more necessary to enhance accuracy and inspection speed performances of the systems. In the years to follow it is expected that the development of rail integrity inspection techniques based on ultrasonic guided waves will bring significative results both in terms of accuracy of the detection and speed of inspection, with the possibility to introduce a continuous monitoring of the health status of the rail.

1.2.2 Track Circuit Parameters Monitoring

Track circuits are used to detect the presence of a train in a specific section (stretch) of track. Their operating principle is quite simple: a carrier is injected at one end of the considered section and its presence at the other end is checked. A train (or vehicle) is detected when its axis short-circuit the rails, thus impeding to the current to reach the receiver. The one described is only the basic operating principle of the track circuit:

[3] Systems ensuring a safe train routing.

[4] Possible causes of derailments are: broken rails, track in poor geometric conditions, obstacle on the track, over-speeding in corners, breakages in the wheelset, etc,

[5] Other precursors to accidents include track buckles, signals passed at danger, wrong-side signaling failures, broken wheels and axles.

[6] When in the planned maintenance intervals or by interrupting the traffic in days and hours when this is not too harmful for the railway operation.

by modulating the carrier, in fact, it is possible to operate a track-train communication[7] Moreover, the rails are part of the electric circuit which constitutes the track circuit. Track circuits therefore enable to detect a complete rail breakage if it interrupts the electrical continuity: in this case, the track circuit results occupied (as if a train were in section). So, when a track circuit becomes unduly occupied, before claiming it as a fault of the system, it is necessary to inspect the stretch of track to assess the cause of the occupation: it may be due to a vehicle, an obstacle o a broken rail. Only once the problem is fixed, the normal train operation can be resumed. Whenever the railway line is electrified, track circuits may be covered by the return of traction current from locomotives to electric substations[8] These currents are usually intense, and it is necessary to introduce additional electrical components to the circuitry of the track circuit to separate the two current paths, that of traction and that of track circuit. It is also necessary to say that the track circuits, being laid in open air, are exposed to environmental and atmospheric conditions, which may affect their operation [119].

For the reasons expressed so far, performing safety critical operations, track circuits require a precise and careful calibration, both during commissioning and during maintenance [119]. It should also be possible to adjust the track circuit parameters also when the track is under normal operation. That calls for the use of a low-cost albeit accurate portable device, which can be used to quantify the real performances of the system. This instrument should encase all the functionalities made available by much less compact, portable and economic devices, easing the task of field personnel. Alstom is developing a portable device, called *track circuit contactless signal analyzer* (TCCSA) which, through the connection to a laptop and collaborating with a purpose-written software, is able to measure and plot the behavior of different quantities (voltages and current) which describe the operation of the track circuit. The different features and parameters extracted from the measured signals can not only be analyzed by the operator, but also fed into a machine learning software capable to classify the received signals and to assess the health status of the track circuit components (rails included). The system is capable to work also when trains in commercial service are passing on the section of tracks under analysis; moreover, thanks to this, it is also possible to monitor the effects of passing trains on the track circuit parameters.

1.3 Organization of the Manuscript

This manuscript is organized in 4 parts. In the following it is given the description of the methodological approach followed in the writing of each part and a brief description of the content of each chapter.

1.3.1 Part I, Background

Part I aims at giving the basis for the discussion to follow about rails, their defects and the techniques used for their detection. It gives also a theoretical introduction about the mechanical properties of solids, the characteristics of elastic waves in solids and of ultrasonic guided waves. It is hypothesized that the reader hasn't a specific background into the mentioned specific matters (a part from a basic knowledge of physics): therefore, the treatment starts from the base properties.

Chapter 2, *Rail Defects*, introduces a basic knowledge about the track, the rails, and delves into the description of the defects that may induce to a series damage of the rail.

Chapter 3, *Existing Rail Defect Detection Techniques*, gives an overlook about the principal techniques, in use and in development, for rail defect detection.

Chapter 4, *Mechanical Properties of Solids and Elastic Waves*, introduces the mechanical properties of solids and the elastic waves: these are general physics knowledge useful to the following discussion.

Chapter 5, *Ultrasonic Guided Waves*, treats in reasonable detail ultrasonic guided waves; it is given as an example the analytical treatment of the behavior of ultrasonic guided waves in plates.

Chapter 6, *Semi Analytical Finite Element (SAFE) Method*, describes one of the algorithms that can be used to describe the behavior of ultrasonic guided waves in a waveguide, the *semi analytic finite element* (SAFE) method.

Chapter 7, *Characteristics of Guided Waves in Rails*, examines the peculiar characteristics of ultrasonic guided waves propagation in rails.

Chapter 8, *Piezoelectric Transducers and Piezoelectric Effect*, depicts the operation of piezoelectric transducers, introducing the piezoelectric effect and then describing their equivalent electrical models.

[7] Useful to inform the engineer about the aspect of the signals that he will encounter in advance, or about the maximum speed admitted.

[8] In electrified lines one of the traction circuit poles is constituted by an overhead wire or by a third rail; the other pole carrying the return current is constituted by the rails.

1.3.2 Part II, Ultrasonic Guided Waves Applications

In part II are described the applications of ultrasonic guided waves.

Chapter 10, *Existing Applications of UGW Rail Defect Detection Techniques*, aims at analyzing and comparing the most innovative techniques available in literature for the detection of defects in rails using ultrasonic guided waves. Two are the main approaches on which scholars have focused in literature, one regards a mobile inspection system and one a fixed monitoring system. Their characteristics, implementation, performances, advantages and disadvantages are described in detail.

Chapter 11, *Introduction to the Experimental Activity*, introduces to the performed experimental activity, declaring the objectives and sub-goals and describes a possible implementation of a diagnostic system, based on the outcomes of Chapter 10.

Chapter 12, *Qualitative Results About UGW Propagation in a Waveguide*, describes one of the possible approaches that is possible to follow when beginning to study the behavior of ultrasonic guided waves in a waveguide. Starting from specific objectives, are depicted the measurement set-ups and methods used to carry out the study; are then given and critically analyzed the obtained results.

Chapter 13, *Validation of Qualitative Results on UGW Propagation in a Waveguide*, describes the process of validation of the results obtained in the previous chapter. The various aspects involved in UGW propagation are targeted and individually analyzed.

Chapter 14, *Conclusions and Future Developments*, draws some conclusions over the performed activities and illustrates some possible future developments.

1.3.3 Part III, TCCSA - Track Circuit Contactless Signal Analyzer

Part III regards the *track circuit contactless signal analyzer* (TCCSA). Also in this case the approach followed is to give a description of the system to a reader which has no, or very little, specific knowledge in the railway field, but having a background in electronics and telecommunications. Being this topic somewhat (but not entirely) disconnected from the topic of rail damage detection, it has been decided to put all the treatment about track circuits into a specific part of the document.

Chapter 15, *Introduction to the TCCSA*, gives a very brief introduction about the TCCSA.

Chapter 16, *The Track Circuit*, introduces the basic notions about the operation, use and design of track circuits, describing the main typologies available and the relative electrical models.

Chapter 17, *Literature Review on Projects Regarding Track Circuits*, makes ax excursus in the literature about the projects involving track circuits. Two are the areas of interest: the methods for the detection of track circuit faults by analyzing their operative features, and the methods to detect the presence of defects in rails by using track circuits.

Chapter 18, *The Track Circuit Contactless Signal Analyzer - TCCSA*, describes the TCCSA, the implementation of different functionalities, and delivers a possible connection with the rail integrity diagnostics by means of ultrasonic guided waves.

1.3.4 Part IV, Indices and Bibliography

In part IV are given the indices of figures and tables, and it is included the bibliography.

Part I

Background

Chapter 2

Rail Defects

The railroad track system is very complex, and involves many interactions [6]: a damage can occur anytime and anywhere. The development of a rail defect is due to uncontrolled and random processes or events [8]. The prediction of the crack growth rates and defect size at failure can be strongly dependent on multiple parameters [87]. If not detected in time, a defect can lead to rail failure, in some cases without any previous indication.

2.1 Track and Rail Basics

2.1.1 Track

The track is composed by *rails* which are fixed to *sleepers* (or *ties*) by means of *fasteners*. Sleepers are used to connect trasversally the two rails forming the track, ensuring the gauge[1] and to discharge the applied loads to the underlying substructure. The track is supported by a layer of crushed stone, named *ballast*, which traps sleepers in order to hold in position the railway track, sustain elastically and uniformly track loads and to ensure hydraulic drainage [18]. In the following, we will focus on rails.

2.1.2 Rail

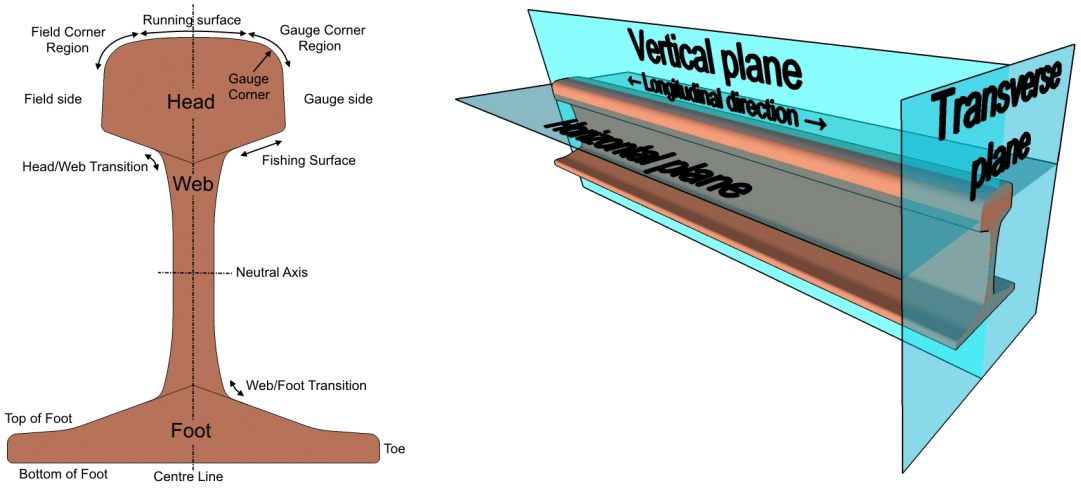
The rail is a steel profile of suitably shaped section that *support* and *guide* railway vehicles. By means of rails, a regular, uniform and low rolling resistance way is constructed. Rails are classified on the basis of their weight per linear unit of distance (kilograms per meter for the international standard, pounds per yard for the imperial standard) [18]. Different rail types (and hence weights) are employed; linear masses of around 60–70 kg/m are considered to be sufficient for most applications [14].

The rail section and nomenclature is illustrated in Fig. 2.1. The vertical direction goes from the head to the foot passing through the web; the longitudinal direction is normal respect to the transverse section.

The rail cross-section is formed by different areas, defined in ref. [20] as:

- *Running surface* is the area on the top of the railhead, which makes contact with the wheel tread;
- *Gauge side* is the face of the rail facing the other rail of the track;
- *Gauge corner region* is the top corner on the gauge side of the rail, which makes contact with the wheel throat region;
- *Gauge corner* is the single point in the gauge corner region which its tangent is at 45° to the horizontal (with or without cant);
- *Field side* is the face of the rail facing the area external to the track;
- *Field corner region* is the top corner of the field region of the rail;
- *Fishing surface* is the region at the bottom of the railhead which makes contact with fishplates, used to join the rail bars;
- *Rail centre line* is the vertical centre of the rail section;

[1] Track gauge: the distance between the inner edges of the heads of the rails composing the track, measure 14 mm below the rolling surface. The international standard track gauge is 1435 mm [18].



(a) (b)

Figure 2.1: Rail nomenclature. In (a) is represented the transverse section [20], in (b) are represented the reference planes.

- *Rail head* is the region of the rail cross-section that is above the extensions of the fishing surfaces towards the rail centre line;
- *Rail foot* is the region of the rail cross-section that is below the extensions of the top of foot surfaces to the rail centre line;
- *Rail web* is the region of the rail cross.section between rail head and foot;
- *Neutral axis* is the vertical distance of the rail at which the Second Moment of area of the section above is the same as the section below: this is the point in which, in bending, there is no tension or compression;
- *Top of rail foot* is the region on the top of the rail foot which makes contact with rail fasteners or insulating biscuits;
- *Bottom of rail foot* is the region on the bottom of the rail foot which makes contact with sleeper plates or rail pads;
- *Toe of foot* is the edge region of the rail foot.

Rails are often installed with their vertical axis having an inclination of $1/20$ or $1/40$ (depending on the country) towards the gauge side. This enables to have a perfect coupling with vehicles' wheels, which have the same inclination. This enables:

- to reduce the winding motion in tangent track, thanks to axis recall action;
- to have an optimal cornering, since wheel and axis coupling act as differential [18].

In curved tracks, to counteract centrifugal forces, and hence to raise travel speed in corners, the track is usually superelevated, i.e. the outer rail is raised respect to the inner rail. The rail closer to the center of the curve (the inner rail) is often referred as the *low rail*; the rail more distant to the center of the curve (the outer rail) is often referred as the *high rail*.

Rails in the past were joined together using fishplates; this procedure has been substituted with welding, leading to long (some kilometers) stretches of rails welded together forming a continuous welded rail.

2.1.3 Rail Loads

According with ref. [20], vehicle's wheels apply to the rail different types of loads, named in the following.

Vertical loading it is applied by the wheel tread to all rails under normal operation. It is made up of three components.

- *Static* or *nominal component*: is the gross weight of the vehicle divided by the number of wheels. In curved tracks, depending on the track superelevation and operating speeds, the loads applied between high and lower rail can be quite different (∼20-30%).

- *Dynamic component*: the increase over the static load occurring as the vehicle speed increases, caused by vehicle dynamics.
- *Impact component*: the increase over the static load and dynamic components occurring either when a wheel bumps over short vertical rail irregularities or if a wheel contains wheel flats or is of round.

Lateral loading it is applied by the wheel flange to the high rail in curved tracks because of curving forces.

- In relatively sharp curves (600-800 m in radii) it is relatively stable throughout the curve.
- In shallow curves or tangent track it can be occasionally very high when vehicles exhibit a high lateral dynamic behavior, such as hunting.

Creep forces are generated at the wheel/rail contact patch, by the very localized action of the wheel rolling on the rail.

- In longitudinal direction creep forces are generally obtained when traction is applied or when entering in a sharp curve.
- In lateral direction creep forces are generally obtained when the wheelset oscillates laterally on the rail.

Often both longitudinal and lateral creep forces are produced at the same time.

The contact patch usually has an elliptical shape and is relatively small: 10-12 mm long, 5-8 mm wide. It supports the whole wheel load.

2.1.4 Rail Stresses

According with ref. [20], the rail undergoes several bulk stresses, named in the following.

Bending stresses rail bending can be caused by a range of factors:

- *Applied vertical wheel loads*: the rail head bends vertically between the rail supports or on the web support, causing respectively tensile longitudinal stresses in the rail foot and on the fishing surface.
- *Applied lateral wheel loads*: the rail head moves laterally relative to the foot, causing tensile vertical stresses in the rail web and increasing thensile longitudinal stresses on the fishing surface on the field side.
- V*ertical loads applied at some distance from the rail centre line*: the rail is put in torsion, causing additional tensile vertical stresses in the rail web and tensile longitudinal stresses on the fishing surface.

Thermal stresses in long or continuously welded rails, longitudinal thermal expansion or contraction when the temperature respectively rises or falls respect to the welding temperature, thermal stresses are induced.

- When the rail expands, the rail undergoes compressive longitudinal stresses, and this can cause rail buckling.
- When the rail contracts, tensile longitudinal stresses are established, and this can cause development of transverse rail defects.

Residual stresses occur in prime rails because of manufacturing process. Their characteristics are very variable, including very high tensile vertical stresses in the rail web due to head hardening or aluminothermic welding, or tensile longitudinal stress on top of the rail head and the foot, and compressive longitudinal stresses in the web and the side of the head, all caused by head hardening.

2.2 Classification of Rail Defects

In literature, different classifications of rail defect exist. According to ref. [7] a rail track defect could be either of geometry or structural type. Track geometry defects are generated from the geometry conditions of the track; track structural defects (*rail defects*), indicate ill-conditioned structural parameters, including the rail [6].

We focus our attention on track *structural* defects, which can be divided in three broad categories:

1. *rail manufacturing defects*,
2. *improper use or handling of rails defects*, and

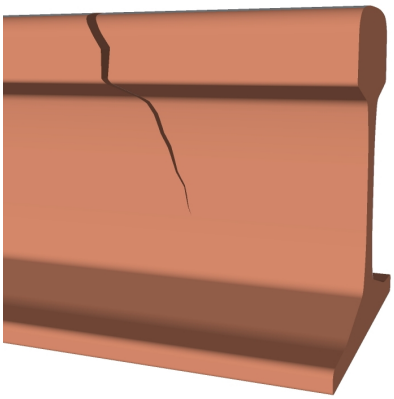


Figure 2.2: 3D representation of transverse defects.

3. *rail wear and fatigue defects.*

This classification will be exploited in the following, keeping in mind that it is difficult to classify a defect as due to improper use or to fatigue, since the two types of defects can be related to each other in some cases [8].

2.2.1 Rail Manufacturing Defects

Rail manufacturing defects usually comprise inclusions or wrong local mixings in the rail steel that generate localized stresses under operative load which could trigger the rail failure process [5, 9]. Today, great attention is being assured during fabrication and finished product quality testing to avoid that defective rails enter the market [14]. These procedures have lowered the occurrence of rail faults due to manifacturing process, but still some defects can pass undetected [8]. Two types of damages can be conducted to this category [5, 10, 11]:

Transverse defects (TD), a progressive fracture occurring in the railhead parallel to the transverse direction [5], see Fig. 2.2. Some examples of transverse defects are *tache ovale* or *shatter cracks*, internal defects which initiate near the centre of the railhead because of the presence of excessive levels of hydrogen in rail steel and that grow transversally [20].

Longitudinal defects (LD), an internal progressive fracture propagating longitudinally in the rail, see Fig. 2.3. There are two sub-categories: *vertical* and *horizontal* split heads [5].

- Vertical split heads generally form near the centre line of the rail and can be of considerable length ($>$ 0.5-1 m). They usually initiate at elongated seams, inclusion stringers, or highly segregated regions, and progress vertically thanks to heavy axle loads in association with impact loads or extreme wear conditions. The exact mechanism for which the crack extend along the rail is still unclear [20].
- Horizontal split heads initiate on the field side of the rails, 10-15 mm below the running surface and grow horizontally across and along the railhead. The initiation causes are the same of vertical split heads, but the actual stress conditions producing the crack initiation and growth are not known, but may entail a shear stress component [20].

2.2.2 Improper Use or Handling of Rails Defects

Improper use or handling of rails defects are generally caused by spinning of train wheels on rails (e.g. *wheelburn* defect, visible in Fig. 2.4) or sudden train brakes [5].

Wheelburns form on the running surface of the rails, and are very similar in appearance to squats (see Section 2.2.3). However, while squats can form at discrete locations, wheelburns always occur in pairs on both rails of the same track. Wheelburns are caused by the continuous slipping of locomotive wheels on the rails when the longitudinal creepage reaches saturation. The slipping wheel increases the temperature near the surface of the rail to very high values. The subsequent fast cooling cause the rail material to transform in a hard and brittle martensite phase, with associated spalling and flow of material, creating a depression. Because of this, transverse defects can be initiated, and their grow rate can be increased thanks to the higher impact loads caused by the depression.

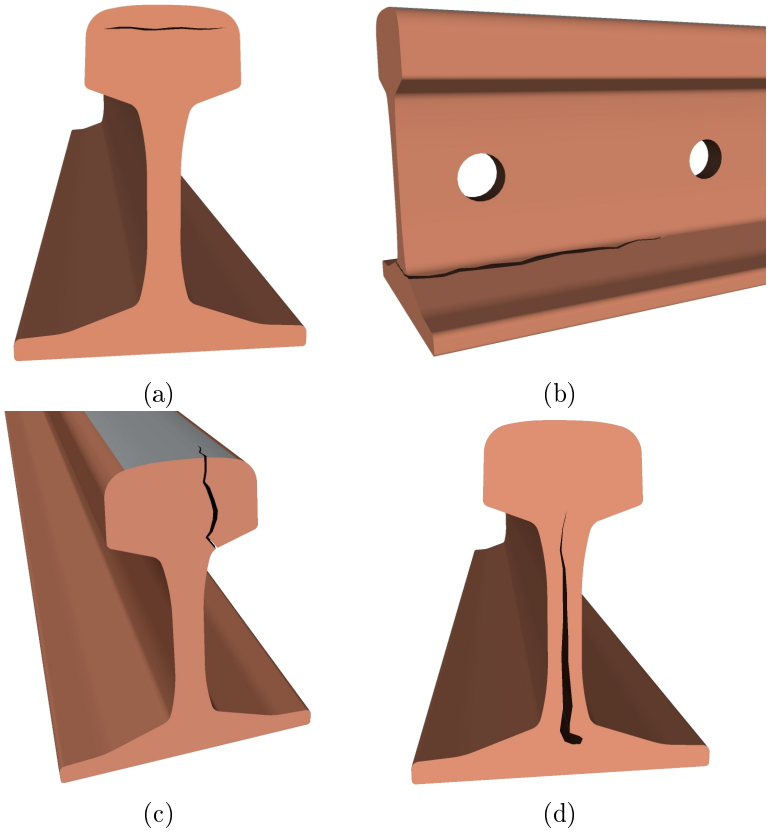


Figure 2.3: 3D representation of longitudinal defects. (a) and (b) are horizontal fissures, (c) and (d) are vertical fissures.

2.2.3 Rail Wear and Fatigue Defects

Rail wear and fatigue defects are due to wearing mechanisms of the rolling surface and/or to fatigue. There are three common defects [5]:

1. *corrugation*,
2. *rolling contact fatigue* (RCF) damage, and
3. *bolt-hole cracks*.

Corrugation is correlated to the wearing of the railhead [12]. Wear occurs whenever a wheel rolls over a rail and transmits a tangential force; it develops usually rather slowly unless sliding between wheel and rail or high tangential forces. Corrugation is constituted by cyclic (wave-like), generally vertical, irregularities on the top of the railhead [20]. Wear is related to plastic flow, that is a bodily deformation without loss of material; it results from high loading of the rail surface normal and tangential to the wheel/rail contact area [17]. Wear is thought to be the cause of short-pitch corrugation (30-90 mm wavelength), while plastic

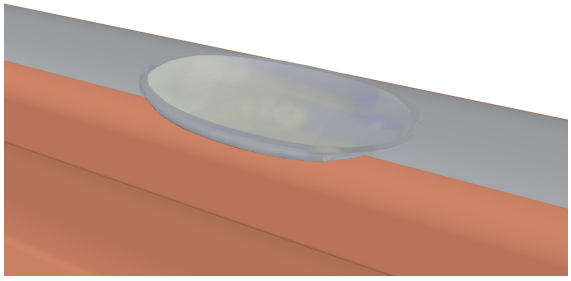
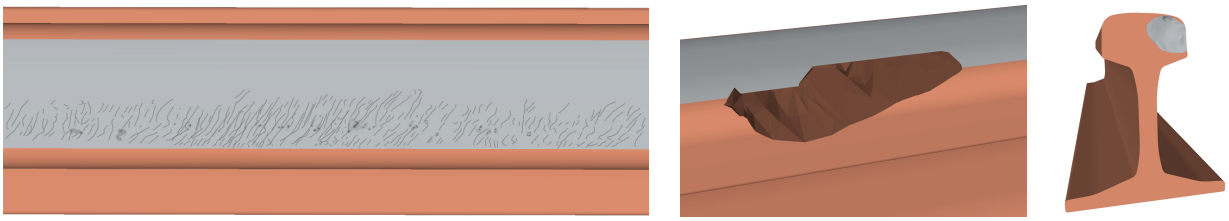
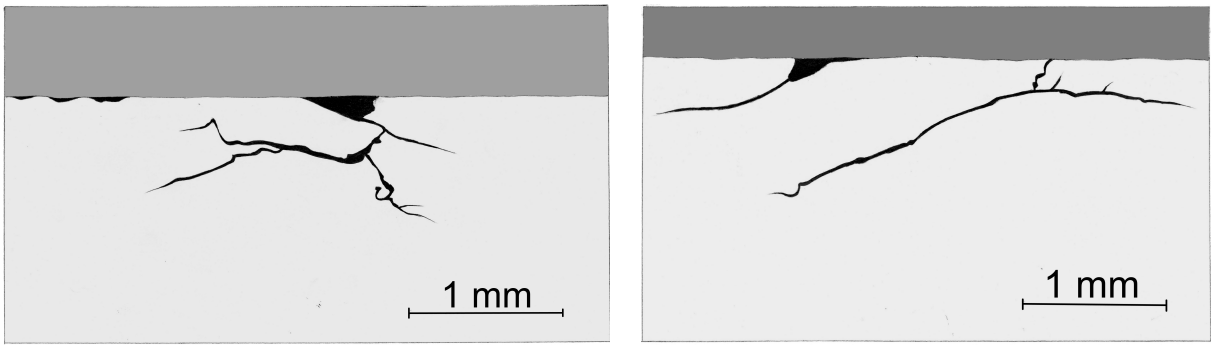


Figure 2.4: 3D representation of a wheelburn defect.



(a) (b) (c)

Figure 2.5: 3D representation of RCF caused defects. (a) Gauge corner checking and spalls, which can cause (b) shelling or (c) transverse defects.



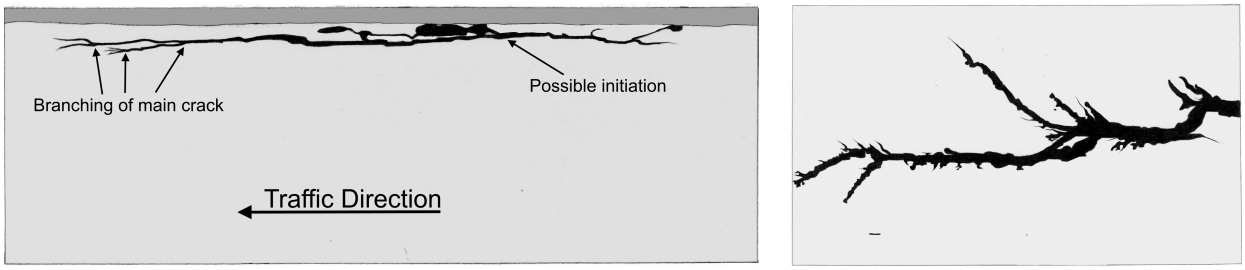
(a) (b)

Figure 2.6: Initial stages of development of checking cracks and spalls. (a) Transverse section of the rail, (b) longitudinal section of the rail. It is possible to note how the crack grows vertically into the rail head in a first instance and then levels out. .

flow is thought to be the cause of long-pitch corrugation (>300 mm wavelength) [20]. Corrugation does not compromise rolling safety, but has adverse effect on both track elements and rolling stock by increasing noise emission, loading and fatigue [13].

RCF damages are much more severe from the point of view of the structure integrity, as they could lead to complete rail failure [14, 15, 16]. RCF, as the name suggests, is due to the contact stresses between the rolling wheels and the rail surface [17]. The occurrence of RCF increases with increasing the train speed or weight [5], and is more often found in the outer rail of curved tracks [17]. RCF explicit itself in different forms or types of defects.

- *Gauge corner checking* (GCC): this is a surface condition occurring mainly on the high rails in sharper curves, or also in shallower curves or in tangent tracks when the vehicle tends to exhibit hunting behavior [20]. GCC has the appearance (see Fig. 2.5(a)) of many closely-spaced cracks on the running surface and gauge corner of the rail (*headchecks*). The angle of the crack mouths is related to the angle of the force that initiated the crack [17]. Independently of any material defect, the fatigue crack initiates on (or very close to) the rail running surface [15] and develops into the rail at an angle of about 20° respect to the surface [17].
- *Shelling* is an internal defects that initiates at a depth of 2-8 mm below the gauge corner region of the rails. It does not form as regularly as GCC. Initially it appears like a dark spot in the gauge corner region of the rail. Shelling cracks develop on a horizontal or longitudinal plane consistent with the shape of the rail on the gauge corner. As visible in Fig. 2.6, the cracks can continue to grow in a longitudinal direction on that plane at an angle of about 10°-30° respect to the surface [20]. During this phase, the crack grows through the layer of compressive residual stress near the running surface. At the lower edge of this layer, the crack tends to turn, giving rise to different defects [17].
  - If the crack turns along the rail surface, it could join with another crack [17], detaching a piece of metal and forming a *shell* [20].
  - If the crack turns down into the rail material, it develops subsequently under bending and other bulk stresses, giving rise to *transverse defects* [17]. As already stated, transverse defects may also directly initiate from irregularities in the steel (inclusions) and grow without the need for a prior shelling defect. Because of their internal nature, transverse defects cannot be visually detected: regular ultrasonic rail inspection is needed [20].



(a) (b)

Figure 2.7: Sub-surface cracking associated with squat defect. (a) Longitudinal section seen from the side of the rail, (b) longitudinal section seen from the top of the rail.

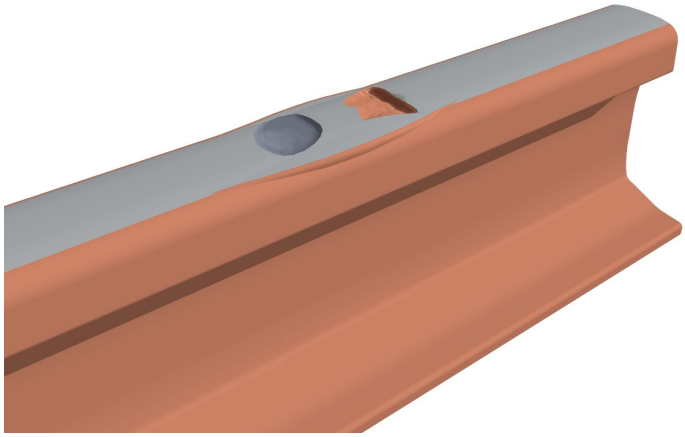


Figure 2.8: 3D representation of a large crown squat.

The conditions in which the crack branches are not yet fully understood [17].

- *Running surface checking* or *flaking* occurs on the running surface of low and/or high rails. Initially it appears as a mosaic or snakeskin-like pattern on the railhead. The cracks can then grow and produce *spalling*. It is a rather continuous effect along the rail length [20].
- *Squats* are surface or near-surface initiated defects and can be of two types.
  - *initiated on the crown or ball of the railhead*: these are the more common type, and appear as dark spots or "brushes" on the running surface. Each of such squats consists of two main sub-surface cracks: a longer and fringed one leading one propagating in train travel direction, and a trailing one propagating in the other direction. This is depicted in Fig. 2.7.
  - initiated by gauge corner checking cracks: they grow laterally spreading towards the crown of the railhead.

  Both types develop mainly in shallower curves and tangent track, and in hardened rails, as discrete defects or as closely-spaced multiple defects.

  Historically, squats have been classified as RCF-caused defects, and this may be true for GCC-initiated squats. But the crown-initiated squats (see Fig. 2.8) are caused by a "white etching", a hard and brittle layer in rails. White etching arises because of micro-slips of traction wheels on rails, which cause adiabatic shear stress between rails and wheels: this phenomenon is very similar in nature, but not in degree, to wheelburns.

  Secondary or minor sub-surface cracks may grow and become transverse defects, that could be masked by main superficial cracks to traditional ultrasonic inspection. Squats can also increase the vertical impact wheel loading, deteriorating both rails and wheels [20].

The difference in initiation characteristics of the checking and shelling defects (surface or sub-surface) is primarily due to the influence of the traction coefficient[2] on the resultant shear stresses. Lower values (high axle loads) induce shelling; higher values (lower axle loads or in sharp curves or when vehicles exhibit adverse dynamic) induce checking [20]. Also checking cracks can be capable to advance into the railhead

[2]The ratio of tangential to normal forces T/N.

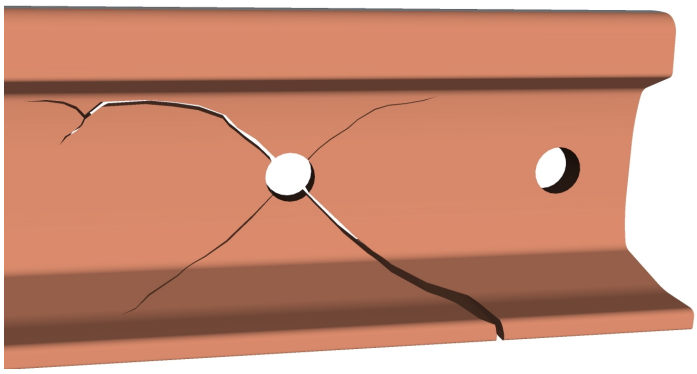


Figure 2.9: 3D representation of bolt-hole cracks.

[17] under conditions of low axle load and high speeds. Contamination of the rail surface with fluids can accelerate the propagation of checking cracks into longer and deeper cracks [20], following then to a process similar to that of shelling cracks. At a some depth the crack turns, with the risk of the development of a transverse defect [17].

RCF causes problems to traditional ultrasonic inspection procedures, since the earlier shallow-crack development phase can shadow downward-turning cracks [8].

To counteract RCF, the rail can be prematurely removed in the most severe cases or grinded to remove the surface initiated crack. The running surface can also be lubrificated, but it is believed that [14] fluid entrapment in metal is one of the most common causes for speeding up the surface initiated crack growth. Also higher strength rail steels can be used to increase the allowable shear stress limits, reducing wear and plastic flow [20].

Bolt-hole cracks happens in joined tracks; as visible in Fig. 2.9, they originate on the closer surface to the bolt-hole and propagate with a $\pm 45^{\circ}$ angle from the vertical until the web-railhead junction. The principal cause of these defects it is believed to be the fretting fatigue caused by the bolt shank against the hole surface [5].

2.2.4 Welded Rails

To enhance riding confort and reduce noise pollution caused by wheels bumping on rail joints, it is preferable to have a long continuous stretch of rail. It can be obtained through rail welding, and takes the name of *continuous welded rail*.

Rail welding can be of two types:

- *flash welding*, in which the two extremities of the rails to be welded are heated by means of high electrical currents to reach the plastic state and then put in contact until they are joined together;
- *aluminothermic welding*, in which melted aluminum added with iron oxide is poured in the space between the extremities of the two rails to be welded [19].

In welded rails, therefore, points which need careful attention are welds (see Fig. 2.10), because of weld internal defect that can affect the structural integrity and fatigue performance, bringing to rapid failures [23].

Moreover, welded rails act as a long metal bar subjected to thermal stresses, i.e. elongation and contraction because of rising and falling temperature.

- When the temperature rises, the metal expands, the rails elongate: there is the risk of track buckling.
- When the temperature falls, the metal contracts, the rails shrink: there is the risk of a brittle rail failure, also in welds.

Both of these effects give a high risk of train derailment. To attenuate thermal expansion, the rails are welded together under conditions that simulate rather high ambient temperature, so that rails usually work in tension. This reduces the risk of buckling in the hot summer days, but raises the risk of rail breakage when the temperature is low, exalting the need for careful inspection [14].

Figure 2.10: Thermite weld, in this case joining two different-profile rails: 50E5 (UNI-50) on the right, 60E1 (UNI-60) on the left.

2.2.5 Defect Growth

The failure through cracks of metallic structures goes through three phases:

1. *crack initiation*,
2. *crack growth*, and then
3. a subsequent *rail breakage*.

The first two phases require the accumulation of loading cycles over a period of time. It is necessary to avoid that the crack growth reaches the third phase, so actions are needed to detect a crack when it is still in the first stages of its growth [14].

2.2.6 Analytical and Statistical Models

The structural degradation of rails, in particular due to RCF, can be modeled. Such models can be used to estimate the expected level of degradation in given portions of the rail network on the basis of service conditions and inspection intervals. Fatigue models can be very accurate, accounting for a wide range of applied loads. The problem is that the input data on these loads are or unavailable or unreliable, making difficult to asses the accuracy of these models.

The classification between defects due to improper use or fatigue is often unclear, since the two types of defects can be related to each other in some cases. Railway engineers developed a *rail defect management* (RDM) scheme to take into account the failures attributable to the two types of defects mentioned and maximize the efficiency of maintenance RDM is based on statistical information obtained from failure reports (location and type of damage, route, date, type of rail, etc.). On the basis of these statistics (occurrence of failure, causes, frequency), those about network usage and defect occurrence, the maintenance schedule can be optimized [14]. Since the defect development is due to uncontrolled and random processes or events, this process is difficult to be represented by statistical models [8]. This has resulted in the improvement of current rail inspection technologies to find cracks at earlier stages [6].

Chapter 3

Existing Rail Defect Detection Techniques

3.1 Difference Between Testing and Monitoring Approaches

Two strategies can be adopted to assess the structural health of a structure:

- *structural health monitoring* (SHM) and
- *nondestructive testing* (NDT)

Structural health monitoring consists in constantly monitoring the structure by means of a fixed diagnostic infrastructure, so the health evaluation is made in real time. Nondestructive testing diagnostic equipment is applied to the structure only for the time required to asses the structural properties of the item to be tested in order to find defects. The main differences between the two approaches are listed in Table 3.1.

3.2 Rail Defect Detection Techniques

Different NDT and SHM rail inspection and monitoring techniques exist and are used nowadays. These can be classified as follows [23]:

1. *Eddy Currents-based methods*,
2. *ultrasonic techniques*,
3. *visual inspection*,
4. *thermal techniques*,
5. *Radiographic techniques*.
6. *Acoustic emission techniques* [8],
7. *Track circuits*.

In most cases the instrumentation is embarked on a vehicle (train, diagnostic vehicle), which performs its measurements while moving on tracks. Track circuits, instead, are distributed along the railway lines and the track itself is part of the system.

Table 3.1: Differences between SHM and NDT. Table taken from [21].

SHM	**NDT**
Online evaluation	Offline evaluation
Condition-based maintenance	Time-based maintenance
Determine fitness for service and remaining useful time	Find existing damage
Less cost and labor	More cost and labor
Baseline required	Baseline not available
Environmental data compensation methods required	

3.2.1 EC-Based Methods

EC-based methods include *eddy current* (EC) inspection and *magnetic flux leakage* (MFL).

Eddy current inspection method measures the material response to an induced electromagnetic field [29]. Two coils are used: one, being fed by an alternating current, induces a time-varying magnetic field near the surface of the railhead, thus causing eddy currents to be generated just below the surface of the test object. When a near-surface or surface defect (such as RCF, wheelburns, short-wave corrugations) is present, it disturbs the eddy currents, causing fluctuations in the secondary magnetic field sensed by the receiving coil: its impedance changes [8]. The method is non-contact, but very sensitive to probe lift-off from the surface of the railhead; moreover, it is characterized by the skin depth, i.e. by the depth of penetration in the material which is capable the AC current [23].

An improvement with respect to EC technique is the *alternating current field measurement* (ACFM) technique, patented by TSC Inspection Systems. This techniques allows the lift-off to be increased with respect to EC probes (5 vs 2 mm), thanks to the fact that the signal strength diminishes with the square (and not with the cube) of the lift-off [23].

Magnetic flux leakage consists of magnetizing the test object through permanent magnets or DC alternating currents and scanning its surface with a flux-sensitive sensor [30]. In the presence of a defect, the magnetic flux (usually contained into the magnetized test object) leaks into air and can be measured [31]. MFL is good for superficial or near-superficial transverse fissures like RCF, but it is not suitable for longitudinal fissures, and it is strongly dependent on scanning speed [24]. MFL can be used as a complement to traditional ultrasonic inspection [8].

The inspection speed of these systems is limited (75 km/h for EC testing, 35 km/h for MFL) with respect to those used by revenue service trains [23].

3.2.2 Ultrasonic Techniques

Ultrasonic techniques comprise *conventional bulk wave ultrasonics*, *phased arrays*, and *ultrasonic guided waves* (UGW) techniques, which can be subdivided in *on-board* and *land-based* systems. An introduction to ultrasonic techniques is given in the following Section 3.3.

3.2.3 Visual Inspection

Visual inspection can be exploited by a human expert walking on tracks seeking for defects (but this is rather subjective), or through automated visual techniques [8]. Automated visual inspection employs high-speed video cameras, optical sensors, and custom-designed algorithms embedded with machine learning to analyze rail surfaces as the train moves over the rail [32, 33, 34, 35]. Test speeds can be as high as the velocity of high-speed trains [8]. As the image resolution increases, the amount of data and the demanded elaboration effort also increases. If the travel speed is high and real-time operation is required, this rises the demand for processing capability, depending on the type of inspection and the expected quality [23]. This technique inspects only the visible surface of the rail [23]: internal defects cannot be detected. Furthermore, the outcomes can change if lighting condition change and no countermeasures have been taken [36]. Visual inspection is suitable to measure the railhead profile and percentage of wear, other surface damages like RCF and corrugation, or missing or poorly positioned track components [8, 37, 38, 39, 40, 41, 42, 43].

3.2.4 Thermal Techniques

Thermal techniques exploit the change in thermal properties of the material in presence of a defect. From a single frame or video sequence of temperature distribution of the specimen's surface the presence of an anomaly can be inferred [23, 48]. IR technology applied to rail inspection presents several challenges, such as the variability of the surface cleanliness, low emission of polished steel, camera issues, optical distortions, maintenance of the correct field position, etc.) [23].

3.2.5 Radiographic Techniques

Radiography it is the only technique able to examine visually the interior of a rail or a weld. This method can detect by means of X-rays[1] cracks, flaws, and thickness reduction with very good detail, but has safety hazard concerns, can be expensive and time consuming. It is often used to investigate the rail on the basis of the

[1] Gamma-rays were used in the past [8].

Figure 3.1: ETR.500 Y 2 "Dia.man.te", high speed (300 km/h) visual inspection EMU (for track and catenary) of Rete Ferroviaria Italiana (RFI), Italian railway's infrastructure manager. Courtesy of © Stefano Paolini.

outcomes of other NDT techniques [23]. Radiography is good at detecting longitudinal defects, but not very efficient in detecting transverse defects [8].

3.2.6 Acoustic Emission Techniques

Acoustic emission systems consists in the analysis of the radiation of acoustic waves in solids emitted when a material undergoes irreversible changes in its internal structure produced by mechanical loading. Acoustic emission is a passive technique in which these waves are sensed through sensitive piezoelectric transducers; the signals are then amplified, processed and analyzed. The wave parameters are used to identify the nature of the source; it is fundamental to separate stress-wave emission from noise, through filtering and advanced signal processing. By means of triangulation and other techniques it is possible to locate and size the crack [8]. In the rail field it is possible to evaluate the noise produced at the wheel-rail interfaced by a test train during motion. This enables to detect rail breaks, wheelburns, squats, worn rails, track geometric issues, ballast deterioration, etc. [44].

If the signal processing cannot eliminate the amount of external noise, it is possible to use the active approach to AE [45]: *acoustic emission pulsing* (AEP). A transducer generates ultrasonic waves, and the emitted noise spectrum is analyzed since this changes in the presence of a defect. Acoustic emission pulsing is a static inspection technique [8].

3.2.7 Track Circuits

Track circuits are systems used to detect the presence of a train in a stretch (section) of track by means of an electrical current. At one extremity of the track section,a transmitter is placed, at the other a receiver. When no trains are present, the electrical signal flows through the rails from transmitter to receiver; the train wheels, being conductive, short-circuit the pair of rails: the signal is no more able to reach the receiver, the section is occupied. As a side effect, the same happens in the presence of a complete rail failure, since the electrical continuity is interrupted; this results in an occupied section, even if there are no trains between transmitter and receiver. The problem is that track circuits detect a rail breakage only if it interrupts the electrical continuity of the circuit, so they are unsuited to detect early-stage cracks and their extension does not cover all the rail network [49]. For additional information about track circuits, see Chapter 16.

3.3 Ultrasonic Rail Defect Detection Techniques

Ultrasonic rail defect detection techniques are divided in three main categories [23]:

1. *conventional ultrasound*,
2. *phased arrays*,
3. *guided waves ultrasound*.

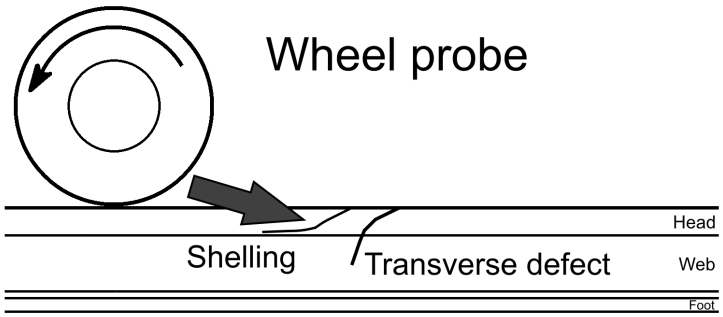


Figure 3.2: Transverse defect masked by the presence of a shelling during conventional ultrasonic testing.

Figure 3.3: "us1-Galileo," conventional ultrasonic measure train of Rete Ferroviaria Italiana (RFI), Italian railways infrastructure manager. Courtesy of © Benedetto Sabatini.

3.3.1 Conventional Ultrasonic Techniques

Ultrasonic rail inspection can be performed manually using a portable device, or by using a test vehicle. In conventional ultrasonic techniques, the railhead is scanned by means of ultrasonic beams generated by probes and transmitted into the rails. Defects are detected by measuring reflected or scattered energy using a collection of transducers [8]. The presence of a defect, along with its location and severity, is inferred from the reflections' amplitude and their arrival time [24]. The transducers are contained within a liquid-filled tire, to increase signal quality by reducing the acoustic mismatch between rail steel and air. For the same reason, water is sprayed between the ultrasonic probes (wheel) and the rail [8]. Since defects can be located in various parts of the rail, the energy is transmitted at various incident angles [47] and the ultrasonic transducers are positioned to look also across the railhead for longitudinal defects [46].

Although ultrasonic testing is capable of inspecting the whole railhead [25], it has several drawbacks.

1. *Limited ultrasonic inspection car speed* [14]: the practical achievable test speeds can be as low as 15 km/h [8], forcing the inspection to be carried out outside commercial trains operation period [5].

2. *Shallow crack shadowing* [14, 11]: small shallow cracks can shade much more severe and deeper cracks by reflecting the ultrasonic beams [5], as shown in Fig. 3.2.

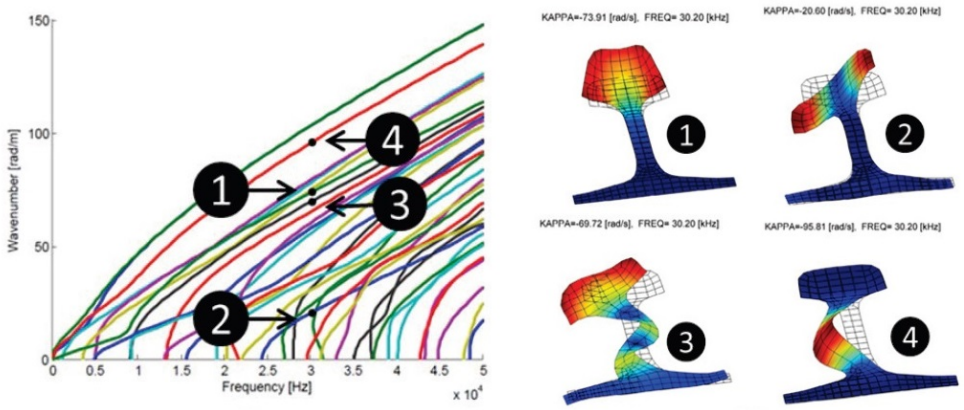
3. *False defect detection* [5], which slows down inspection operations.

The Italian railway infrastructure manager's ultrasonic testing vehicle "us1-Galileo" is shown in Fig. 3.3.

Ultrasonic test trains perform relatively well in detecting deep surface breaking ($>$ 4 mm in depth) and internal defects in rail head and web. The performance in foot analysis or in aluminothermic welds is relatively poor [23].

3.3.2 Phased Arrays

Phased arrays use multiple ultrasonic elements and electronic time delays to create beams by constructive and destructive interference. Phased arrays provide ultrasonic beams that can be steered, scanned, swept, and



(a) (b)

Figure 3.4: Dispersion curves (a) and mode shapes (b) of a rail waveguide. Some candidate modes for long range damage detection are highlighted. Modes 1 and 2 are respectively symmetric and anti-symmetric, both with energy concentrated in the crown of the rail. Mode 3 has energy more evenly distributed across the rail cross section, but with little energy in the foot. Mode 4 is concentrated in the web of the rail. Pictures taken from [55].

focused electronically [28]. This technique represents an improvement with respect to the single-element US transducer, offering more flexibility at the expense of the complexity data processing; this could lower the inspection speed [23].

3.3.3 Guided Waves

In an effort to overcome some of the issues associated with ultrasonic wheel inspections, some research groups are investigating ultrasonic guided waves for rail inspections. Different types of guided waves can propagate in any bounded medium: for a theoretical outline, see Section 5.

The test can be done according to the two modes listed below

Pulse-echo mode: the transducer transmits the guided wave along the structure; the presence of defects or other structural features is inferred by analyzing the returning echoes [22].

Pitch-catch mode: a transmitter generates a guided wave in the waveguide and a receiver measures the incoming wave. If a defect reflects part of the energy, the received signal is weaker than in pristine waveguide conditions; hence, the presence of the defect can be inferred. [62]

The advantages of ultrasonic guided waves are the following [26]:

- *guided waves propagate along, rather than across the rail*, and are thus ideal for detecting the critical TDs, but relatively insensitive to parallel fissures [56];
- *guided waves increase the inspection coverage*, thereby relaxing limits on the achievable speed (if mounted on-board of a moving vehicle);
- *guided waves can travel underneath shelling and still interact with internal defects*, since they penetrate a finite depth of the rail surface;
- *guided waves can penetrate alumino-thermic welds*, hence potentially targeting weld cracks and discontinuities, since they travel in the mid-frequency range, between 20 kHz and 1 MHz.

There is a downside in guided waves:

- *Their complicated propagation behavior* cannot be easily managed [27]. In fact, iltrasonic guided waves, especially in complex waveguides such as rails, exhibit
  - a *multi-mode* character (many modes can propagate simultaneously) and
  - a *dispersive* character (the propagation velocity depends on the frequency) [22].

  Several tens of different wave types can occur in the frequency range, and the dispersion curves are very complicated [27], as it can be inferred from Fig. 3.4.

Chapter 4

Mechanical Properties of Solids and Elastic Waves

4.1 Mechanical Properties of Solids

It is useful to summarize the mechanical properties of solids, so that the physical background of rail breakage mechanisms and the theory of ultrasonic guided waves can be easily understood. In the following subsections, a brief summary about the most meaningful properties of solids is presented. This section is mainly inspired by ref. [63].

4.1.1 Elastic Deformation

All solid bodies deform under the action of external mechanic solicitations: mechanical properties rule the response of the material to such stresses.

When a material is stressed by means of a force (a *load*) of small intensity, it undergoes a small *deformation* of entity proportional to the applied load. The deformation is nullified when the load ceases its action. The proportionality between deformation and load constitutes the *Hooke's law*[1].

If a bar of length l and section of area S is considered, the following definitions can be given:

- *specific* or *unitary load* σ as the ratio between the force $\mathbf{F}$ applied orthogonally to a surface and the surface area itself S,

$$\sigma = \frac{F}{S} \qquad \left[\frac{N}{m^2}\right]; \tag{4.1}$$

- *specific deformation* or *unitary linear elongation* ε as the ratio between elongation which the solid undergoes and its length (dimensionless),

$$\varepsilon = \frac{\Delta l}{l}. \tag{4.2}$$

The ratio between specific load σ and unitary elongation ε for load values such that the two quantities are proportional is defined as *Young modulus* or *elasticity modulus* E:

$$E = \frac{\sigma}{\varepsilon} \qquad \Rightarrow \qquad \varepsilon = \frac{1}{E}\sigma \qquad \Rightarrow \qquad \frac{\Delta l}{l} = \frac{1}{E}\frac{F}{S}. \tag{4.3}$$

The Young modulus represents the relationship between linear elongation Δl and applied force F as a function of geometric parameters of the solid (length l and base area S) and its elastic characteristics (E). Generally speaking, in metals the values of Young modulus decrease with temperature[2].

Because of traction, the bar not only stretches, but also changes its section. If r is a transverse dimension (e.g. for a cylindrical bar) it is found that, remembering (4.1), (4.2) and (4.3):

[1] $F = -kx$: if a force is applied to an extremity of a spring whose other extremity is fixed, the spring itself undergoes an elongation Δx, reaching an equilibrium configuration, with Δx proportional to F. The elongation causes in the spring an elastic reaction.

[2] Until a phase change happens, i.e. the material crystalline structure changes.

$$\frac{\Delta r}{r} = -\nu \frac{\Delta l}{l} = -\nu\,\varepsilon = -\nu \frac{\sigma}{E}, \tag{4.4}$$

where ν is a dimensionless quantity representing the *Poisson's coefficient*, ranging from 0 (cork) to 0.5 (rubber). Eq. (4.4) holds for small values of σ. During traction, the volume of the bar does not diminish: it remains constant if $\nu = 0.5$ and grows if $\nu < 0.5$. On the contrary, under compression, the volume of the bar never grows.

4.1.2 Plastic Deformation

Above the critic load value, depending on the material, a permanent deformation occurs, named *plastic deformation*; this is not reversible when the load is removed. The critical load value after which plastic deformation occurs is defined *enervation specific load* σ_s. If the applied load grows further, it is reached the *breakage unitary load* σ_u and the solid breaks.

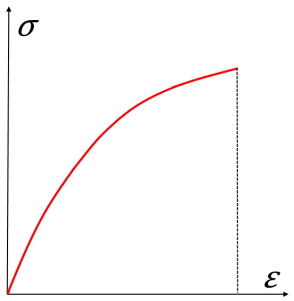


Figure 4.1: Load-elongation curve.

The product between specific load σ and specific elongation ε has the dimensions of an energy per unit volume ($^J/_{m^3}$). The area subtended by the curve σ-ε in Fig. 4.1 has the significance of an energy absorbed by the material per unit volume and it is usually denoted as *tenacity modulus*. The material *ductility* is an index of the plastic deformation before breakage, and can be expressed by the specific elongation (or reduction) developed under load.

A characteristic property of solid bodies is the *elastic hysteresis*: the relation between specific load σ and specific elongation ε is not univocal for bodies that undergo plastic deformation, but depends on the amount of plastic deformation. This is described by the *hysteresis cycle* on the σ-ε plane, as shown in Fig. 4.2(a); the mechanical energy absorbed during the whole cycle is completely dissipated being transformed into heat. It is possible to go back to the initial state $\sigma = 0$ and $\varepsilon = 0$ by going through several cycles of hysteresis of reducing area, as shown in Fig. 4.2(b). By repeating a loading cycle in which the applied load once exceeds the specific load and then is zeroed, it is possible to considerably increase the material elasticity limit: this process is named *hardening*. Through a wide temperature heating (*annealing*) it is possible to reverse this process, bringing back the material to its original properties.

4.1.3 Creep

To explain creep, consider a solid parallelepiped glued on two opposite faces to two rigid plates, one fixed at the bottom, and the other one at the top able to move in a parallel way with respect to the first one. As shown in Fig. 4.3, by applying a force **F** to the movable plate, a creep is observed on the upper face of the solid with respect to the lower one, measurable through the angle θ. The linear relation

$$\sigma = G\,\theta \tag{4.5}$$

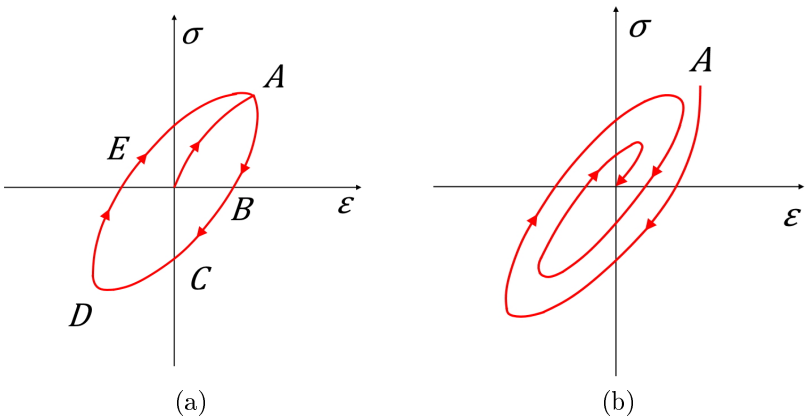


Figure 4.2: Elastic properties of a material. (a) Hysteresis cycle; (b) hardening.

hold between the specific load $\sigma = \frac{F}{S}$ and θ; here, G is the *rigidity modulus* or *shear modulus*, a material-specific parameter characteristic of such deformation.

At the application of **F** the solid reacts elastically and an equilibrium is obtained when $\theta = \frac{\sigma}{G}$; when **F** is removed, the body comes back to its original configuration. The discussion about the elasticity limit applies also tor creep.

Between elastic constants, for small values of deformation (i.e., small loads), the formula

$$G = \frac{E}{2\,(1+\nu)} \tag{4.6}$$

holds.

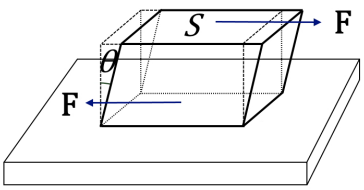


Figure 4.3: Solid parallelepiped that, under the action of force **F** shows creep.

4.1.4 Torsion

To explain torsion, consider the cylindrical bar shown in Fig. 4.4 of length l and base radius r. Its upper extremity is fixed; a couple of forces with momentum **M** is applied to the lower face. **M** is parallel to the bar axis, which is thus subjected to *torsion*.

The value of momentum **M** required to be able to rotate the lower base of an angle θ with respect to the upper one is given by

$$M = \frac{\pi}{2}\, G\, \frac{r^4}{l}\, \theta = k\,\theta. \tag{4.7}$$

The torsion effect strongly depends on transverse dimensions: with a very tin bar (like a wire), a measurable deviation is obtained even if a small momentum is applied.

The elastic reaction of the bar is manifested by means of a momentum $-\mathbf{M}$ whose modulus is proportional to the torsion angle, being $-k\,\theta$.

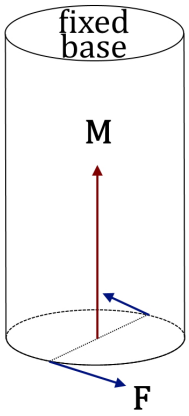


Figure 4.4: Cylindrical bar which under the couple of forces with momentum **M** shows torsion.

4.1.5 Pressure and Uniform Compression

The ratio between the force **F** applied orthogonally to a surface and the surface area itself S, previously named specific load, is the *pressure*

$$p = \frac{F}{S}. \tag{4.8}$$

For now pressure is considered only in the scope of a particular way to deform a body, the uniform compression. It consists of the application of an uniform pressure on all the surfaces of the object.

If V is the volume of the body in correspondence to the external pressure p, when a pressure variation Δp is applied, a volume variation ΔV is observed; this is given by

$$\frac{\Delta V}{V} = -\frac{1}{\beta}\,\Delta p, \tag{4.9}$$

in which β is the *isothermal compressibility module*[3]; β is measured in $^N/_{m^2}$ (the same units as those of the pressure). The parameter β is related to other elastic properties by means of the equation

$$\beta = \frac{E}{3\,(1-2\,\nu)}. \tag{4.10}$$

Solids and liquids have a very little compressibility (hence β is very high), but for gases it can be shown that β it is equal to the pressure itself (so they are very compressible).

[3]The volume variation is measured in experimental conditions such that the temperature of the body is unchanged during compression.

4.1.6 Hardness

Hardness is defined as the resistance of a body to penetration from another body shaped like a small sphere or a tip. It is possible to put in relation a material hardness to its mechanical resistance; since it is quite an empiric concept, different methods exist to measure hardness (Brinell's, Rockwell's, Vicker's). For metallic alloys, a linear relationship exists between Brinell's hardness and breakage specific load.

Numerically, hardness is generally measured as a ratio between the applied force and the resulting footprint area leaved on the sample.

4.2 Elastic Waves

This section is based on refs. [63] and [64], with some contributions from other specified authors when needed.

4.2.1 Definition

Elastic waves propagate in a material medium: their propagation is due to the interactions between atoms and molecules of the medium itself. Propagation never implies effective matter transport: atoms and molecules oscillate around an equilibrium position: the effective quantities transported are energy and momentum. *Transducers* are based on this effect to detect the presence of a pressure wave.

Formally, a wave can be recognized as a variation in the equilibrium condition of a field describing a property of the physical system[4]. A sound wave is a perturbation of pressure and density equilibrium state of a gas. An elastic wave in a bar is a local deformation of a bar section, representable with a displacement from equilibrium and with a variation of pressure propagating along the bar itself.

The field perturbation produced by a source which propagates in space can be represented by the *wave function* $\xi(x, y, z, t)$. The symbol ξ can represent either a displacement of a system element from an equilibrium position happening in all elastic waves, or a pressure variation Δp, or a density variation $\Delta\rho$, or also a force F or a power P.

4.2.2 Plane Waves

A particular situation is represented by *plane waves*, described by the function $\xi(x, t)$. This function is spatially one-dimensional and depends only on the spatial coordinate x as well as time. The waves are named "plane" since the perturbation in a given instant t_0 assumes the same value $\xi(x_0, t_0)$ in all the points of the plane of equation $x = x_0$, orthogonal to the propagation axis x, and therefore parallel to the plane y, z.

Limiting only to plane waves, the wave function satisfies always the differential equation

$$\frac{\partial^2\xi}{\partial x^2} = \frac{1}{v^2}\frac{\partial^2\xi}{\partial t^2} \qquad \text{or} \qquad \frac{\partial^2\xi}{\partial t^2} = v^2\frac{\partial^2\xi}{\partial x^2} \tag{4.11}$$

named *equation of plane waves* or *d'Alembert equation*, and the coefficient v^2 is the square of *propagation speed*.

It can be proved that the functions solution of eq. (4.11) can be of any type, but the argument of ξ has to contain variables x and t in the form of linear combination.

The most general solution of eq. (4.11) is

$$\xi(x, t) = \xi_1(x - v\,t) + \xi_2(x + v\,t)\,. \tag{4.12}$$

The physical significance of solutions in eq. (4.12) lies in the fact that they represent, because of their structure, a *propagation phenomenon* along the x axis with speed v. The value ξ_0 assumed by the function ξ at the instant t_0 at the position x_0, $\xi_0 = \xi(x_0 - v\,t_0)$, can be found in every following instant $t > t_0$ in the point x satisfying the condition

$$x - v\,t = x_0 - v\,t_0 \qquad \text{or} \qquad x = x_0 + v\,(t - t_0)\,, \tag{4.13}$$

a relation which expresses a rectilinear motion along x axis with velocity v. Similarly, $\xi(x + v\,t)$ represents a quantity propagating at speed v in negative direction along the x axis. In both cases it is about a *rigid translation*, the function never changes shape: this shape invariance during propagation is a characteristic of any function solving eq. (4.11).

The *overlap* of two waves is still a wave that can be obtained through summation in any instant and any point. Two waves, even if they act together, remain independent: one is not modified by the presence of the other one. Waves satisfying eq. (4.11) obey to the *overlap principle*. Anyway, the overlap of two or more waves in a space region can give rise to the insurgence, under particular conditions, of *interference phenomena*.

[4] The word *field* denotes a physical quantity that can be defined in every instance of time in any point of the space.

4.2.3 Elastic Waves in a Solid Bar

Until now, a constant force or momentum was applied to a solid body to determine deformation characteristics. These are time-invariant phenomena: after a transitory phase during which the force is applied, the deformation of each bar element is constant with time.

Consider, instead, the application of an impulsive excitation to the extremity of the bar. Since the bar is free to move, it starts moving with speed

$$v = \frac{J}{m} \tag{4.14}$$

in which J is the force impulse module and m is the bar mass. Because of the elastic properties of the bar, a second phenomenon is observed: the initial stretch of the bar, compressed by the impulsive force, subsequently expands compressing the following stretch, and so on. After a given amount of time, the perturbation applied to one extremity has arrived at the other one: an elastic perturbation has propagated along the bar; its speed of propagation depends on the physical characteristics of the material in which the bar is realized.

Consider, as shown in Fig. 4.5, a bar element at distance x from the excitation extremity whose length is dx: if the bar has a circular transverse section, it is a cylinder having base area S (that of the section) and height dx. On the two basis of this small cylinder are acting the forces $F(x)$ and $F(x+dx)$, exercised respectively from the stretch of bar lying on the left (towards the exciting extremity) and from that on the right of the considered element. In the propagation process, the force F is not constant along the bar, but varies both along the x axis and with time. Under the action of the aforementioned forces, each section changes position (albeit slightly). In the following

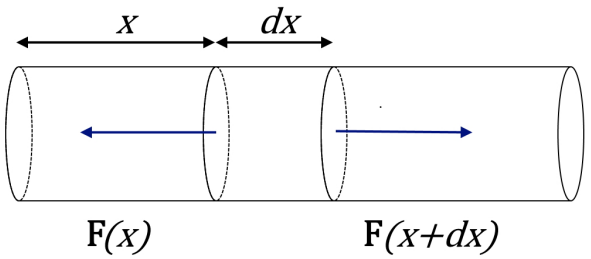


Figure 4.5: Bar element undergoing force F.

- $\xi(x,t)$ is a function describing the movement from the initial position in correspondence of abscissa x and time t;
- $\xi(x+dx,t)$ is a function describing the movement from the initial position at the same instant but in correspondence of $x+dx$.

The length of the small cylinder, which was dx, becomes

$$x + dx + \xi(x+dx,t) - x - \xi(x,t) = dx + d\xi, \tag{4.15}$$

since

$$\xi(x+dx,t) = \xi(x,t) + d\xi \tag{4.16}$$

The relative elongation of the element is $\frac{d\xi}{dx}$ or, better, $\frac{\partial\xi}{\partial x}$ (we are considering only a variation with respect x and not with respect time).

If we consider the equation describing the relationship between the linear elongation Δl and the applied force F, as a function of the geometric parameters of the solid (length l and base area S) and its elastic characteristics (E is the material elastic modulus, or Young modulus), i.e.

$$\frac{\Delta l}{l} = \frac{1}{E}\frac{F}{S}, \tag{4.17}$$

we can then substitute in eq. (4.17) the relative elongation $\frac{\partial\xi}{\partial x}$ for the relative elongation $\frac{\Delta l}{l}$ obtained from the Young modulus, achieving

$$\frac{\partial\xi}{\partial x} = \frac{1}{E}\frac{F}{S} \qquad \Rightarrow \qquad F(x) = E\,S\,\frac{\partial\xi}{\partial x}. \tag{4.18}$$

Based on the last result, the resultant of the forces working on the element can be expressed as

$$F(x+dx) - F(x) = \frac{\partial F}{\partial x}dx = E\,S\,\frac{\partial^2\xi}{\partial x^2}dx \tag{4.19}$$

On the other side, the motion of the small cylinder of mass

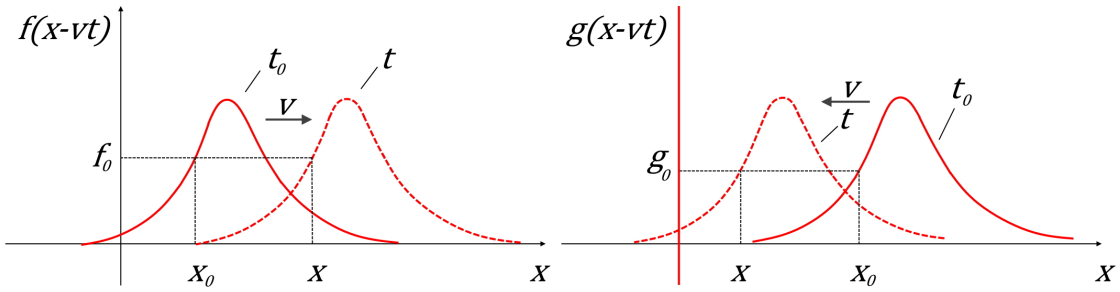


Figure 4.6: Most general solutions of D'Alembert equations, forward-propagating f and backward-propagating g in two different instants.

$$dm = \rho\, S\, dx \tag{4.20}$$

in which ρ is the bar material density, happens with acceleration

$$a = \frac{\partial^2 \xi}{\partial t^2}; \tag{4.21}$$

then, thanks to the Newton's law $F = m\, a$, we have

$$E\, S\, \frac{\partial^2 \xi}{\partial x^2} dx = \rho\, S\, dx\, \frac{\partial^2 \xi}{\partial t^2}. \tag{4.22}$$

Finally, we obtain the equation

$$\frac{\partial^2 \xi}{\partial t^2} = \frac{E}{\rho}\, \frac{\partial^2 \xi}{\partial x^2} = v^2 \frac{\partial^2 \xi}{\partial x^2} \tag{4.23}$$

in which v, the speed of propagation, is defined as

$$v = \sqrt{\frac{E}{\rho}} \tag{4.24}$$

and has an implicit dependence on temperature.

The solution of eq. (4.23) describes the propagation of the perturbation $\xi\,(x, t)$ along the bar. The solutions of eq. (4.23) are all and only the functions such that the variables x and t appear with a linear combination $x - v\, t$ or $x + v\, t$ in their argument i.e.

$$f\,(z) = f\,(x - v\, t) \qquad g\,(z') = g\,(x + v\, t)\,. \tag{4.25}$$

It is possible to prove that the most general solution of eq. (4.23) can be expressed as

$$G\,(x, t) = f\,(x - v\, t) + g\,(x + v\, t) \tag{4.26}$$

where

- $f\,(x - v\, t)$ represents propagation with velocity v in the positive direction of axis x;
- $g\,(x + v\, t)$ represents propagation with velocity v in the negative direction of axis x instead.

Therefore, eq. (4.23) is the *plane wave differential equation* or *D'Alembert equation*, and $f\,(x - v\, t)$ and $g\,(x + v\, t)$ are plane waves describing propagation along x with speed v in positive and negative direction, respectively. Hence, it is possible to sum the previous results up saying that along the bar a displacement wave propagates with a speed v which depends on density ρ and Young modulus E of the material.

Together with displacement ξ, from equilibrium position, also the force F propagates along the bar. From eqs. (4.18) and (4.23) it is easily inferred that

$$\frac{\partial^2 F}{\partial t^2} = E\, S\, \frac{\partial^2}{\partial t^2}\, \frac{\partial \xi}{\partial x} \qquad \Rightarrow \qquad \frac{\partial^2 F}{\partial t^2} = v^2 \frac{\partial^2 F}{\partial x^2}. \tag{4.27}$$

Along the bar a force wave propagates; better, remembering that in compression and torsion processes the ratio $\frac{F}{S}$ is meaningful, this is a pressure wave.

Both the displacement $\xi\,(x \mp v\, t)$ and the force $F\,(x \mp v\, t)$ describe waves propagating along and parallel to the x axis. Waves having such characteristics are said *longitudinal waves*.

4.2.4 Considerations About Waves

A propagation phenomenon happens when the spatial and temporal evolution of a physical quantity follows an equation such as eq. (4.11). The solution function propagates along the x axis by keeping the same shape.

The above-mentioned results show that, while usually the result of $\mathbf{F} = m\,\mathbf{a}$ describe the motion of a body having mass m, therefore describing a matter movement, in some other cases it predicts a phenomenon in which the matter, albeit still moving, stands still on average. In such cases, only a local perturbation propagates in a progressive way from one element to the next: there is no net movement of matter. This is also shown in Fig. 4.7. In such a context the wave is the motion of a perturbation and not a mass particle motion, even if both the motion types are based on the Newton's law.

As already shown, an elastic wave can be composed of two components, two waves, which propagate independently one from each other.

- In *longitudinal waves* the variation of the propagating quantity happens parallel to the propagation direction; it can be demonstrated that its propagation speed is given by

$$v_l = \sqrt{\frac{E}{\rho}}; \tag{4.28}$$

- In *transverse (shear) waves* the variation of propagating quantity is orthogonal to the propagation direction; it can be demonstrated that its propagation speed is given by

$$v_s = \sqrt{\frac{G}{\rho}}. \tag{4.29}$$

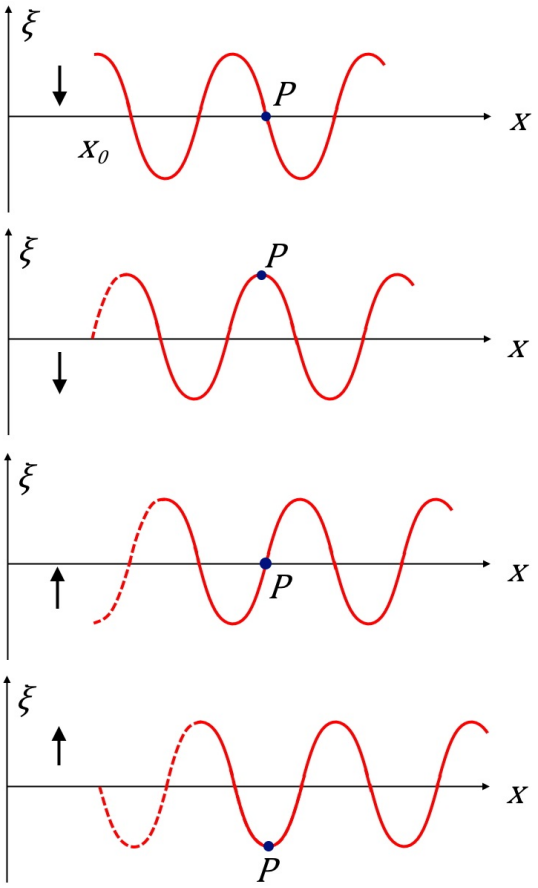


Figure 4.7: Demonstration that the wave is the motion of a perturbation and not of a particle.

It can be demonstrated that the propagation speed of longitudinal waves v_l is always greater than those of shear waves v_s, and that the inequality

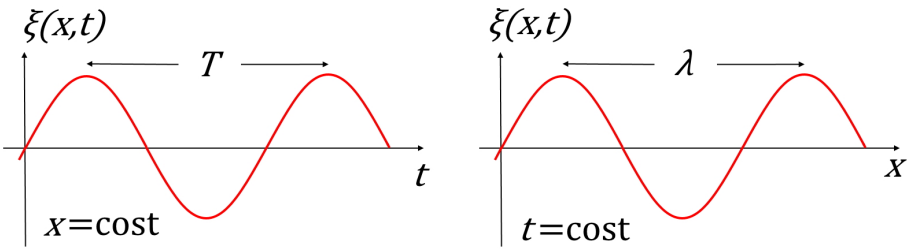
$$v_l > v_s\sqrt{\frac{4}{3}} \tag{4.30}$$

always holds [65].

The propagation characteristics depend on the elastic properties, therefore on the microscopic properties of the mean.

- In an indefinite fluid there is no shear resistance; the only elastic property is compression: only longitudinal waves can exist.
- In a bar, both longitudinal and shear waves can exist. Longitudinal waves can be generated by a compression mechanism, while shear waves are generated by soliciting an extremity of the bar orthogonally to its axis and propagate by means of creep. The bar, fixed at one extremity and solicited at the other with a variable torsion momentum transmits also torsional waves, which propagating speed is again $v_t = \sqrt{\frac{G}{\rho}}$.

When a wave propagates, there is also a transport of momentum and energy, provided by the external solicitation generating the phenomenon. Momentum and energy are then found along the propagation direction as the result of local interactions between contiguous particles. The transmission of energy and momentum happens initially between the vibration source and the contiguous element of the mean, and then from every element to the next. The fact itself that an element previously undisturbed enters into oscillation demonstrates the transport of energy and momentum by the wave. Another proof is the detection of wave arrival by means of an external detector, having a transducer capable to measure and translate into electric signal the vibrations of the mean. If the solicitation of the source is impulsive, the local oscillations due to the wave passage can be considered as a transitory phenomenon; if the solicitation is periodic, a forced oscillation phenomenon is obtained.



(a) (b)

Figure 4.8: Period (a) and wavelength (b).

4.2.5 Harmonic Plane Waves

A particular type of plane wave is the harmonic plane wave, whose wave function is:

$$\xi(x,t) = \xi_0 \sin(k\,x - \omega\,t), \qquad \xi(x,t) = \xi_0 \cos(k\,x - \omega\,t) \tag{4.31}$$

in which ξ_0 is the wave *amplitude*, k is the *wave-number*, ω is the *angular frequency* of the harmonic wave.

Given v the propagation speed at which the wave propagates, the angular frequency is related to the wave-number by:

$$\omega = k\,v \qquad \left[\frac{rad}{s}\right]. \tag{4.32}$$

The interval between two peaks of the harmonic oscillation is the *period*, see Fig. 4.8(a):

$$T = \frac{2\pi}{\omega} \qquad [s]. \tag{4.33}$$

The distance between two peaks of the harmonic oscillation (spatial periodicity) is the *wavelength*, see Fig. 4.8(b):

$$\lambda = \frac{2\pi}{k} \quad [m] \qquad \Rightarrow \qquad \left(k = \frac{2\pi}{\lambda} \quad \left[\frac{1}{m}\right]\right). \tag{4.34}$$

Temporal and spatial period are related through the propagation speed v by

$$\lambda = v\,T \tag{4.35}$$

therefore, in the time T, the sinusoid proceeding at speed v advanced by a distance λ, which therefore assumes the significance of the distance traveled by an harmonic wave in a period.

Velocity and wavelength are related by the *frequency* f, the inverse of the period:

$$f = \frac{1}{T} = \frac{\omega}{2\pi} \qquad \left[Hz = \frac{1}{s}\right], \tag{4.36}$$

so that

$$\lambda\,f = v. \tag{4.37}$$

In eq. (4.31) it is worth introducing the parameter δ, the argument value for $x = 0$ and $t = 0$ and for every other pair of x, t such that $k\,x - \omega\,t = 0$. The complete argument of the function is

$$\phi(x,t) = k\,x - \omega\,t + \delta \tag{4.38}$$

named *phase* of the harmonic wave.

4.2.5.1 Frequency-Based Characterization of Elastic Waves

Elastic and sound waves are characterized by their frequency f, see Table 4.1.

Table 4.1: Frequency classification of elastic waves.

Frequency range	Classification
$f < 20\,Hz$	infrasounds
$20\,Hz < f < 20\,kHz$	human audible sound
$20\,kHz < f < 10^9\,Hz$	ultrasounds
$f > 10^9\,Hz$	hypersounds

4.2.6 Energy Propagation and Wave Intensity

As already stated above, the propagation of a field describing a wave is paired with the propagation of energy. To obtain an expression for the energy flux related to harmonic plane wave propagation, we can consider a bar of density ρ, elasticity module E and section S,; this flux propagates at speed

$$v = \sqrt{\frac{E}{\rho}}. \tag{4.39}$$

The longitudinal displacement wave given by

$$s = A\,\sin\left(k\,x - \omega\,t\right). \tag{4.40}$$

The force exerted on a bar element by the leftmost element is the same in eq. (4.18) in Subsection 4.2.3; the power is given by

$$P = \mathbf{F}\cdot\frac{\partial s}{\partial t} = -F\,\frac{\partial s}{\partial t} = -E\,S\,\frac{\partial s}{\partial x}\,\frac{\partial s}{\partial t}. \tag{4.41}$$

By deriving (4.40), employing (4.39) and (4.32), and then averaging the mean power

$$P_m = \frac{1}{2}\,\rho\,\omega^2\,A^2\,v\,S \tag{4.42}$$

is obtained.
If the the mechanic energy volume density

$$w_\tau = \frac{1}{2}\,\rho\,\omega^2\,A^2 \tag{4.43}$$

is defined, we can write that

$$P_m = w_\tau\,v\,S. \tag{4.44}$$

By dividing eq. (4.44) by S, we obtain a quantity depending only on energy density and propagation speed, denoted as *wave intensity*. It represents the mean value of the energy passing through an orthogonal section to the propagation direction per unit of time and per unit of area. The intensity can be expressed as:

$$I = \frac{1}{S}\left(\frac{dU_{\text{mech}}}{dt}\right)_m = \frac{P_m}{S} = w_\tau\,v = \frac{1}{2}\,\rho\,\omega^2\,A^2\,v \tag{4.45}$$

In the examined mechanical waves the energy, and then the energy density, is proportional to ω^2 and to A^2: therefore the intensity of such waves is proportional to the square of the angular frequency (and hence of the frequency) and on the square of the wave amplitude.

The origin of the energy flux characterized by the intensity is in the wave source: there a mechanical work is accomplished the same that then appears as energy transported with the wave.

Integral energy transport is always a limit case: either because of internal friction and because of energy transfer to the environment, the transmitted energy becomes smaller as the wave proceeds. This process is denoted with *absorption*.

4.2.7 Multidimensional Waves

Until now a plane wave has been defined as a wave propagating in a single direction. If the propagation direction coincides with x axis and the wave is harmonic, its wave function is $\xi = \xi_0\,\sin\left(k\,x - \omega\,t\right)$. Its value is constant on every plane orthogonal to the propagation direction: on such planes is therefore constant the phase $\phi = k\,x - \omega\,t$: all the points on that plane are in phase among themselves. When it is said that the wave propagates along x axis, it is not intended that the phenomenon is localized on the points of the x axis: conversely all the points

having the same x coordinate show the same phenomenon. In a bar or in a finite-section gas column, all the points belonging to the same section S orthogonal to x axis oscillate in phase.

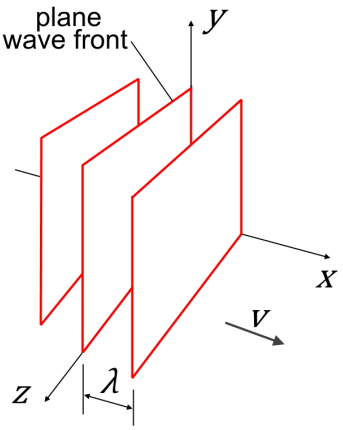


Figure 4.9: Wave fronts of a plane wave.

It is defined *wave front* a surface on which at a given instant the phase is constant, visible in Fig. 4.9. For a plane wave the wave front is a plane or a portion of a plane. The plane wave front propagate with the propagation speed of the wave v, covering a distance equal to the wavelength λ in a time equal to the period T. If two consecutive wave fronts, by having among them a phase difference of 2π, are distant from each other λ.

To characterize the propagation direction of a plane wave, depending on the source shape, but not on the coordinate system, the propagation vector $\mathbf{k}$ is introduced. As shown in Fig. 4.10, it has modulus $k = \frac{2\pi}{\lambda}$ and direction and towards equal to those of $\mathbf{v}$. Named $\mathbf{r}$ the vector radius which locates the point P in a given plane wave front, it is given

$$\mathbf{k} \cdot \mathbf{r} = k\, r\, \cos\theta = k\, x, \tag{4.46}$$

and the wave function can be written as

$$\xi = \xi_0\, \sin\left(\mathbf{k} \cdot \mathbf{r} - \omega\, t\right). \tag{4.47}$$

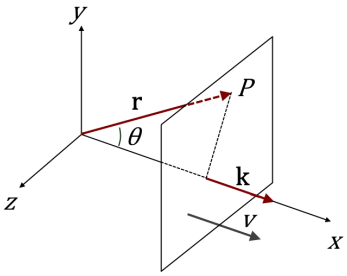


Figure 4.10: Propagation vector $\mathbf{k}$.

The invariance of the scalar product $\mathbf{k}\cdot\mathbf{r}$ ensures that eq. (4.47) constitutes the general expression of an harmonic plane wave, independently on the chosen coordinate system for analytical description. The propagation direction is that of vector $\mathbf{k}$, the wave fronts are the planes orthogonal to that direction, locus of points in which the phase $\mathbf{k} \cdot \mathbf{r} - \omega\, t$ is constant.

In a generic Cartesian coordinate system

$$\mathbf{k} = k_x\, \mathbf{u}_x + k_y\, \mathbf{u}_y + k_z\, \mathbf{u}_z, \qquad \mathbf{r} = x\, \mathbf{u}_x + y\, \mathbf{u}_y + z\, \mathbf{u}_z, \tag{4.48}$$

eq. (4.47) becomes

$$\xi = \xi_0\, \sin\left(k_x\, x + k_y\, y + k_z\, z - \omega\, t\right), \tag{4.49}$$

with

$$k_x^2 + k_y^2 + k_z^2 = k^2 = \frac{\omega^2}{v^2}. \tag{4.50}$$

In such a coordinate system, the *general equation of plane waves* is

$$\frac{\partial^2 \xi}{\partial x^2} + \frac{\partial^2 \xi}{\partial y^2} + \frac{\partial^2 \xi}{\partial z^2} = \frac{1}{v^2}\, \frac{\partial^2 \xi}{\partial t^2}. \tag{4.51}$$

Through eq. (4.50) it is possible to verify that (4.49) is the solution of eq. (4.51). If the propagation environment is bi-dimensional, for example the plane x, y, eq. (4.51) becomes

$$\frac{\partial^2 \xi}{\partial x^2} + \frac{\partial^2 \xi}{\partial y^2} = \frac{1}{v^2}\, \frac{\partial^2 \xi}{\partial t^2}. \tag{4.52}$$

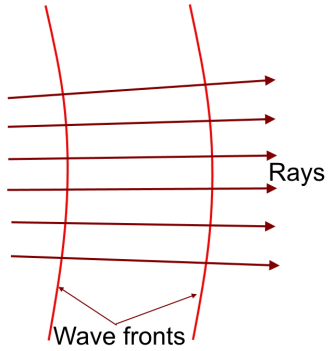


Figure 4.11: Non-plane wave fronts.

In case of plane waves, eqs. (4.51) and (4.52) can be always brought back, by means of a coordinate transformation, to the typical form of a plane wave (4.11).

However, the interest in eqs. (4.51) and (4.52) stays in the fact that they admit solutions different from those of a typical plane wave. In the three-dimensional case the solutions of (4.51) are also waves having spherical or cylindrical wave front: the former are produced by sources having spherical symmetry or point-like, the latter by a source distributed along a line. The wave front can be spherical only if the propagation velocity of the perturbation is the same in all directions; in case of cylindrical symmetry, the velocity must be the same at least in all the orthogonal directions to the axis on which the source is located. Such type of non-plane wave fronts is illustrated in Fig. 4.11. Circular waves in two dimensions are obtained from a point-like source (at the limit) if the propagation speed does not depend on the angle. An isotropic mean certainly satisfies such conditions. The symmetry of the wave front is not necessarily paired with the intensity symmetry.

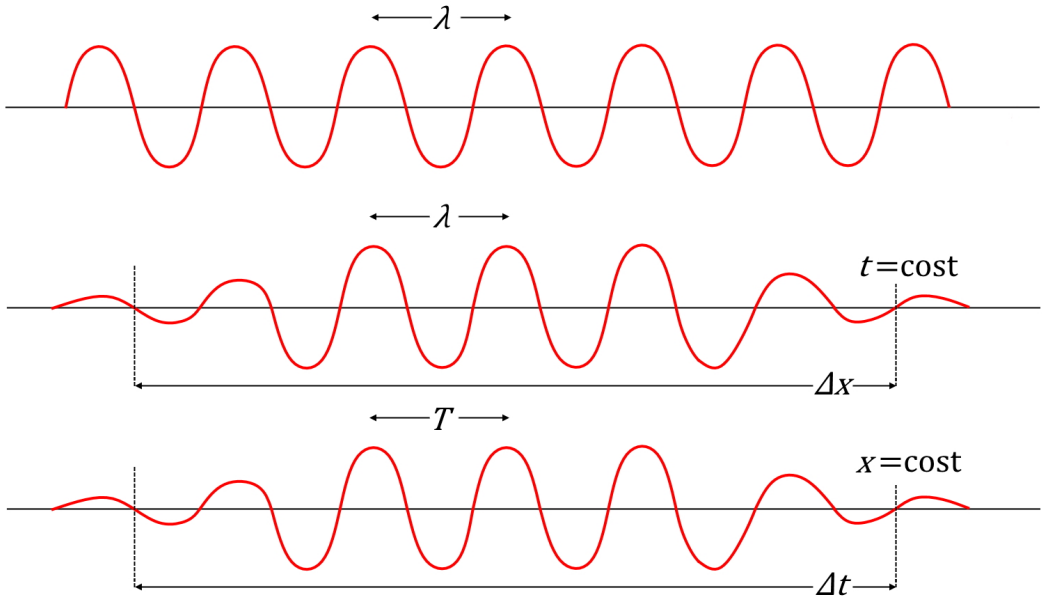


Figure 4.12: Wave packet length and duration.

The notion of plane, spherical, cylindrical, or whatever other shape wave front brings with itself the notion of *radius*, that represents the line orthogonal to the wave front in a given point representing, in that point, the direction of propagation of the wave and of its energy.

4.2.8 Characteristics of Mechanical Plane Waves in Relation to the Mean

The analysis of wave properties evidences that some depend on the source, like frequency, amplitude, and intensity distribution in the various directions, while other depends on the mean, like the propagation speed.

The property of the waves to be longitudinal or transverse depends on the source, and also on the mean.

4.2.8.1 Absorption

The experimental verification in a material mean of the constancy (plane waves) or decrease of the intensity with the laws $1/r^2$ (spherical waves) or $1/r$ (cylindrical waves), shows that the intensity decreases quicker than expected. This can be attributed to the *absorption of energy* carried by the wave by the traversed and other eventual nearby means. Mean absorption is due to internal friction phenomena. Measures bring to the conclusion that, on a thickness dx traversed between x and $x + dx$, the intensity decrease is proportional to the intensity in x and the thickness dx:

$$dI = -\alpha\, I\,(x)\; dx. \tag{4.53}$$

The negative sign appeaering in the last equation is due to the fact that it represents a decrease and α is a characteristic quantity of the mean, named *absorption coefficient*, and characterizes the rapidity of decrease. The distance $x_0 = \frac{1}{\alpha}$ is named *length of absorption*, and measures the distance between two points such that $\frac{I(x_2)}{I(x_1)} = \frac{1}{e} \simeq 0.368$.

Given that the amplitude varies as $\sqrt{I}$, the wave function for an absorbing mean assuming a constant α is:

$$\begin{aligned}
&\text{plane waves} && \xi = \xi_0\, e^{\frac{\alpha\, x}{2}}\, \sin\left(k\, x - \omega\, t\right)\\
&\text{spherical waves} && \xi = \xi_0\, \frac{e^{\frac{\alpha\, r}{2}}}{r}\, \sin\left(k\, r - \omega\, t\right)\\
&\text{cylindrical waves} && \xi = \xi_0\, \frac{e^{\frac{\alpha\, r}{2}}}{\sqrt{r}}\, \sin\left(k\, r - \omega\, t\right).
\end{aligned} \tag{4.54}$$

4.2.9 Wave Packets

Generally speaking, all sources emit waves through processes having finite duration Δt: therefore, in reality a wave always has temporal and spatial finite extensions, even if it has an harmonic behavior. It is defined *wave*

packet or *pulse* a wave having finite duration Δt and length Δx; this is illustrated in Fig. 4.12; Δt and Δx are related by the equation $\Delta x = v\,\Delta t$ if v is the propagation speed of the wave. The number of oscillations (cycles) N contained in the packet is determined by

$$\Delta x = N\,\lambda, \qquad \Delta t = N\,T. \tag{4.55}$$

The wavenumber k and angular frequency ω are expressed as

$$k = \frac{2\pi}{\lambda} = \frac{2\pi\,N}{\Delta x}, \qquad \omega = \frac{2\pi}{T} = \frac{2\pi\,N}{\Delta t}. \tag{4.56}$$

If we admit that N is not completely fixed, but has a given indetermination expressed with $\Delta N = 1$, the relations for k and ω can be derived as:

$$\Delta k = \frac{2\pi}{\Delta x}, \qquad \Delta\omega = \frac{2\pi}{\Delta t} \qquad \Rightarrow \qquad \Delta f = \frac{1}{\Delta t}, \tag{4.57}$$

that can be rewritten as

$$\Delta k\,\Delta x = 2\pi, \qquad \Delta\omega\,\Delta t = 2\pi, \qquad \Delta f \Delta t = 1 \tag{4.58}$$

These relation evidence the substantial difference between harmonic wave and harmonic wave packet. While the former has well-defined wavelength λ and frequency f allowing to describe it completely, in the latter it is present a *band of frequencies* $\Delta f = \frac{1}{\Delta t}$ in the neighborhood of frequency f and an *interval of wavenumbers* $k = \frac{2\pi}{\Delta x}$ in the neighborhood of $k = \frac{2\pi}{\lambda}$. The wider Δx and Δt are, the narrower these bands result (for Δx and Δt tending to infinity, an harmonic wave is obtained). For short packet durations and lengths, wavelengths and frequencies contained are significantly different from λ and f. For motivations related to wave interference phenomena, Δt and Δx are respectively named *coherence time* and *coherence length*.

Relations similar to eq. (4.58) are valid for wave packets of any shape and type if the intervals Δk and $\Delta\omega$ are correctly defined. The physical significance of eq. (4.58) lies in relating temporal and spatial duration to k and f intervals containing substantial contributions from harmonic waves composing the wave packet. Narrower is their length, larger is the spectra of wavenumbers and frequencies necessary to represent the packet.

It is possible to demonstrate that eq. (4.58) actually constitutes a lower limit, and that the most correct forms are:

$$\Delta k\,\Delta x \geq 2\pi, \qquad \Delta\omega\,\Delta t \geq 2\pi, \qquad \Delta f \Delta t \geq 1. \tag{4.59}$$

4.2.10 Phase and Group Velocities

The discussion on wave packets and relations (4.58) in Subsection 4.2.9 is based on the application of the Fourier's integral to a non-periodic function in space and time. Since the function has finite and generally short duration and length, this allows to limit the band of important harmonic components. It is observed that propagation properties does not intervene in the computation.

In this subsection we investigate the consequence of packet propagation in space after being produced by the source, and the effects of the various harmonics. Substantially two possibilities arise:

1. the propagation speed of the perturbation *does not* depend on its frequency;
2. the propagation speed of the perturbation *is function* of its frequency;

4.2.10.1 Non-Dispersive Mean

If the propagation speed does not depend on the propagation frequency, it is said that the propagation happens in a *non-dispersive mean*. In such conditions, all the harmonic components of the wave packet have the same velocity; therefore their sum, the packet, travels with that speed, and its shape does not change. In $\omega = k\,v$, v is a constant which does not depend on ω and neither on k: ω and k are simply proportional.

Examples of non-dispersive propagation are:

- elastic waves in a gases (sound waves);
- elastic waves in a solid bar, in a tight rope, in a tight membrane;
- electromagnetic waves in the vacuum.

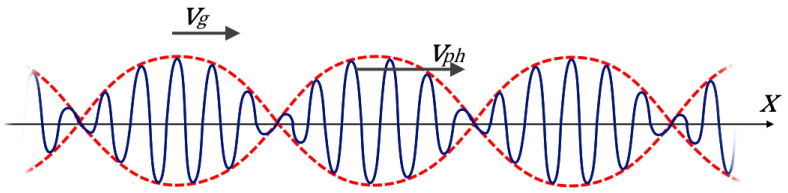


Figure 4.13: Phase and group velocity.

4.2.10.2 Dispersive Mean

When the propagation speed depends on perturbation frequency, ($v(f)$ or $v(\lambda)$), the propagation happens in a *dispersive mean*: a *dispersion phenomenon* occurs. In this case, the packet components, which have different angular frequencies, propagate at different speeds, and the shape of the packet changes with propagation: the waves disperse. The velocity of the packet can not anymore be identified with that of the components; moreover the problem is to determine the propagation speed of the packet. This is essential since the effective phenomena is the packet itself and the energy of the wave propagate with its speed, for example. The packet speed is named *group velocity*, intending the group of waves which form the packets; the speed of the single components is named *phase velocity*.

Example of dispersive propagation:

- electromagnetic waves in material means or in conductive cavities;
- waves on the surface of a liquid.

To exemplify the concept of group velocity, consider the simple case of a packet formed by only two harmonic waves, of equal amplitude A, wavenumbers k_1 and k_2, angular frequencies ω_1 and ω_2:

$$\xi(x,t) = A\,\sin(k_1\,x - \omega_1\,t) + A\,\sin(k_2\,x - \omega_2\,t)\,; \tag{4.60}$$

if we define

$$\Delta k = k_1 - k_2, \qquad k_m = \frac{k_1 + k_2}{2}, \qquad \Delta\omega = \omega_1 - \omega_2, \qquad \omega_m = \frac{\omega_1 + \omega_2}{2} \tag{4.61}$$

and transform $\xi(x,t)$ according to sum trigonometric rules, we obtain

$$\xi(x,t) = 2\,A\,\cos\left(\frac{\Delta k}{2}\,x - \frac{\Delta\omega}{2}\,t\right)\,\sin(k_m\,x - \omega_m\,t)\,. \tag{4.62}$$

A representation of the last function at a given instant is shown in Fig. 4.13.
The wave average phase velocity is:

$$v_{ph} = \frac{\omega_m}{k_m} \tag{4.63}$$

which is almost equal to that of the component waves, $v_1 = \frac{\omega_1}{k_1}$ and $v_2 = \frac{\omega_2}{k_2}$, if $\omega_1 \simeq \omega_2$ and $k_1 \simeq k_2$.

The amplitude of the wave

$$\xi_0 = 2\,A\,\cos\left(\frac{\Delta k}{2}\,x - \frac{\Delta\omega}{2}\,t\right) \tag{4.64}$$

is not constant and presents a wave-like structure $f(x - v\,t)$, with speed

$$v_g = \frac{\Delta\omega}{\Delta k} = \frac{\omega_1 - \omega_2}{k_1 - k_2} = \frac{k_1\,v_1 - k_2\,v_2}{k_1 - k_2}. \tag{4.65}$$

The speed v_g is the propagation speed of the amplitude, and it is the velocity of the packet composed by the two waves: in the considered case it is therefore the group velocity.

Substantially, the relative motion of a wave respect to the other produces the overlap shown in Fig. 4.14. The inside rapidly-oscillating curve changes during motion, but its envelope keeps the same shape propagating with speed v_g.

A real packet emitted by a source has finite duration and contains a high number of components, which can be determined through Fourier analysis. In a dispersive mean the phase velocity of the single components depend on k and $\omega = k\,v$ is still valid, but it is written as

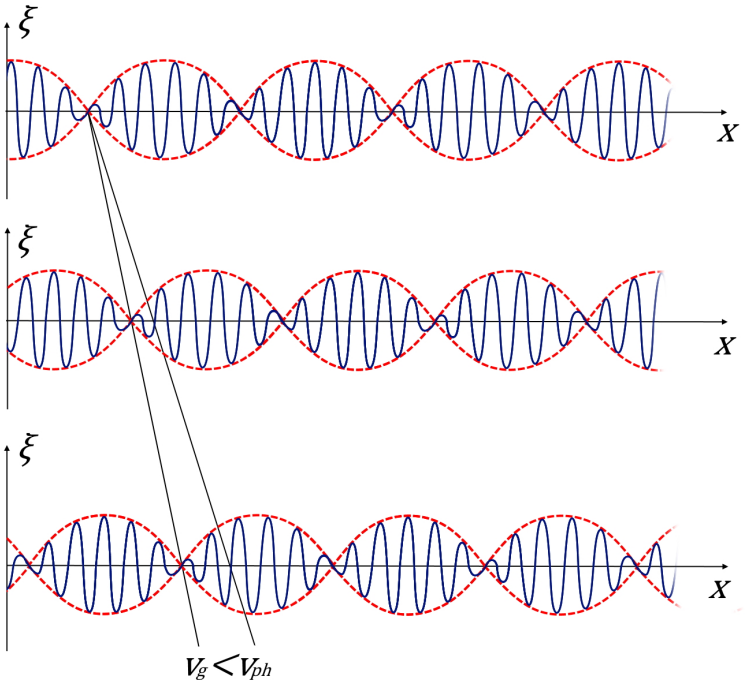


Figure 4.14: Relative motion of a wave respect to another.

$$\omega(k) = k\, v_{ph}(k) \tag{4.66}$$

and it is named *dispersion relation* of the mean under exam. The definition of group velocity is

$$v_g = \frac{d\omega}{dk}. \tag{4.67}$$

By using eq. (4.66) and making use of relations $\frac{dk}{k} = -\frac{d\lambda}{\lambda} = \frac{df}{f}$, it can be shown that:

$$v_g = v_{ph} + k\,\frac{dv_{ph}}{dk} = v_{ph} - \lambda\,\frac{dv_{ph}}{d\lambda} = v_{ph} + f\,\frac{dv_{ph}}{df}. \tag{4.68}$$

The group velocity can be greater or lower than v_{ph} depending on the sign of the derivative of v_{ph} respect to the considered variable. Eq. (4.68) applies also to the case of non-dispersive mean, in which $v_g = v_{ph}$ since $\frac{dv_{ph}}{dk} = 0$.

A consequence of the dependency on k of the phase velocity is that the packet structure can change during propagation. The group velocity itself could depend on k, depending on the shape of the function $v_{ph}(k)$, and hence could be variable within the packet. The existence itself of a band Δk implies a band Δv_g. For a packet of 0 width, after a time t its band is $t\,\Delta v_g$ wide: therefore, in a first approximation it is possible to conclude that a packet, emitted with width Δx the time $t = 0$, at the time t it is $\Delta x + t\,\Delta v_g$ wide.

4.2.11 Doppler Effect

When a wave source S and a wave detector R are in relative motion, the perceived frequency by the detector generally results different from the own frequency of the source: this phenomenon is named *Doppler effect*. The Doppler effect is non-symmetric and it is due to the fact that the wave propagation speed seen by the detector is not the same seen at the source.

4.2.12 Acoustic Impedance, Reflection and Transmission Coefficients

4.2.12.1 Impedance

The concept of *impedance* Z can be also defined for sound waves, similarly as any other transmission mean. For a mechanical mean or system in which an elastic perturbation propagates, the impedance can be defined as the ratio between the force applied in a point F and speed of that point:

$$Z = \frac{F}{v} = \frac{F}{\frac{\partial \xi}{\partial t}}. \tag{4.69}$$

For longitudinal waves the relation between stress and deformation is given by eq. (4.18). Therefore, the exciting force applied in the longitudinal direction to a rod or a bar of section S can be expressed as

$$F = S\,E\,\frac{\partial \xi}{\partial x} = \frac{S\,E}{v}\,\frac{\partial \xi}{\partial t} \tag{4.70}$$

and therefore, by using the definition of mechanical impedance and that of propagation speed[5], we obtain the impedance[6]

$$Z = \frac{S\,E}{v} = \rho\,S\,v = S\sqrt{\rho\,E} \qquad \left[\frac{kg}{s}\right]. \tag{4.71}$$

The characteristic impedance Z_0 is given as the impedance per unit area

$$Z_0 = \frac{Z}{S} = \sqrt{\rho\,E}. \tag{4.72}$$

For transverse waves a similar relation holds; for a tight rope, the impedance is given as

$$Z_s = \sqrt{\rho T}. \tag{4.73}$$

Some observations can be made: the impedance of a material is a constant and depends only on the mechanic characteristics of the material: density, Young modulus or tension. It does not depend on position, instant of time or, for an ideal medium, frequency of the incident solicitation. This is not true for real media.

4.2.12.2 Reflection, Transmission and Modal Conversion

The existence of media having different impedance consequently leads to the concepts of *transmission* and *reflection* when a monochromatic elastic plane wave impinges on their surface of separation. The two elastic media, hereinafter 1 and 2 have impedance $Z_1$and Z_2. The mechanism is the same seen in optics or telecommunications.

When a wave traveling in media 1 impinges with angle θ_1 on the interface with media 2, two waves can be produced:

- the *reflected wave* remains in medium 1; the reflection angle is the opposite of the incident one ($= -\theta_1$);
- the *refracted wave* is transmitted in medium 2 but with a usually different angle θ_2.

Transmission and reflection angles are related by means of the *Snell's law* to the phase speeds of waves in the two media, v_1 and v_2 (referring to medium 1 and 2, respectively):

$$\frac{\sin\theta_2}{\sin\theta_1} = \frac{v_2}{v_1}. \tag{4.74}$$

It is useful to define an a*ngle of total reflection* or *critical angle*: if the incident angle is greater, no refracted wave is produced:

$$\theta_{crit} = \arcsin\left(\frac{v_1}{v_2}\right). \tag{4.75}$$

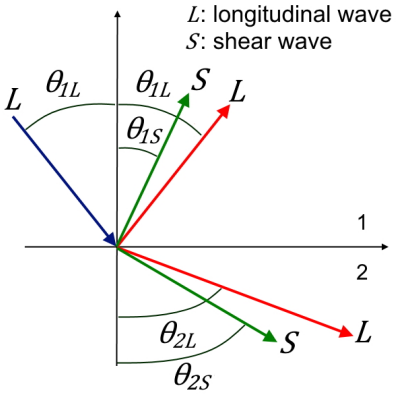


Figure 4.15: Snell's law and mode conversion. Picture from [21].
$v_{1L}\sin\theta_{2L} = v_{2L}\sin\theta_{1L}$
$v_{1L}\sin\theta_{2S} = v_{2S}\sin\theta_{1L}$

The intensity of the impinging wave is distributed between the transmitted and reflected wave; respectively two coefficients, $\mathcal{T}$ and $\mathcal{R}$, quantify the amount of energy that is transferred between the two waves [66]:

$$\mathcal{R} = \left(\frac{Z_1 - Z_2}{Z_1 + Z_2}\right)^2 \qquad \mathcal{T} = \frac{4\,Z_1\,Z}{(Z_1 + Z_2)^2}. \tag{4.76}$$

[5] $v = \sqrt{\frac{E}{\rho}} \quad \Rightarrow \quad E = \rho\,v^2$.

[6] In the *mks* system it is measured in mechanical ohms $\left(\frac{kg}{s}\right)$.

These coefficients are such that, denoting with I the intensity of the impinging wave, I_1 the intensity of the reflected wave and with I_2 the intensity of the transmitted wave, it is achieved:

$$\begin{array}{l} I_1 = \mathcal{R}\, I \\ I_2 = \mathcal{T}\, I \end{array} \qquad \text{with} \qquad \begin{array}{l} I_1 + I_2 = I \\ \mathcal{R} + \mathcal{T} = 1. \end{array} \tag{4.77}$$

Modal conversion arises since the propagation speed in the two media is different, and since the propagation speed of longitudinal and transverse elastic waves is different. Because of modal conversion, the incident wave (being also only longitudinal or shear) emerges as a mixed wave (containing both longitudinal and shear components), see Fig. 4.15. The wave type does not vary (for symmetry reasons) only if the incident angle is null or, in case of a random incident angle, if the resulting wave is transverse and has vibrations parallel to the separation surface [65].

Chapter 5

Ultrasonic Guided Waves

This chapter is mainly based on ref. [21] with contributions from other authors when needed.

5.1 Types of Ultrasonic Waves

As seen in Subsection 4.2.5.1, an elastic wave is defined ultrasound or ultrasonic wave when its frequency is comprised between 20 kHz and 1 GHz. There are two types of ultrasonic waves [50]:

1. *bulk (fundamental) waves*;
2. *guided waves*.

Bulk propagation refers to waves that encounter no boundaries, like waves propagating in an infinite media or in medias in which boundaries does not affect wave propagation; guided waves, instead, relies on propagation in bounded media, like plates, rods or tubes.

A scheme illustrating the various types of guided waves is depicted in Fig. (5.1).

5.1.1 Comparison Between Bulk and Guided Waves

In Fig. 5.2 bulk and guided waves inspection methods are introduced. From the figure, it is possible to note that the coverage area of guided waves is massively increased respect to the local region covered by bulk waves. The whole thickness of the specimen can be inspected for a wide span from a single probe position: to cover the same area, the traditional bulk wave transducer needs to be moved around the whole surface of the object.

The angle beam method could be used in bulk wave evaluation with waves reflecting inside the boundaries of the structure: the adopted frequency influences the generation of a bulk or guided wave. It is generated a guided wave when the wavelength of the wave is greater than the structural thickness of the object used as a waveguide. To obtain a long wavelength, a low frequency must be used. For comb transducer inspection, the probe spacing is the wavelength λ of the signal used as excitation.

Guided waves propagation is strongly dependent on the excitation frequency and the impinging angle of waves on the boundaries of the structure under exam: therefore, a complex wave interference phenomenon occurs.

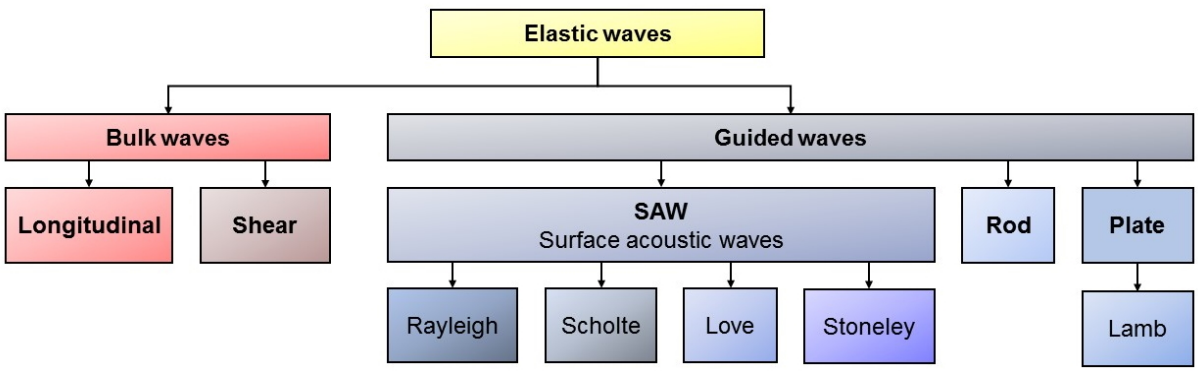


Figure 5.1: Classification of guided waves.

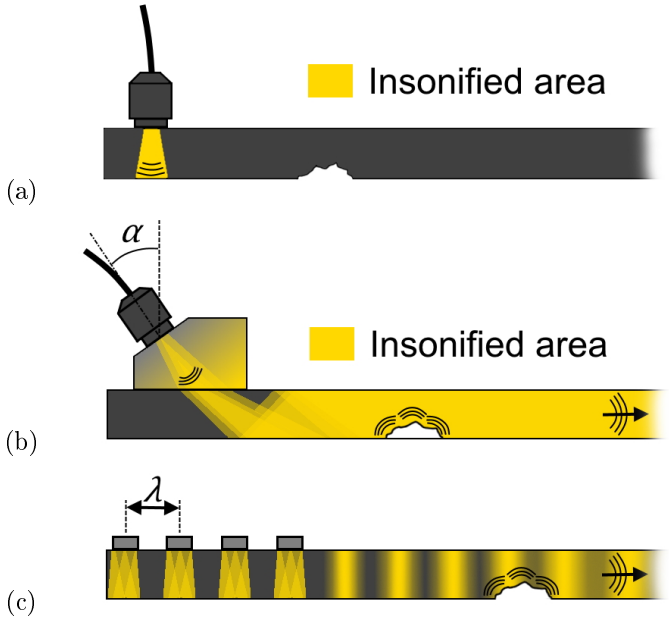


Figure 5.2: Comparison of traditional ultrasonic bulk wave inspection method having normal-beam excitation (a) with two different guided wave inspection methods: one (b) using angle-beam excitation, and the other (c) deploying comb excitation.

Table 5.1: General comparison between bulk and guided waves characteristics [21].

	Bulk waves	**Guided waves**
Phase velocities	Constant	Function of frequency
Group velocities	Same as phase velocities	Generally not equal to phase velocities
Pulse shape	Non dispersive	Generally dispersive

The advantages of ultrasonic guided waves analysis techniques are:

- capacity to inspect over long distances from a single probe position without the need to scan the whole object;
- greater sensitivity;
- capacity to inspect hidden, or buried, or underwater, or encapsulated structures;
- cost-effectiveness thanks to simple and rapid inspection.

In Table 5.1 a comparison between bulk and guided waves characteristics is given.

The type of mean matters when speaking of wave propagation:

- an *isotropic* material has properties which are independent on direction;
- *anisotropic* materials properties are dependent on the direction: it is the case of composite materials.

In Table 5.2 the key differences between anisotropic and isotropic media are given by having in mind guided waves propagation [21].

5.2 Ultrasonic Guided Waves

To allow guided wave propagation, boundaries are required: boundary conditions therefore must be satisfied. Even the surface of a very thick structure is a boundary: the structure can be considered as an half-space or a semi-infinite media. In this case, Rayleigh surface waves can propagate over the surface of such thick structure if their wavelength is very small compared to the thickness of the structure. Guided waves can also propagate in many different kinds of waveguides, such as plates, rods, tubes, multilayered structures, etc.. In this case, *interference phenomena* arise from the waves bouncing back and forth inside the waveguide itself.

Table 5.2: Differences between isotropic and anisotropic media in relation to guided waves [21].

	Isotropic	**Anisotropic**
Wave velocities	Not function of launch direction	Function of launch direction
Skew angles	No	Yes

As bouncing happens, *mode conversion* occurs: each time a wave impinges on an interface, reflection and refraction of shear and/or longitudinal waves happens. Depending on the angle and frequency, constructive, destructive or intermediate interference takes place, leading to hundreds of solutions of constructive interference points, leading to guided wave packets which travel in the structure [51], the *modes*. A whole set of incident angles and frequencies that have led to a set of interference points could represent the solutions to the guided wave problem of that waveguide. These constructive interference points can be plotted to produce a wave velocity *dispersion curve* relating phase velocity to frequency. This is possible since incident angles and phase velocity are related by the Snell's law [21]. Each waveguide has its own unique set of phase velocity dispersion curves; every curve in the set is related to a mode of propagation. Dispersion curves are used to display all the kinds of waves and modes that can propagate inside the waveguide, and are useful to design and analyze an ultrasound guided waves NDT or SHM system [51].

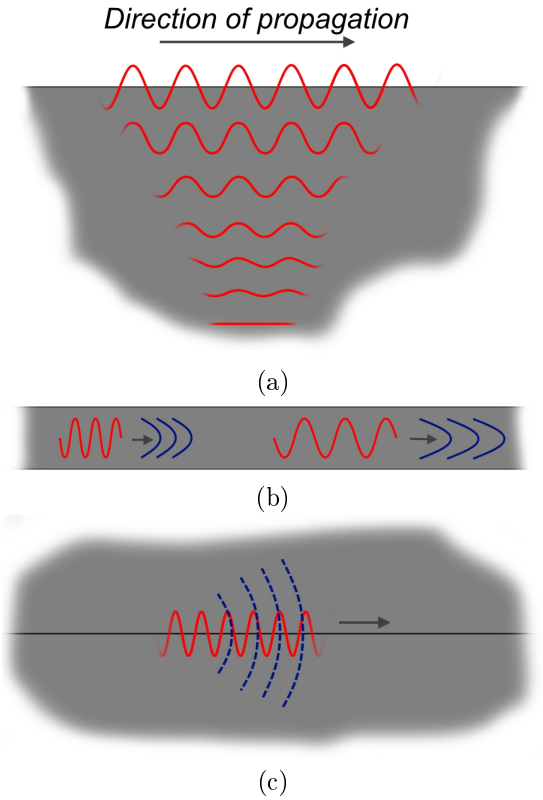


Figure 5.3: Schematic representation of wave in plates: (a) Rayleigh (surface) waves; (b) Lamb waves; (c) Stonely waves.

To solve a guided wave problem it is useful to search for a variable having the appropriate sensitivity in the given problem to solve. By selecting a particular point in the dispersion curve it will be possible to design the ultrasonic transducer capable to excite that point. The problem arises because of the existence of phase velocity and frequency spectra: as frequency is swept across the frequency axis, multiple modes and oscillating frequencies will be generated, making difficult to obtain a precise excitation [21]. Another important feature to consider is the wave structure of the selected mode (*mode shape*): it shows its in-plane displacement, out-of-plane displacement, or actual stress distribution, which usually varies (often quite significantly) across the thickness of the structure. This knowledge is useful to establish the maximum penetration power for structures, or to establish the maximum sensitivity to a defect located in a specified area of the test specimen. The ability to select a precise dispersion curve at specific phase velocity and frequency affects the penetration power and sensitivity of a guided waves ultrasound system [51].

When an ultrasonic wave propagates through a medium, ultrasonic attenuation is caused by a loss of energy in the ultrasonic wave (and other reasons). Some factors affecting the amplitude and waveform of the ultrasonic wave are the ultrasonic beam spreading, energy absorption, dispersion, nonlinearity, transmission at interfaces, scattering by inclusions and defects, Doppler effect, etc. [50]. Attenuation is a factor to keep in mind when designing ultrasonic guided waves systems, especially if they must operate in the long range.

5.3 Waves in Plates

A common example to illustrate the analytical procedures adopted to solve the problem of Lamb wave propagation in solid materials is that involving a traction-free, homogeneous, and isotropic plate. The same procedures can be used to achieve governing equations and dispersion curves (phase velocity vs frequency) of a countless number of guided wave problems involving bars, tubes, multiple layers, and anisotropic media. An alternative technique to obtain dispersion curves is constituted by *finite elements* (FE) simulations; of particular relief when the cross-section of the waveguide is constant along the propagation direction is the *semi-analitical finite element* (SAFE) method, presented in Chapter 6.

Bulk and guided waves are fundamentally different, but both are governed by the same set of partial differential wave equations. The main difference, mathematically, is that guided waves need to satisfy also physical boundary conditions; the trouble is that these make analytical solution difficult to solve, in some cases impossible. In guided wave propagation, generally, an infinite number of modes can be associated to a single wave problem, conversely to the finite number found in bulk propagation (primarily longitudinal shear, shear perhaps being horizontal or vertical). A finite body can therefore support an infinite number of different guided wave modes.

Some solved guided wave problems are those related to Rayleigh, Lamb, and Stonely waves.

- *Rayleigh waves* are free waves on the surface of a semi-infinite solid. On the boundary, traction force has to vanish, and the waves must decay with depth. See Fig. 5.3(a).
- *Lamb waves* are waves of plane strain occurring in a free plate. At the upper and lower surface of the plate the traction force has to vanish; by varying entry angle and frequency different mode structures occur, from point to point. See Fig. 5.3(b).
- *Stonely waves* are free waves that occurring at the interface between two media: here, continuity of traction and displacement must be ensured, and a radiation condition needs to be satisfied. See Fig. 5.3(c).

5.3.1 Governing and Boundary Equations

Recalling from the theory of elasticity that, given u the three-dimensional displacement vector:

- the 3 equations of motion are given by:

$$\sigma_{ij,j} + \rho\, f_i = \rho \ddot{u}_i \tag{5.1}$$

  in which $i, j = 1, 2, 3$ is the axis index, σ is the stress, ρ the material density, and f is the force component on the given axis;

- the 6 independent strain displacement equations are:

$$\varepsilon_{ij} = \frac{1}{2}\left(u_{i,j} + u_{j,i}\right) \tag{5.2}$$

- the 6 independent constitutive equations for isotropic materials are:

$$\sigma_{ij} = \lambda\, \varepsilon_{kk}\, \delta_{ij} + 2\,\mu\, \varepsilon_{ij} \tag{5.3}$$

  in which λ is the Lamè constant, μ the shear modulus, and δ is the Kronecker delta.

The first two equations, (5.1) and (5.2), are valid for any continuous medium; the medium type is specified through eq. (5.3).

By eliminating stress and strain factors from these equations, it is obtained:

$$\mu\, u_{i,jj} + (\lambda + \mu)\; u_{j,ji} + \rho\, f_i = \rho \ddot{u}_i \tag{5.4}$$

which contain only particle displacement and the governing partial differential equation of the displacement. These equations does not account for boundary conditions, which are needed to have a well-posed problem if the domain is finite. Boundary conditions take the form of prescribed tractions and/or displacements on the boundaries of the domain of interest. The general forms of such boundary conditions are:

- on surface displacements u:

$$u\,(x,t) = u_0\,(x,t)\,; \tag{5.5}$$

- on surface tractions t:

$$t_i = \sigma_{ji}\, n_j \tag{5.6}$$

  in which n_j is j-th component of the normal to the surface;

- on a mixed boundary condition:

$$\begin{cases} u(x,t) = u_0(x,t) & \text{on surface } S_1 \\ t_i = \sigma_{ji}\, n_j & \text{on surface } S_2 \end{cases} \tag{5.7}$$

in which S_1 and S_2 are the two surfaces of the plate.

5.3.2 Free Plate Problem

The plate geometry and notation used hereinafter is depicted in Fig. 5.4. In addition to governing equations (5.4), the boundary conditions for such problem are of type of eq. (5.6): the surfaces at coordinates $x_3 = \frac{d}{2} = h$ and $x_3 = -\frac{d}{2} = -h$ are considered traction-free. The ultrasonic excitation is given in a point in of the plate: when the excitation energy impinges on the upper and lower plate surfaces mode conversion happens. The various reflected wave propagate and impinge on the boundaries over and over: from superposition, and hence interference, wave packets arise: these are commonly known as *guided wave modes* in the plate. It is possible to predict how many different modes can be produced in the plate by knowing the entry angle and the frequency of the excitation.

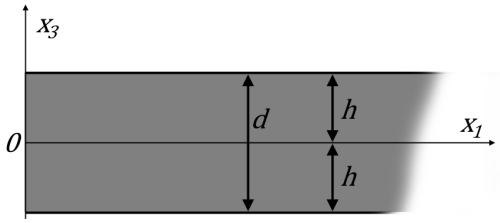


Figure 5.4: Geometry and notation of the free plate problem.

By decomposing the displacement field vector according to Helmholtz, and by substituting the result into eq. (5.4), two uncoupled wave equations are achieved. By assuming plane strain:

- governing longitudinal waves:

$$\frac{\partial^2 \phi}{\partial x_1^2} + \frac{\partial^2 \phi}{\partial x_3^2} = \frac{1}{v_L^2}\frac{\partial^2 \phi}{\partial t^2}; \tag{5.8}$$

- governing shear waves:

$$\frac{\partial^2 \psi}{\partial x_1^2} + \frac{\partial^2 \psi}{\partial x_3^2} = \frac{1}{v_S^2}\frac{\partial^2 \psi}{\partial t^2}. \tag{5.9}$$

The assumption of plane strain is not the most general for such problem, but enables to simplify the analysis. As a result of this assumption, it is possible to express the displacements and stresses in terms of potentials:

$$u_1 = u = \frac{\partial \phi}{\partial x_1} + \frac{\partial \psi}{\partial x_3} \tag{5.10}$$

$$u_2 = v = 0 \tag{5.11}$$

$$u_3 = w = \frac{\partial \phi}{\partial x_3} + \frac{\partial \psi}{\partial x_1} \tag{5.12}$$

$$\sigma_{31} = \mu\left(\frac{\partial u_3}{\partial x_1} + \frac{\partial u_1}{\partial x_3}\right) = \mu\left(\frac{2\,\partial^2 \phi}{\partial x_1\,\partial x_3} - \frac{\partial^2 \psi}{\partial x_1^2} + \frac{\partial^2 \psi}{\partial x_3^2}\right) \tag{5.13}$$

$$\sigma_{33} = \lambda\left(\frac{\partial u_1}{\partial x_1} + \frac{\partial u_3}{\partial x_3}\right) + 2\,\mu\,\frac{\partial u_3}{\partial x_3} = \lambda\left(\frac{\partial^2 \phi}{\partial x_1^2} + \frac{\partial^2 \phi}{\partial x_3^2}\right) + 2\,\mu\left(\frac{\partial^2 \phi}{\partial x_3^2} - \frac{\partial^2 \psi}{\partial x_1\,\partial x_3}\right) \tag{5.14}$$

where λ and μ are Lamè constants.

The analysis can be carried on by assuming that eqs. (5.8) and (5.9) have infinite plane harmonic wave solutions to in the form:

$$\phi = \Phi(x_3)\, e^{j(k\,x_1 - \omega\,t)} \tag{5.15}$$

$$\psi = \Psi(x_3)\, e^{j(k\,x_1 - \omega\,t)}. \tag{5.16}$$

These solutions represent traveling waves in direction x_1 and standing wave in direction x_3. This phenomenon is named *transverse resonance* and can be exploited to solve the problem. The fixed distributions along the transverse direction is yet unknown.

By substituting the solutions ϕ and ψ into eqs. (5.8) and (5.9) yields to equations governing the unknown functions ϕ and ψ, which the solution is:

$$\phi = \Phi(x_3) = A_1 \sin(p\,x_3) + A_2 \cos(p\,x_3) \tag{5.17}$$
$$\psi = \Psi(x_3) = B_1 \sin(q\,x_3) + B_2 \cos(q\,x_3) \tag{5.18}$$

where

$$p^2 = \frac{\omega^2}{v_{L^2}} - k^2 \qquad \text{and} \qquad q^2 = \frac{\omega^2}{v_{S^2}} - k^2 \tag{5.19}$$

Having these solutions, from eqs. (5.10) to (5.14) it is possible to obtain displacement and stress equations. To simplify, by omitting the exponential term $e^{j(k\,x_1 - \omega\,t)}$, the results are:

$$u_1 = \left[j\,k\,\Phi + \frac{d\Psi}{dx_3} \right] \tag{5.20}$$
$$u_3 = \left[\frac{d\Phi}{dx_3} + j\,k\,\Psi \right] \tag{5.21}$$
$$\sigma_{31} = \mu \left(2\,j\,k\,\frac{d\Phi}{dx_3} + k^2\,\Psi + \frac{d^2\Psi}{dx_3^2} \right) \tag{5.22}$$
$$\sigma_{33} = \left[\lambda \left(-k^2\,\Phi + \frac{d^2\Phi}{dx_3^2} \right) + 2\,\mu \left(\frac{d^2\Phi}{dx_3^2} - j\,k\,\frac{d\Psi}{dx_3} \right) \right] \tag{5.23}$$

Since the field variables involve sines (respectively cosines) of argument x_3, which are odd (respectively even) functions about $x_3 = 0$, it is possible to split the solution int two sets of modes:

- *symmetric modes;*
- *antisymmetric modes.*

The displacement motion along the x_1 direction will be symmetric respect to the plane mid-plane if u_1 contains cosines; it will be antisymmetric if u_1 contains sines. For displacements in the x_3 direction, the reverse is true.

- Symmetric modes:

$$\Phi = A_2 \cos(p\,x_3) \tag{5.24}$$
$$\Psi = B_1 \sin(q\,x_3) \tag{5.25}$$
$$u = u_1 = -j\,k\,A_2 \cos(p\,x_3) + q\,B_1 \cos(q\,x_3) \tag{5.26}$$
$$w = u_3 = -p\,A_2 \sin(p\,x_3) - j\,k\,B_1 \sin(q\,x_3) \tag{5.27}$$
$$\sigma_{31} = \mu \left[-2\,j\,k\,p\,A_2 \sin(p\,x_3) + \left(k^2 - q^2\right) B_1 \sin(q\,x_3) \right] \tag{5.28}$$
$$\sigma_{33} = -\lambda \left(k^2 + p^2\right) A_2 \cos(p\,x_3) - 2\,\mu \left[p^2\,A_2 \cos(p\,x_3) + j\,k\,q\,B_1 \cos(q\,x_3) \right] \tag{5.29}$$

  for these modes the wave structure across the plate thickness is symmetric for u and antisymmetric for w.

- Antisymmetric modes:

$$\Phi = A_1 \sin(p\,x_3) \tag{5.30}$$
$$\Psi = B_2 \cos(q\,x_3) \tag{5.31}$$
$$u = u_1 = j\,k\,A_1 \sin(p\,x_3) - q\,B_2 \sin(q\,x_3) \tag{5.32}$$
$$w = u_3 = p\,A_1 \cos(p\,x_3) - j\,k\,B_2 \cos(q\,x_3) \tag{5.33}$$
$$\sigma_{31} = \mu \left[\frac{2\,j\,k\,p\,A_1 \cos(p\,x_3)}{\left(k^2 - q^2\right) B_2 \cos(q\,x_3)} \right] \tag{5.34}$$
$$\sigma_{33} = -\lambda \left(k^2 + p^2\right) A_1 \sin(p\,x_3) - 2\,\mu \left[p^2\,A_1 \sin(p\,x_3) - j\,k\,q\,B_2 \sin(q\,x_3) \right] \tag{5.35}$$

  for these modes the wave structure across the plate thickness is symmetric for w and antisymmetric for u.

The separation between symmetric and antisymmetric waves is not general: it is only possible when the symmetry of the waveguide allows propagation to happen along the symmetry axis.

The application of the boundary conditions allows to calculate the value of A_1, A_2, B_1, and B_2 constants and the dispersion equations, that are still unknown. The traction-free boundary conditions imply that, in case of plane strain:

$$\sigma_{31} = \sigma_{33} = 0 \qquad \text{at} \qquad x_3 = \pm\frac{d}{2} = \pm h. \tag{5.36}$$

The type of mode (i.e. symmetric or not) influences the resulting displacement, stress and strain fields. By applying the boundary conditions, a homogeneous system in two equations for the appropriate two constants is achieved. The two constants are:

- A_2 and B_1 for symmetric modes;
- A_1 and B_2 for antisymmetric modes.

To avoid the trivial solutions, it is required that the determinant of the matrix of coefficients of the homogeneous equations vanishes. From eq. (5.36) it is obtained:

$$\frac{\left(k^2 - q^2\right)\sin\left(q\,h\right)}{2\,j\,k\,p\,\sin\left(p\,h\right)} = \frac{-2\,\mu\,j\,k\,q\,\cos\left(q\,h\right)}{\left(\lambda\,k^2 + \lambda\,p^2 + 2\,\mu\,p^2\right)\cos\left(p\,h\right)} \tag{5.37}$$

and after some manipulation:

$$\frac{\tan\left(q\,h\right)}{\tan\left(p\,h\right)} = \frac{4\,k^2\,q\,p\,\mu}{\left(\lambda\,k^2 + \lambda\,p^2 + 2\,\mu\,p^2\right)\left(k^2 - q^2\right)}. \tag{5.38}$$

Since, from the definition of v_L and of v_S it is given that:

$$\rho\,v_S^2\left[\left(\frac{\omega}{v_S}\right)^2 - 2\,k^2\right] = \rho\,v_S^2\left(q^2 - k^2\right) = \mu\left(q^2 - k^2\right), \tag{5.39}$$

by substituting this into eq. (5.38), some transcendental equations are obtained, the *Rayleigh–Lamb frequency relations*:

- for symmetric modes:

$$\frac{\tan\left(q\,h\right)}{\tan\left(p\,h\right)} = -\frac{4\,k^2\,p\,q}{\left(q^2 - k^2\right)^2} \tag{5.40}$$

- for antisymmetric modes:

$$\frac{\tan\left(q\,h\right)}{\tan\left(p\,h\right)} = -\frac{\left(q^2 - k^2\right)^2}{4\,k^2\,p\,q} \tag{5.41}$$

Equations of this nature are known as *dispersion relations*. They can be used to determine the velocity (or velocities) of propagation of a wave at a given frequency into the plate: the product frequency-thickness (fd or $f\,h$) assumes here particular importance. Although the equations look simple, they can be solved only by numerical methods.

5.3.2.1 Numerical Solution of Rayleigh-Lamb Frequency Equations

The wavenumber k relates the circular frequency ω (and hence the wavelength λ) to the phase velocity of the Lamb wave mode v_{ph}, $k = \frac{\omega}{v_{ph}}$, $v_{ph} = \lambda\frac{\omega}{2\,\pi}$. Equations (5.40) and (5.41) can be considered as relating

- the circular frequency ω to the wavenumber k of the Lamb wave modes, resulting in the *frequency spectrum*, or
- the phase velocity v_{ph} to the frequency ω , resulting in the *dispersion curves*.

For any given frequency, an infinite number of wavenumbers k will satisfy eqs. (5.40) and (5.41); of these, a finite number will be real or purely imaginary, while infinitely many will be complex.

Some particular insight can be achieved by considering various regions of the Rayleigh–Lamb equations for k compared with $\frac{\omega}{v_L}$ or $\frac{\omega}{v_S}$ [72]:

- Region 1 is for $k > \frac{\omega}{v_S}$; here $v_{ph} < v_S$, and therefore:

$$\frac{\tanh(q'\,h)}{\tanh(p'\,h)} = \left\{ \frac{4\,p'\,q'\,k^2}{\left(k^2 - q'^2\right)^2} \right\}^{\pm 1}; \tag{5.42}$$

  $p = j\,p'$, $q = j\,q'$, $p'^2 = -p^2$, and $q'^2 = -q^2$; the exponent $+1$ is for symmetric and -1 is for antisymmetric modes.

- Region 2 is for $\frac{\omega}{v_S} > k > \frac{\omega}{v_L}$; here $v_S < v_{ph} < v_L$ and therefore

$$\frac{\tan(q'\,h)}{\tanh(p'\,h)} = \pm \left\{ \frac{4\,p'\,q\,k^2}{\left(k^2 - q^2\right)^2} \right\}^{\pm 1}; \tag{5.43}$$

- Region 3 is for $k < \frac{\omega}{v_L}$; here $v_{ph} > v_L$ and eqs. (5.40) and (5.41) are unaltered.

The time-harmonic factor $e^{j(k\,x - \omega\,t)}$ is a key element to understand the physical behavior of dispersion curves. Since eqs. (5.40) and (5.41) are function of k and ω, and since k might assume complex values, this can influence the dispersion relations.

In the wavenumber k it is possible to distinguish its real and imaginary parts, $k = k_{Re} + j\,k_{Im}$; therefore, the time-harmonic factor becomes $e^{j(k\,x - \omega\,t)} = e^{j(k_{Re}\,x - \omega\,t)} e^{-k_{Im}\,x}$. There are three possible signed values for k_{Im}, each one having an influence on the wave characteristics and hence a physical interpretation:

- $k_{Im} < 0$, the waves amplitude grows exponentially with distance: they have not been physically observed;
- $k_{Im} = 0$, the waves propagate with no damping;
- $k_{Im} > 0$, the waves amplitude decay exponentially with distance: *evanescent waves*, they tend to disappear.

It can be concluded that, to supply information about propagating waves in the simple unloaded plate problem, only real values for k are needed. Therefore, being only interested in the real solutions of these equations, representing the undamped propagating modes of the structure, eqs. (5.40) and (5.41) can be rewritten so that they take on only real values for real or pure imaginary wavenumbers k:

- for symmetric modes:

$$\frac{\tan(q\,h)}{q} + \frac{4\,k^2\,p\,\tan(p\,h)}{\left(q^2 - k^2\right)^2} = 0 \tag{5.44}$$

- for antisymmetric modes:

$$q\,\tan(q\,h) + \frac{\left(q^2 - k^2\right)^2 \tan(p\,h)}{4\,k^2\,p} = 0 \tag{5.45}$$

The numerical solution of eqs. (5.44) and (5.45) is now relatively simple and described in ref. [21].

In Fig. 5.5, sample dispersion curves for an aluminum plate are presented, depicting modes A0, S0, A1, S1, A2, S2, etc. (A stays for antisymmetric and S stays for symmetric). Note in the figure the limiting value for the phase velocity $c_p = v_{ph}$ as fd value becomes large. Moreover it is also possible to note that

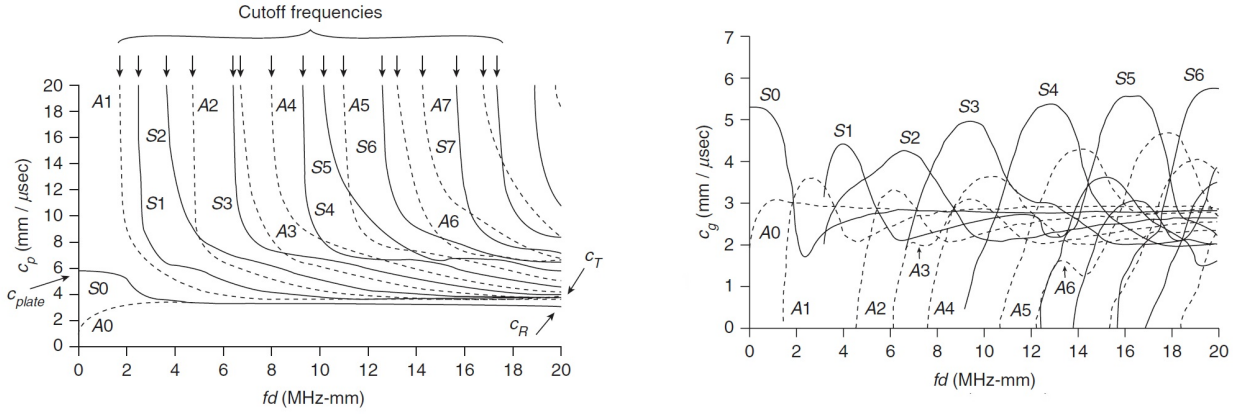
- A0 and S0 modes converge to the Rayleigh surface velocity c_R;
- the other modes converge to the shear (transverse) velocity $c_T = v_S$.

5.3.2.2 Group Velocity

From the phase velocity v_{ph} it is possible to obtain the group velocity v_g by remembering eq. (4.67), $v_g = \frac{d\omega}{dk}$, by substituting $k = \frac{\omega}{v_{ph}}$ in that equation yields

$$v_g = d\omega \left[d\left(\frac{\omega}{v_{ph}}\right) \right]^{-1} = d\omega \left[\frac{d\omega}{v_{ph}} - \omega \frac{dv_{ph}}{v_{ph}^2} \right]^{-1} = v_{ph}^2 \left[v_{ph} - \omega \frac{dv_{ph}}{d\omega} \right]^{-1} \tag{5.46}$$

and by using $\omega = 2\,\pi\,f$:



(a) (b)

Figure 5.5: Dispersion curves for a traction-free aluminum plate. Picture taken from [21], with original notation. (a) phase velocity $c_p = v_{ph}$ vs frequency-thickness product fd; (b) group velocity c_g vs frequency-thickness product fd. $c_{plate} = E^{1/2}\left[\rho\left(1-\nu^2\right)\right]^{-1/2}$ is the plate velocity; c_R is the Rayleigh surface velocity, and the shear velocity is $c_T = v_S$.

$$v_g = v_{ph}^2\left[v_{ph} - (fd)\,\frac{dv_{ph}}{d\,(fd)}\right]^{-1} \tag{5.47}$$

where fd is the product between frequency and thickness.

On the basis of the derivative $\frac{dv_{ph}}{d(fd)}$, some observations can be made:

- if $\frac{dv_{ph}}{d(fd)} = 0$, $v_g = v_{ph}$;
- if $\frac{dv_{ph}}{d(fd)} \to \infty$ (i.e. at cutoff), $v_g \to 0$.

5.3.2.3 Wave Structure Analysis

In the discussion of the free plate problem, it has been stated that the fd product assumes a particular importance: as it varies, it influences the modes wave structure.

- The symmetric mode is not a simply in-plane vibration mode: by moving along the mode, it is changed the ratio between in-plane and out-of-plane displacement. Noteworthy are the changes on the outside surface of the structure.
- The antisymmetric modes is not a mode with only out-of-plane displacement values.

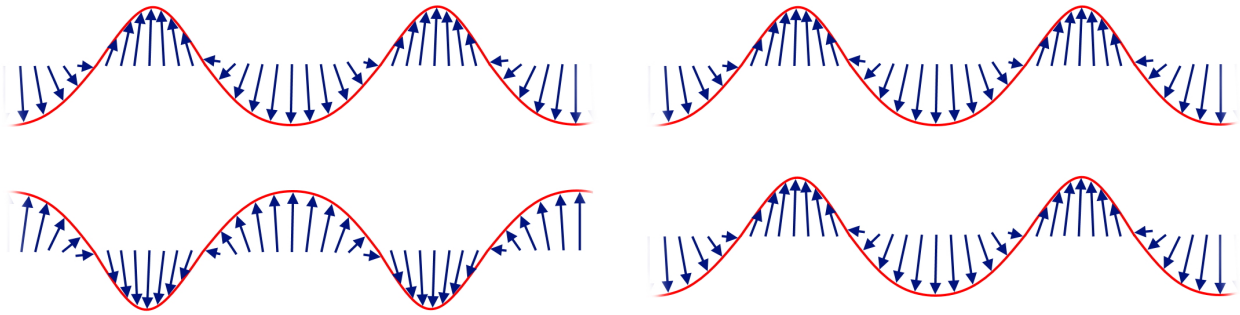
The exploitation of wave structure can increase the wave penetration power along a waveguide. Improved sensitivity to certain defects can be obtained as a result of controlling in-plane or out-of-plane impingement at a certain location across the thickness of the structure.

5.3.2.4 Compressional and Flexural Waves

In Figs. 5.6 and 5.7 are shown highly exaggerated schematic "snapshots" of the displacement vector field distribution on the surface of a plate representing the shape of the modes.

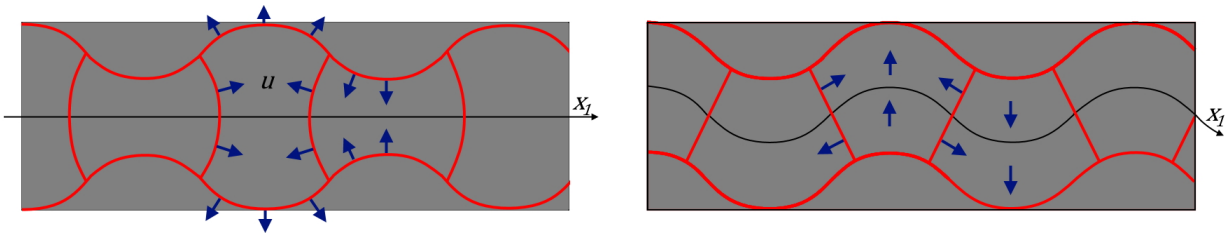
- Symmetric mode waves are defined *compressional*;
- Antisymmetric mode waves are termed *flexural*.

The particle displacements shown are the vector sums of the in-plane u and out-of-plane w particle displacement components.



(a) (b)

Figure 5.6: Compressional and flexural wave particle displacement schematics. Compressional waves (a) are found in symmetric plates; flexural waves (b) are found in antisymmetric modes.



(a) (b)

Figure 5.7: Symmetric (a) and antisymmetric (b) modes cross-section.

5.3.2.5 Lamb Wave Propagation Problem Results Interpretation

The interpretation of the outcomes of the wave propagation problem can be challenging. Even if the theoretical approach has been carefully formulated and a detailed numerical solution has been developed, a great deal of study and attention is required when it comes to use the obtained curves in a practical sense. Mode selection, isolation and control are fundamental aspects, as well as the subject of generation and reception of specific modes and frequencies chosen for a particular problem.

Chapter 6

Semi Analytical Finite Element (SAFE) Method

6.1 Introduction

The *semi-analytical finite element* method (SAFE) is a popular approach to solve wave propagation problems in waveguides, alternative to more traditional methods. The waveguide cross-section is discretized and in the wave propagation direction is adopted an analytical solution. A system of linear equations can be constructed on the basis of a variational scheme; the unknowns are constituted by the frequency and the wavenumber. SAFE can be used to solve problems of wave propagation in waveguides having complex cross-section, in which analytical solution is difficult or impossible to obtain [21]. It allows to obtain velocity, attenuation, mode shapes of the modes of the given waveguide [67]. SAFE methods is superior to *finite elements* (FE) methods when analyzing waveguides that are infinitely long in one dimension [21] and have an arbitrary but constant cross-section. This enables to analytically represent in an exact way only one or two dimensions of the waveguide [68], i.e. only a 2-D mesh of the waveguide cross-section is required, and this allows to reduce computational cost [21]. The wave propagation along the waveguide is described by a complex exponential function, while across the cross-section conventional finite element interpolation functions are employed. Additionally, SAFE enables to compute the waveguide harmonic forced response [68].

6.2 Formulation of Semi-Analitical Finite Element Method

In the following, the description of SAFE algorithm given in ref. [68] is followed; in some occasions, contributions from other sources is added and quoted.

The displacement distribution along the waveguide is described by complex exponential which represent the wave motion

$$\begin{aligned} u\,(x,y,z,t) &= u(x,y)\cdot e^{-j(\kappa z-\omega t)} \\ v\,(x,y,z,t) &= v(x,y)\cdot e^{-j(\kappa z-\omega t)} \\ w\,(x,y,z,t) &= j\cdot w(x,y)\cdot e^{-j(\kappa z-\omega t)} \end{aligned} \tag{6.1}$$

where, in the notation adopted by Gavrić in ref. [69], z is the coordinate corresponding to the direction along the waveguide, κ is the wavenumber and ω is the frequency.

The waveguide strain ε and its strain energy k can be separated into terms that are independent, linearly dependent, or quadratically dependent on the wavenumber:

$$\varepsilon\,(x,y,z,t) = \varepsilon(x,y)\cdot e^{-j(\kappa z-\omega t)} \tag{6.2}$$

in which

$$\varepsilon(x,y) = \varepsilon_0(x,y) + \kappa\,\varepsilon_1(x,y) = \begin{bmatrix} \frac{\partial u}{\partial x} \\ \frac{\partial v}{\partial y} \\ \kappa\,\omega \\ \frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} \\ -j\,\kappa\,v + j\frac{\partial w}{\partial y} \\ -j\,\kappa\,u + j\frac{\partial w}{\partial x} \end{bmatrix}; \tag{6.3}$$

therefore,

$$\varepsilon_0(x,y) = \begin{bmatrix} \frac{\partial u}{\partial x} \\ \frac{\partial v}{\partial y} \\ 0 \\ \frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} \\ j\frac{\partial w}{\partial y} \\ j\frac{\partial w}{\partial x} \end{bmatrix}, \qquad \varepsilon_1(x,y) = \begin{bmatrix} 0 \\ 0 \\ \omega \\ 0 \\ -j\,v \\ -j\,u \end{bmatrix}. \tag{6.4}$$

It is possible to demonstrate that the strain energy is given by:

$$k = k_0 + \kappa\, k_1 + \kappa^2\, k_2 \tag{6.5}$$

in which

$$k_0 = \varepsilon_0^* c\, \varepsilon_0 \qquad k_1 = \varepsilon_0^* c\, \varepsilon_1 + \varepsilon_1^* c\, \varepsilon_0 \qquad k_2 = \varepsilon_1^* c\, \varepsilon_1 \tag{6.6}$$

where c is the elasticity matrix and * indicates the complex conjugate transpose.

Elemental mass M and stiffness matrices K are obtained by applying a conventional finite element discretization to the terms above. The system equations of motions for the waveguide are obtained by assembling the elemental mass and stiffness matrices

$$M\,\ddot{u} + \left[\kappa^2 \cdot K_2 + \kappa \cdot K_1 + K_0\right]\, u = f, \tag{6.7}$$

where the displacement vector u contains nodal displacements in the x, y, and z directions; the force vector f contains the corresponding forces.

By implementing eq. (6.7) with an identity, it is obtained

$$\begin{bmatrix} K_0 - \omega^2 M & 0 \\ 0 & -K_2 \end{bmatrix} \begin{Bmatrix} u \\ \kappa\, u \end{Bmatrix} + \kappa \begin{bmatrix} K_1 & K_2 \\ K_2 & 0 \end{bmatrix} \begin{Bmatrix} u \\ \kappa\, u \end{Bmatrix} = \begin{Bmatrix} f \\ 0 \end{Bmatrix}, \tag{6.8}$$

so that it can be written in the standard form

$$A\,\overline{u} - \kappa\, B\,\overline{u} = \overline{f}, \tag{6.9}$$

where A and B are related to the dynamic stiffness and mass matrices of the waveguide [67].

The modes supported by the waveguide at a specific frequency may be found from the free vibration problem

$$A\,\overline{u} - \kappa\, B\,\overline{u} = 0. \tag{6.10}$$

by setting the frequency and then solving the complex eigenvalue problem. The eigenvalues can either be real or complex, and have positive or negative sign [67]:

- a real eigenvalue indicates a propagating wave;
- a complex eigenvalue indicates an evanescent wave.

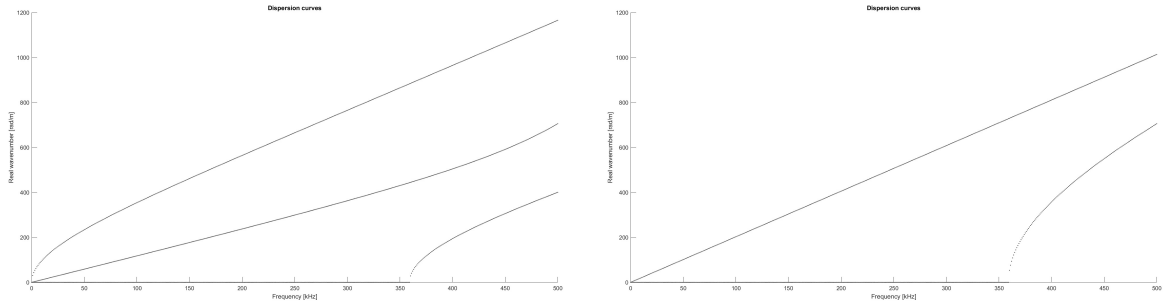
To achieve the dispersion characteristics for a wide frequency span it is therefore necessary to solve the eigenvalue problem at several frequency steps.

From the eigenvalues are gathered the wavenumbers κ, while in the eigenvectors ψ are contained the mode shapes ϕ and the corresponding wavenumber in the form $\psi = \{\varphi, \kappa\,\varphi\}^T$, $\psi = \{\phi, \kappa\,\phi\}^T$. The wavenumbers of interest in guided wave inspection are only those of propagating waves, and hence only those are plotted, see Fig. 6.1.

To compute the group velocity it can be demonstrated (see ref. [70]) that it is possible to use the equation

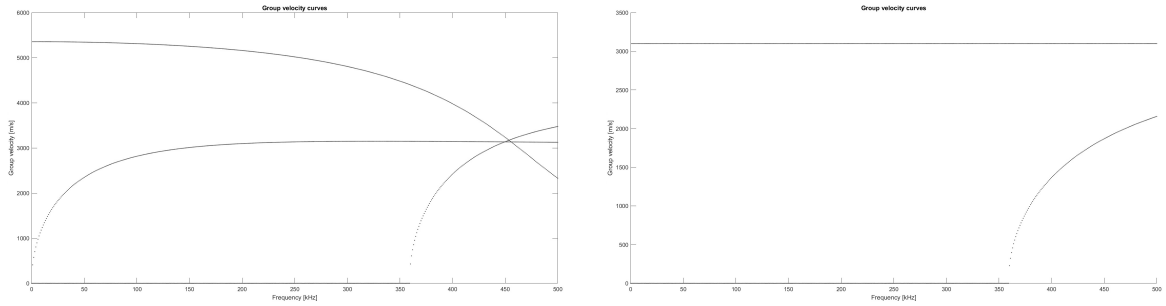
$$v_g = \frac{\partial \omega}{\partial \kappa} = \frac{1}{2\,\omega} \frac{\varphi^T\, (K_1 + 2\,\kappa\, K_2)\, \varphi}{\varphi^T\, M\, \varphi}. \tag{6.11}$$

Also the group velocity can be plotted, see Fig. 6.2.



(a) (b)

Figure 6.1: Plots of wavenumber vs frequency dispersion curves for an aluminum plate (40 mm wide, 4.32 mm thick) simulated through SAFE software (GUIGUW, see ref. [73]). (a) Lamb waves; (b) Shear waves.



(a) (b)

Figure 6.2: Plots of group velocity vs frequency dispersion curves for an aluminum plate (40 mm wide, 4.32 mm thick) simulated through SAFE software (GUIGUW, see ref. [73]). (a) Lamb waves; (b) Shear waves.

6.3 Separation of Modes

The dispersion characteristics plot in Fig. 6.1 comprises a set of frequency–wavenumber points rather than a family of curves (dispersion curves) in which the wavenumber is a function of the frequency. In fact, wavenumbers at adjacent frequency are not directly related. Therefore, if one is interested in a particular mode of propagation, it is thus necessary to extract at each frequency the relative wavenumber.

To separate the modes, it is possible to rely on the orthogonality property of the mode shapes ψ:

$$\psi_\tau^T(\omega)\, B\, \psi_S(\omega) = 0 \qquad \psi_\tau^T(\omega)\, B\, \psi_\tau(\omega) \neq 0. \tag{6.12}$$

If small frequency steps are taken, the mode shapes will not change significantly between steps: a mode shape at one frequency would almost be B-orthogonal to those at the next step. It is then possible to compute the B-orthogonality of the real mode shapes at frequency step k to those at frequency step $k+1$:

$$\Theta = \psi^T(\omega_k)\, B\, \psi(\omega_{k+1}) \tag{6.13}$$

If no crossing happens between the dispersion curves in the frequency interval, the largest terms in matrix $\boldsymbol{\Theta}$ will be the diagonal terms; otherwise, the presence of an off-diagonal term that is larger than the corresponding diagonal term indicates crossing.

6.4 Material Damping in SAFE

It is possible, see ref. [71], to extend the SAFE method to take into account waveguide damping. For such materials, the elasticity matrix is composed by two components:

- a real matrix, which contains the storage modulus;
- an imaginary matrix, which contains the loss modulus.

If it is made the simplifying assumption that the loss modulus are proportional to the storage modulus (hysteretic model), the stiffness matrices for the damped waveguide simply have an additional imaginary component, which is proportional to the real matrix previously derived:

$$\overline{K}_0 = (1 + j\,\gamma)\ K_0 \qquad \overline{K}_1 = (1 + j\,\gamma)\ K_1 \qquad \overline{K}_2 = (1 + j\,\gamma)\ K_2 \tag{6.14}$$

in which K_0, K_1, K_2 are the undamped stiffness matrices, $\overline{K}_0$, $\overline{K}_1$, $\overline{K}_2$ are the damped ones, and γ is the coefficient of proportional hysteretic damping.

Chapter 7

Characteristics of Guided Waves in Rails

The rail, as a waveguide, has a complex profile, so an exact analytical model does not exists. Therefore, it is difficult to compute accurately the dispersion characteristics of guided waves: a form of finite elements method is required [22]. As seen, a frequently used method is *semi-analytical finite element method* (SAFE). It is worth to say that dispersion curves of a rail depend on material properties as well as the rail geometry, neither of which are known to a sufficient degree of accuracy.

Dispersion curves of a rail exhibit mode repulsion and mode crossing, and it is difficult to distinguish between the two, see Fig. 7.1. The mode shapes of rails, having a symmetric profile (see Fig. 2.1), are either symmetric or anti-symmetric. A mode is symmetric when its energy is distributed symmetrically in the waveguide cross-section (in the case of rail, it is equally distributed between field and gauge side, the areas separated by the vertical plane). A mode is antisymmetric when its energy is concentrated non-symmetrically in the waveguide (only on one side of the rail). It has been shown [52] that symmetric and antisymmetric modes can cross each other while the modes within the symmetric and antisymmetric families do not appear to cross. Furthermore, the introduction of even a small asymmetry appears to lead to repulsion forces that prevent any mode crossings. This is very useful knowledge in the selection of a propagating mode for rail integrity inspection. wavenumber vs frequency curves and the concentration of energy for different modes are visible in Fig. 3.4.

Different modes (symmetric or antisymmetric) have energy concentrated in the head, web or foot of the rail. This introduces the possibility to locate defects in a particular portion of the rail cross-section by selecting the most appropriate mode [22]. In [54], Ryue et. al. proved that[1]:

- A localized head bending wave travels primarily through the railhead in the vertical direction above 15 kHz.
- At the side of the railhead, the primary wave measured is a lateral bending wave which has global deformation including the web and rail head in the lateral direction.
- The first-order web bending wave propagates dominantly through the web above 10 kHz.
- The waves propagating along the foot do not travel long distance because of the rail foundation which has a large damping effect.

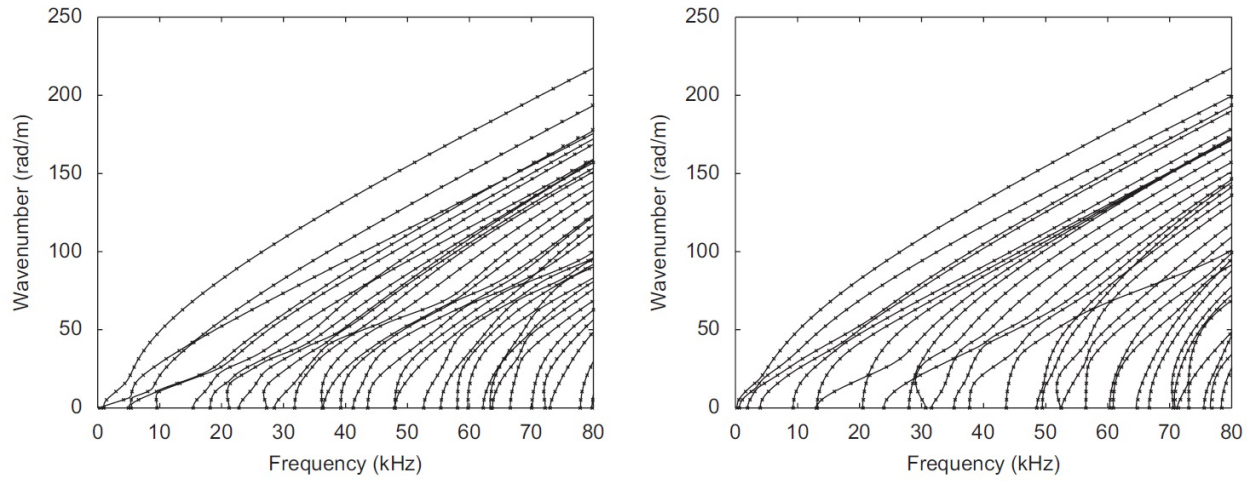
Some concerns can be raised about the influence of ties, fasteners, and absorbing pads. At high frequencies (such as those of ultrasonic guided waves) their presence does not influence the dispersion curves [54]: only decay rates are affected [53]. Clips are generally much more flexible than rail pads, and the contribution of sleepers and ballast is limited to the low-frequency region below 1 kHz [72]. Rail pads mainly influence the decay rates at high frequencies, but not the dispersion curves, since their damping loss factor is generally much higher than that of the rail itself. Welds slightly reflect waves traveling through the web of the rail [54].

Two different mechanism could drive the energy loss from the rails [53]:

- internal dissipation due to material damping;
- acoustic emission from the rail surface, but this effect is predicted to have a smaller effect for most of the frequency range of interest (below 80 kHz).

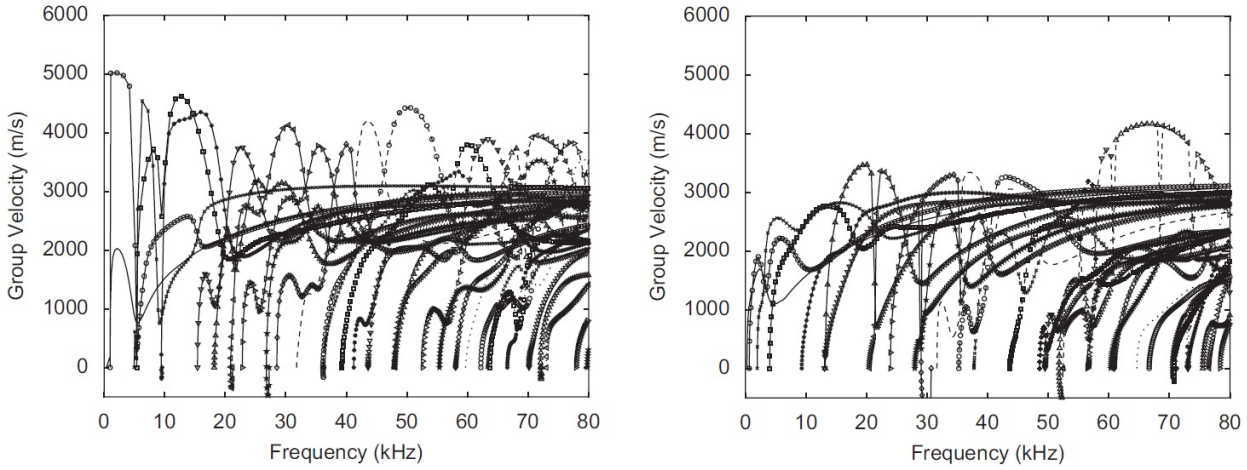
The measured minimum decay rates in the given frequency ranges have been measured by Ryue et. al. in [53], being:

[1] On the track structures with relative properties they have tested, composed by UIC60 (60E1) rail of about 32 m in length, mounted with Pandrol Fastclips, 10 mm studded rubber rail pads and concrete sleepers surrounded by ballast.



(a) (b)

Figure 7.1: Dispersion curves predicted in ref. [54] for supported rail (a) for vertical and symmetric longitudinal waves, (b) for lateral, torsional and antisymmetric longitudinal waves. Picture taken from [54].



(a) (b)

Figure 7.2: Group velocities predicted in ref. [54] through SAFE analysis for supported rail; are shown in (a) vertical and symmetric longitudinal waves and in (b) lateral, torsional and antisymmetric longitudinal waves. Picture taken from [54].

- about 0.04 dB/m at the underside of the railhead between 22 and 40 kHz;
- about 0.04 dB/m at the side of the railhead between 22 and 35 kHz;
- about 0.05 dB/m at the middle of the web between around 20 and 40 kHz.

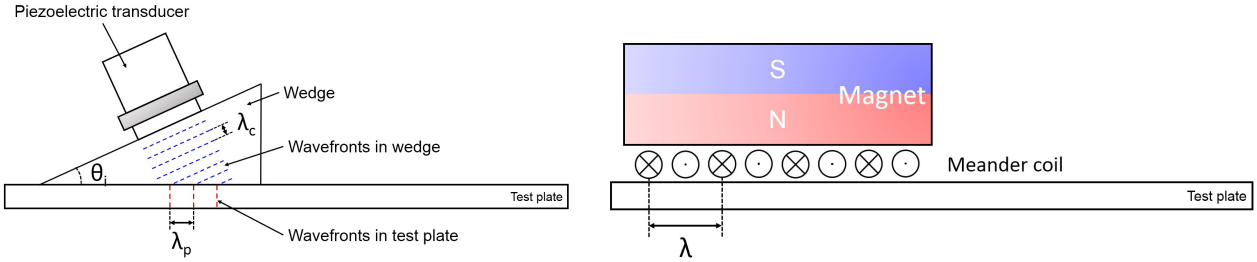
Therefore, the most effective frequency range for long range wave propagation along railway tracks[2] is between 20 and 40 kHz. The tests demonstrated that an amplitude range of more than 50 dB is measurable in rail vibration. Therefore, by assuming a 50 dB level attenuation, the maximum detectable propagating distances are:

- about 1.2 km in the railhead;
- about 1.0 km in the rail web.

Trains are well capable in exciting rail vibration even at high frequencies.

Experimental studies shown in ref. [54] have evidenced that (see Fig. 7.2) group velocities of propagating modes in rails tend to approach a fixed value as frequency increases, possibly the shear wave speed (around

[2]Given the test track considered.



(a) (b)

Figure 7.3: Schematic diagram of excitation of guided waves by (a) piezoelectric transducers or (b) EMAT.

3100 m/s for a bulk wave in steel). Longitudinal waves group velocities instead approach a speed of around 5000 m/s. Some modes show, in the proximity of cut-on frequency, a negative group velocity: for a forward-propagating wave, its carried energy travels backward and vice versa.

Three main problems are experienced when working with ultrasonic guided waves [22]: *dispersion*, *coherent noise*, *temperature or material properties changes*.

1. *Dispersion:* if different modes are excited, they travel at different frequency-dependent velocities in both directions. As a consequence, each mode take a different time to travel along the waveguide, compromising the spatial resolution (i.e. a reflector can generate multiple echoes, or it is difficult to infer the distance from the reflector on the basis of travel time). A countermeasure can be dispersion compensation.

2. *Coherent noise:* this noise affects the same frequency band asthe signal of interest. It has two main surces:

   (a) the excitation and reception of unwanted modes;

   (b) the transmission of waves in the wrong direction along the waveguide and the reception of echoes from that direction.

   To control coherent noise is therefore necessary to excite and sense only the selected modes. This can be accomplished by using an appropriate transducer and excitation signal, and by suppressing unwanted modes.

3. *Changes in temperature or in material properties with age:* that, even if small or unimportant, will affect the dispersion relationships. This is a factor to be taken into account when comparing signals received in different instances of time.

7.1 Excitation of Guided Modes

A mode is most efficiently excited when the harmonic force applied by a transducer is well coupled to the displacement associated with the mode. Therefore, a transducer should be possibly placed on the rail surface where the displacement in the mode shape is high; the polarization of mode shape and transducer must be equal. These considerations are identical also for efficient reception [56]. Mode control is achieved by choosing an appropriate transducer and excitation signal [22].

Common transducer alternatives are, see Fig. 7.3, *piezoelectric transducers* and *electro-magnetic acoustic transducers* (EMATs).

- The piezoelectric transducer is oriented at an angle θ_t with respect to the structure and directs compression waves towards the structure via a coupling medium. The wavelength of the wave generated in the structure λ_p is related to the wavelength of compression waves in the coupling medium λ_c by $\lambda_p = \lambda_c / \sin\theta_i$, where θ_i is the angle of incidence. The phase velocity in the structure c_p is given by $c_p = v / \sin\theta_i$, where v is the velocity of compression waves in the coupling medium. Therefore, a mode at a particular excitation frequency can be excited by orienting the transducer to the appropriate angle.

- EMAT generates a wave in the structure via the Lorentz force and/or magnetostriction. The phase velocity c_p is given by $c_p = f\,\lambda$, where f is the frequency of the driving signal and λ is the meander coil spacing (the wavelength imposed by the EMAT). Direction control can be achieved by employing a second coil overlapping the first, but displaced from it along the structure by a quarter wavelength.

The size of the transducer and the excitation signal drive the degree of modal selectivity obtained. The transducer size influences, in EMATs, the effective wavelength bandwidth or, in piezoelectric transducers, the effective phase velocity bandwidth; the excitation signal governs the frequency bandwidth. If the transducer has a diameter of 3 to 5 times the wavelength, satisfactory mode control is achieved: in long-range testing an array of point sources is better suited with respect to a monolithic transducer. To improve the signal to coherent noise ratio (extraction of only the desired input mode-received mode combination, while rejecting others) a fundamental contribution is given by signal processing [22].

7.2 Selection of Guided Modes

As shown in the dispersion curves of Fig. 3.4(a), many propagating modes are excited in rails in the frequency range of interest (from 25 kHz to 45 kHz). However, not all of these modes are suitable to long-range propagation, because of attenuation (e.g. in the foot of the rail) or since they are highly dispersive [55].

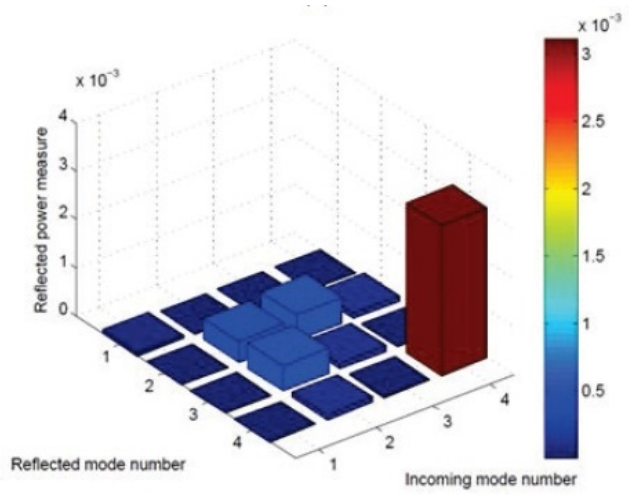


Figure 7.4: Reflection map for a thermite weld with a 6 mm thick weld cap. Picture taken from [55].

As already stated above, a mode is most sensitive to defects at positions in the rail cross section where the energy in the mode shape is most concentrated. To select the mode best suited to detect a defect in pulse-echo operation the following considerations must be kept in mind. When a uni-modal wave impinges on a defect (or other feature), a portion of its energy is reflected, but this is no more uni-modal because of modal conversion. Therefore, it is necessary to obtain reflection coefficient data for each combination of incident and reflected modes, thereby constructing a matrix of reflection coefficients for each defect or feature. A convenient way of representing the information contained in such a matrix is through a color or grayscale map of the amplitudes of each element, i.e. a visual classification tool [56]. An example of such maps is shown in Fig. 7.4.

For example, in ref. [55] it is shown that a symmetric mode with energy concentrated in the railhead (Mode 1 of Fig. 7.4) is well suited to distinguish between cracks in the railhead and welds. The mode with energy concentrated in the web (Mode 4) is able to detect welds and damages in the rail web, but cannot be exploited to differente the two. Detecting cracks at thermite welds (using a land-based approach, see section 10.4) or in the foot of the rail may be very difficult.

Chapter 8

Piezoelectric Transducers and Piezoelectric Effect

The following section is mainly based on ref. [57]. The *piezoelectric effect* consists in the generation of electric charge by a crystalline material udergoing a stress. *Piezoelectricity* manifests itself through the appearance of an electric charge on the surface of a piezoelectric object in response to a pressure. The effect can be found in crystals (such as quartz), artificially polarized man-made ceramics or polymers. It is said that piezoelectric material possess ferroelectric properties.

In ref. [58] Meissner describes a simple but explanatory model of the piezoelectric effect. In this model, illustrated in Fig. 8.2(a), a quartz crystal is represented by means of groups of one silicon (Si) and two oxygen (O_2) atoms which succeed one another along an helix. Each crystal cell is composed by three atoms of silicon and six of oxygen (oxygen is lumped in pairs). Four positive charges are carried by each silicon atom, while two negative charges are carried by each oxigen atom (therefore, the couple carries four charges): consequently, under no-stress conditions, the quartz cell is electrically neutral. When an external force F_x is applied along the x-axis, a deformation is introduced in the hexagonal lattice. For example, in Fig. 8.2(b), a compressing force is applied, and then the atoms are shifted provoking an electrical unbalance along the y-axis. An electric positive charge is formed on the surface side near to the silicon atom, while a negative charge is formed ot the surface side near to the oxygen pair. If the crystal is stretched along the x-axis, a charge of opposite polarity is developed along the y-axis, result of a different deformation (see Fig. 8.2(c)).

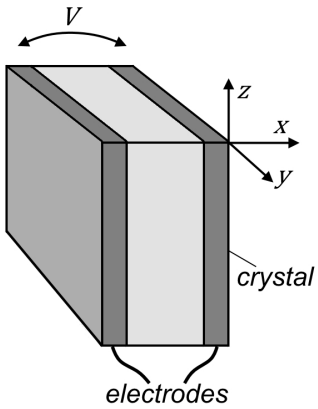
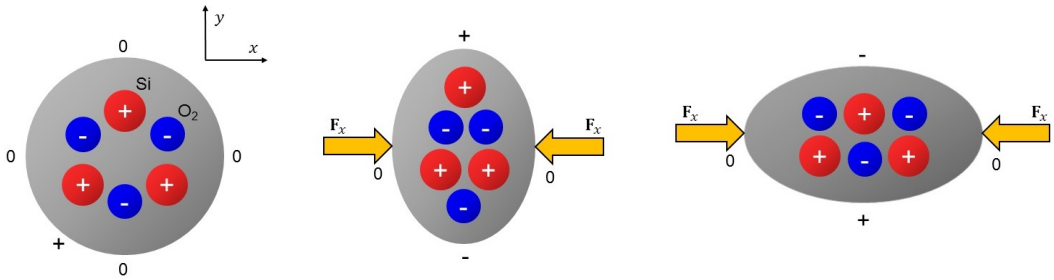


Figure 8.1: Piezoelectric transducer formed by applying two electrodes to the opposite faces of a poled crystalline material. Picture taken from [57].

Two conductive electrodes can be used to pick up the electric charge at the opposite sides of the crystal, as shown in Fig. 8.1. The piezoelectric sensor resembles a capacitor, being constituted by a dielectric material (the piezoelectric crystalline material) sandwiched between two conductive plates (the electrodes). The capacitor is charged in response to the application of strain in the piezoelectric material: a potential difference V appears



(a) (b) (c)

Figure 8.2: Piezoelectric effect in quartz crystal, view along the z-axis. (a) No stress condition; (b) compression in the x-direction; (c) stretching along the x-direction. Picture taken from [57].

between the two capacitor terminals. So, a piezoelectric sensor is a direct converter of mechanical stress into electricity; moreover, the piezoelectric effect is a reversible physical phenomenon. Therefore, if voltage is applied to the piezoelectric transducer, a mechanical strain (i.e., a deformation) is produced.

Chapter 9

Discrete Fourier Transform Applications

In this chapter it will be given a brief outline on the theory of the *discrete fourier transform* (DFT) and of some applications that will be used during the experimental activity. This Chapter is mainly based on refs. [59] and [60].

An alternative Fourier representation can be developed for finite duration sequences, the discrete Fourier transform. The DFT is a discrete sequence itself, corresponding to samples of the signal Fourier transform equally spaced in frequency. It is useful since efficient algorithms exist for its computation. One possible way to present the DFT is to focus on the relationship between periodic sequences and finite-length sequences. Firstly is considered the application of Fourier series representation of periodic sequences results to the representation of finite-length sequences. It will be demonstrated that the Fourier series representation of the periodic sequence corresponds to the DFT of the finite-length sequence [59].

9.1 Discrete Fourier Series: Representation of Periodic Sequences

We start by considering a sequence $\tilde{x}[n]$ which is periodic with period N, so that $\tilde{x}[n] = \tilde{x}[n + rN]$ for any n and r integer value. Such a sequence can be represented by a Fourier series corresponding to a sum of harmonically related complex exponential sequences. These are complex exponentials with frequencies that are integer multiples of the fundamental frequency $(2\pi/N)$ associated with the periodic sequence $\tilde{x}[n]$. They are expressed in the form

$$e_k[n] = e^{j(2\pi/N)kn} = e_k[n + rN], \tag{9.1}$$

where k is an integer. The Fourier series representation can be then expressed in the form

$$\tilde{x}[n] = \frac{1}{N} \sum_k \tilde{X}[k] e^{j(2\pi/N)kn}. \tag{9.2}$$

The Fourier series representation of a discrete-time signal with period Nrequires only N harmonically reltated complex exponentials and can thus take the form

$$\tilde{x}[n] = \frac{1}{N} \sum_{k=0}^{N-1} \tilde{X}[k] e^{j(2\pi/N)kn}. \tag{9.3}$$

To obtain the sequence of Fourier series coefficients $\tilde{X}[k]$ from the periodic sequence $\tilde{x}[n]$ it is possible to exploit the orthogonality of the set of complex exponential sequencees. The result is

$$\tilde{X}[k] = \sum_{k=0}^{N-1} \tilde{x}[n] e^{-j(2\pi/N)kn}. \tag{9.4}$$

The sequence $\tilde{X}[k]$ is periodic with period N for any integer k: $\tilde{X}[k + N] = \tilde{X}[k]$. The Fourier series coefficients can be interpreted to be a sequence of finite length or as a periodic sequence defined for all k, as in eq. (9.4). An advantage of interpreting the Fourier series coefficients $\tilde{X}[k]$ as a periodic sequence is that for the Fourier series representation of periodic sequences there is then a duality between the time and frequency domains. Equations (9.3) and (9.4) are then the analysis-synthesis functions of the *discrete Fourier series* (DFS) representation of a periodic sequence. For sake of convenience, by defining the quantity

$$W_N = e^{-j(2\pi/N)}, \tag{9.5}$$

it is possible to express the functions as

$$\begin{array}{rl} \text{Analysis equation:} & \tilde{X}[k] = \sum_{k=0}^{N-1} \tilde{x}[n] W_N^{kn} \\ \text{Synthesis equation:} & \tilde{x}[n] = \frac{1}{N} \sum_{k=0}^{N-1} \tilde{X}[k] W_N^{-kn} \end{array} \tag{9.6}$$

and it is possible to introduce the notation

$$\tilde{x}[n] \overset{DFS}{\longleftrightarrow} \tilde{X}[k] \tag{9.7}$$

to signify the relationship between the analysis-synthesis functions [59].

9.2 Fourier Transform of Periodic Signals

It is useful to incorporate the DFS representation of periodic signals within the framework of the Fourier transform. The Fourier transform of a periodic signal can be interpreted as an impulse train in the frequency domain in which the value of each impulse is proportional to the sequence DFS coefficients. Therefore, if $\tilde{x}[n]$ is periodic with period N, the corresponding DFS are $\tilde{X}[k]$, and hence the Fourier transform of $\tilde{x}[n]$ is defined to be the impulse train

$$\tilde{X}\left(e^{j\omega}\right) = \sum_{k=-\infty}^{\infty} \frac{2\pi}{N} \tilde{X}[k] \, \delta\left(\omega - \frac{2\pi k}{N}\right), \tag{9.8}$$

in which δ is the Kronecker delta.

$\tilde{X}\left(e^{j\omega}\right)$ has the necessary periodicity with period 2π since $\tilde{X}[k]$ is periodic with period N and the impulse are then spaced at integer multiples of $(2\pi/N)$ [59].

9.2.1 Fourier Transform Sampling

The relationship between an aperiodic sequence with Fourier transform $X\left(e^{j\omega}\right)$ and the periodic sequence for which the DFS coefficients correspond to the equally frequency spaced samples of $X\left(e^{j\omega}\right)$ is important when discussing the discrete Fourier transform and its properties.

Consider an aperiodic sequence $x[n]$ with Fourier transform $X\left(e^{j\omega}\right)$; assume that by sampling $X\left(e^{j\omega}\right)$ at frequencies $\omega_k = 2\pi k/N$ is obtained the sequence $\tilde{X}[k]$

$$\tilde{X}[k] = X\left(e^{j\omega}\right)\Big|_{\omega=(2\pi/N)k} = X\left(e^{j(2\pi/N)k}\right). \tag{9.9}$$

Since the Fourier transform is periodic in ω with period 2π, the resulting sequence is periodic in k with period N. Therefore $\tilde{X}[k]$ *could* be the sequence of DFS coefficients of a sequence $\tilde{x}[n]$. To obtain that sequence, it is possible to substitute $\tilde{X}[k]$ obtained by sampling into the synthesis equation in eq. (9.6)

$$\tilde{x}[n] = \frac{1}{N} \sum_{k=0}^{N-1} \tilde{X}[k] W_N^{-kn}. \tag{9.10}$$

Since it has not been made any assumption about $x[n]$ other than the Fourier transform exists,

$$\tilde{X}\left(e^{j\omega}\right) = \sum_{k=-\infty}^{\infty} x[m] \, e^{-j\omega m}, \tag{9.11}$$

by substituting eq. (9.11) into eq. (9.9) and then replacing the resulting expression for $\tilde{X}[k]$ into eq. (9.10), it is obtained

$$\tilde{x}[n] = \sum_{k=-\infty}^{\infty} x[m] \, \tilde{p}[n-m] \tag{9.12}$$

in which $\tilde{p}[n-m]$ can be seen to be the Fourier series representation of a periodic impulse train, i.e.

$$\tilde{p}[n-m] = \frac{1}{N} \sum_{k=0}^{N-1} W_N^{-k(n-m)} = \sum_{r=-\infty}^{\infty} \delta[n-m-rN]. \tag{9.13}$$

Therefore, the periodic sequence $\tilde{x}[n]$ can be seen as

$$\tilde{x}[n] = x[n] * \sum_{r=-\infty}^{\infty} \delta[n - rN] = \sum_{r=-\infty}^{\infty} x[n - rN] \tag{9.14}$$

where $*$ denotes aperiodic convolution. $\tilde{x}[n]$ is therefore the periodic sequence resulting from the aperiodic convolution of $x[n]$ with a periodic unit-impulse train. That is to say that $\tilde{x}[n]$ corresponding to $\tilde{X}[k]$ obtained through sampling $X\left(e^{j\omega}\right)$ is formed from $x[n]$ adding together an infinite number of shifted replicas of $x[n]$, being the shifts positive and negative integer multiples of period of the sequence $\tilde{X}[k]$, N.

This discussion highlights several important concepts useful in the following. Samples of the Fourier transform of an aperiodic sequence $x[n]$ can he thought of as DFS coefficients of a periodic sequence $\tilde{x}[n]$ obtained through summing periodic replicas $x[n]$. If $x[n]$ has finite length and if a number of equally spaced samples greater than (or equal to) the length of $x[n]$ of its Fourier transform are taken, then the Fourier transform is recoverable from these samples. Equivalently, $x[n]$ is recoverable from $\tilde{x}[n]$ through the relation

$$x[n] = \begin{cases} \tilde{x}[n] & 0 \le n \le N-1, \\ 0 & \text{otherwise.} \end{cases} \tag{9.15}$$

A direct relationship between $X\left(e^{j\omega}\right)$ and its samples, i.e., an interpolation formula for $X\left(e^{j\omega}\right)$ can be derived, but it is not necessary to know eq. (9.14), which can also be represented by a DFS; alternatively, given the sequence of Fourier coefficients $\tilde{X}[k]$ it is possible to find $\tilde{x}[n]$ and then to use eq. (9.15) to recover $x[n]$. When the Fourier series is used in this way to represent finite-length sequences, it is called the *discrete Fourier transform* (DFT). It is important to keep in mind that the representation through samples of the Fourier transform is in effect a representation of the finite-duration sequence by a periodic sequence [59].

9.3 Discrete Fourier Transform: Fourier Representation of Finite Duration Sequences

Consider a sequence $x[n]$ of N samples such that $x[n] = 0$ outside the range $0 \le n \le N-1$. As seen, to each finite-length sequence of length N it is always possible to associate a periodic sequence

$$\tilde{x}[n] = \sum_{r=-\infty}^{\infty} x[n - rN]. \tag{9.16}$$

The finite-length sequence can be recovered from $\tilde{x}[n]$ by using

$$x[n] = \begin{cases} \tilde{x}[n] & 0 \le n \le N-1, \\ 0 & \text{otherwise.} \end{cases} \tag{9.17}$$

The DFS coefficients of $\tilde{x}[n]$ are samples spaced in frequency by $2\pi/N$ of the Fourier transform of $x[n]$. Since $x[n]$ has finite length N, there is no overlap between the terms $x[n - rN]$ for different values of r. Thus $\tilde{x}[n]$ can also be expressed as

$$\tilde{x}[n] = x[(n \text{ modulo } N)] = x[((n))_N]. \tag{9.18}$$

Equation (9.18) is equivalent to eq. (9.16) only when the length of $x[n]$ is less or equal to N.

We have seen that the sequence of DFS coefficients $\tilde{X}[k]$ of $\tilde{x}[n]$ is itself a periodic sequence with period N. To maintain a duality between the time and frequency domains it is possible to choose the Fourier coefficients to be associated with a finite-duration sequence to be a finite-duration sequence corresponding to one period of $\tilde{X}[k]$. This finite-duration sequence $X[k]$ will be referred as the discrete Fourier transform (DFT). The DFT is related to the DFS coefficients by

$$X[k] = \begin{cases} \tilde{X}[k] = X[((k))_N] & 0 \le n \le N-1, \\ 0 & \text{otherwise.} \end{cases} \tag{9.19}$$

By remembering the relation between $\tilde{X}[k]$ and $\tilde{x}[n]$ expressed in eq. (9.6) and remembering that the summations there involve only the interval between zero and $N-1$, it follows that

$$X[k] = \begin{cases} \sum_{k=0}^{N-1} \tilde{x}[n] W_N^{kn} & 0 \le n \le N-1, \\ 0 & \text{otherwise,} \end{cases} \tag{9.20}$$

$$x[n] = \begin{cases} \frac{1}{N}\sum_{k=0}^{N-1} \tilde{X}[k] W_N^{-kn} & 0 \le n \le N-1, \\ 0 & \text{otherwise.} \end{cases} \tag{9.21}$$

The analysis and synthesis equations can be synthesized as follows

$$\begin{array}{rl} \text{Analysis equation:} & X[k] = \sum_{k=0}^{N-1} \tilde{x}[n] W_N^{kn} \\ \text{Synthesis equation:} & x[n] = \frac{1}{N}\sum_{k=0}^{N-1} \tilde{X}[k] W_N^{-kn} \end{array} \tag{9.22}$$

and it is possible to introduce the notation

$$x[n] \overset{DFT}{\longleftrightarrow} X[k] \tag{9.23}$$

to signify the relationship between the analysis-synthesis functions.

While recasting eq. (9.6) into eq. (9.22), the inherent periodicity has not been eliminated: the DFT $X[k]$ is equal to samples of the periodic Fourier transform $X\left(e^{j\omega}\right)$ as with the DFS, and if the synthesis equation is evaluated for n values outside the interval $[0, N-1]$ the result will be not zero, but a periodic extension of $x[n]$ [59].

9.4 Fourier Analysis of Signals using the Discrete Fourier Transform

One of the major applications of the DFT is the analysis of the frequency content of continuous-time signals. The basic steps in the application of DFT to a continuous-time signal $s_c(t)$ are:

- *Anti-alias lowpass filter*, to eliminate or minimize the effect of antialiasing in the later conversion of the signal into a discrete-time sequence; the filter transfer function is $H_{aa}(j\Omega)$. It is produced the filtered signal $x_c(t)$.
- *Continuous-to-discrete-time conversion*, producing the sequence of samples $x[n]$.
- *Windowing*, i.e., multiplication of the sequence of samples $x[n]$ by a finite-duration window $w[n]$, useful to produce a finite-length sequence $v[n]$ for DFT elaboration.
- At last, *DFT analysis*, producing the vector of DFT coefficients $V[k]$.

The windowing is required since in many cases the signal $s_c(t)$, and hence the sequence $x[n]$ is very or indefinitely long, and the application of a finite-duration window to achieve a finite-duration sequence prior to DFT computation is necessary.

The conversion of the filtered continuous-time signal $x_c(t)$. into a sequence of samples $x[n]$ is represented by a periodic replication and a frequency normalization process in the frequency domain

$$X\left(e^{j\omega}\right) = \frac{1}{T}\sum_{r=-\infty}^{\infty} X_c\left(j\frac{\omega}{T} + j\frac{2\pi r}{T}\right). \tag{9.24}$$

If the digital signal $x[n]$ is achieved through analog-to-digital conversion (ADC), which introduces a quantization error.

Windowing, achieved by multiplying $x[n]$ with a finite-duration window $w[n]$ is required since the input of the DFT must have a finite duration: it is therefore produced a finite-length sequence $v[n] = w[n]\,x[n]$. The effect of windowing in the frequency domain is a periodic convolution

$$V\left(e^{j\omega}\right) = \frac{1}{2\pi}\int_{-\pi}^{\pi} X\left(e^{j\theta}\right) W\left(e^{j(\omega-\theta)}\right) d\theta. \tag{9.25}$$

If the window $w[n]$ has a constant value when it is nonzero, we speak of a *rectangular window*. Other types of windows exist, tapered at the edges, such as the Bartlett, Hamming, Hanning, Blackman, Kaiser, etc. For now, it is sufficient to observe that the convolution of $W\left(e^{j\omega}\right)$ with $X\left(e^{j\omega}\right)$ will tend to smooth sharp peaks and discontinuities in $X\left(e^{j\omega}\right)$.

The DFT of the windowed sequence $v[n] = w[n]\,x[n]$ is $V[k]$

$$V[k] = \sum_{k=0}^{N-1} v[n]\, e^{-j(2\pi/N)kn}, \qquad k = 0, 1, \ldots, N-1, \tag{9.26}$$

and it is assumed that the window length L is less than or equal to the DFT length N. $V[k]$ corresponds to equally spaced samples of the Fourier transform of $v[n]$

$$V[k] = V\left(e^{j\omega}\right)\Big|_{\omega=2\pi k/N'}. \tag{9.27}$$

Since the spacing between the DFT frequencies is $2\pi/N$, and the relationship between the normalized discrete-time frequency variable and the continuous-time frequency variable is $\omega = \Omega T$, the DFT frequencies correspond to the continuous-time frequencies

$$\Omega_k = \frac{2\pi k}{NT}. \tag{9.28}$$

9.4.1 Time-Dependent Fourier Transform and the Spectrogram: Fourier Analysis of Nonstationary Signals

In practical application of sinusoidal signal models, often the properties of the signal (amplitude, frequency, phase) change with time. Therefore, such signals cannot be described by a single DFT estimate: as a result it has been introduced the concept of the *time-dependent Fourier transform* (TDFT), also referred as the short-time Fourier trasnsform (STFT).

The STFT of a discrete-time signal $x[n]$ is defined as

$$X[N,\lambda) = \sum_{m=-\infty}^{\infty} x[n+m]\, w[m]\, e^{-j\lambda m}, \tag{9.29}$$

where $w[m]$ is again the windowing function. As a matter of fact, the mono-dimensional sequence $x[n]$, function of a single discrete variable n, is converted into a bi-dimensional function of the discrete-time variable n, and of the variable λ in the frequency domain, which is continuous[1]. Note that the STFT is periodic in λ with period 2π: therefore only the values of λ in an interval 2π are considered. The eq (9.29) can be interpreted as the Fourier transform of the shifted signal $x[n+m]$ as seen through the window $w[m]$. The window has a stationary origin: as n varies, the signal crosses the window so that for each n value a different portion of the signal is viewed.

The main purpose of the window in the STFT is to limit the length of the sequence over which the DFT has to be performed. In this way, the spectral characteristics are reasonably stationary for the entire duration of the window. The faster the signal characteristics change, the shorter the window must be: as a consequence, it can be shown that the frequency resolution decreases. On the other hand, however, the resolution in the time domain increases. Consequently, the choice of window length constitutes a trade-off between resolution in the time domain and in the frequency domain.

The STFT concept is widely used to obtain spectral estimates of discrete-time non-stationary signals, where it is useful to analyze the trend of the spectral components over time [59]. The *spectrogram* can be defined as an intensity plot (usually expressed in decibel, dB) of the short-time Fourier transform (STFT) magnitude. For the computation of the spectrogram the STFT windows windows are usually allowed to overlap in time [60].

9.4.2 The Periodogram: Fourier Analysis of Stationary Random Signals

In this Section it is studied the estimation of the power spectrum (or the *power density spectrum*) of a random signal using the DFT. The interest is in obtaining an unbiased content estimator: this analysis is very difficult, and only approximate results are achievable [59].

The *periodogram* is based on the definition of the power spectral density (PSD) [60]. Assume that the input signal $s_c(t)$ to the DFT processing is a stationary random signal. The anti-alias filter create a new stationary random signal having a band-limited spectrum; hence, the signal can be sampled without aliasing producing a discrete-time random signal $x[n]$. Its power density spectrum $P_{xx}(\omega)$ is proportional the power density spectrum $P_{ss}(\Omega)$ of the continuous time signal $s_c(t)$ over the bandwidth of the anti-aliasing filter

$$P_{xx}(\omega) = \frac{1}{T} P_{ss}\left(\frac{\omega}{T}\right), \qquad |\omega| < \pi, \tag{9.30}$$

if the cutoff frequency of the anti-aliasing filter is π/T and that T is the sampling period. A reasonable estimate of $P_{xx}(\omega)$ will give a reasonable estimate of $P_{ss}(\Omega)$. When the signal undergoes windowing, a finite length L of samples of $x[n]$ are selected into $v[n] = w[n]\, x[n]$, whose Fourier transform is

[1] It is used λ to differentiate the variable in the frequency of the STFT from the frequency variable ω of the DFT; the mixed notation $X[N,\lambda)$ is used to remember that n is a discrete variable, while λ is continuous.

$$V\left(e^{j\omega}\right)=\sum_{n=0}^{L-1} w\left[n\right] x\left[n\right] e^{-j\omega n}. \tag{9.31}$$

An estimate of the power spectrum can be given by the quantity

$$I\left(\omega\right)=\frac{1}{LU}\left|V\left(e^{j\omega}\right)\right|^{2}, \tag{9.32}$$

where the constant U anticipates a need for normalization to remove bias in the spectral estimate. If $w\left[n\right]$ is a rectangular window sequence, the estimator for the power spectrum is called the *periodogram*, if the window is not rectangular, it is called the *modified periodogram*. The periodogram has some basic properties of the power spectrum (non negative and real and even function of frequency for real signals). The periodogram is in fact the Fourier transform of the aperiodic correlation of the windowed data sequence [59]. In the limit as L goes to infinity, the expected value of the periodogram equals the power spectral density of the random stationary signal $x\left[n\right]$ [60].

Explicit computation of the periodogram can be carried out only at discrete frequencies: if the DTFT of $w\left[n\right]x\left[n\right]$ is replaced by its DFT, are obtained samples at the DFT frequencies $\omega_k=2\pi k/N$ for $k=0,1,\ldots N-1$ by

$$I\left(\omega_k\right)=\frac{1}{LU}\left|V\left[k\right]\right|^{2}, \tag{9.33}$$

where $V\left[k\right]$ is the N-point DFT of $w\left[n\right]x\left[n\right]$.

If a random signal has a nonzero mean, its power spectrum has an impulse at zero frequency. If the mean is relatively large, this component will dominate the spectrum estimate, causing low-amplitude, low-frequency components to be obscured by leakage. Therefore, in practice often it is computed the *sample mean*

$$\hat{m}_x=\frac{1}{L}\sum_{n=0}^{L-1} x\left[n\right], \tag{9.34}$$

and the resulting estimate is subtracted from the random signal before computing the power spectrum estimate (i.e., it is evaluated $x\left[n\right]-\hat{m}_x$). Although the sample mean is only an approximate estimate of the zero-frequency component, subtracting it from the signal often leads to better estimates at neighboring frequencies [59].

9.5 Two-Dimensional Fourier Transform

Fourier transform can be generalized to higher dimensions, an example is constituted by the application to two-dimensional signals $f\left(x,y\right)$, functions of a 2D space defined over an x-y plane.

To explain the meaning of the two-dimensional Fourier transform, it is possible to reference the Fourier transform of a continuous, aperiodic signal. The results can be generalized also for other cases (discrete, periodic signals).

$$F\left(u,v\right)=\iint_{-\infty}^{\infty} f\left(x,y\right)e^{-j2\pi(xu+yv)}dx\,dy \tag{9.35}$$

$$f\left(x,y\right)=\iint_{-\infty}^{\infty} F\left(u,v\right)e^{j2\pi(xu+yv)}du\,dv \tag{9.36}$$

The inverse transform represents the spatial function $f(x,y)$ as a linear combination of complex exponentials $e^{j2\pi(xu+yv)}$ with complex weights $F(u,v)$. The complex weight can be represented in polar form as

$$F\left(u,v\right)=F_r\left(u,v\right)+jF_i\left(u,v\right)=\left|F\left(u,v\right)\right|e^{j\angle F(u,v)} \tag{9.37}$$

or in terms of its amplitude $\left|F\left(u,v\right)\right|$ and phase $\angle F\left(u,v\right)$ as

$$\begin{cases}\left|F\left(u,v\right)\right|=\sqrt{F_r\left(u,v\right)^2+F_i\left(u,v\right)^2}\\ \angle F\left(u,v\right)=\tan^{-1}\left[F_i\left(u,v\right)/F_r\left(u,v\right)\right]\end{cases}. \tag{9.38}$$

The complex exponential can be represented as

$$e^{j2\pi(xu+yv)}=e^{j2\pi\omega(xu/\omega+yv/\omega)}=e^{j2\pi\omega(\vec{\mathbf{r}}\cdot\vec{\mathbf{n}})}, \tag{9.39}$$

where $\omega = \sqrt{u^2 + v^2}$, $\vec{\mathbf{n}} = \left(\frac{u}{\omega}, \frac{v}{\omega}\right)$ is the unit vector along direction (u, v) and $\vec{\mathbf{r}} = (x, y)$ is a vector along the direction (x, y) in the 2D spatial domain. The inner product $(\vec{\mathbf{r}} \cdot \vec{\mathbf{n}})$ represents the projection of a spatial point $\vec{\mathbf{r}} = (x, y)$ onto the direction of $\vec{\mathbf{n}}$. Since all points (x, y) on a straight line perpendicular to the direction of $\vec{\mathbf{n}}$ have the same projection, $e^{j2\pi(xu+yv)} = e^{j2\pi\omega(\vec{\mathbf{r}}\cdot\vec{\mathbf{n}})}$ represents a planar sinusoid in the x-y plane along the direction $\theta = \tan^{-1}(v/u)$ (i.e., $\vec{\mathbf{n}}$) with frequency $\omega = \sqrt{u^2 + v^2}$.

Therefore, the 2DFT of a signal $f(x, y)$ can be written as

$$f(x, y) = \int_{-\infty}^{\infty}\int |F(u, v)| e^{j\angle F(u,v)} e^{j2\pi(xu+yv)} du\, dv = \int_{-\infty}^{\infty}\int |F(u, v)| e^{j(\angle F(u,v)+2\pi\omega(\vec{\mathbf{r}}\cdot\vec{\mathbf{n}}))} du\, dv \tag{9.40}$$

which represents a signal $f(x, y)$ as a linear combination (integration) of infinite 2D spatial planar sinusoids $F(u, v)$ of frequency $\omega = \sqrt{u^2 + v^2}$, direction $\vec{\mathbf{n}} = \tan^{-1}(v/u)$, amplitude $|F(u, v)| = \sqrt{F_r(u, v)^2 + F_i(u, v)^2}$ and phase $\angle F(u, v) = \tan^{-1}[F_i(u, v)/F_r(u, v)]$.

It can be demonstrated that 2D DFT can be carried out by 1D transforming all the rows of the 2D input signal $f(x, y)$ arranged in a 2D matrix and then 1D transforming all the columns of the resulting matrix. The order of the steps is not important [61].

Part II

Ultrasonic Guided Waves Applications

Chapter 10

Existing Applications of UGW Rail Defect Detection Techniques

10.1 Systems for Rail Defect Detection Based on Ultrasonic Guided Waves: Classification

Substantial research efforts have been devoted to the study of systems exploiting ultrasonic guided waves for the detection of rail defects. More specifically, the following two broad categories of systems have been been investigated:

1. *On-board systems* [62, 74, 75, 76, 79, 80] - The measurement equipment is embarked on board a moving vehicle (e.g., an inspection cart), allowing for scanning the head of the rails on which it is riding. The detection of defects in the scanned rails is achieved through a pitch–catch mechanism. Initial implementations were based on an active approach, in which a source of ultrasounds was used to inject UGW into the rails, and sensing was accomplished through an array of transducers. More recent implementations adopt a passive approach, in which the ultrasonic excitation is generated by the rolling wheels of the moving vehicle. The approaches based on on-board systems have been promoted, among others, by the *Experimental Mechanics & NDE Laboratory of the University of California*, San Diego, USA.

2. *Land-based systems* [82, 83, 86, 84, 85, 92, 87, 89, 90, 93, 99, 100, 101, 94, 98, 97, 102, 91] - The employed equipment is placed on the ground near to the railway infrastructure, and its actuators and sensors are attached to the rails under test. The detection of defects in the rails is achieved through a pulse-echo approach. A lot of contributions about this topic have been published by the Sensor Science and Technology Department of the Council for Scientific and Industrial Research, South Africa. The ultrasonic broken rail detection (UBRD) system has been developed by this institution and put in use on the Oreline in South Africa to detect the presence of a fully broken rail; the system is evolved to implement an early defect detection system. In recent years, important results in this research field have been achieved by two Chinese research groups. In fact, significant contributions about the best methods to excite and detect propagation modes in a rail, and about feasible methods for rail defect location have been published by various researchers working at the Beijing Jiaotong University (*School of Mechanical, Electronic, and Control Engineering*, and *Key Laboratory of Vehicle Advanced Manufacturing, Measuring and Control Technology*). Moreover, an electronic system able to efficiently excite ultrasonic guided waves in rails has been developed by various researchers working at the Xi'an University of Technology (*Department of Electronic Engineering*).

In the remaining part of this chapter, both categories of systems are taken into consideration. In particular, the architecture of these systems is described in Section 10.2. Sections 10.3, 10.4 and 10.5 are devoted to the implementation of on-board approaches, to that of the UBRD system, and to the evolution of the UBRD system towards early defect detection and monitoring. In Section 10.6, commercial projects like RailAcoustic (by Enekom), or university studies, like those accomplished by the Beijing Jiaotong University or the Xi'an University of Technology are summarized. Finally, various results about the performance provided by each of the considered systems are illustrated in Section 10.7.

10.2 Systems for Rail Defect Detection Based on Ultrasonic Guided Waves: Architecture

In the technical literature, various architectures have been proposed for on-board or land-based systems for rail defect detection based on UGWs. In this section, some essential features of the available technical options are illustrated.

10.2.1 On-Board Systems

Over the years, *active* and *passive* detection strategies have been proposed for on-board systems. The main characteristics of the systems employing such strategies can be summarized as follows:

An active rail inspection system is able to detect surface-breaking cracks and internal defects located in the railhead. It is based on UGWs and air-coupled contactless probing. Both the transmission of test signals and their reception are accomplished by means of air-coupled piezoelectric transducers arranged in a pitch–catch configuration. The use of a laser transmitter has been also investigated [62]; however, this option has been judged too costly and is potentially hard to maintain. In an active system, receivers are arranged in pairs and placed at the two sides of the transmitter: this allows for implementing a differential detection scheme. The system output, after data analysis, is represented by the so-called *damage index* (DI), exhibiting peaks in correspondence of discontinuities (i.e., of potential defects) in rails. In recent years, the use of laser technology for both exciting UGWs in rails and sensing them has been investigated [77, 78]. In this case, non-ablative laser sources are used to generate a guided tensional wave system in the rail; the response to such a system can be sensed by means of a rotational laser vibrometer.

A passive rail inspection system aims at estimating the acoustic transfer function of rails; the excitation is represented by the rolling wheels of the moving instrumented train [79]. When a train is in motion, its rotating wheels generate a continuous dynamic excitation of the rail. Such an excitation is uncontrolled, non-stationary, and difficult to characterize [79]. Therefore, in this case, the challenge to be faced is to generate a stable rail transfer function that is not affected by the excitation [80]. The presence of a discontinuity in the rail can be inferred from the estimated transfer function, using again the DI indicator.

10.2.2 Land-Based Systems

The UBRD is a *broken rail detection* (BRD) system, designed to continuously detect full breaks in rails [82]. Its transducers are tied to the rail and allow for monitoring a span of up to 1 km of track in pitch–catch mode. This system has been already installed on 841 km of tracks in the high-haul Oreline network in South Africa [83, 86]. The system architecture is represented in Fig. 10.1; its operation is based on a simple transmit–receive confirmation protocol. As long as a reliable reception of the ultrasonic signal generated by the transmitter is possible, the integrity of the rail is verified; otherwise, an alarm is set [82, 84].

The developers of the UBRD system are studying how UGWs can be exploited to detect and monitor a growing defect, and how specific guided modes can be excited in a rail. The use of techniques for the excitation of guided waves different from those based on piezoelectric transducers have been investigated. For example, the use of *electromagnetic acoustic transducers* (EMATs) has been taken into consideration [92]. In the meantime, the reflection characteristics of defects and their dependence on time and environmental conditions are under evaluation [102, 87, 89, 90].

In principle, UGWs have the potential to monitor a long span of a waveguide using permanently installed transducers in pulse–echo mode. The same transducer array is used to generate the ultrasonic wave and then to measure the echo generated by the wave reflection on a defect. A continuously welded rail is a perfect example of such a waveguide and, as a matter of fact, defects (and welds) act as reflectors for guided waves. A system based on this approach is able to detect the *presence* of a defect, and to estimate its *magnitude* and *location*. These results are achieved by monitoring the echoes coming from defects (defect presence), measuring their amplitude (defect magnitude) and the time-of-flight required by the transmitted wave to travel back and forth from the transducer to the defect (defect location) [83].

An important advantage offered by land-based solutions is represented by the fact that monitoring can be accomplished in near real-time, so that defect growth can be analyzed over time and an alarm set only when a crack assumes a potentially critical size [102, 87].

Further details about the architecture of the UBRD system are provided in Section 10.4. In Section 10.5 it is described its evolution towards an early defect detection system. Other projects or studies about land-based systems are illustrated in Section 10.6.

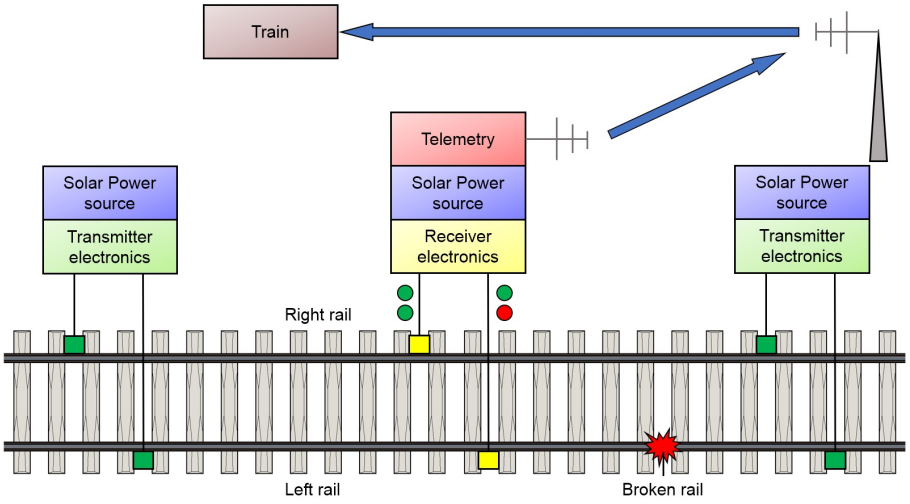


Figure 10.1: UBRD system concept [83].

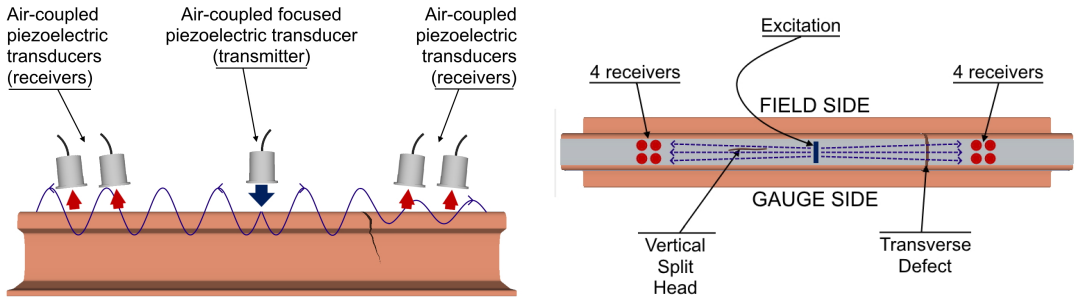


Figure 10.2: Schematic of the non-contact defect detection system based on ultrasonic guided waves: (a) side view and (b) top view. The two red lines indicate the possible locations of a *vertical split head* and a *transverse defect* [75].

10.3 Implementation of On-Board Systems

In this section, the implementation of on-board systems is described in detail; both active and passive approaches are considered.

10.3.1 Active Approach

10.3.1.1 Hardware Configuration

The configuration adopted in the active on-board system is shown in Fig. 10.2: a single focused transmitter and four pairs of receivers are fixed to a frame which is in turn attached to a cart [62, 74, 76]. The lower active surface of the transducer is separated from the upper surface of the rail; their distance is known as *lift-off* and is on the order of a few centimeters for both the transmitter and the receivers. The transmitter feeds the top of the rail surface with a narrowband toneburst signal (a Hanning-windowed sinusoid); it has a tunable repetition rate (representing the frequency of the excitation and related to spatial resolution). This results in a (repeated) multimodal guided wave that propagates symmetrically with respect to the excitation line. The receivers detect the guided waves leaking into the surrounding air; for this reason, they are tilted according to Snell's law.

Table 10.1: Features extracted from the received time-domain waveforms x_1 and x_2; these are recorded by two receivers forming a symmetric pair [75].

Feature number	Feature content
1	$\max\left(\frac{\max\lvert x_1\rvert}{\max\lvert x_2\rvert}, \frac{\max\lvert x_2\rvert}{\max\lvert x_1\rvert}\right)$
2	$\max\left(\frac{\mathrm{Pkpk}\lvert x_1\rvert}{\mathrm{Pkpk}\lvert x_2\rvert}, \frac{\mathrm{Pkpk}\lvert x_2\rvert}{\mathrm{Pkpk}\lvert x_1\rvert}\right)$
3	$\frac{RMS(x_2)}{RMS(x_1)}$

10.3.1.2 Defect Detection Principles

The pair of receivers placed at the two sides of the transmitter are employed in a differential detection scheme; this compares the strength of the signal received at both sides. If an internal flaw is present on one side of the transmitter or on the other one, the transmitted wave is scattered and the outputs of the receiver pair are unbalanced. Some strength-related metrics are extracted from the received waveforms to generate a "feature supervector"; in practice, this is obtained by combining the "feature vectors" provided by the two receivers. Thesemetrics are listed in Table 10.1, in which the symbols $PkPk(x)$ and $RMS(x)$ denote the *mean peak-to-peak* and *root-mean-square*, respectively, of the quantity x [75].

10.3.1.3 Signal Processing

The received signal needs to be properly processed (and, in particular, denoised) to overcome the inherently low *signal-to-noise ratio* (SNR) of air-coupled piezoelectric sensors [62]. A *multivariate outlier analysis* (MOA) is used to compensate for the natural signal variations expected during the test; the final result is represented by a *damage index* (DI) computed at each test position. Defects are detected by an high value of DI, i.e., by an unusually large discordancy index from the "normal" baseline of the rail. The DI is represented by the Mahalanobis squared distance discordancy metric D_ζ of a MOA [75], which is defined as

$$D_\zeta = (\mathbf{x}_\zeta - \bar{\mathbf{x}})^T \cdot \mathbf{K}^{-1} \cdot (\mathbf{x}_\zeta - \bar{\mathbf{x}}), \tag{10.1}$$

where $\mathbf{x}_\zeta$ is the potential outlier vector, $\bar{\mathbf{x}}$ is the mean vector of the baseline, $\mathbf{K}$ is the covariance matrix of the baseline and $(\cdot)^T$ represents the transpose operator. A new observation is classified as an outlier if the corresponding value of D_ζ is higher than a proper threshold, previously established [74].

To reduce the probability of false alarms due to the occurrence of isolated noise-related high peaks in the DI trace, system redundancy can be exploited [75]. When scanning a defect, the presence of a real crack should result in multiple peaks in the observed trace. Therefore, a peak is detected by the system only when a given number of DI values exceeds the above-mentioned threshold.

10.3.1.4 Reverberation of Airborne Signals Caused by an Acoustic Mismatch

The presence of reverberations of the airborne waves between the transmitter and the top of the excited rail has been found. This phenomenon is caused by the unavoidably high acoustic impedance mismatch between the air medium and the solid boundaries of the rail steel and of the piezoelectric transducer. The strong intensity of the reverberations affects the detection of acoustic waves traveling through the rail steel associated with the presence of rail defects. Although time gating can effectively separate the signal of interest, a lower SNR is observed when the repetition rate (i.e., the frequency of the excitation) is sufficiently high such that the airborne reverberations from a previous excitation overlap with the waves of a new excitation. This limits the test speed to ~1.6 km/h (~2.4 km/h when a slightly worse SNR is acceptable) because of the required spatial resolution. The most effective solution developed for this problem until now consists of inserting a sponge between the transmitter and the rail to attenuate the airborne reverberations. The main drawback of this solution is represented by the fact that the sponge has friction with the rail, and this raises the noise level affecting the received waves as the train speed increases; in practice, a reliable signal detection is possible up to 24 km/h [76]. Some test results are illustrated in Section 10.7.1.

10.3.2 Passive Approach

10.3.2.1 Hardware Configuration

In this case, as illustrated in refs. [80] and [81], the sensing head is made of two receivers, denoted A and B separated by a known distance; both are sensitive only to waves propagating unidirectionally (from left to right in Fig. 10.3). The sensing head is mounted on a beam rigidly connected to the front-axle of the test car; similarly as in the active approach, contactless probing is employed. The sensors are tilted with respect to the rail surface (according to Snell's law) to best capture the leaky surface waves propagating in the railhead. The orientation of receivers ensures directional sensing of the waves excited by the wheels located on only one side of the arrays (front end), with virtually no sensitivity to waves propagating in the opposite direction (i.e., to reflections or waves excited from the wheels located to the other side of the arrays' back end).

Data are continuously recorded for the entire duration of the test [80]. Tachometer logic pulse marking the spatial position of the test car and GPS positioning are recorded, in addition to high-speed camera videos of the tested rail to verify the presence of visible discontinuities when the prototype detects an anomaly [79].

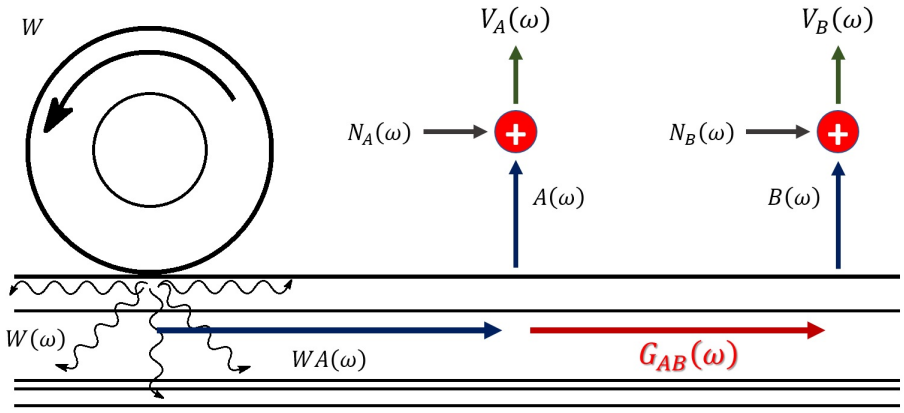


Figure 10.3: Representation of the linear wheel-rail interaction in the frequency domain [80].

10.3.2.2 Defect Detection Principle

In this case, defect detection requires the estimation of the time domain *Green's function* $G_{AB}(t)$ between the two receivers A and B, without prior knowledge of the source excitation spectrum. This function represents the response of the test object measured a B from an impulse excitation at A. If the frequency domain Green's transfer function $G_{AB}(\omega)$ can be estimated, discontinuities in the rail (such as defects) can be detected by means of a procedure similar to that illustrated for the pitch–catch ultrasonic guided-wave approach. In fact, discontinuities can be perceived as a change in the structural impulse response, $G_{AB}(t)$, obtained from $G_{AB}(\omega)$ through an inverse Fourier transform. In terms of defect detection, as long as the reconstructed Green's function is stable during a test run, successful passive structural inspection becomes possible.

The estimation of the Green's function is based on the computation of the ensemble averaged *cross-power spectrum* between the output and the excitation, normalized to a modified ensemble averaged *auto-power spectrum* of the excitation. The following assumptions are made:

- the system is linear and piecewise stationary, so that the statistics of the excitation $W(\omega)$ do not change in the observation interval;
- the noise terms are uncorrelated, have a zero DC component, and tend to zero on average, thus enabling for separating the useful signal component from noise.

The function $G_{AB}(\omega)$ is estimated by computing the deconvolution of the cross and auto power spectra. In fact, the spectra of the responses at A and B can be expressed as

$$V_A(\omega) = W(\omega) \cdot WA(\omega) \cdot A(\omega) + N_A(\omega) \tag{10.2}$$

and

$$V_B(\omega) = W(\omega) \cdot WA(\omega) \cdot G_{AB}(\omega) \cdot B(\omega) + N_B(\omega), \tag{10.3}$$

respectively; here, $V_A(\omega)$ ($V_B(\omega)$) is the response measured at receiver A (B), $W(\omega)$ is the wheel excitation spectrum, $WA(\omega)$ is the transfer function of the rail between the wheel and the transducer at A, $A(\omega)$ ($B(\omega)$) is the frequency response of the receiving sensor A (B), $N_A(\omega)$ ($N_B(\omega)$) is the uncorrelated noise originating from the environment at A (B) (see Fig. 10.3).

Based on eqs. (10.2) and (10.3), it can be shown that [80]

$$\langle \text{Cross Power} \rangle = |W(\omega)|^2 \, |WA(\omega)|^2 \, G_{AB}(\omega) \tag{10.4}$$

and

$$\langle \text{Auto Power} \rangle = |W(\omega)|^2 \, |WA(\omega)|^2, \tag{10.5}$$

in which the symbol $\langle\rangle$ represents the *ensemble average* operator.
By using eqs. (10.4) and (10.5), the frequency Green's Function between the two receivers is computed as

$$\frac{\langle \text{Cross Power} \rangle}{\langle \text{Auto Power} \rangle} = \frac{|W(\omega)|^2 \, |WA(\omega)|^2 \, G_{AB}(\omega)}{|W(\omega)|^2 \, |WA(\omega)|^2} = G_{AB}(\omega). \tag{10.6}$$

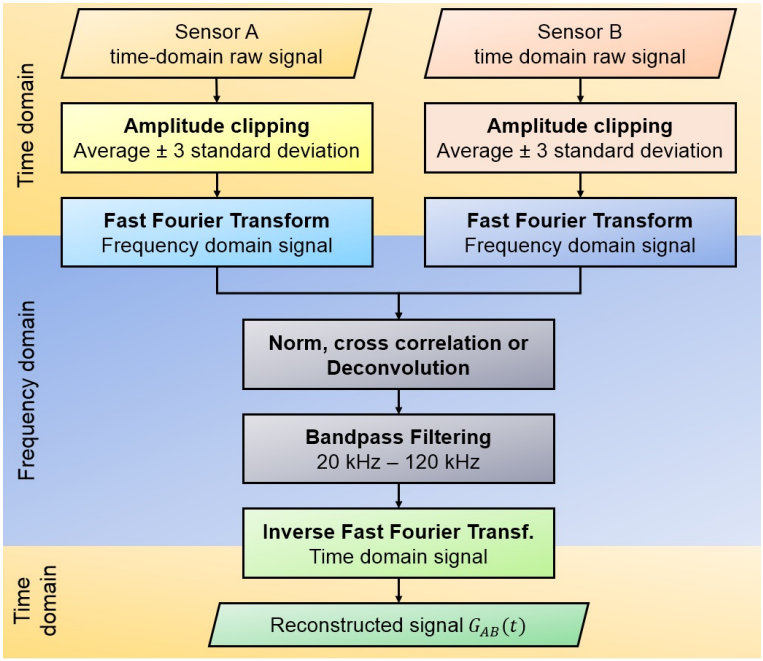


Figure 10.4: Signal processing steps accomplished in the passive reconstruction of the transfer function $G_{AB}(\omega)$ [79].

Then, the time domain Green's function can be evaluated by taking the inverse Fourier transform from the left-hand side of the last equation, i.e., as

$$G_{AB}(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} G_{AB}(\omega)\, e^{j\omega t}\, d\omega. \tag{10.7}$$

The last integral is computed in a approximate fashion through an *inverse fast Fourier transform* (IFFT).

10.3.2.3 Data Processing

The employed data processing is sketched in Fig. 10.4. In each run, the recordings from receivers A and B are first amplitude-clipped to within the ± 3 standard deviations to mitigate the effects of isolated spikes in the passive reconstruction of the impulse response $G_{AB}(t)$. Then, the resulting signals undergo *fast Fourier transform* (FFT) processing, so that cross and auto power spectra and the transfer function $G_{AB}(\omega)$ can be computed. Finally, the time domain Green's function $G_{AB}(t)$ is evaluated by

1. averaging the function $G_{AB}(\omega)$ over four sensor pairs, and
2. accomplishing band-pass filtering and IFFT processing.

An outlier analysis is implemented in order to compute the DI related to the strength of the reconstructed transfer function; this issue has already been discussed in Section 10.3.1.3, when describing the active method for rail defect detection. The computation of the DI allows for normalizing the available data, thus mitigating the normal (baseline) data variability occurring in each run [79, 80].

10.3.2.4 Trade-Offs

The impulse response $G_{AB}(t)$ is the result of the constructive interference of the wave modes continuously generated by the wheel excitation and propagating in the rail between the receivers. The constructive interference and, hence, the rate of convergence of the passively reconstructed transfer function (or, equivalently, its SNR) benefit from signal averaging. For this reason, time windows involving long recordings are preferred. Note, however, that, since both the transmitters and the receivers are moving along the test specimen (rail), the stationarity of the reconstructed transfer function can be guaranteed only in a fixed position. For this reason, a good trade-off needs to be achieved between the long recording time required by the averaging process and the stationarity (related to spatial localization) of the transfer function that calls for shorter observations. This explains why the test speed has to be properly selected: the higher the velocity is, the shorter the recording time is, ensuring a sufficiently accurate spatial localization [79].

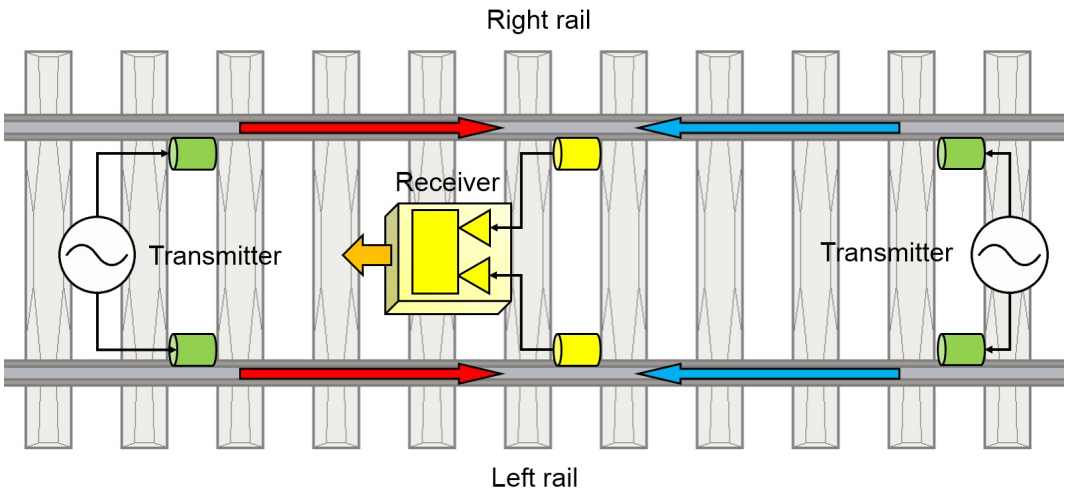


Figure 10.5: Block diagram of the UBRD [84].

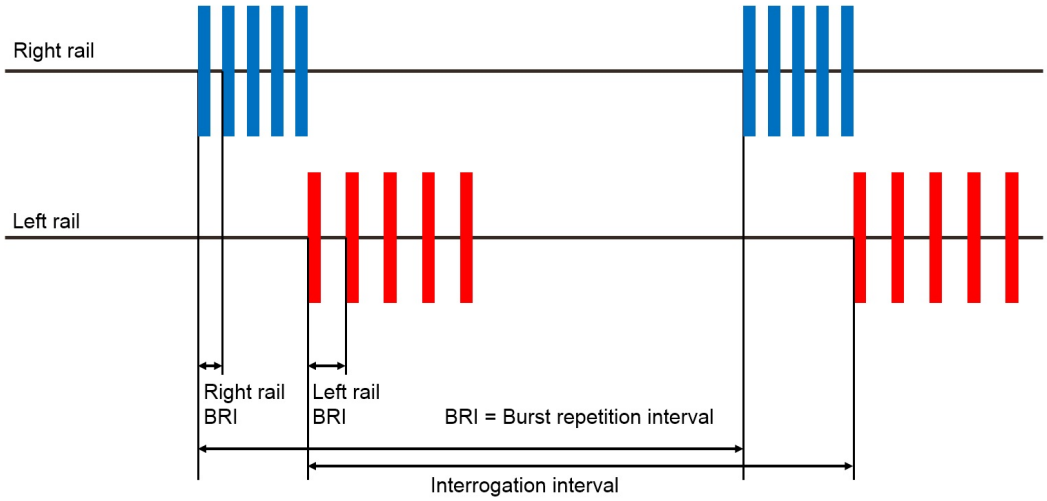


Figure 10.6: Burst injection scheme [82].

Some test results are illustrated in Section 10.7.2.

10.4 Land-Based Systems Implementation - Premise: Ultrasonic Broken Rail Detector (UBRD)

In this section, we focus on the implementation of land-based systems. Since the topic is wide, the description of these systems is provided in three consecutive sections, each focusing on specific technical issues, as already mentioned at the end Section 10.2.

10.4.1 UBRD Hardware Configuration and Generated Signals

Transmitters and receivers of the UBRD system are interleaved, as illustrated in Fig. 10.5. Every receiver is able to acquire signals coming from both directions of the rail; to identify the exact orientation, the transmitted signals consist of different burst sequences depending on the transmitter. In fact, as it can be easily inferred from Fig. 10.6, each transmitted sequence consists of a number of pulses spaced by a given *burst repetition interval* (BRI) and is repeated at a specific *interrogation interval* (II). This avoids the overlap at one receiver of burst trains coming from adjacent transmitters for extended periods. For such reason, receivers are also individually configured to recognize specific pulses as arriving from either the up or down direction depending on the settings of adjacent transmitters. Note also that each signal does not reach a receiver far from its transmitter because of the attenuation due to the propagation medium [82, 84].

10.4.2 Defect Detection Principle and Employed Signal Processing Method

The transducers of each transmitter generate an acoustic signal that propagates through the rail; this signal travels in both directions and can reach receivers located at a significant distance from the transmitter. The received signals are filtered, amplified, and processed. Rail integrity between the transmitter and each receiver is confirmed as long as an acceptable signal is received. If a clean break occurs in the considered stretch of rail, the corresponding receiver will no longer be reached by the transmitted signal and will generate an alarm [82, 84]. Receivers detect valid signals on the basis of specific criteria concerning signal frequency, burst length, burst repetition interval, and burst train continuity.

Severe continuous noise (such as that generated by an approaching train) at a receiver will affect signal detection, regardless of the efficiency of the employed detector. Under such circumstances, the receiver will generate the message "train in section", will stop processing its input signals, and will remain idle until the noise intensity decreases; then, it will return to its normal conditions.

This system is commercialized under the name "RailSonic - ultrasonic broken rail detector". The system is not classified "Fail Safe", although in ref. [82] it is claimed that many fail safe principles have been incorporated in order to correctly detect most of equipment failures; this allows for correctly identifying these events and are, at the same time, to avoid false alarms.

10.4.3 UBRD Updates

To update the UBRD, a new piezoelectric transducer has been developed, thanks to numerical modeling and measurement techniques. This improvement and the use of new digital signal processing techniques have been incorporated into a new version of the UBRD system that allows for doubling the distance between transmit and receive stations [83].

Yuan et al. have proposed a method to integrate the UBRD system with a *train detection* (TD) system [85]; the resulting system is able to distinguish the following three different track conditions; *free*, *busy* and *broken rail*. This is made possible by extracting, from the ultrasonic guided waves, the following three features:

- *the RMS value* and *the energy* in the time-domain;
- *the frequency component with the highest amplitude* in the frequency domain.

In addition, *temporal* and *spatial dependencies* of the signals are taken into consideration. High-level signal processing techniques, such as deep learning algorithms, are also employed to deal with varying environmental situations and conditions. The suitability of this approach has been assessed through experimental tests.

In ref. [91] the same authors as ref. [85] have proposed to exploit the *variational mode decomposition* (VMD) algorithm to de-noise and reconstruct ultrasonic guided wave signals. The VMD method is a quasi-orthogonal signal decomposition technique for non-recursively decomposing a multi-component signal into a finite number of compactly band-limited *intrinsic mode functions* (IMFs). The VMD algorithm is used to decompose ultrasonic guided wave signals into the fundamental, harmonics, interharmonics and non-stationary disturbances, and can be used in conjunction with the method developed in ref. [85].

10.5 Land-Based Systems Implementation - Evolution: Early Rail Defect Detection Capability

An experimental system able to monitor an operational rail track by means of ultrasonic guided waves has been proposed in ref. [102]. In this section, the prototype and the methodologies adopted to process and analyze the available data are illustrated. Moreover, various technical problems and the solutions developed to solve them are briefly described.

10.5.1 Introduction

Numerous types of defects appear in rail tracks, and multiple parameters affect the prediction of crack growth rates and of the defect size at failure [87]. In principle, the *baseline subtraction* method could be used to detect the insurgence of a defect over time. Baseline subtraction is based on the idea of feeding a system in a given condition with a known excitation, measuring its response, and storing it as a baseline. Each and every modification in the system reflects in a change in its response to the same excitation. Therefore, in principle, the occurrence of any change potentially due to a defect can be detected by comparing each new response to the baseline. In practice, however, it is not easy to establish if such a change is caused by a defect or by a variation of other parameters affecting the system response to the considered excitation. In fact, it should be always

kept in mind that, in harsh environments like rail tracks, any variation observed in the propagation conditions of ultrasonic signals can also originate from changes in *environmental and operating conditions* (EOCs); these conditions include, for instance, temperature, train passing, or maintenance operations [83].

The continuous monitoring of a rail track can provide better performance results in terms of defect detection probability than a single inspection, though it is influenced by the EOC. In this case, the most important technical challenges are:

- the development of a permanently installed UGW-based monitoring system for rail defect detection at a reasonable cost per kilometer of track;
- the development of a defect detection technique able to reliably operate in presence of real EOCs.

Of course, if a suitable facility is unavailable, tests can be performed on operative tracks. The problem is that damaged sections of rails are removed as soon as defects are detected; furthermore, rails must not be damaged or altered. Consequently, it is impossible to test the available prototypes on real cracks: their performance is assessed by using welds or glued masses, both acting as reflectors for guided waves [87].

Different problems must be addressed in order to achieve an effective automatic monitoring system for rail damage detection. In particular, the following specific technical issues should be investigated:

- the behavior of a defect growing over time;
- the influence of the selected transducers on the quality of the received signal;
- the influence of the changes in rail properties on the propagation of guided waves and the parameterization of these changes;
- the influence of time varying EOCs on wave propagation and the methods for compensating for these variations;
- the identification of the defect detection algorithm and the use of proper methods for assessing its performance.

10.5.2 Monitoring Set-Up

In this section are analyzed in detail the hardware and software of the system proposed in ref. [102].

10.5.2.1 Selected Modes for Guided Wave Propagation

Numerical simulations based on the *semi-analytical finite element algorithm* (SAFE) have allowed for identifying a guided wave mode suited for the detection of defects in rails. This mode propagates with relatively small attenuation and dispersion, and it is strongly reflected by transverse defects in the railhead. It can be excited by a transducer attached under the railhead. The identified mode, used for both transmission and reception, is a symmetric mode that has motion in the vertical axis while propagating along the longitudinal direction moreover with the energy mainly concentrated in the railhead [102].

10.5.2.2 Monitoring System Hardware and Software

Two piezoelectric transducers, forming an array and attached under the railhead, are employed to perform pulse-echo measurements [87]; the sandwich piezoelectric transducers have been designed to effectively excite the selected mode of propagation around 35 kHz [102]. The distance between the axes of the transducers has to be equal to one quarter of the wavelength of the excitation signal.

Two specific technical problems have been faced in system design [89]. The first problem is represented by the slightly different resonant frequencies of the two transducers; this is due to the tolerances characterizing the manufacturing process. Moreover, the transducer performance may change differently over time and, in the presence of temperature variations, thus causing asymmetric changes in the measured reflections and replicas. A possible solution to this problem is to independently scale the reflections from the positive direction and from the negative one, so that the difference in their amplitudes can be compensated for; this requires the estimation of the attenuation in both directions. The second problem is represented by the presence of multiple replicas originating from reflections; phased array processing can be exploited to solve this problem.

The experiments were performed on a section of tracks consisting of rails joined by aluminothermic welds. Transducers and the artificial defects mentioned in Section 10.5.2.3 have been placed far from the joints.

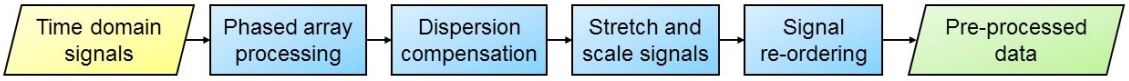


Figure 10.7: Pre-processing steps of the acquired signals [102].

10.5.2.3 Behavior of a Defect During Time

The behavior of a defect over time can be analyzed by using a deteriorating defect emulator, such as a glued mass. The reflection of a glued mass deteriorates quickly over time, thanks to the stresses caused by trains and to the progressive corrosion of the rail under the glue. A monotonically increasing defect, similar to a growing crack, can be simulated if the reflected signals recorded during the time window starting at the installation of the defect emulator and ending at its detachment are time reversed [87]. The size of the mass representing the crack has to be selected to provide a realistic reflection, such as a small transverse crack in the railhead. Reflections from rail welds can provide a useful reference during testing; furthermore, if these welds can be detected, it is believed that a crack can be found well before it reaches a critical size [103].

10.5.3 Signal Pre-Processing

The employed excitation signal is a Hanning-windowed tone burst signal whose center frequency is equal to 35 kHz. The measurements are acquired by exciting one piezoelectric transducer, acquiring the response of both transducers of the array, and then repeating the process by exciting the other transducer. This cycle is repeated k times, and the average of the k acquired measurements is stored as a new signal in the measurement set, which consists of four time domain signals.

Noise from passing trains has to be avoided since it would overwhelm the useful signal. For this reason, the system first senses the presence of train noise and proceeds with the acquisition only in its absence; otherwise, it delays its acquisition. In the observation interval m, distinct sets of measurements are recorded.

Before defect detection or monitoring, the acquired data must be pre-processed to compensate for unwanted effects. The pre-processing steps are illustrated in Fig. 10.7 and described in the following sub-sections. Data pre-processing involves

1. *phased array processing*;
2. *dispersion compensation*;
3. *signal stretching and scaling*;
4. *signal reordering*.

These steps are required to discriminate the direction of the reflections, to reduce the influence of dispersion and to compensate for some of the changes occurring in EOC, respectively. Signal reordering accounts for the use of a mass as a defect emulator: the mass, in fact, behaves in a reversed time order with respect to a crack (see Section 10.5.2.3).

10.5.3.1 Acquired Signals

The impact of noise in the received signal can be reduced by band-pass filtering, making easier the task of detecting the reflections in the steps to follow. The center frequency and bandwidth of the filter can be set to be the same of the excitation signal. During pre-processing, the signals are converted to the frequency domain.

10.5.3.2 Phased Array Processing

An array of transducers allows for exciting a specific mode in a desired direction of propagation, even if the array is composed of two transducers only. This result is obtained by feeding both transducers with excitation signals characterized by a small difference in delay or phase. The phase difference depends on the wavelength of the propagation mode selected at the transmission frequency and on the axial distance between the two transducers. The same idea can be exploited at the receive side; in fact, if a phase shift is applied to the signal acquired by a transducer, the energy associated with the desired mode and originating from the desired direction can be captured. It is worth noting that a perfect cancellation is impossible, since other modes characterized by different wavelengths are transmitted and received in both directions. This problem can be mitigated by increasing the overall number of transducers forming the array.

Phased array processing can be applied to the full matrix of signals acquired from an array of transducers, as explained by Wilcox in ref. [110]. The acquired time signals are converted to the frequency domain and, after processing, are converted back to the time domain. When the first (second) transducer is excited, the responses acquired by both transducers are stored in the first (second) column of the *measured displacement matrix*:

$$\mathbf{V}(\omega) = \begin{bmatrix} V_{11}(\omega) & V_{12}(\omega) \\ V_{21}(\omega) & V_{22}(\omega) \end{bmatrix}; \tag{10.8}$$

here, $V_{ij}(\omega)$ is the spectrum of the response of the i-th transducer of the array (with $i = 1$ and 2, if the array consists of two transducers only), in response to the excitation injected through the j-th transducer (with $j = 1$ and 2, if the array consists of two transducers only).

The analysis of different combinations of transmitted and received propagation modes requires the computation of mode shape matrices. Each column of the mode-shape matrix refers to the displacement of a mode at each of the transducer locations. The mode shape matrices for an array of transducers in a 1D waveguide (such as a rail) contains both mode shape and phase information according to the axial position of the associated transducer, for each mode in every possible direction [110]. Since it may happen that the modes selected for receive processing are different from those chosen for transmit processing, two mode shape matrices are required, one for reception and the other for transmission (denoted $\mathbf{R}(\omega)$ and $\mathbf{T}(\omega)$, respectively). If the same modes are selected for reception and transmission, then $\mathbf{R}(\omega) = \mathbf{T}(\omega)$. The coefficients of these matrices are computed through a SAFE analysis of the rail.

If the mode shape matrices $\mathbf{R}(\omega)$ and $\mathbf{T}(\omega)$, and the measured displacement matrix $\mathbf{V}(\omega)$ (10.8) are known, the contributions of the different modes to the received signal can be computed. The spectrum $\alpha(\omega)$ of the overall response is evaluated as

$$\alpha(\omega) = \mathbf{R}(\omega)^{-1}\, \mathbf{V}(\omega)\, \mathbf{T}(\omega)^{*^{-1}} = \begin{bmatrix} \alpha_{r_1t_1}(\omega) & \alpha_{r_1t_2}(\omega) \\ \alpha_{r_2t_1}(\omega) & \alpha_{r_2t_2}(\omega) \end{bmatrix}; \tag{10.9}$$

here, the spectrum $\alpha_{r_1t_1}(\omega)$ ($\alpha_{r_2t_2}(\omega)$) describes the transmission and reception in the forward (backward) direction, whereas, $\alpha_{r_2t_1}(\omega)$ ($\alpha_{r_1t_2}(\omega)$) the transmission in the forward (backward) direction and the reception in the backward (forward) direction. Note that

- $\alpha_{r_1t_1}(\omega)$ is ideally equal to zero;
- the presence of reflections from features[1] in the forward (backward) direction are inferred from $\alpha_{r_2t_1}(\omega)$ ($\alpha_{r_1t_2}(\omega)$);
- proper calibration factors can be introduced to compensate for the different sensitivities of the two transducers (see Paragraph 10.5.2.2).

10.5.3.3 Dispersion Compensation

Even if a mode characterized by a small dispersion is selected, the effects of this phenomenon can be appreciable in the presence of a long propagation range. Dispersion compensation can be performed by exploiting the algorithm developed by Wilcox [104] and the dispersion data evaluated by means of the SAFE method in the considered scenario. The compensation process allows also for converting the received signals from the time domain to the distance domain. The correctness of this conversion can be verified by comparing the location of the welds on field and that inferred from the reflections appearing in the compensated signals.

10.5.3.4 Signal Stretching and Scaling

Time varying EOCs influence the acquired signals, even in a defect-free rail: for this reason, compensation techniques not affecting the influence of defects on the processed signal are required. The group velocities of guided wave modes are influenced by temperature variations; in fact, the reflections appear to be shifted in time (or distance) if measurements are acquired at a different temperature. This effect can be compensated for by stretching the dispersion-compensated signals, in time or in distance domain [102].Since dispersion compensation converts time domain signals in the distance domain, stretching in [102] has been performed in the last domain.

The influence of temperature variations cannot be represented only through a phase shift, since the wave envelope can also be distorted. A piecewise linear stretch can be exploited to compensate for the envelope distortion; in doing so, the knowledge about the location of the welds appearing along the considered span of rails is used. If this method is adopted, the distorted envelope is stretched in a way that the reflection peaks corresponding to the welds are aligned with their known locations. In this phase, however, the phase

[1] Defects, welds and any other element reflecting back guided waves towards the array of transducers.

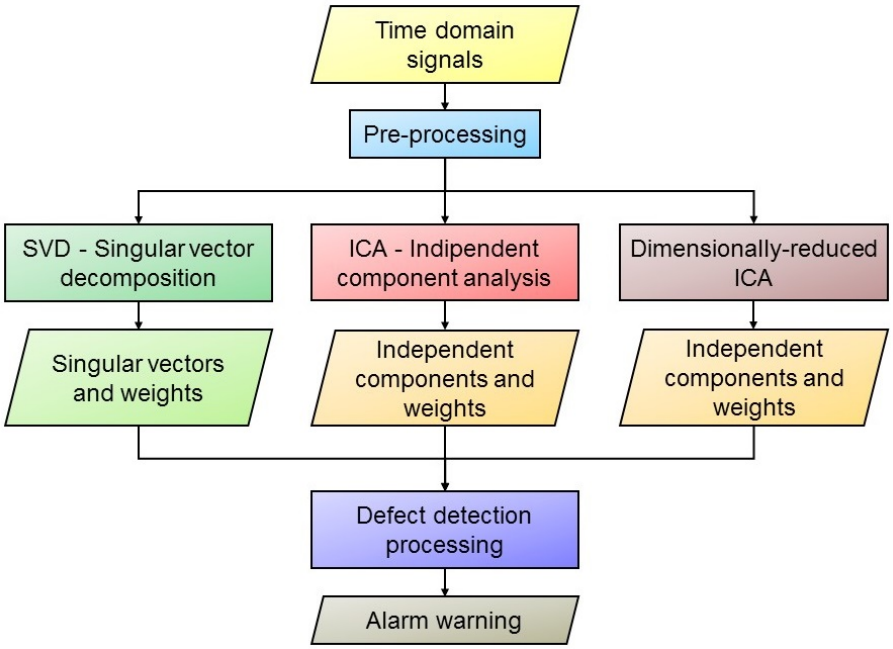


Figure 10.8: Processing steps in defect detection.

information is discarded. The received signal might also be affected by variations in their amplitude during the monitoring period. These can originate from a change in the sensitivity of transmitters or in the attenuation of the rail with temperature, or from the rail itself gradually sinking into the ballast. To compensate for this effect, signal scaling based on the peaks originating from weld reflection can be performed. Guided waves are attenuated during propagation; for this reason, distant weld reflections appear smaller than those due to welds closer to the transducer array. To compensate for this effect, an energy-based normalization can be applied to each signal envelope by adopting the technique proposed by Moustakidis et al. in ref. [105]. The energy attenuation employed in normalization is computed by exploiting a moving average technique.

The phase of the compensated signals is discarded, and only their envelope remains. Unluckily, the envelope is affected by a large *direct current* (DC) component, affecting the performance of the defect detection algorithms (and, in particular, of the ICA; see Section 10.5.4). To reintroduce phase information, the wave envelopes of scaled signals are multiplied by a single-frequency sine wave. However, the employed method for defect detection is based only on the analysis of signal envelopes [102].

10.5.3.5 Signal Reordering to Simulate the Monotonic Growth of a Defect

The reflection of a glued mass deteriorates quickly over time because of the stresses caused by trains and of the progressive corrosion of the rail under the glue. A monotonically increasing defect, similar to a growing crack, can be simulated if the reflected signals recorded during the time window starting at the installation of the defect emulator and ending at its detachment are time reversed [87]. A steady decrease in the amplitude of the reflection due to the artificial defect amplitude is expected, with some fluctuation due to temperature changes. Test results illustrated in refs. [87, 102] have proven that the reflection peak due to an artificial defect does not have a simple correlation with temperature; no satisfactory explanation for this behavior has been provided, although it is known that measurements may have been affected by the complicated resonant conditions producing the large reflections [87].

10.5.4 Defect Detection

The processing steps accomplished in defect detection are summarized in Fig. 10.8. Because of the EOCs-induced variations in the propagation conditions of UGW signals, the simple baseline subtraction is not effective [102]. Some promising results in structural health monitoring of plates and pipes have been obtained by means of two unsupervised machine learning techniques, namely *singular value decomposition* (SVD) [106] and *independent component analysis* (ICA) [107]. These techniques outperform baseline subtraction in the detection of simulated defect signatures superimposed on measured data from a pipe subject to time varying EOCs [95]. Therefore, it makes sense to apply these techniques to rails too.

Ideally, if EOCs are perfectly compensated for, the signal obtained from pre-processing can be seen as the additive contribution of two or more distinct components:

- one resembling the baseline ultrasonic signal containing weld reflections, with a constant weight over the duration of the experiment;
- other components resembling the ultrasonic signature of the defect, with a weight increasing over time and corresponding to the deterioration of the defect [87].

10.5.4.1 Data-Driven Defect Detection Techniques

The pre-processed measurements are stored in the $m \times n$ matrix $\mathbf{X}$; here m represents the overall number of performed measurements over the monitoring time, whereas n is the number of samples acquired in each measurement. The matrix $\mathbf{X}$ can be factored as

$$\mathbf{X} = \mathbf{AC}. \tag{10.10}$$

The matrix $\mathbf{C}$ contains the additive components of the acquired signal (baseline and defect signatures) in the distance domain, whereas the matrix $\mathbf{A}$ is the corresponding weights in the observation interval.

The singular value decomposition (SVD) can be exploited to evaluate the factorization expressed by eq. (10.10) [106]. In fact, applying this technique to the real data matrix $\mathbf{X}$ produces

$$\mathbf{X} = \mathbf{USV}^T, \tag{10.11}$$

where $\mathbf{U}$ and $\mathbf{V}$ are the left and right singular vector matrices, respectively, and $\mathbf{S}$ is the diagonal singular value matrix. The columns of the matrix $\mathbf{U}$ are related to the slow signal behavior evolution over the monitoring period, while the columns of the matrix $\mathbf{V}$ are related to the fast signal behavior evolution over the same observation interval. $\mathbf{S}$ allows for quantifying the amount of information contained in the associated left and right singular vectors. In fact, the main diagonal of $\mathbf{S}$ collects the singular values, sorted in descending order. The size of the given dataset can be reduced by discarding small singular vectors, since these convey a negligible amount of information.

The independent component analysis (ICA) technique is employed to extract the additive components of a multivariable signal and analyze their relative trends. More specifically, this technique allows for decomposing the matrix $\mathbf{X}$ into a given number of independent components forming the columns of the matrix $\mathbf{C}$ in eq. 10.10; such components have minimal statistical correlation. In addition, the matrix $\mathbf{A}$ contains the weights of the independent components. In our case, the ICA method can be used to separate the contributions of different sources. In fact, we are interested in separating the contribution of the defect signatures from the system background, including the baseline signal and EOC-related variations. To efficiently implement ICA, the iterative FastICA algorithm can be used. This algorithm is fed by a normalized version of the data set $\mathbf{X}$, and computes the principal values and associated components of the input data by calculating the eigenvalues and eigenvectors, respectively, of the covariance matrix $\mathbf{XX}^T$. Repeatable results are generated if the initial value is selected for the vector of weights.

If the presence of data overfitting is detected while using the ICA algorithm, dimensionality reduction techniques, like the *principal component analysis* (PCA) and can be employed to mitigate this problem [88]. The combination of the ICA algorithm with principal component analysis (PCA) is called *ICA with dimension reduction* in ref. [87]. During ICA processing, the eigenvalues (principal values) are ordered according to their magnitude, and only the largest values and their associated components are kept.

Some test results are shown in Section 10.7.4.

10.5.5 Adaptive SAFE Model for Rail Parameter Estimation

Temperature is not the only EOC variation affecting signal propagation. For instance, the propagation environment of ultrasonic guided waves in rails is modified by any change occurring in rail geometry and in the parameters of the employed materials. Compensating for these additional effects allows for achieving better performance.

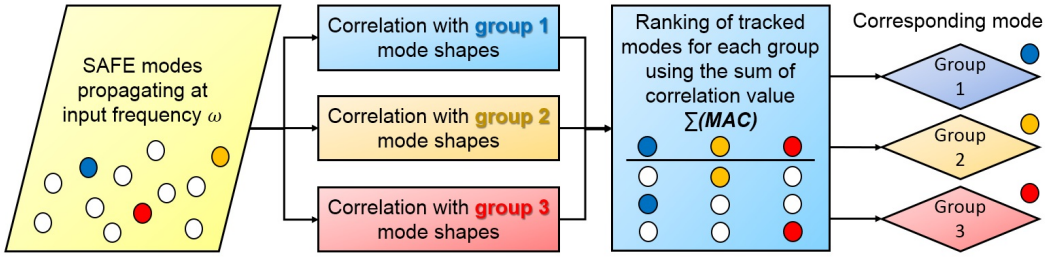


Figure 10.10: Representation of the mode shape tracking process that employs selected mode shape groups [90].

10.5.5.1 Problem Statement

Since the performance of any monitoring system that makes use of ultrasonic guided waves is influenced by some characteristics of the propagating environment, it would be useful to estimate these characteristics from the signals acquired by the system itself. In fact, the dispersion curves of a rail depend on the properties of its material, as well as on its geometry, neither of which are known with a sufficient level of accuracy. Moreover, the rail geometry changes over time because of wear and regular maintenance operations (including rail grinding of the crown) [86].

10.5.5.2 Possible Solutions

Material and geometric parameters are implicit in the computation of the dispersion curves. For this reason, inferring these parameters from such curves requires solving an inverse problem through an iterative approach. Solving this inverse problem becomes easier if the corresponding forward problem, consisting of the computation of dispersion curves, can be solved efficiently and if the number of rail parameters to be estimated can be reduced.

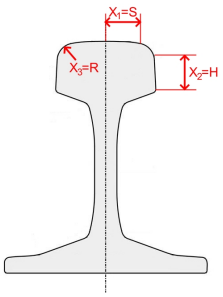


Figure 10.9: Representation of the parameters useful to identify the presence of a worn rail.

The use of the *semi-analytical finite element* (SAFE) method for efficiently solving the forward problem is investigated in refs. [86] and [90]. In these manuscripts, a set of parameters describing the geometry and material properties of a worn rail has been determined. This set is employed to develop a SAFE model for computing the propagation characteristics of worn rails. Wear and grinding of the railhead can be represented through the three geometric parameters shown in Fig. 10.9. An adaptive mesh model has also been developed by means of parametric equations to represent material removal. This allows an efficient automatic modification of the geometry and the mesh of the considered rail.

The use of the SAFE method requires the knowledge of different properties of the employed material, such as its *elastic modulus* E and its *density* ρ, or its *Poisson's ratio* ν. The evaluation of the dispersion curves of a waveguide through this method requires solving an eigenvalue problem. Material properties are usually unknown to an acceptable accuracy, and their value is dependent on environmental conditions. In the case of an isotopic elastic medium, it is shown that E and ρ can be combined in a single parameter, namely the *longitudinal speed of sound in the medium*, $c = \sqrt{E/\rho}$. Therefore, only two parameters, c and ν, are required instead of three (i.e., E, ρ, and ν): this allows to save computation time. Computational efficiency can be further improved by solving the eigenvalue problem in terms of the parameter $\beta = \omega/c$, i.e. of the ratio between the frequency ω and the longitudinal speed of sound c. Thus, only a single set of dispersion curves needs to be computed for each value of the couple (c, ν): if c changes, but ν remains constant, the dispersion curves need only to be scaled according to the variation of the parameters β.

Experimental results have proven that various propagating modes could be detected on field. This has allowed for achieving a proper tracking of mode shapes computed through the SAFE method by using different material and rail geometric properties. The *modal assurance criterion* (MAC) has been adopted to accomplish the tracking task. More specifically, the mode shapes produced by the SAFE method in the presence of different input parameters have been correlated with three reference groups of the mode shapes identified manually on the basis of field measurements, as illustrated in Fig. 10.10 [90].

Finally, it is worth mentioning the technique developed in ref. [90] allows for identifying which set of simulated dispersion curves best fits the experimental spectrograms computed in the absence of prior knowledge about the distance of the reflectors from the transducer. Experimentation results on the field have proven that this technique performs well.

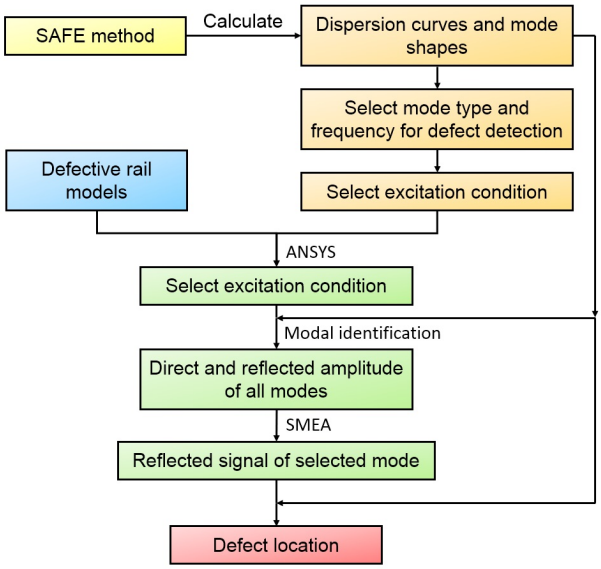


Figure 10.11: Flow chart describing the defect location algorithm described in ref. [94].

10.5.5.3 Conclusions

The estimation technique illustrated above has been developed specifically for the considered application; hence, the involved inverse problem appears to be well-posed. Further research work is required to establish which conditions need to be satisfied to ensure that this technique works, or to determine when the associated inverse problem becomes ill-posed. The inverse problem has been solved by simply finding the SAFE model that provides the best fit in the sample space. The use of a response surface to interpolate the sample space should also be investigated.

10.6 Land-Based Systems Implementation - Other Projects

Other land-based systems and products based on similar working principles as RailSonic by IMT-CSIR have been or are being developed. A brief introduction to these systems is provided in this section.

10.6.1 RailAcoustic by Enekom

The Broken Rail Detection System - RailAcoustic by ENEKOM is a system based on vibrations and not on ultrasonic guided waves. In ref. [93] it is claimed that the operation of this system is exclusively based on electro-mechanical methods. A transmitter injects a vibration into the rail; this excitation is detected simultaneously by a receiver near to the transmitter and by a second receiver placed 2 km apart. The received signal is de-noised thanks to a synchronization signal, and the status of the rail is inferred from the difference between transmitted and received vibration signals. The receiver near to the transmitter allows the system to check if the excitation signal matches some specifications. Enekom claims that:

- the modules of this system can be installed or dismantled without any damage to the involved rails;
- this system works reliably and responsively under all weather conditions.

Successful performance testing has been done on the Konya-Ankara high-speed railway.

10.6.2 Technical Contributions Provided by the Beijing Jiaotong University

A research group working at the Beijing Jiaotong University is investigating the propagation characteristics of rails and, in particular, the most suitable modes for detecting their defects [99, 100, 101]. They do so by analyzing the diagrams representing the strain energy distribution and the phase velocity dispersion curves of different propagation modes. Moreover, the researchers of the above-mentioned institution are trying to discover the best way to excite a specific mode or, alternatively, the best way to detect a given mode in a multi-mode scenario. As a matter of fact, their investigation has led to the development of a method for locating rail defects

[94]; this method is called the single modal extraction algorithm (SMEA) and is briefly described in the next paragraph.

10.6.2.1 Single Modal Extraction Algorithm

The single modal extraction algorithm is described in ref. [94]. It relies on the observation that:

- the vibration displacement measured at any point on the rail is the superposition of the contributions of all the propagation modes at that point;
- the vibration displacement due to each mode at any point can be extracted from the total displacement of the point.

A flow chart describing this method for defect location is shown in Fig. 10.11. The principles on which this method is based can be summarized as follows. The primary task of defect location is to select the mode, the frequency, and the excitation conditions under which defect detection can be accomplished. The excitation response of the rail is analysed through ANSYS[2], that allows to simulate a three-dimensional model of the rail which also includes some defects. This allows to evaluate the vibration displacements of a series of points on the rail, so providing a realistic representation of the signals observed in an experiment. Information about the reflected signals corresponding to the selected modes are provided by the results of the modal identification and SMEA. Defect location is based on the estimation of the group velocity and propagation time of the reflected modes.

10.6.3 Technical Contributions Provided by the Xi'an University of Technology

A research group working at the Xi'an University of Technology has provided relevant contributions to the development of the electronics employed to drive the piezoelectric transducers that excite ultrasonic guided waves in rails. Their specific contributions include:

- The design of an high-voltage pulser achieving an improved transmission efficiency [98].
- The development of a tracking method able to estimate the optimal excitation frequency, i.e. the one maximizing the received signal power [97].
- The use of Baker coded UGW signals for enhancing the SNR of the received signal and the development of an adaptive peak detection algorithm [108].

As far as the last point is concerned, it is worth mentioning that Baker coding, albeit simple, is characterized by a autocorrelation with lower sidelobes than those observed in the autocorrelation of binary codes having the same length; moreover, its simplicity enables minimizing hardware complexity at the transmit side. It is also worth mentioning that the proposed peak detection algorithm offers the advantage of a low energy consumption and achieves better performance than other more sophisticated techniques, like those based on the discrete wavelet transform [109].

10.7 Performance Analysis

The target of a rail diagnostics system is to detect most of the defects affecting rails, while minimizing the number of false alarms [76]. Its performance can be assessed through the so called *receiver operating characteristic curves* (ROC), which represent the trade-off between the *probability of detection* (PD) and the *probability of false alarm* (PFA) achievable by varying the threshold level adopted in defect detection. A related performance indicator is the *area under the receiver operating characteristic curve* (AUC), that provides an overall indication of the "goodness" of detection. In fact, AUC = 1 means that perfect detection is achieved for the given threshold (i.e., PD = 1 and PFA = 0); conversely, if AUC = 0, detection is jeopardized by false alarms, so that PD = 0 and PFA = 1.

In this section, various numerical results referring to the on-board and land-based systems described in the previous sections are illustrated.

10.7.1 On-Board Active System: Performance Analysis

Some representative DI traces acquired along a test track and at various testing speeds by means of the prototype of on-board active defect detection system described in Paragraph 10.3.1 are illustrated in Fig. 10.12; note that

[2] ANSYS: Engineering simulation software.

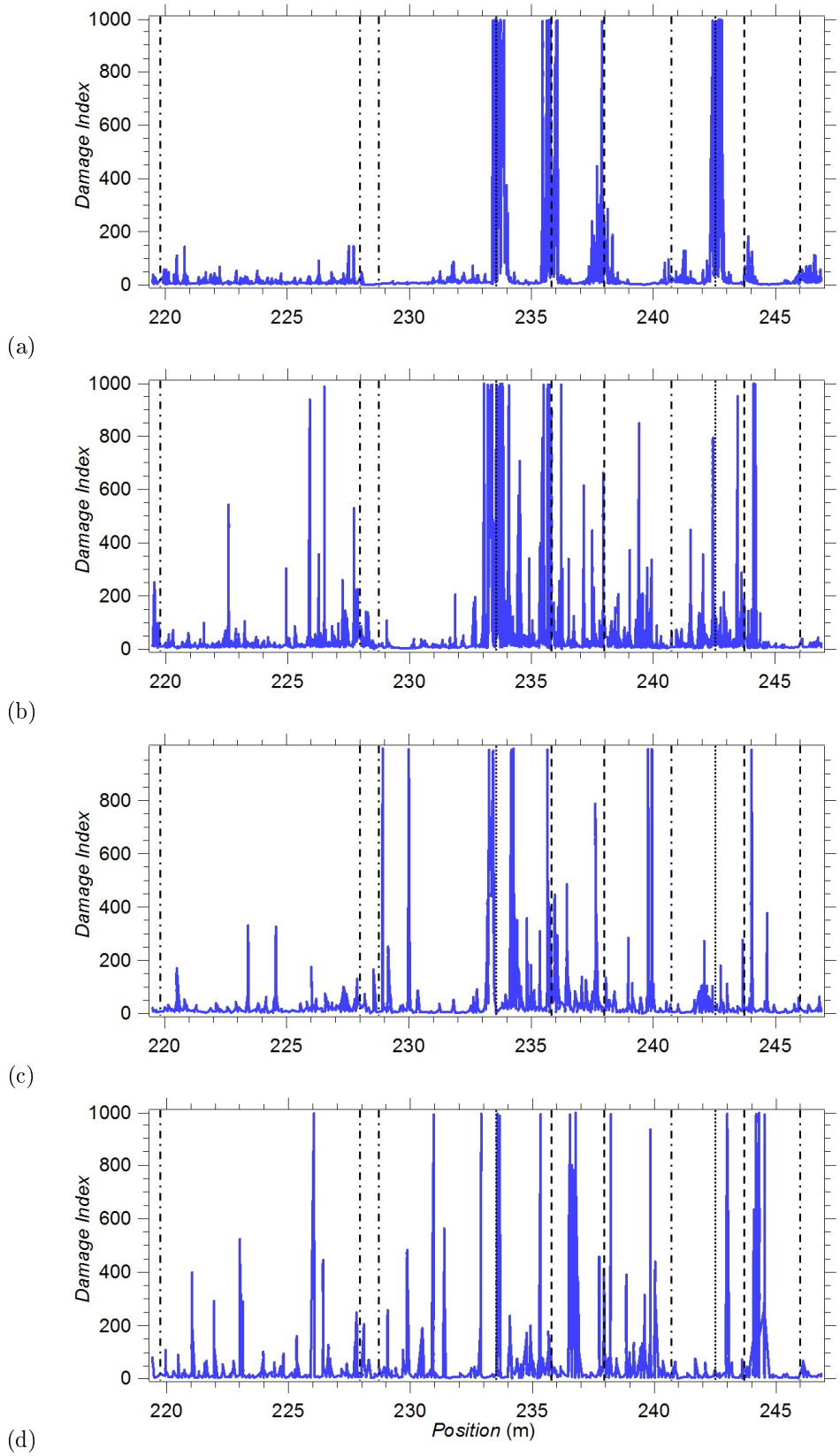


Figure 10.12: Example of damage index traces collected during tests (dashed lines = locations of defects; dashed-dotted lines = locations of welds; dotted lines = locations of joints). Testing speeds: (a) 1.6 km/h; (b) 8 km/h; (c) 16 km/h; (d) 24 km/h.

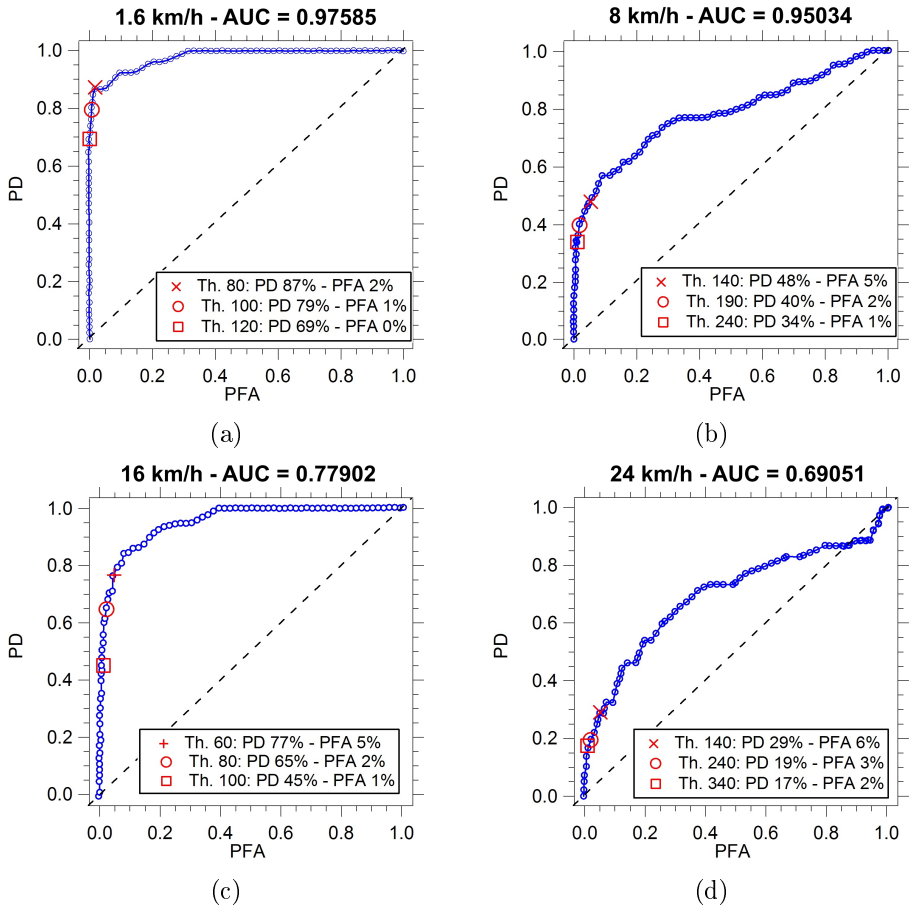


Figure 10.13: Cumulative ROC curves for runs executed at four different speeds: (a) 1.6 km/h; (b) 8 km/h; (c) 16 km/h; (d) 24 km/h. The results associated with three different values of the detection threshold are identified by specific labels.

the SNR of the DI traces degrades as the train speed increases. The plots of Fig. 10.12 are generated on the basis of the data available in ref. [76].

The performance of the above mentioned prototype, expressed in terms of ROC, are illustrated in Fig. 10.13. The plots of Fig. 10.13 are generated on the basis of the data available in ref. [76].

The probabilities of detection computed for four different PFAs on the basis of the cumulative curves shown in Fig. 10.13 are listed in Table 10.2. From Fig. 10.13 and Table 10.2 it is easily inferred that the performance of the considered defect detection system gets worse as the speed at which the test is accomplished increases [76]. Moreover, the experimental campaigns accomplished for this system have proven that it is sensitive to both transversetype defects and mixed-mode cracks (vertical split heads or compound fractures). However, its performance is limited by the fact that air-coupled ultrasound transduction in steel suffers from a loss of energy due to the large impedance mismatch between air and steel, both in transmission and in reception. This explains the low SNR characterizing received signals [62] and limits the test speed [76].

Table 10.2: Significant PD and PFA values extracted from the cumulative ROC curves of Fig. (10.13) [76].

Test speed	**PD (%) achievable for specific PFAs**				**AUC (%)**
(km/h)	**0% PFA**	**1% PFA**	**5% PFA**	**10% PFA**	
1.6	65.38	76.92	86.54	92.31	97.58
8	7.08	47.79	78.76	85.84	95.03
16	0.00	29.23	47.69	56.92	77.90
24	1.92	9.62	28.85	32.69	69.05

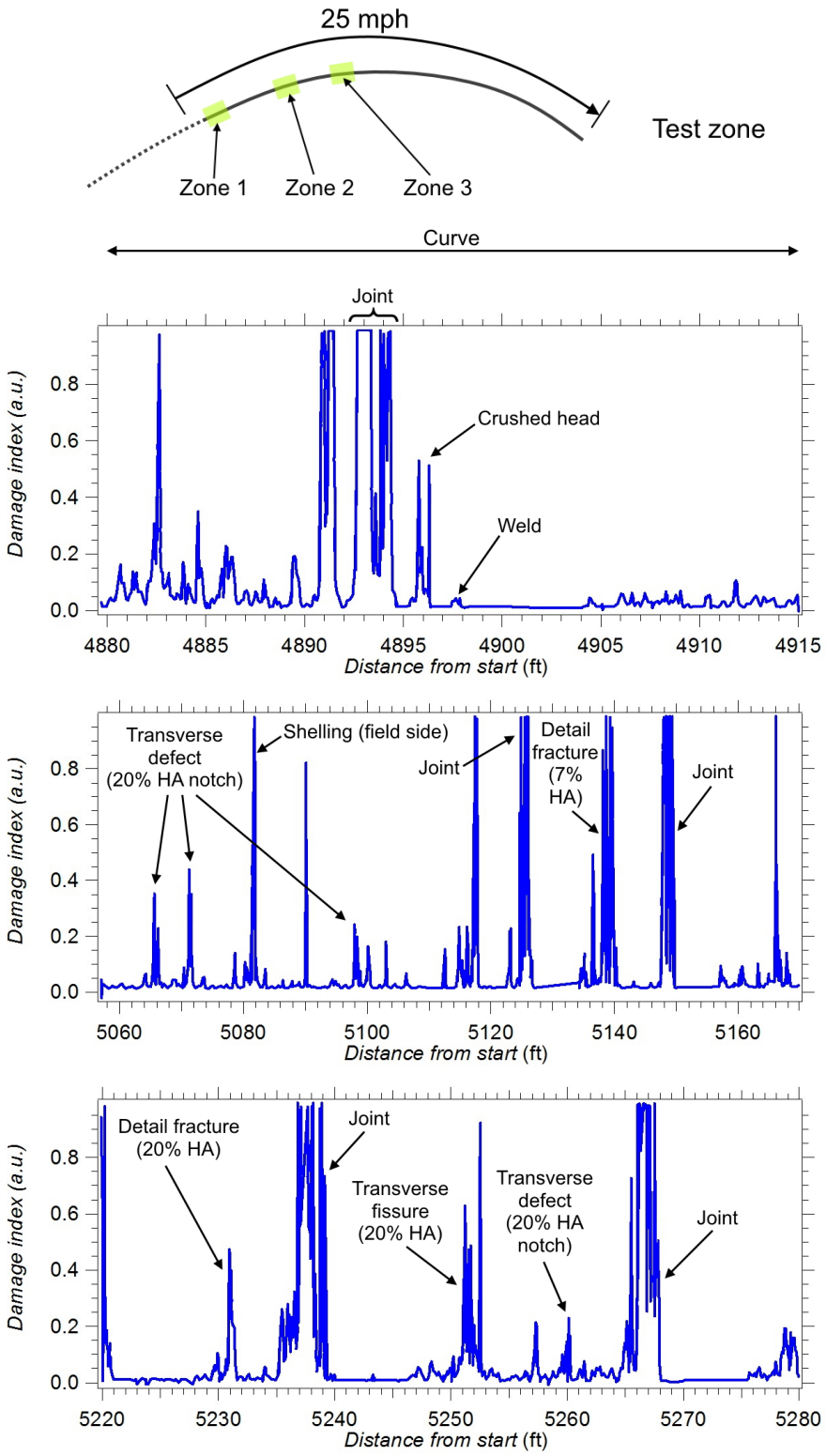


Figure 10.14: Damage index traces referring to the RDTF test track at 25 mph (40 km/h) in three selected test zones.

Table 10.3: Compromise values between test speeds, recording time window and SNR of the passively reconstructed transfer function achieved in field tests.

Test Speed	Recording Time Window	SNR of Reconstructed Transfer Function
30 mph (48.28 km/h)	∼15.25 ms	∼12
50 mph (80.46 km/h)	∼9.15 ms	∼9
80 mph (128.75 km/h)	∼5.7 ms	∼4.5

10.7.2 On-Board Passive System: Performance Analysis

Some DI traces obtained for the on-board passive defect detection system described in Paragraph 10.3.2 are illustrated in Fig. 10.14. The plots of Fig. 10.14 are generated on the basis of the data available in ref. [80]. Additional results illustrated in ref. [81] have evidenced that

1. The zones in which the received signal was strong enough for defect detection were primarily localized in sections of curved tracks, where wheels were flanging, so generating stronger excitation signals.

2. As speed decreases, a larger portion of the run becomes sub-optimal, since the presence of a defect cannot be easily detected.

3. At high speeds (say, higher than 96 km/h), an optimal source excitation energy is generated; this allows to achieve a stable Green's function.

4. Poor results were found in areas where the source excitation energy was too low to be detected by the receivers, e.g. in most of the tangent portion of the track. affecting the performance of the the defect detection algorithms (and, in particular, of the ICA; see Section 10.5.4).

The performed tests did not allow for assessing the true repeatability of the achieved results; for this reason, further tests have been planned to investigate the reliability of the considered system [80]. Moreover, it has been found that, as long as acoustic excitation is detectable, the strength of the received signal does not play an important role in the reconstruction stability. However, instability is found when the signal intensity is comparable to the noise floor of the receiver sensors.

Based on the considerations illustrated in Section 10.3.2.4 concerning the conflicting needs of selecting a proper test speed and achieving an accurate localization, some experiments have been made; the obtained results are illustrated in ref. [79]. In these experiments, various lengths for the recording time have been tested when the prototype was running at three different test speeds (30 mph, 50 mph, and 80 mph, corresponding to 48.28 km/h, 80.46 km/h, and 128.75 km/h, respectively). Such results prove that, in general, the SNR increases when the recording time increases; however, this leads to progressively losing spatial localization and to generating a non-stationary transfer function. Moreover, the SNR decreases as test speeds get larger. The duration of the time window has been selected in a way to achieve an accuracy equal to 8 in. (corresponding to 20.3 cm) in spatial localization at the various test speeds. The values evaluated for this duration are listed in Table 10.3.

Additional work is in progress to quantify and improve the reliability of the defect detection prototype by

- determining the optimal value of the threshold for the DI;

- minimizing the effects of poor signal reconstruction;

- improving the strength of the source excitation signal at low speeds [81].

10.7.3 Land-Based Ultrasonic Broken Rail Detection System Test Results and Performance Analysis

The performance of the UBRD system version 4 is illustrated in ref. [96]. After having solved some initial problems (equipment failures and GSM network reliability in some areas), a preliminary analysis of this system has been accomplished. In the first two months of operation, four broken rails and approximately four cracks have been detected before complete fracture occurred. It is believed that the system has prevented at least one derailment, the cost of which is similar to the cost of installing the system on the entire 840 km line. Detailed performance results have not been published yet.

10.7.4 Land-Based UBRD System with Early Rail Defect Detection Capability: Performance Analysis

Various experimental results referring to the land-based UBRD system with early defect detection capability are analyzed in ref. [102]. As already mentioned in Section 10.5.3.5 the reordered set of pre-processed data has been processed by three algorithms (namely, SVD, ICA, and dimensionally-reduced ICA), whose performance in defect detection has been evaluated. Since the outcomes generated by these algorithms are highly dependent on the input dataset, we do not illustrate here an accurate analysis of the achieved results. We comment briefly on the performance achieved by such algorithms.

It is expected that, when a monotonically growing defect is found, its weight in ICA or SVD also grows monotonically. To establish if a weight exhibits a monotonic increase, the *Mann-Kendall* (M-K) test can be exploited, as suggested in ref. [95]. This test is based on computing the difference between the number of increments and that of decrements between each pair of data points forming the input sequence; normalizing this difference with respect to the overall number of data points produces the normalized test statistic, denoted Z_{mk}. If the value of Z_{mk} is greater than 1.96, the probability of a monotonic trend is equal to 0.95; larger values of Z_{mk} indicate that this probability is even higher.

The M–K test has been performed on the weights computed by all the considered algorithms.

10.7.4.1 Defect Detection in the Presence of Large Defects

From test results, it can be inferred that:

- The SVD, ICA, and dimensionally-reduced ICA techniques are able to separate the signature of artificial defects from the baseline.
- Their performance is appreciably affected by the EOCs and the multi-mode nature of ultrasonic guided waves propagation in rails. As far as the last issue is concerned, it is important to stress that mode conversion is one of the possible causes of phantom reflections appearing when ultrasonic guided waves impinge on a defect; the multimode nature of propagation may cause the defect signature to spread over different components. Moreover, the EOCs can have a different influence on each mode.
- In general, the independent components produced by the ICA algorithm appear to be less noisy than the singular vectors generated by the SVD technique.
- The presence of relatively large monotonic trends is confirmed by the analysis of the M–K test scores over the component weights.
- A good match between the components achieving a large M–K test score and the components related to the defect signature is found in the SVD, ICA, and dimensionallyreduced ICA techniques.

10.7.4.2 Defect Detection in the Presence of Small Defects

The performance of the SVD, ICA, and dimensionally-reduced ICA algorithm was tested also by considering a dataset composed only of the measurements corresponding to the initial evolution of the defect. When the defect appears, it is supposed to reflect only a small amount of the impinging UGW. Therefore, the corresponding dataset is composed of the measurements in which the peak of reflection caused by the defect is small in amplitude. In this case:

- From the outcomes of SVD analysis, it has been inferred that the presence of an artificial defect cannot be easily detected. Moreover, selecting the singular values characterized by the highest M–K score does not guarantee defect detection; furthermore, in these conditions, the defect cannot be located by analyzing the components of the singular vectors.
- The analysis of the results generated by the ICA have proven that the defect signature is spread over multiple independent components, whose associated weights exhibit a monotonic increase, proportional to defect growth.

All of these results have led to the conclusion that ICA with dimension reduction performs better than the SVD algorithm in the presence of EOCs when attempting to detect the growth of small defects [102].

10.7.4.3 Concluding Analysis on the Evolved UBRD System

The cancellation of unwanted modes through phased array processing can be improved by using a larger number of transducers in the array employed for transmission and reception. The resulting system is able to accurately estimate the distance and direction of reflectors along the considered span of rails. However, since EOCs are not exclusively due to temperature variations, an important role in the achievable performance is played by techniques for compensating EOCs other than temperature. Defect detection has been performed through SVD and ICA; test results have led to the conclusion that ICA with dimension reduction performs better than SVD. Nevertheless, it has been found that the signature of a given defect spreads over different independent components; this might be due to the multi-modality of defect reflection signature. The last issue deserves further investigation.

Based on the results obtained in ref. [102], it can be stated that the detection of relatively small defects in the railhead is possible even if an array consisting of two transducers only is used and that, despite this positive result, the development of an automatic monitoring system, capable of autonomous operation with a low rate of false alarms, requires substantial effort. For this reason, the authors of that manuscript suggest that, during the design of an autonomous machine learning algorithm, the decision to stop trains or to ask for an on field inspection should be made by an expert able to interpret its output. A possible workflow for the development of a defect detection algorithm is also suggested: this could start from the analysis of the size of M–K scores, looking for values characterized by a significantly larger trend than the others. Then, it can be verified if the associated independent component resembles a single reflection from a defect. An appropriate level of alarm could be assigned on the basis of magnitude of the considered weight and its rate of increase. In any case, decision thresholds for these steps should be selected only after analyzing a large dataset.

10.8 Discussion and Conclusions

In this Chapter our attention has focused on the diagnostic technique based on ultrasonic guided waves. This technique is expected to solve the problem of the high-speed (or quasi-real-time) inspection of both the surface and the inner part of rails. In our analysis, two possible approaches, namely an on-board approach and a land-based approach, have been taken into consideration. In the on-board approach, the ultrasonic instrumentation is installed on an inspection car for high-speed scanning. In the land-based approach, instead, the ultrasonic equipment is attached to the rail in a fixed position for quasi-real-time monitoring. Moreover, in our analysis of the on-board approach, two different technical methods have been described: an active method and a passive one. In the on-board active method, ultrasonic guided waves are generated and sensed by the instrumentation installed on the inspection vehicle. On the other hand, the on-board passive method exploits the ultrasonic waves generated by the rail-wheel interaction of the moving train; consequently, the instrumentation placed in the lead car of the train only measures the response of the rail. A land-based system for ultrasonic broken rail detection is already active in South Africa; the system is being evolved to detect early-stage cracks too. A comparison between the technique making use of ultrasonic guided waves and other techniques described in Chapter 3, Existing Rail Defect Detection Techniques, is given in Table 10.4.

Table 10.4: NDT techniques for rail defect detection.

Eddy Current (EC)	Manual inspection, on-board, non-contact.	Surface defects and near-surface internal railhead defects.	$\leq$ 70 km/h	Reliable detection of surface defects; adversely affected by grinding marks. It is sensitive to probe lift-off.	Mature technology
Magnetic Flux Leakage	On-board, non-contact.	Surface defects and near-surface internal railhead defects. Unsuitable for longitudinal fissures.	$\leq$ 35 km/h	Reliable detection of surface and near-surface defects in the railhead. Cracks smaller than 4 mm are not detected. Performance degradation observed at high speeds.	Mature technology, research aims at increasing its scanning speed

Technique	Approach	Defects detected	Scanning speed	Performance	Open problems
Visual inspection	Manual inspection, on-board, non-contact.	Surface breaking defects, rail head profile, corrugation, missing parts, defective ballast. No internal defects.	From 4 to 320 km/h	Reliable detection on the surface of the rail and track. Internal cracks are not detected. Sensitive to lighting conditions (these require proper compensation).	Mature technology
Radiography	Manual system	Welds and all types of known internal defects.	Static	Reliable detection of internal defects in rails and welds that are difficult to inspect by means of other techniques. Some transverse defects can be missed. Safety hazard, costly and time-consuming.	Mature technology
Conventional ultrasound	Manual inspection, on-board, contact.	Surface defects, rail head, web and foot internal defects.	< 70 km/h in tests; in practice around 15 km/h	Reliable manual inspection, but rail foot defects can be missed. In high-speed inspection, surface defects smaller than 4 mm, as well as internal defects (especially in the rail foot) can be missed. Limit on the maximum speed achievable. Surface shallow RCF defects can mask severe internal cracks.	Mature technology
Active on-board guided waves ultrasound	Active, on-board, non-contact.	Surface defects and rail head web and foot internal defects. Best for transverse and mixed mode defects.	≤ 16 km/h laser, ≤ 24 km/h air-coupled piezoelectric	Reliable on-board system able to detect surface and internal defects. Affected by sensors lift-off variations. Difficult to be deployed at high speeds. The employed laser is costly and hard to maintain.	In case of piezoelectric excitation, the airborne signal reverberation limits the achievable test speed because of acoustic mismatch between air and steel.
Passive on-board guided waves ultrasound	Passive, on-board.	Surface defects and rail head web and foot internal defects. Best for transverse and mixed mode defects	> 96 km/h	The system reliably detects defects when the signal strength is high, i.e. where the rail-wheel interaction is stronger (e.g., in curves). As long as the quality of the received signal is good, its strength is not so important. System under development.	Quantification and improvement of defect detection reliability by selecting appropriate thresholds. Improvement of source excitation signals at low speeds or in tangent tracks.

Technique	Approach	Defects detected	Scanning speed	Performance	Open problems
Ultrasonic broken rail detector	Active, land-based	Broken rails	Fixed system (1 km sections), quasi-real-time operation	The system is designed to reliably detect clean breakages in rails. It is already working in operating conditions.	Mature technology. Evolution to reliably detect defects in rails and to increase section length.
Land-based guided waves ultrasound rail defect detection	Active, land-based	Transverse and mixed mode internal defects	Fixed system (1 km sections), quasi-real-time operation	The technology is under development. The aim is to detect and monitor a growing defect, and raise an alarm when it reaches a critical size.	Study of the behavior of a defect during time, compensation of resonant transducers variations, compensation of varying environmental and operating conditions which modify the propagation of guided waves.

10.8.1 Advantages and Disadvantages of the Considered Systems

The systems reviewed in this Chapter present both advantages and disadvantages. Due to the approaches followed being so different, a direct comparison is impossible. However, an attempt can be made to assess their pros and cons by considering their main characteristics, range of application, and ease of development.

The main disadvantage of the active on-board inspection system is its limited inspection speed; in these terms, the system does not achieve a significant improvement over the conventional ultrasonic inspection systems. On the contrary, if the passive on-board system is considered, the opposite problem is found: in fact, the system is able to provide good results only when the diagnostic vehicle is moving at high speed or along curved tracks. This limitation can be circumvented by exploiting the ability of this system to reconstruct the waveguide transfer function (which allows defect detection) without the perfect knowledge of the excitation signal. An alternative to on-board approaches is constituted by land-based systems, such as the UBRD and the UBRD evolved to recognize small-scale defects. The main feature of the simple UBRD system, the detection of complete rail breaks, is already implemented in the track circuits, which are well known to railway engineers. The advantage of the UBRD system is that it does not perform critical operations in terms of railway traffic safety (such as train detection); therefore, it might not be subject to stringent safety constraints, unlike track circuits. The main advantages of the evolved UBRD system are represented by its ability to identify defects before a complete rail breakage and the possibility of accomplishing a continuous monitoring of the health status of rails. The last feature is not provided by on-board systems that, because of their nature, call for a periodic inspection of the infrastructure. It should be also kept in mind that a continuous monitoring system allows for acquiring a large number of measurements having limited accuracy. In fact, the process of searching for a defect, identifying it, and monitoring its evolution is repeated continuously over a long period. Basically, it does not matter that a defect is identified and reported at its onset, when it still has a small size. What really matters is that the presence of a defect is signaled and its evolution kept under control over time, so that maintenance operations for its removal can be planned before they become dangerous for the safety of railway traffic. A periodic inspection system, depending on the frequency of its tests, needs to acquire accurate measurements. A defect must be detected in time, when it is still small in size, so that its evolution can be monitored, inspection after inspection, before it reaches a critical state. This is needed since any defect can exhibit a sudden evolution.

Other advantages and disadvantages to be taken into consideration are those related to the practical usage of the considered systems. If the health of the rails of an entire railway network has to be monitored, a land-based system must be installed on all its tracks; this could be time-consuming, laborious, and expensive. In addition, system maintenance might require a relevant effort, especially when some of its devices have been installed in remote areas, which are hard to reach. Another relevant problem is represented by the fact that its equipment, being physically attached to the rails, could represent an obstacle during any operation of infrastructure renovation. On the contrary, if an on-board system is employed, it is only needed to equip an

adequate number of diagnostic vehicles, and make them travel frequently and regularly throughout the network. This means that fewer devices, albeit more complicated, need to be built, installed, and maintained. Moreover, their regular maintenance may take place in workshops and not on the field, perhaps in remote areas. On the other hand, the costs associated with carrying out special diagnostic runs have to be paid; moreover, these runs could prevent the circulation of commercial service trains in certain time intervals, reducing the capacity of the network. If a system achieving a good balance between performance, installation, and maintenance costs is developed, these problems can be overcome. In fact, a part of the fleet of commercial service trains could be equipped with such a system, allowing such trains to perform diagnostic inspections while performing their normal duties.

Another relevant issue to be considered is the complexity of system testing during its development. An on-board system requires the execution of high-speed tests on field using railway vehicles in a relatively early stage of its development period. For this reason, it is required to have (a) a test vehicle capable of sustaining the envisioned inspection speeds, (b) the possibility of accessing an infrastructure suitable for tests, and finally (c) qualified personnel to drive and route the diagnostic vehicle on the selected infrastructure. All this might be quite costly. From this perspective, a land-based system has very different requirements. In fact, its initial development does not require access to particular infrastructures other than a sufficiently long stretch of rail. This is true at least until it is needed to test the system in its effective operating conditions. Even in this case, however, it is sufficient to have access to an adequate stretch of track where the system can be installed. Once the system has been set up, the execution of tests does not require to run dedicated trains or to employ specialized personnel.

10.8.2 Future Developments

Various possible developments of the systems described in this Chapter have been already mentioned in the previous sections. In summary, as far as the passive on-board system is concerned, research efforts should aim at improving its performance in defect detection at low speeds. Moreover, it is necessary to quantify and improve the reliability of the developed prototype in defect identification. On the other hand, to achieve a reliable operation of land-based systems, a fundamental research problem needs to be solved. In fact, it is needed to acquire a deeper understanding on the evolution over time of the defect response to UGW solicitations. In addition, the influence of variable EOCs on such a response must be studied in detail. Some improvement in defect identification and localization based on unsupervised machine learning techniques is also foreseeable.

10.8.3 Conclusions

All of the ultrasonic systems described in the technical literature still suffer from various technical problems, related to ultrasonic propagation and sensing in rail, to the influence of environmental and operating factors, etc. Therefore, even if various technically relevant results have already been obtained in this field, substantial research efforts are still needed. The strong interest in the above-mentioned problems is motivated by the increasing attention paid to railway transport safety in recent years.

Chapter 11

Introduction to the Experimental Activity

In this and in the following chapters the experimental activity carried on for the study of the propagation of ultrasonic guided waves in rails will be described.

The long-term objective is the design of a system to diagnose the structural integrity of rails on field which makes use of UGW. Nevertheless, given the finite amount of time available (further reduced by the spread of Covid-19 pandemic in Italy) for the study, starting from the ground-up, it is impossible to accomplish the full design of a complete diagnostic system. This is even more true if it is considered the complexity of UGW propagation -as seen in Part I, Background, and the fact that different research groups with ages of experience have not yet designed a diagnostic system with satisfactory performances, as illustrated in Chapter 10. The task is not impossible, but it requires knowledge, time and effort. In the short-term, therefore, the final target of the activity has been scaled-down and redefined.

The main task of the experimental activity is to build a *toolbox*, a set of knowledge and tools useful to familiarize, study and work with ultrasonic guided waves. This set of knowledge and tools will be useful to the design of any diagnostic or communication system making use of UGW, not only focused on the detection of defects in rails. Bearing in mind the long-term target, the short-term objective was then subdivided into smaller sub-goals, that are summarized as follows:

1. The first sub-goal of the activity is to build a background knowledge on ultrasonic guided waves propagation and their applications. It is needless to remark the importance of the study of UGW propagation in solid waveguides theory when starting the design which makes use of that theory. Some more words can be said about the critical analysis of the existing applications of UGW in rails. By knowing what has already been done and studied, it will belater possible not only to establish some goals for the short-term experimental activity, but also to define a viable architecture of the diagnostic system as a whole, targeting some aspects that need further investigation. The results of this first sub-goal have been presented in Part I, Background (theory) and Chapter 10, Existing Applications of UGW Rail Defect Detection Techniques (study of existing UGW applications in rails).

2. The second sub-goal of the activity is to establish which results about the propagation of UGW in simple waveguides and in rails are achievable with the available instrumentation. These results, if any, can be analyzed qualitatively. This sub-goal represents -in fact- a feasibility study about the possibility to use UGW to detect a defect (i.e., a discontinuity) in a waveguide or to determine the characteristics of UGW propagation in waveguides. The experimental activity, moreover, enables to focus the aspects which need more investigation or require a different or better instrumentation. The activity carried to accomplish the second sub-goal is described in Chapter 12, Qualitative Results About UGW Propagation in a Waveguide.

3. The third sub-goal of the activity is to develop a method to validate the qualitative results obtained while accomplishing the second sub-goal. The correctness of the methods and procedures used is verified by breaking down the problem into its constitutive parts, which are analyzed and evaluated individually. In this way it is also possible to refine the procedures envisioned on the basis of the accumulated knowledge, and to repeat the experiments. The activity carried to accomplish the third sub-goal is described in Chapter 13, Validation of Qualitative Results on UGW Propagation in a Waveguide.

By combining the outcomes of the mentioned sub-goals, it is possible con build the toolbox useful to deal with UGW propagation problems, which is the initial short-term target. Some conclusions and an outline of possible future developments are given in Chapter 14, Conclusions and Future Developments.

Before giving the description of the experimental activities, it is useful to make some hypotheses on the possible implementation of a rail integrity diagnostics system based on UGW propagation. These are based on the outcomes of the first sub-goal of the activity.

11.1 Possible Implementation of a Diagnostic System

On the basics of the outcome of the analysis on the existing applications of rail diagnostic systems and by knowing the available resources available it is possible to outline a possible way to implement a rail diagnostic system. The considerations made at the end of the analysis of existing UGW applications for defect detection (see Chapter 10) provide guidelines for the definition of a possible diagnostic system. It has been chosen the *land-based* approach, i.e. with sensors applied to the rails. The reasons are different:

- The staff of Alstom Bologna, which is a UNIMORE partner in this research activity, has considerable experience in track safety systems, in particular in track circuits.
- A land-based system does not require installation on a railway vehicle. In essence, we save ourselves the design (and certification) of a support frame adequate to withstand the stresses of high-speed traveling. Obviously the system to fix the transducers to the rail of a land.based diagnostics system will also have to be designed and certified, but it is assumed that the process will be simpler.
- A land-based system does not have to deal with the difficulties associated with receiving ultrasonic waves that, once emitted by the rails, propagate into air before reaching the receivers. In fact, the transducers are solidly bound to the rails; therefore, the design of the system is, in fact, simplified.
- With a land-based system it is not necessary to develop or adopt systems for locating the inspection vehicle and thus the defect along the railway network (GPS systems or high frame rate video cameras).
- A land-based system is simpler and cheaper to test than an on-board system. In fact, it is sufficient to find a section of rail long enough to carry out the tests, no matter if it is possible or not to run a railway vehicle on it.
- With a land-based system you are not forced to find a test facility to carry out test runs. To perform test runs are required additional skills for the conduct of the vehicles. Both the test facility and the personnel to conduct the test vehicles are not commonly (and cheaply) available. In reality RFI makes available in the Bologna San Donato test circuit a track including some known defects, but in case it would be necessary to make arrangements with RFI to carry out the tests.

Ultimately, a land-based system is (theoretically) easier to design and test in the field. However, it also has a rather important flaw: to cover the entire network with a land-based system, it would be necessary to install the system on all the tracks that compose it, for their entire length. Even assuming that the attenuation of the signals is low enough enabling UGW to cover large distances, in order to obtain a sufficient level of localization of defects, transmitters and receivers cannot have an excessive spacing. This essentially involves the installation of a large number of field devices (which must be powered) and a network (wired or not) that allows data to be sent to the central command post. An installation of this kind is possible if and only if the cost of installation and management of each individual unit is particularly low, also taking into account the benefits offered. In fact, already today there are systems that allow to identify a complete breakage of the rail in real time: the track circuits. These, however are not so widespread also because of their cost. An on-board system, if installed on many rolling stock in commercial service, can cover the whole network, as already mentioned when talking about the systems proposed in the literature. To the detriment of land-based systems, there is the intention of infrastructure managers in the near future to limit the installation of fixed safety systems such as track circuits, or to replace them with other radio-type systems (ERTMS, in particular). A land-based system in the near future may not be appreciated as it involves, by its very nature, the widespread installation of devices attached to the track throughout the entire network. Despite these possible problems, it is believed that developing a land-based system is still worthwhile. It provides an immediate advantage that an on-board system, however widespread, does not offer: the possibility of identifying in real time a serious defect or a break in a rail, and of monitoring the development of defects at regular (or continuous) intervals.

In any case, the design of a finished system is still a long way off. The basic notions that will be developed in this experimental activity will be valid both for the development of a land-based and on-board system. Should conditions change in the short to medium term, it will still be possible to change approach.

Chapter 12

Qualitative Results About UGW Propagation in a Waveguide

As already mentioned in Chapter 11, the second sub-goal of the activity, after the analysis of the existing applications of UGW propagation in rails for diagnostic purposes, is to establish which are the characteristics of the UGW propagation in simple waveguides and rails. This represents a feasibility study about the use of UGW to detect defects in rails and to gather knowledge about the propagation characteristics of a waveguide using commonly available instrumentation such as function generators and oscilloscopes.

This experimental activity was carried out in the specialized laboratories of the *advanced research center on electronic systems* "Ercole de Castro" (ARCES) of the University of Bologna.

The above mentioned task consists on determining if ultrasonic guided waves can be generated and detected in a waveguide by exploiting the instrumentation available in the laboratory involved two different experiments, with increasing levels of complexity. Two experiments were accomplished to explore the different areas of the problem to be tackled:

1. The first experiment aims at determining the effective possibility to detect the presence of a defect in a waveguide by analyzing the received UGW signal.

2. The second experiment aims at showing the dispersion curves of a rail.

The setup, the analysis procedures involved and the outcomes of each of these experiments are described in the following sections. In particular, the first experiment is illustrated in Section 12.2, whereas the second one in Section 12.3.

12.1 Available Instrumentation

To carry out the experimental activity in the involved laboratory, the following basic instrumentation was available:

- Agilent 33220A waveform generator;

- Tektronix DPO-3014 oscilloscope;

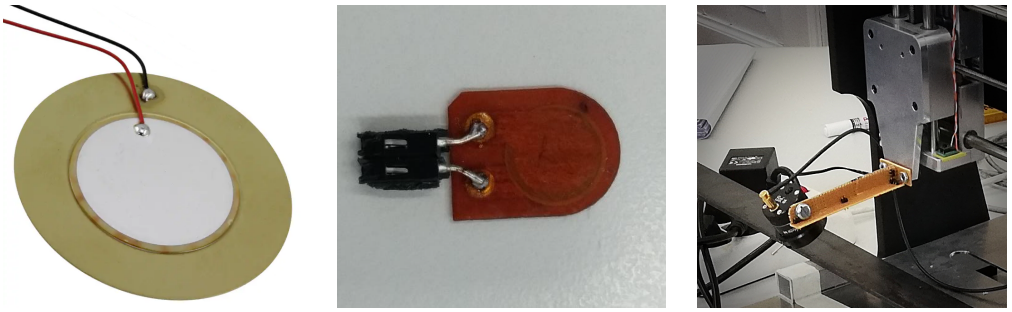
- Tegam 2350 signal power amplifier.

To generate and sense the ultrasonic guided waves in the waveguides during the experiments, the following transducers (represented in Fig. 12.1) was available in the laboratory:

- Murata 7BB-41-2L0 piezoelectric transducers;

- DuraAct P-876.SP1 piezoelectric transducers;

- Microacoustic BAT-1 ultrasonic air probe with corresponding Microacoustic Q-AMP signal preamplifier.



(a) (b) (c)

Figure 12.1: Transducers (a) Muata 7BB-41-2L0 piezoelectric transducer; (b) DuraAct P-876.SP1 piezoelectric transducer; (c) Microacoustic BAT-1 ultrasonic air probe.

12.2 Study of the Effects Caused by the Presence of a Defect in a Waveguide

Before setting up more complex and costly experiments, it is necessary to establish whether the presence of a defect in a waveguide produces effects that can be measured through the available instrumentation. Once the presence of these effects has been verified, it is necessary to carry out an analysis (in a first instance qualitative) of the influence that the defects and other structural characteristics of the waveguide have on the UGW propagating inside it.

The study just described can therefore be divided into three types of analysis that share the same measurement set-up:

1. The first type of analysis aims at determining if a defect in the waveguide produces effects on ultrasonic signals. This is done by comparing the signals collected in two different propagation conditions in the waveguide, that is to say, in the presence or absence of the defect. The comparison between the various signals collected takes place not only in the time domain, but also in the joint domain of frequencies and time. This analysis is described and the results analyzed in Section 12.2.4.

2. The second type of analysis has the purpose of verifying the possibility to trace a correspondence between the propagation modes that show up in the measured signals time-frequency analysis and among those predicted by the waveguide SAFE simulation (SAFE analysis theory is illustrated in Chapter 6). This can be achieved by superimposing the simulated dispersion curves on the signals represented in the joint domain of time and frequency. In the case of a good correspondence of the curves it can be asserted that the SAFE analysis is able to correctly estimate the propagation behavior of UGW in the given waveguide. This analysis is described and the results analyzed in Section 12.2.5.

3. The third type of analysis aims at investigating how the defect and other waveguide structural elements affect UGW propagation. This refers, in particular, to the effects of reflections that are expected to be produced not only by the defect, but also by the ends of the waveguide. These reflections are expected to cause constructive and destructive interference that shapes the behavior of signals that will be measured. Since the signals, by reflecting over discontinuities, cover progressively longer paths, thus taking more and more time to travel through them, it is possible to check whether the expected delay of the signal front arrival corresponds to the actually measured one. This verification can be done by knowing the length of the various non-direct propagation paths. The time that the signals take to propagate at the propagation velocity of modes calculated by SAFE simulations can be computed. By knowing these information it is thus possible to obtain the delay curves of the modes for each path, which can be superimposed to the representation of the signal in the time-frequency domain. If the correspondence between theoretical and measured delays exists, the effects that the various elements of the waveguide have on the propagating UGW can be determined. This analysis will be described and the results analyzed in Section 12.2.6.

Carrying out these analyses, although introductory and qualitative, also allows the development and calibration of some hardware and software tools useful in the subsequent analysis and development phases.

12.2.1 Measurement set-up

To carry out the mentioned analyses,a steel bar whose dimensions are indicated in Fig. 12.2 has been used. The bar has a slightly oblique cut in an asymmetrical position, closer to one of the two ends (as shown in Fig. 12.4).

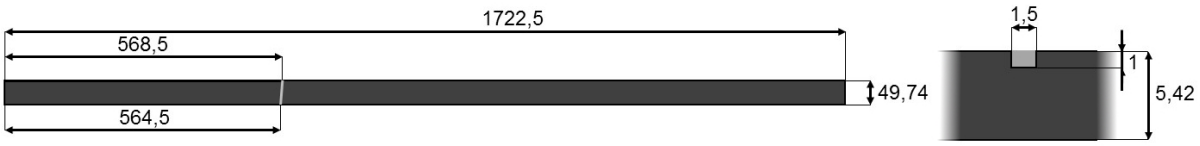


Figure 12.2: Dimensional drawing of the bar used during the experiments, on the right, seen from above; on the left, seen from the side, with detail on the dimensions of the cut. Measurements are expressed in millimeters.

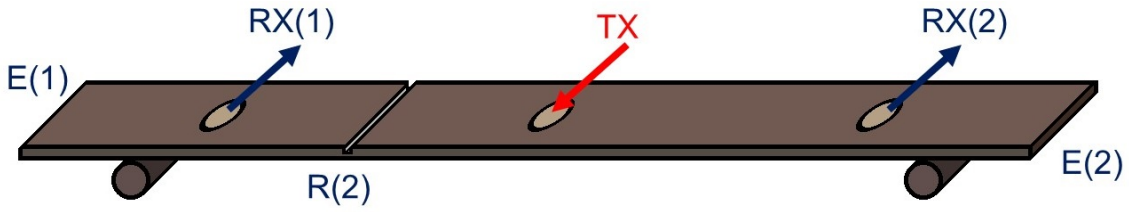


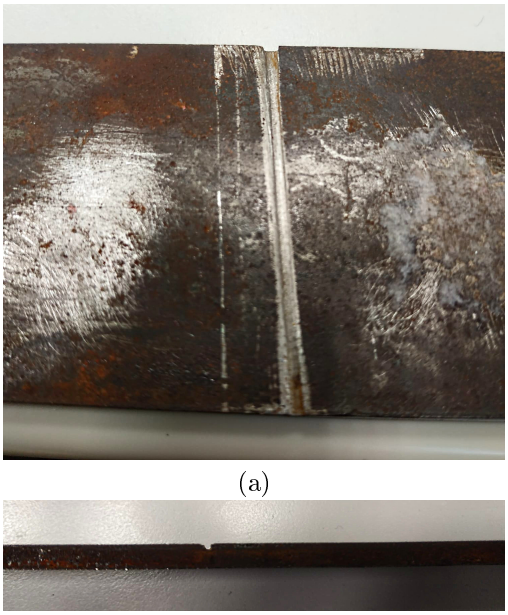
Figure 12.3: Schematic representation of the bar used for the qualitative experiments. TX is the transmitter placed in central position, RX(1) and RX(2) are the receivers placed in a paracentral position; E(1) and E(2) denote the extremities, while R(2) denotes the defect in the bar.

The asymmetrical position of this defect in the wave guide makes it possible to compare the two possible propagation conditions, i.e. in the presence and absence of defects, under the same environmental conditions. To carry out this comparison, in fact, it was decided to place a piezoelectric transducer (Murata 7BB-41-2L0) in the center of the bar with the function of ultrasonic wave generator, and two other identical transducers in a symmetrical paracentral position to measure the received signals. The adopted arrangement is shown in Fig. 12.3.

To allow easier but secure fixing and removal of the transducers, these were attached to the bar by means of bi-adhesive tape, helped by neodymium magnets placed above each piezoelectric transducer. During the execution of the experiments, the bar was placed on two metal supports.

All other conditions being equal, this measurement set-up ensures that:

- transmission and reception of the signals at the same instant (i.e., there are no ambient temperature variations);
- equal distance between transmitter, receiver and other elements that affect the propagation path, such as the ends of the bar (E(1) and E(2) in Fig. 12.3);
- equal propagation distance for the two paths.



(a)

(b)

Figure 12.4: Detail of the defect machined in the bar, from above (a) and from the side (b).

The only observable difference in the two path is constituted by the fact that one of the two is affected by the presence of the defect. The eventual differences between the transducers performance are not yet taken into account yet.

During the execution of the experiment, the measuring instruments were connected to each other and to the transmitters as in the diagram of Fig. 12.5. The excitation signal is sent simultaneously to the transducer TX and to an oscilloscope input; function generator and oscilloscope are synchronized via the trigger signal. Each receiver is connected to a free channel of the oscilloscope. In Fig. 12.6 it is included a photo taken during the preparation of the measurement set-up.

12.2.2 Choice of the Excitation Pulse and Acquisition Parameters

In this phase of qualitative analysis, we were not interested in soliciting the propagation of a specific waveguide mode. On the contrary, the interest at that moment was to excite as many modes as possible in order to evaluate

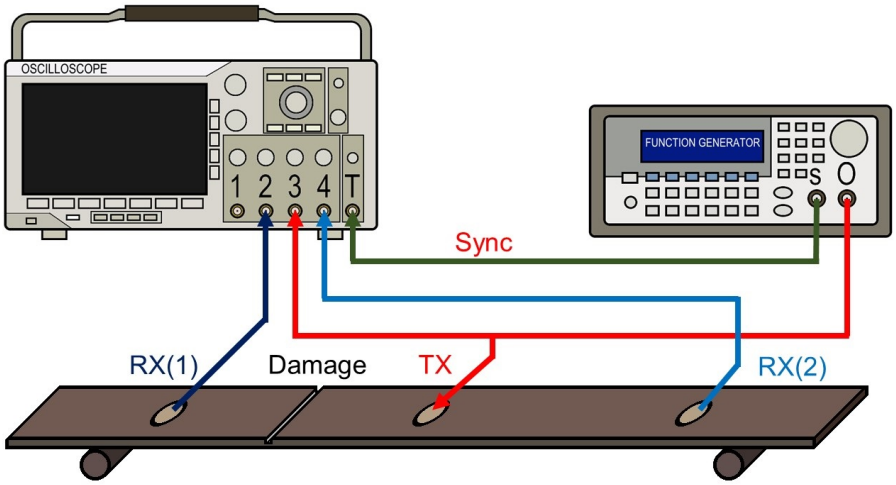


Figure 12.5: Schematic of the connections between instruments (function generator on the right, oscilloscope on the left) and transducers (TX is the transmitting transducer, RX(1) is the receiving transducer on the damaged side, RX(2) is the transducer on the pristine side).

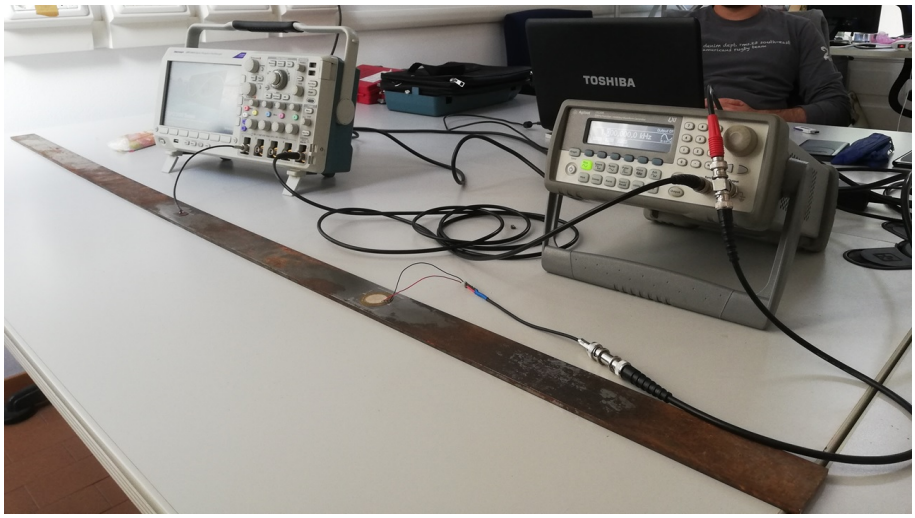


Figure 12.6: Picture of the bar and of the instrumentation used in our qualitative experiments.

the overall behavior of the system. Only on the basis of the results obtained in this phase it was possible proceed with the identification of the modes that can be considered most suitable for the propagation and identification of the defects of interest, and then to study how to exclusively excite them.

If the excitation of as much mode as possible is desired, a wide-band pulse should be used, such as a unipolar rectangular pulse. This pulse is able to transfer a good amount of energy to the modes that propagate to the waveguide. In any case, the signal that can be received is not particularly intense: it follows that the *signal-to-noise ratio* (SNR) of the received signal is not particularly high. A well-known method for limiting the effects of noise is to acquire the signal several times (under the same propagative conditions) and then perform an averaging over of all the collected sequences. Attenuation therefore comes into help: by transmitting a given number of pulses, separated by a time distance sufficient to allow the intensity of the propagating signal to fade below the noise level, it is possible to obtain a number of sequences of received signals that are not correlated and that can be averaged. The sequence produced from averaging is supposed to be a faithful representation of the propagation conditions of the system as a whole, free from the effects of added noise, be it mechanical vibrations and or electronic noise.

The signal generated by the function generator and transmitted through the piezoelectric transducer is a unipolar rectangular pulse of 10 V in amplitude, duration 19.53 μs and repeated (in burst mode) every 500 ms. The sampling rate was 25 MS/s and the stored sequence was the result of an average of 32 sequences. The selected sampling rate represented a good compromise between the size of the collected data and an accurate representation of the signal by taking into account the Nyquist's law.

Table 12.1: Coordinates of the elements along the waveguide.

Configuration	E(1) [mm]	R(2) [mm]	RX(1) [mm]	TX [mm]	RX(2) [mm]	E(2) [mm]	Distance TX-RX
RX(1) 7.5 cm before the defect	0	567	644	860	1086	1720	220
RX(1) 2.5 cm before the defect			594		1136		270
RX(1) 2.5 cm after the defect			544		1186		320
RX(1) 7.5 cm after the defect			494		1236		370

12.2.3 Collected Signals

To perform the different types of analyses mentioned in Section 12.2, the same data set has been always used. With the positioning of the transducers described in the previous sections, the aim was to acquire the signals that propagate both on the damaged and on the pristine side of the waveguide. To the comparison of the propagation in the presence or absence of the defect, the study on the effects of the different distance between the receiving and defect transducers has been also added. In fact, there are four different distances between transmitter and receivers (TX, RX(1) and RX(2) respectively in Fig. 12.3). The naming of the configurations relating to the positions of the receivers will refer, here and in the rest of the chapter, to the position of the receiving transducer RX(1) with respect to the cut R(2), having the point of view of the transmission transducer TX.

The configurations are listed in Table 12.1, where also the position and coordinates of waveguide elements and transducers is made explicit.

In the following Sections we describe in detail the types of analysis mentioned in Section 12.2. First, the analysis procedures applied to the available data are described; then, the obtained results are analyzed.

12.2.4 Comparison Between Collected Data

12.2.4.1 Analysis Algorithm

The first type of analysis described in Section 12.2.2 concerns the comparison of the collected signals in various domains. To carry out this comparison, it has been used a matlab algorithm whose flow diagram is sketched in Fig. 12.7.

First, the data are plotted in the time domain, so that it is possible to make a direct comparison between them. An example is offered in Fig. 12.8.

Secondly, the spectrogram of the data is calculated and plotted. The spectrogram can be defined as a graph of the amplitude (expressed in dB) of the *short-time Fourier transform* (STFT). In short, the STFT is made up of a sequence of FFTs applied to segments of windowed data in which windows overlap over time [60]. That is to say, the spectrogram is a representation of the signal in the time-frequency domain. Additional details are provided in Chapter 9. This type of analysis allows to evaluate how the signal components are distributed over time and frequency. It is also possible to determine what are the effects of interference, dispersion, reflections, modal conversion, etc. affecting the signals.

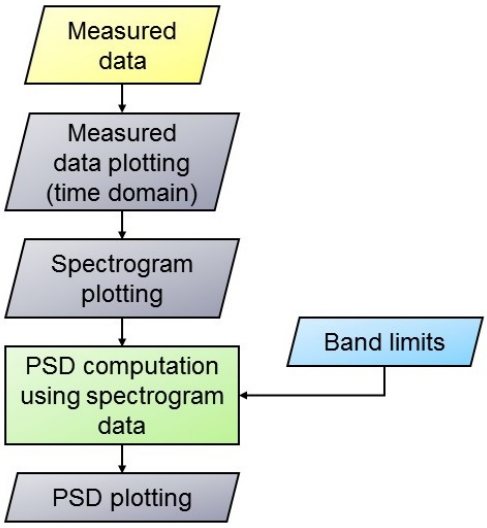


Figure 12.7: Flow diagram of the processing steps implemented in Matlab to compute the power spectral density of the measured signals.

The matlab's `spectrogram` function was used to calculate and plot the spectrograms of the receiving signals. As already mentioned in Chapter 9, there is a trade-off between the resolution of the spectrogram in the frequency domain and the resolution in the time domain. Basically, the `spectrogram` function is able to find this balance on its own, but for optimized results it was preferred to specify the window length and the amount of overlapping. The `spectrogram` function is not only able to plot the spectrogram, but provides also the output values (i.e., the matrix of the STFT values, the frequency value and the value of the instants of time) for later use. The function is also able to calculate the power spectrum of the data, which is provided in the form of a matrix, and this is useful during the data analysis. The power spectrum is calculated as the one-sided *modified periodogram* estimate of the *power spectral density* PSD of each segment (see Chapter 9).

In our analysis, the information contained in the power spectrum is further analyzed. Beyond the morphological analysis of the spectrograms, which will be discussed below, analyzing how the power carried by the received signals is concentrated in the frequency spectrum is an interesting starting point. If the spectrum is divided into frequency bands and it is analyzed how much energy is carried by the signals in those bands, it is possible to understand how the energy of the signal is distributed as the propagation conditions vary. Consequently, it is possible to evaluate the effects produced by a different position of the transducers, or if the presence of a defect causes a significant alteration in the energy distribution. The analysis is crude, but provides a rather immediate indication of the situation.

By plotting the PSD values calculated for each configuration in a bar diagram, the energy distribution under varying conditions can be determined. In this way it is possible to evaluate at a glance all the macro-variations that occur in the various conditions.

12.2.4.2 Analysis of Results

In the analysis of results in the time domain, we refer to Fig. 12.8. At first glance, it is possible to immediately notice the attenuation that the transmitted signal experiences: if the pulse width is 10 V, the peak of the received signal is around 10 mV: this is an expected result. The second observation is the dispersion which the signal suffers from, even where (in the milliseconds after the arrival of the first wave front) it is estimated that the measured signal is not influenced by the main reflections existing in the waveguide. In this case, the dispersion also means that it is not always easy to identify the peaks corresponding to the arrival of a wave front that follows a mode dispersion curve. When the effects of reflections begin to influence the received signal, with their constructive and destructive interference, the process of identifying and explaining the presence and cause of the observable peaks becomes even more difficult. In any case, from the analysis of the results obtained, the difference between the signals collected on the side presenting the defect and those collected on the other side is also quite evident.

The time domain signals, however, do not allow to directly associate the presence of the defect with the existence of particular peaks. In fact, each peak is also determined by the particular conditions of interference that are created in the waveguide area below the receiving transducer. These conditions change with each variation in transducer position.

It is also worth mentioning that, in order to recognize the presence and explain the existence of the various peaks, it is necessary to resort to the knowledge of the relative distances between transmitters, receivers, and the structural elements of the waveguide. This helps to determine the length of the possible propagation paths taking into account the various signal reflections, and therefore the expected succession of the various peaks. Without having this knowledge, analyzing only the received signals as they are, it would be very difficult to recognize all the differences between the two signals collected in the two propagation positions. The differences are in fact minimal and difficult to recognize at first glance.

It is clear that the analysis of the data in the time domain alone does not lead to certain and appreciable results, because of the dispersion and multi-modality that characterize propagation. However, it must be said that the choice of the pulse (broadband) and the length of the waveguide (short) certainly did not positively contribute to the readability of the signals received.

It is fair to say that our goal was precisely to appreciate the effects of dispersion and reflection on the waveguide discontinuities, and to analyze how the signals were distributed in the common domain of time and frequencies. The next steps focus on this aspect.

It is now possible to proceed with a quick analysis of the spectrograms. The analysis is not very detailed now since in the subsequent phases, when the superposition of dispersion curves are evaluated, the spectrograms can be analyzed more appropriately and in detail. Here, we limit ourselves to make some high-level observations on the distribution of the signal as the frequency and time vary, comparing the various configurations of transmitters and receivers in the presence and absence of the defect. We refer to Fig. 12.9.

First of all, some general considerations. As already stated, the spectrogram resolution is a compromise between spatial and temporal resolution. The spectral behavior of the signals appears to be distributed in bands: this is believed to be caused by resonance effects of the transducers or due to attachment to the surface of the bar. The color scale shown on the right, relative to the ratio between spectral power and frequency is not constant, but varies between configuration and configuration to emphasize the distribution of the signal.

By evaluating the first configuration (transducers placed 7.5 cm before the defect) in Figs. 12.9(a,b), it is possible to note that there are no particular differences in the spectral distribution of energy between the signals measured by RX(1) and RX(2). In any case, small variations are perceptible, for example in points (1a) and (1b), and (2a) and (2b). It is possible to note how the "whiskers" in Fig. 12.9(a) go higher in frequency than the Fig. 12.9(b). In the range of frequencies between 100 and 150 kHz it can be seen how the peak (3a) in Fig. 12.9(a), in Fig. 12.9(b) is more delayed and distributed over time (3c). Instead, a transfer of the spectral

(a)

(b)

(c)

(d)

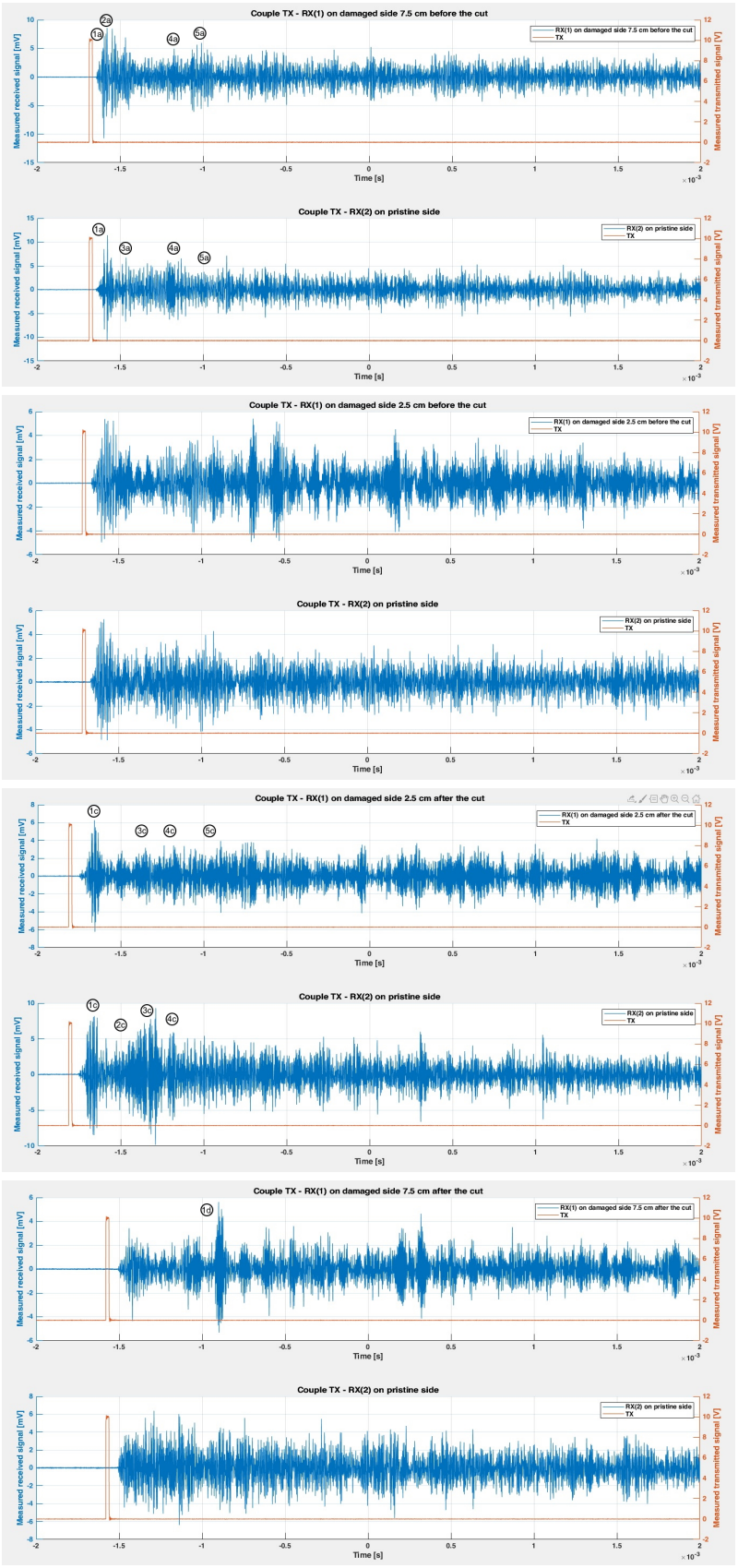
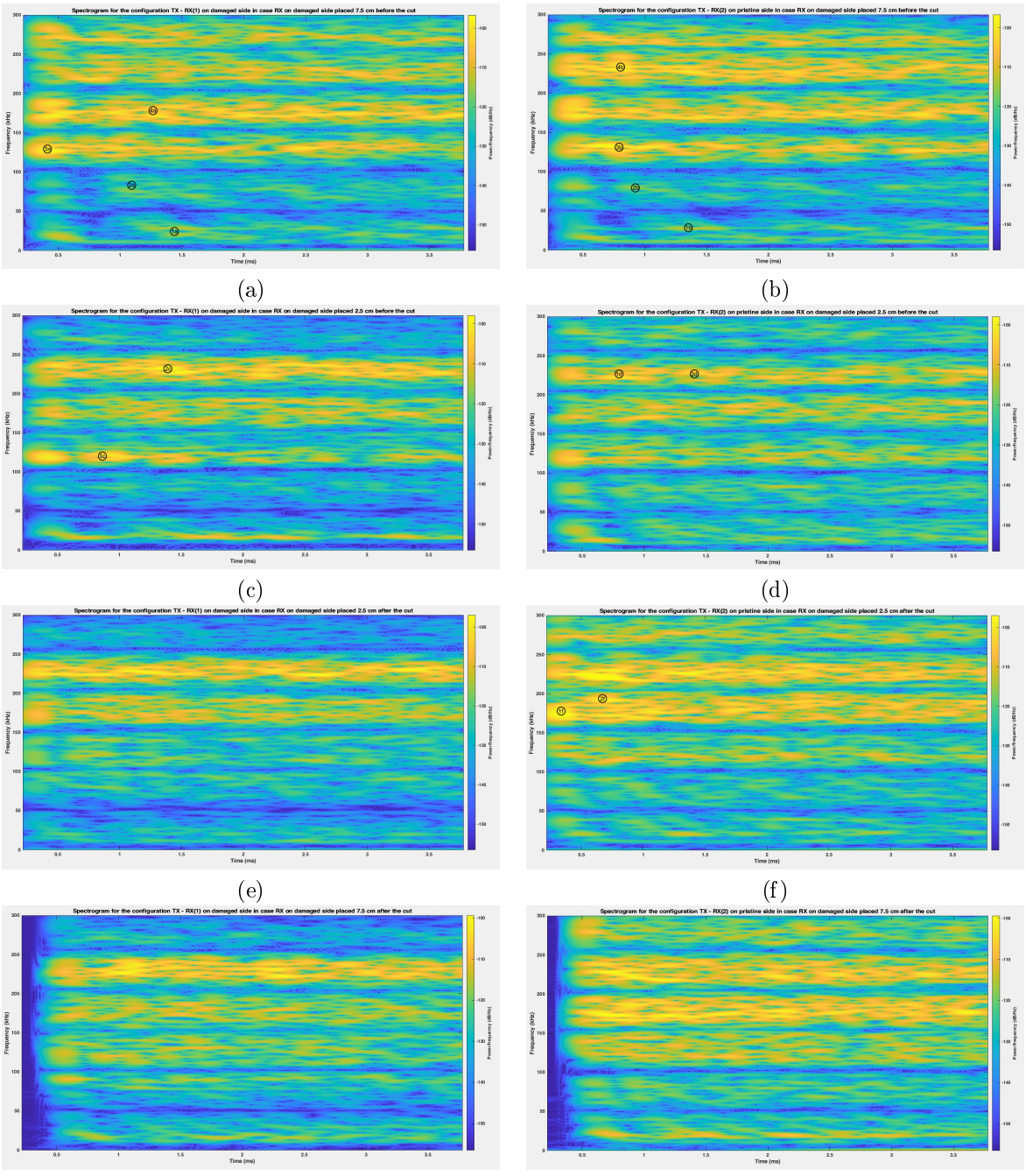


Figure 12.8: Comparison of received signals in the time domain. (a) 7.5 cm before the cut; (b) 2.5 cm before the cut; (c) 2.5 cm after the cut; (d) 7.5 cm after the cut.



(a) (b) (c) (d) (e) (f) (g) (h)

Figure 12.9: Comparison of the spectrograms. (a) and (b) configuration 7.5 cm before the cut; (c) and (d) configuration 2.5 cm before the cut; (e) and (f) configuration 2.5 cm after the cut; (g) and (h) configuration 7.5 cm after the cut. The left column refers to spectrograms plotted from signals measured from the receivers placed on the damaged side (RX(1)); the right column refers to spectrograms plotted from signals measured from the receivers placed on the pristine side (RX(2)).

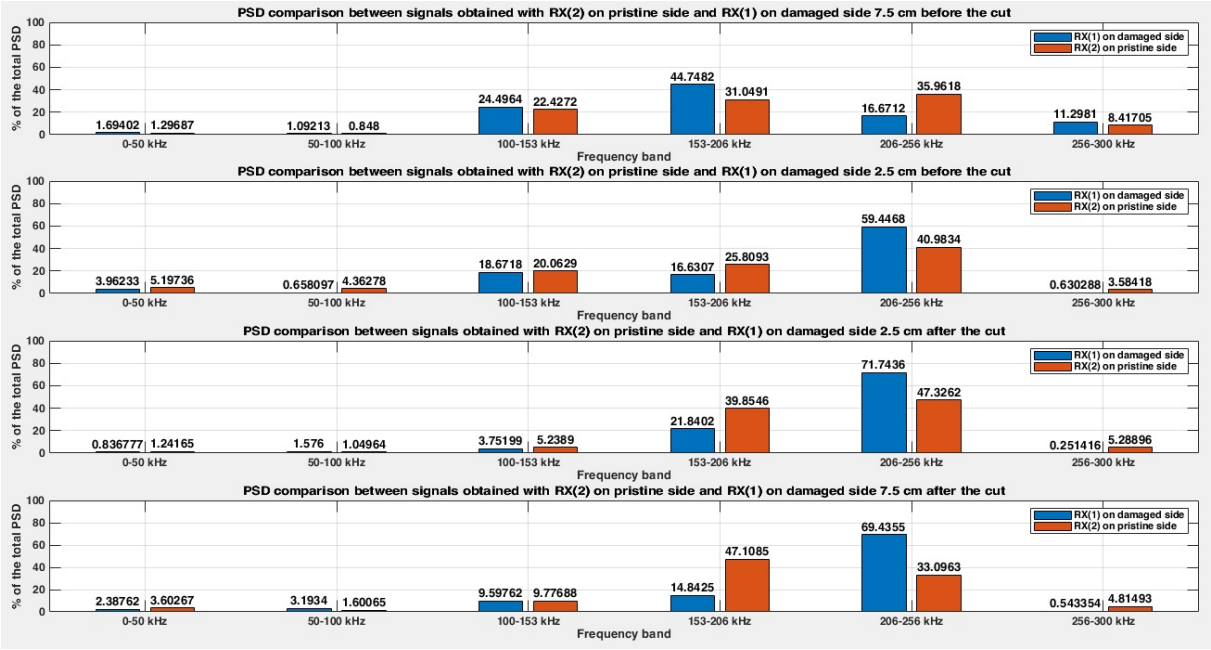


Figure 12.10: Comparison between the PSD contained in different frequency bands of the measured signals for the four different configurations of receiver position respect to the defect position.

content occurs between the 150-200 kHz band and the next, 200-250 kHz, in the presence or absence of the defect (see (4b)).

By evaluating the second configuration, the one in which RX(1) is placed 2.5 cm before the defect, in the graphs of Fig. 12.9(c,d) it is possible to see a different distribution of energy in the frequency range. There is also a different time distribution of the energy, for example the peak (1c), which in the case of the signal measured by RX(1) is in the 100-150 kHz band, in the signal measured by RX(2) it is in the 200-250 kHz band (2d). At the same time the peak (2c) in the signal on the defect side is much less intense in the signal on the pristine side (2d). The fact that in the signal measured by RX(2) there is more energy concentrated at low frequency, highlights some patterns in the lower portion of the graph. In the following analysis phases (Sections 12.2.5 and 12.2.6) the meaning of these patterns will be discussed.

Even in the case of the third configuration with RX(1) placed 2.5 cm after the defect, in the spectrograms in Fig. 12.9(e,f), it is possible to note a more homogeneous distribution of energy over the entire spectrum in the signals measured by RX(2). As already seen by analyzing the data in the time domain, the signal measured by RX(2) is more intense than the signal measured by RX(1), and this can also be seen in the spectrogram which has more intense peaks. In this case, it is interesting to note that in the signal measured by RX(2) there is immediately (1f) a very intense peak in the 150-200 kHz band, absent in the signal measured by RX(1). It can be thought that part of this energy is reflected by the defect before reaching RX(1) and that it overlaps the wave fronts of the slower modes arriving on RX(2) (2f).

Finally, analyzing the last configuration, the one in which RX(1) is placed 7.5 cm after the defect, in the spectrograms in Fig. 12.9(g,h), considerations similar to those made in the previous case can be made. In any case, it is necessary to bear in mind that there are small differences related to the different signal propagation distances.

The last step concerns the analysis of Fig. 12.10 which aims at evaluating how the PSD is distributed in the various bands, depending on the configurations and whether the signals are measured by RX(1) or RX(2).

Evaluating the first configuration (first graph at the top in Fig. 12.10), it is possible to observe how the energy is more uniformly distributed in the frequency spectrum in the signal measured by RX(2). Having stated that a minimum part of the energy content is distributed at frequencies below 100 kHz, the graph shows how the signal energy is concentrated between 100 and 250 kHz, with a preference for the 206-256 kHz band. In the signal measured by RX (1) these proportions change considerably: the signal energy is now concentrated in the 150-206 kHz range and the distribution is not so homogeneous in the 100-256 kHz band.

For the second configuration the situation is different: the energy is once again concentrated between 100 and 256 kHz, but both for the signals collected by RX(2) and those measured by RX(1) its distribution is no longer so uniform within the frequency range. Compared to the previous configuration, a greater proportion of energy is distributed at low frequencies, and less at high frequencies. Both in the case of signals measured

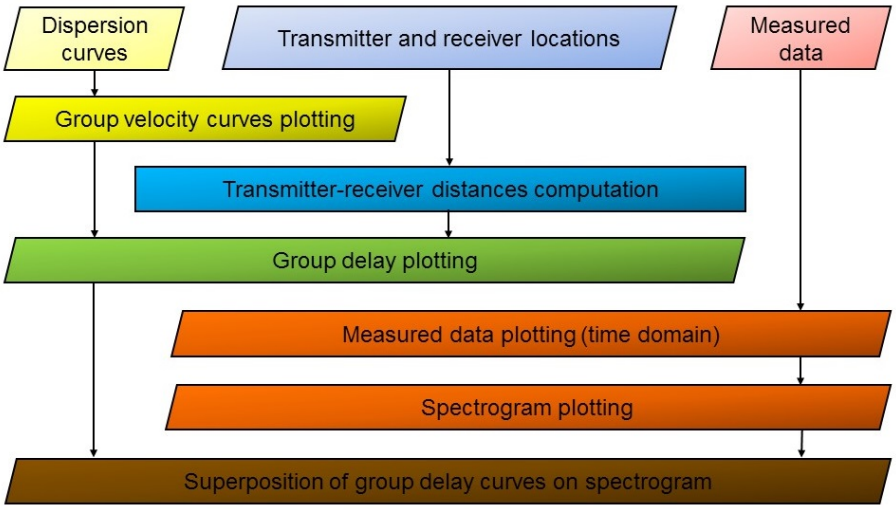


Figure 12.11: Flow diagram of the processing steps implemented in Matlab to plot the superposition of group delay simulated curves to spectrograms.

by RX(1) and RX(2) the energy is concentrated in the 206-256 kHz band. Between the two sides of the bar it changes the proportion of energy included in the aforementioned band, much greater in the case of RX(1) than in that of RX(2).

By evaluating now the third and fourth configuration, those where RX(1) is placed beyond the defect, common observations can be made. The energy is concentrated mainly in the 153-256 kHz range. Furthermore, the energy of the signal measured by RX(1) tends to concentrate more massively in the 206-256 kHz band (significantly more than in the previous cases). The energy of the signal measured by RX(2) is more uniformly distributed in the 153-256 kHz band, with a preference for the 153-206 kHz band in the case of the fourth configuration.

12.2.4.3 Partial Conclusions

The analysis of the plots relative to the distribution of the energy spectral density, of the signals in the time domain and of the spectrograms allow us to state that the presence of a defect does produce sensitive and effects measurable by the available instrumentation. It is also clear that are necessary further and more in-depth investigations to explain the cause of all the described differences. In fact, at this stage of the experimentation, there is no precise method to discriminate between a signal collected in the presence of a defect and one collected in its absence, without knowing a priori the state of the waveguide.

If the target is only to identify the presence of a defect (but not to locate it or to determine its type and size), it may be sufficient to focus on the analysis of the variation of PSD. However, the PSD variation does not change univocally as the relative position of the defect varies, and further analysis would still be needed to determine if that described is a valid method. However, we are not only interested in identifying the mere presence of a defect, but also in locating and identifying it. It is therefore not possible to completely lose the temporal dimension.

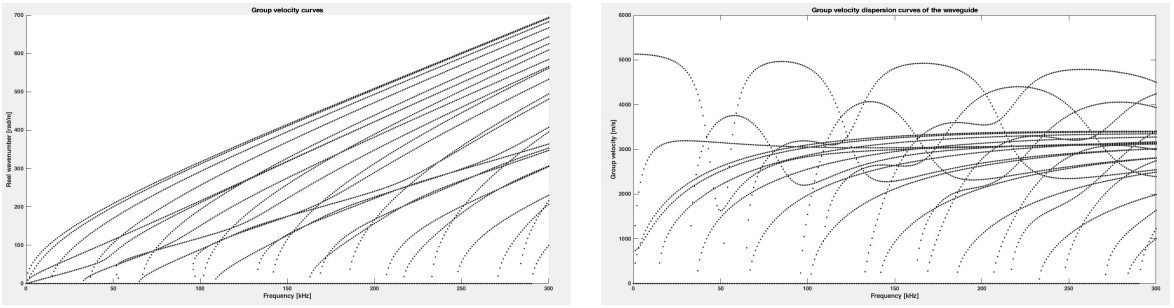
12.2.5 Detection of the Presence of the Propagation Modes Predicted by the SAFE Analysis in the Measured Signals

As already anticipated in Section 12.2, the second type of analysis that will be addressed is related to the search for a correspondence between the propagation modes that propagate and whose effect is possible to measure in practice and those predicted by SAFE simulations.

12.2.5.1 Analysis Algorithm

The matlab algorithm used to analyze the data is skectched in the flow chart of Fig. 12.11.

The spectrograms plotted in Section 12.2.4 actually contain an alternative representation of the dispersion curves. As we have already stated, they represent the signal in the joint time-frequency domain. The points of greatest energy concentration in the spectrogram correspond to the instants of arrival at the receiver of a specific signal spectral component. Basically, in the spectrogram it is possible to see the time taken by each mode to



(a) (b)

Figure 12.12: Dispersion curves of the steel bar used in the qualitative experiments: (a) real wavenumber, (b) group velocity.

travel from the instant in which the pulse has been generated to when the wave packet reaches the receiver, after possibly being reflected on the discontinuities. Once the dispersion curves of the waveguide have been calculated, it is possible to derive group velocity curves of each mode $v(\omega)$. From these, knowing the length of the propagative path s that separates transmitter and receiver, exploiting the laws of motion $t(\omega) = s/v(\omega)$, it is immediate to know the *group delay* $t(\omega)$, that is the time that each mode is supposed to take to travel the path from origin to destination.

The dispersion curves obtained from simulation of the steel bar used for the experiments are illustrated in Fig. 12.12. To compute the illustrated curves the the SAFE method, implemented in the GUIGUW application (*graphical user interface for guided ultrasonic waves*) has been used. GUIGUW is software designed to simulate the propagation properties of guided ultrasonic waves[1]. The software requires the size of the waveguide cross-section and the characteristics of its constituent material. The user can then set the characteristics of the mesh elements (maximum size and distribution), which is then automatically generated by the program. GUIGUW returns not only the dispersion curves, but also the curves of the group velocity, for the frequency range of interest and with the desired resolution. The characteristics of the mesh and the frequency resolution (i.e., the number of points given the frequency range) must be chosen in order to have a good compromise between the accuracy of the simulation and the calculation time[2].

The group delay curves relative to the direct path[3] can then be superimposed on the spectrogram, if one has the foresight to synchronize the instant in which the transmission of the impulse occurs with the delay. In the generated plot, the curves of the group delay should trace the behavior of the points where the spectrogram indicates that the signal is more intense. If the parameters set in the simulation are correct, the overlap should be perfect.

To obtain the dispersion curves, a waveguide having a rectangular section with the dimensions shown in Fig. 12.2 has been considered. The characteristics of the material set in the simulation are visible in Fig. 12.13(a)[4]. The mesh used to divide the cross-section of the waveguide had the characteristics shown in Fig. 12.13(b).

The obtained simulation results, in terms of dispersion curves and group velocity curves, are illustrated in Fig. 12.12. The simulations shows that, in the range of frequencies of interest, a large number (25) of symmetrical and antisymmetric modes propagate, overlapping, intersecting and rejecting. The propagation environment, even for a waveguide with a simple section like the one considered, is particularly rich, and this makes the analysis complicated.

12.2.5.2 Analysis of Results

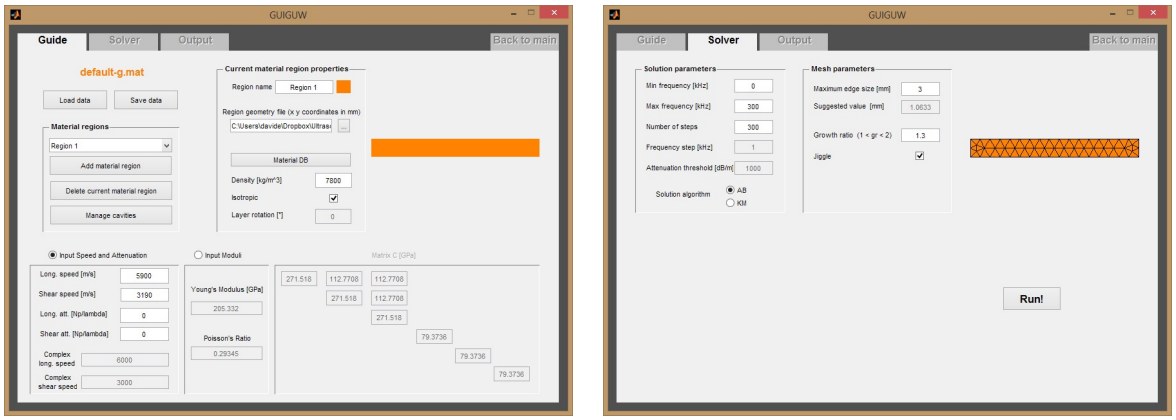
Let us repeat now the analysis of the results configuration by configuration (these were already described in Section 12.2.3). Let us consider for the moment the direct path only between transmitter and receiver when

[1] GUIGUW is SAFE formulation that considers linear waveguides, of infinite length in the direction of propagation of the guided waves, but constant and arbitrary section, immersed in vacuum. The waveguide can be made up of several distinct layers of different materials bonded together along the entire extension of the waveguide. The materials can be elastic or viscoelastic, isotropic or anisotropic, homogeneous at the wavelength scale. The cross-section of the waveguide is divided into a mesh made up of planar SAFE elements. The program is able to exploit the symmetry properties of the waveguide [111].

[2] The accuracy of the solution increases as the number of finite elements (FE) used increases, at the price of an increased computational cost. It is also necessary to say that the adequate number of SAFE elements (that is, their maximum size) to have a faithful result also depends on the number of waves that propagate and on the maximum frequency of interest [111].

[3] The paths involving reflections on the defect or on the extremities of the waveguide are not considered for now.

[4] They refer to a generic steel, since it is not possible to trace the exact composition of the alloy used for the manufacture of the bar; however, this is an acceptable compromise.



(a) (b)

Figure 12.13: GUIGUW screenshot illustrating (a) the input material parameters for the bar used; (b) the mesh and simulation parameters set.

calculating the time taken by the signal to go from the transmitter to the receiver. The direct path is the shortest path, equivalent to the mere distance between the centers of the two transducers.

The group delay curves are shown, for each of the four described configurations, in Fig. 12.14. Depending on the frequency and mode, the signal will undergo a certain dispersion and will arrive at the receiver with a more or less consistent delay compared to the instant of transmission. Considering the configurations, as the distance between transmitter and receiver increases, the signal path will obviously be longer. This not only means that the impulse will take longer to reach the receiver, but also that it will be more dispersed.

The superposition of the group delay curves over spectrograms is shown In Fig. 12.15. Despite the low resolution of the spectrogram, result of the mentioned compromise between resolution in the time and frequency domain, a good overlap between the simulated group delay curves and the behavior of the signal intensity peaks in the spectrogram can be observed[5].

12.2.5.3 Partial Conclusions

As already mentioned, the transmitted signal is reflected on the defects and other discontinuities of the waveguide, and the reflections tend also to overlap on the signal that travels the direct path. This can be guessed by observing the cited figures as the signal has numerous other peaks that are not (if not in part) due to the direct path only. This issue will be discussed in Section 12.2.6. Obviously, small imperfections in the overlap between the simulated curves and those represented in the spectrogram may be due to several reasons, not easily controllable. First of all, the exact alloy with which the bar was manufactured is unknown, so the assumed parameters, although appropriate for steel, may not perfectly fit the waveguide material. Secondly, the effects of the ambient temperature at which the measurements were carried out are not taken into account: the temperature has a certain influence on the propagation speed of the UGW. All these variables, however, are estimated to have an overall limited effect.

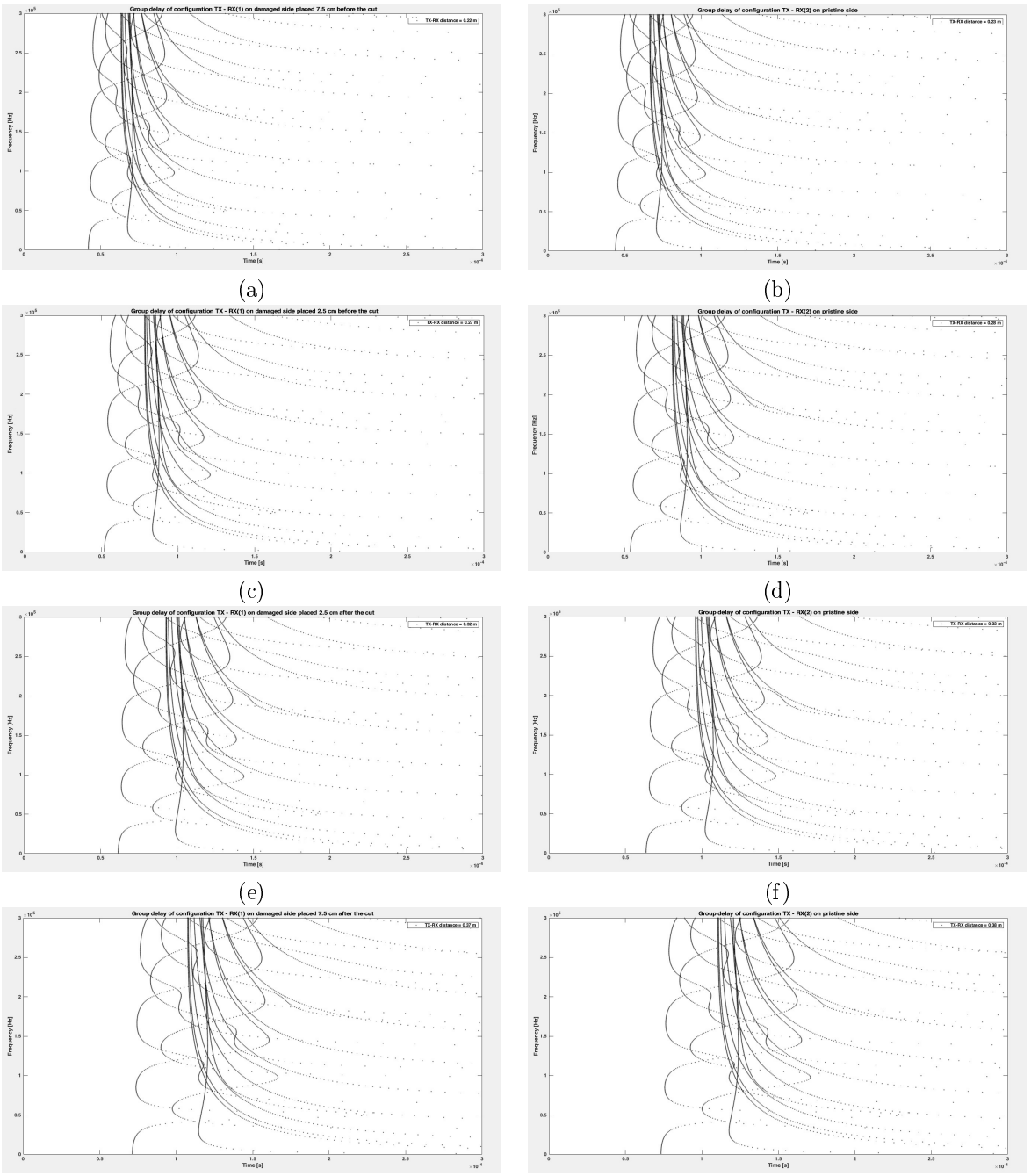
12.2.6 Influence of UGW Reflections on Discontinuities Observable in Measured Signals

As mentioned above, the third type of analysis concerns the evaluation of the effects that the signal, reflecting on the discontinuities, produces on the signal seen by the receivers.

12.2.6.1 Analysis Algorithm

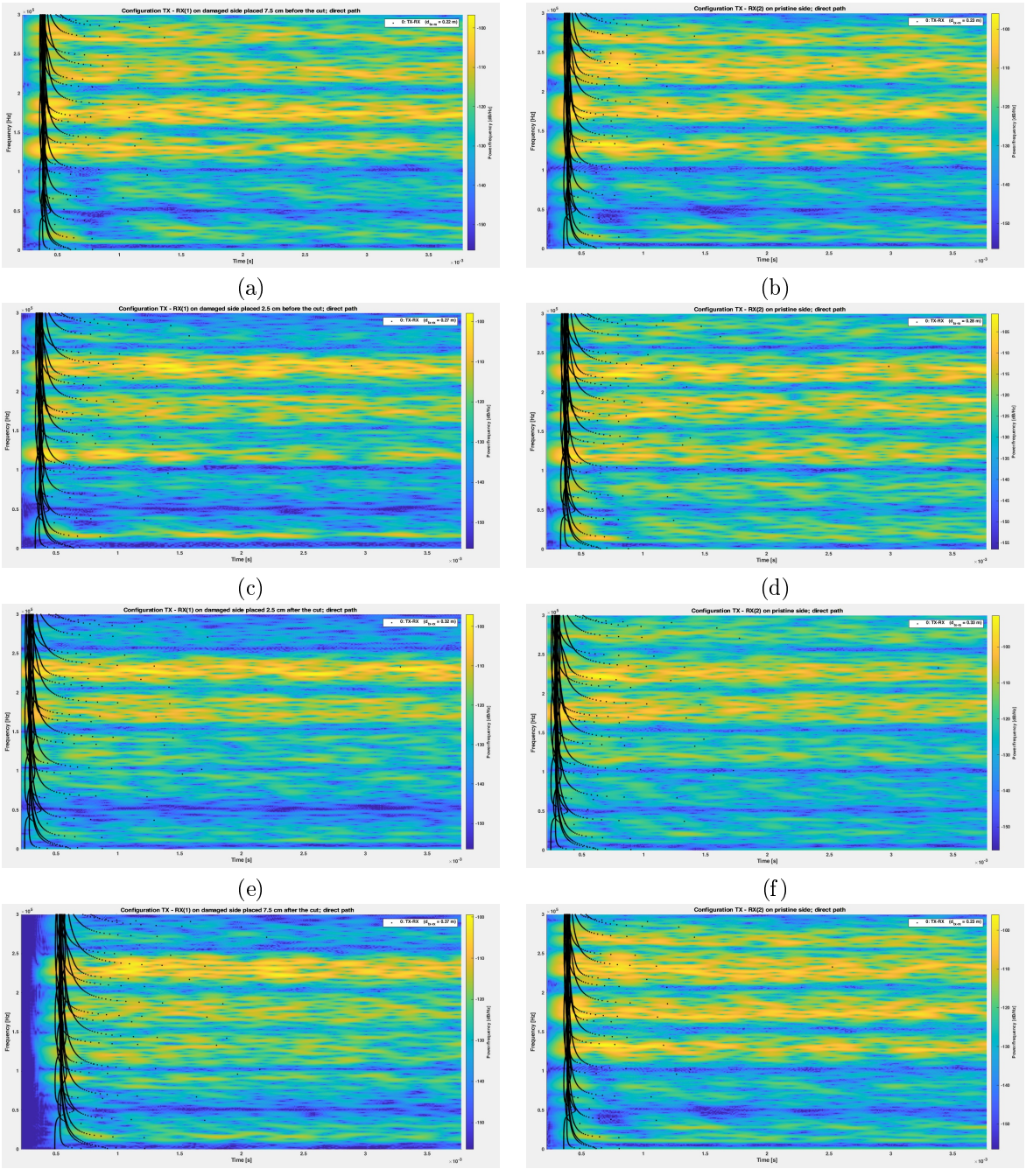
In this case, the same procedure adopted in Section 12.2.5 has been used: the simulated group delay curves were superimposed on the spectrogram. In this case, however, instead of the direct path, it is necessary to consider the indirect paths, i.e. those paths that take into consideration one or more reflections on the discontinuities present in the bar: the cut and the extremities. By knowing the length of these paths, it is possible to calculate and superimpose the curves of the group delay due to each path over the spectrogram, in the same way illustrated

[5]In observing the time axis it is possible to note an inconsistency with respect to that of Figs. 12.14: in that figure the pulse takes much less time to reach the receiver than what is seen in the superimposed curves in the spectrograms. Actually, the time elapsed between the start of the acquisition and the generation of the rectangular pulse is also taken into account, in order to maintain adequate synchronization.



(a) (b) (c) (d) (e) (f) (g) (h)

Figure 12.14: Group delay dispersion curves computed for the receivers placed, on the left column, on the damaged side; on the right column, on the pristine side. (a,b) configuration 7.5 cm before the cut, transmitter-receiver distance = 22 cm; (c,d) configuration 2.5 cm before the cut, transmitter-receiver distance = 27 cm; (e,f)configuration 2.5 cm after the cut, transmitter-receiver distance = 32 cm; (g,h) configuration 7.5 cm after the cut, transmitter-receiver distance = 37 cm.



(a) (b) (c) (d) (e) (f) (g) (h)

Figure 12.15: Group delay dispersion curvessuperimposed to the spectrograms computed from the data obtained from the receivers placed on the damaged side; on the right column, on the pristine side. (a,b) configuration 7.5 cm before the cut, transmitter-receiver distance = 22 cm; (c,d) configuration 2.5 cm before the cut, transmitter-receiver distance = 27 cm; (e,f)configuration 2.5 cm after the cut, transmitter-receiver distance = 32 cm; (g,h) configuration 7.5 cm after the cut, transmitter-receiver distance = 37 cm.

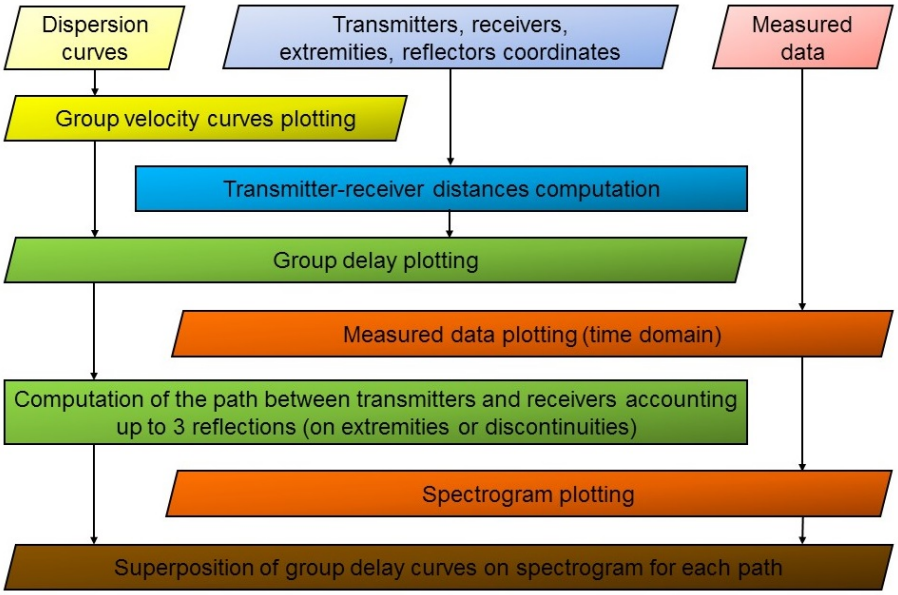


Figure 12.16: Flow diagram of the processing steps implemented in Matlab to plot the superposition of group delay simulated curves to spectrograms.

in Section 12.2.5. Data processing follows the flow chart shown in Fig. 12.16. In this way, it is possible to evaluate how the arrivals at the receiver of the various wave fronts which follow different propagation paths succeed one another, overlapping. This type of analysis is really useful for understanding the behavior of both the time-domain signal and of the spectrogram itself, making it possible to trace the cause of each peak.

The matlab script used to process the data is able to determine the length of all possible propagation paths. The program requires knowledge of the position along the waveguide of

- transmitters;
- receivers;
- extremities;
- defects;

and is able to establish not only the length of each path, but also on which elements the signal is reflected along each path. Theconsidered paths are only those practicable in connecting each transmitter to each receiver on the waveguide. No energy loss calculations are contemplated, so at the moment the effect of the extremities and defects is the same as the computation of path length matters. Although at most 3 reflections are considered (including those on the ends), a rather large set of different paths is still produced, which increases as the number of discontinuities in the bar increases.

The coordinates of transmitters, receivers, extremities and defects for each configuration is summarized in Table 12.1.

Tables 12.2 and 12.3 collect, again for each configuration and for each receiver, all the propagation paths considered in the subsequent analyzes[6]. The distances contained in Table 12.2 are then used to calculate the group delays curves from simulation data in the various cases.

12.2.6.2 Analysis of Processing Results

If the group delay curves evaluated for a rectangular cross-section bar are superimposed on the spectrogram and all the propagation paths are considered, the result is an incomprehensible tangle (see Fig. 12.17 for the configuration with RX(1) placed 2.5 cm before the defect). That representation gives a very clear idea of the complicated propagative environment and the overlap existing between the various reflections. This is furthermore exacerbated by the signal dispersion and the different propagation velocities of the modes. The plots does not help to understand the cause of the peaks that can be noted in the signal.

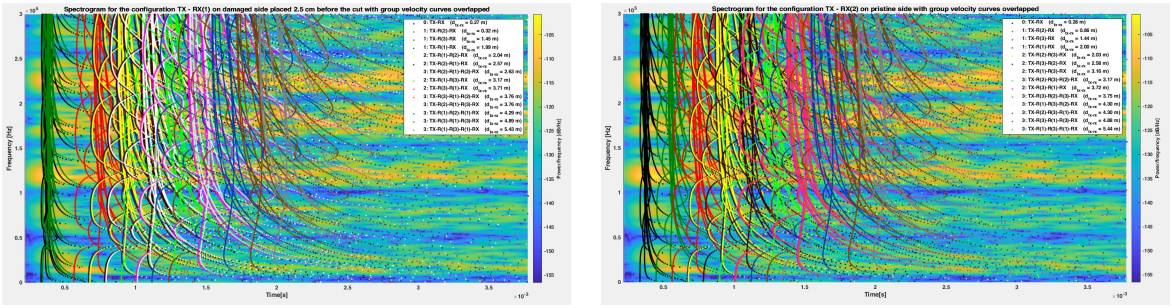
[6]Note that, for ease of implementation, the software in indicating the path does not distinguish between the names of the ends and defects; however, the table provides a transition table between the two notations.

Table 12.2: Path distances for the different configurations and involving up to 3 reflections (RX(1) before the defect).

Configuration	Receiver	Path elements	Path (graphs)	Path length [m]	Number of reflections
RX(1) 7.5 cm before the defect	*RX(1)*	TX - RX	TX - RX	0.22	0
		TX - R(2) - RX	TX - R(2) - RX	0.37	1
		TX - E(1) - RX	TX - R(3) - RX	1.50	1
		TX - E(2) - RX	TX - R(1) - RX	1.94	1
		TX - E(2) - R(2) - RX	TX - R(1) - R(2) - RX	2.09	2
		TX - R(2) - E(2) - RX	TX - R(2) - R(1) - RX	2.52	2
		TX - R(2) - E(2) - R(2) - RX	TX - R(2) - R(1) - R(2) - RX	2.68	3
		TX - E(2) - E(1) - RX	TX - R(1) - R(3) - RX	3.22	2
		TX - E(1) - E(2) - RX	TX - R(3) - R(1) - RX	3.66	2
		TX - E(1) - E(2) - R(2) - RX	TX - R(3) - R(1) - R(2) - RX	3.81	3
		TX - R(2) - E(2) - E(1) - RX	TX - R(2) - R(1) - R(3) - RX	3.81	3
		TX - E(2) - R(2) - E(2) - RX	TX - R(1) - R(2) - R(1) - RX	4.24	3
		TX - E(1) - E(2) - E(1) - RX	TX - R(3) - R(1) - R(3) - RX	4.94	3
		TX - E(2) - E(1) - E(2) - RX	TX - R(1) - R(3) - R(1) - RX	5.38	3
	RX(2)	TX-RX	TX-RX	0.23	0
		TX - R(2) - RX	TX - R(2) - RX	0.81	1
		TX - E(2) - RX	TX - R(3) - RX	1.49	1
		TX - E(1) - RX	TX - R(1) - RX	1.95	1
		TX - R(2) - E(2) - RX	TX - R(2) - R(3) - RX	2.08	2
		TX - E(2) - R(2) - RX	TX - R(3) - R(2) - RX	2.53	2
		TX - R(2) - E(2) - R(2) - RX	TX - R(2) - R(3) - R(2) - RX	3.12	3
		TX - E(1) - E(2) - RX	TX - R(1) - R(3) - RX	3.21	2
		TX - E(2) - E(1) - RX	TX - R(3) - R(1) - RX	3.67	2
		TX - E(2) - R(2) - E(2) - RX	TX - R(3) - R(2) - R(3) - RX	3.80	3
		TX - E(1) - E(2) - R(2) - RX	TX - R(1) - R(3) - R(2) - RX	4.25	3
		TX - R(2) - E(2) - E(1) - RX	TX - R(2) - R(3) - R(1) - RX	4.25	3
		TX - E(2) - E(1) - E(2) - RX	TX - R(3) - R(1) - R(3) - RX	4.93	3
		TX - E(1) - E(2) - E(1) - RX	TX - R(1) - R(3) - R(1) - RX	5.39	3
RX(1) 2.5 cm before the defect	*RX(1)*	TX - RX	TX - RX	0.27	0
		TX - R(2) - RX	TX - R(2) - RX	0.32	1
		TX - E(1) - RX	TX - R(3) - RX	1.45	1
		TX - E(2) - RX	TX - R(1) - RX	1.99	1
		TX - E(2) - R(2) - RX	TX - R(1) - R(2) - RX	2.04	2
		TX - R(2) - E(2) - RX	TX - R(2) - R(1) - RX	2.57	2
		TX - R(2) - E(2) - R(2) - RX	TX - R(2) - R(1) - R(2) - RX	2.63	3
		TX - E(2) - E(1) - RX	TX - R(1) - R(3) - RX	3.17	2
		TX - E(1) - E(2) - RX	TX - R(3) - R(1) - RX	3.71	2
		TX - E(1) - E(2) - R(2) - RX	TX - R(3) - R(1) - R(2) - RX	3.76	3
		TX - R(2) - E(2) - E(1) - RX	TX - R(2) - R(1) - R(3) - RX	3.76	3
		TX - E(2) - R(2) - E(2) - RX	TX - R(1) - R(2) - R(1) - RX	4.29	3
		TX - E(1) - E(2) - E(1) - RX	TX - R(3) - R(1) - R(3) - RX	4.89	3
		TX - E(2) - E(1) - E(2) - RX	TX - R(1) - R(3) - R(1) - RX	5.43	3
	RX(2)	TX-RX	TX-RX	0.28	0
		TX - R(2) - RX	TX - R(2) - RX	0.86	1
		TX - E(2) - RX	TX - R(3) - RX	1.44	1
		TX - E(1) - RX	TX - R(1) - RX	2.00	1
		TX - R(2) - E(2) - RX	TX - R(2) - R(3) - RX	2.03	2
		TX - E(2) - R(2) - RX	TX - R(3) - R(2) - RX	2.58	2
		TX - E(1) - E(2) - RX	TX - R(1) - R(3) - RX	3.16	2
		TX - R(2) - E(2) - R(2) - RX	TX - R(2) - R(3) - R(2) - RX	3.17	3
		TX - E(2) - E(1) - RX	TX - R(3) - R(1) - RX	3.72	2
		TX - E(2) - R(2) - E(2) - RX	TX - R(3) - R(2) - R(3) - RX	3.75	3
		TX - E(1) - E(2) - R(2) - RX	TX - R(1) - R(3) - R(2) - RX	4.30	3
		TX - R(2) - E(2) - E(1) - RX	TX - R(2) - R(3) - R(1) - RX	4.30	3
		TX - E(2) - E(1) - E(2) - RX	TX - R(3) - R(1) - R(3) - RX	4.88	3
		TX - E(1) - E(2) - E(1) - RX	TX - R(1) - R(3) - R(1) - RX	5.44	3

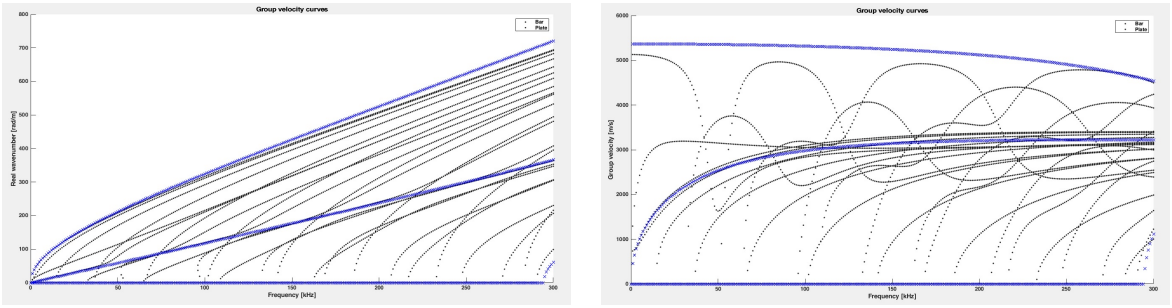
Table 12.3: Path distances for the different configurations and involving up to 3 reflections (RX(1) after the defect).

Configuration	Receiver	Path elements	Path (graphs)	Path length [m]	Number of reflections
		TX - RX	TX - RX	0.32	0
		TX - E(1) - RX	TX - R(3) - RX	1.40	1
		TX - E(2) - RX	TX - R(1) - RX	2.04	1
		TX – E(1) - R(2) - E(1) - RX	TX - R(3) - R(2) - R(3) - RX	2.54	3
		TX - R(2) - E(2) - RX	TX - R(2) - R(1) - RX	2.62	2
	RX(1)	TX - E(2) - E(1) - RX	TX - R(1) - R(3) - RX	3.12	2
		TX – E(2) - E(1) - R(2) - RX	TX - R(1) - R(3) - R(2) - RX	3.17	3
		TX - E(1) - E(2) - RX	TX - R(3) - R(1) - RX	3.76	2
		TX - E(2) - R(2) - E(2) - RX	TX - R(1) - R(2) - R(1) - RX	4.34	3
		TX - E(1) - E(2) - E(1) - RX	TX - R(3) - R(1) - R(3) - RX	4.84	3
		TX - E(2) - E(1) - E(2) - RX	TX - R(1) - R(3) - R(1) - RX	5.48	3
		TX-RX	TX-RX	0.33	0
RX(1) 2.5 cm after the defect		TX - R(2) - RX	TX - R(2) - RX	0.91	1
		TX - E(2) - RX	TX - R(3) - RX	1.39	1
		TX - R(2) - E(2) - RX	TX - R(2) - R(3) - RX	1.98	2
		TX - E(1) - RX	TX - R(1) - RX	2.05	1
		TX - E(2) - R(2) - RX	TX - R(3) - R(2) - RX	2.63	2
		TX - E(1) - E(2) - RX	TX - R(1) - R(3) - RX	3.11	2
	RX(2)	TX - R(2) - E(2) - R(2) - RX	TX - R(2) - R(3) - R(2) - RX	3.22	3
		TX - E(2) - R(2) - E(2) - RX	TX - R(3) - R(2) - R(3) - RX	3.70	3
		TX - E(2) - E(1) - RX	TX - R(3) - R(1) - RX	3.77	2
		TX - E(1) - E(2) - R(2) - RX	TX - R(1) - R(3) - R(2) - RX	4.35	3
		TX - R(2) - E(2) - E(1) - RX	TX - R(2) - R(3) - R(1) - RX	4.35	3
		TX - E(2) - E(1) - E(2) - RX	TX - R(3) - R(1) - R(3) - RX	4.83	3
		TX - E(1) - E(2) - E(1) - RX	TX - R(1) - R(3) - R(1) - RX	5.49	3
		TX - RX	TX - RX	0.37	0
		TX - E(1) - RX	TX - R(3) - RX	1.35	1
		TX - E(2) - RX	TX - R(1) - RX	2.09	1
		TX – E(1) - R(2) - E(1) - RX	TX - R(3) - R(2) - R(3) - RX	2.49	3
		TX - R(2) - E(2) - RX	TX - R(2) - R(1) - RX	2.67	2
	RX(1)	TX - E(2) - E(1) - RX	TX - R(1) - R(3) - RX	3.07	2
		TX – E(2) - E(1) - R(2) - RX	TX - R(1) - R(3) - R(2) - RX	3.22	3
		TX - E(1) - E(2) - RX	TX - R(3) - R(1) - RX	3.81	2
		TX - E(2) - R(2) - E(2) - RX	TX - R(1) - R(2) - R(1) - RX	4.39	3
		TX - E(1) - E(2) - E(1) - RX	TX - R(3) - R(1) - R(3) - RX	4.79	3
		TX - E(2) - E(1) - E(2) - RX	TX - R(1) - R(3) - R(1) - RX	5.53	3
		TX-RX	TX-RX	0.38	0
RX(1) 7.5 cm after the defect		TX - R(2) - RX	TX - R(2) - RX	0.96	1
		TX - E(2) - RX	TX - R(3) - RX	1.34	1
		TX - R(2) - E(2) - RX	TX - R(2) - R(3) - RX	1.93	2
		TX - E(1) - RX	TX - R(1) - RX	2.10	1
		TX - E(2) - R(2) - RX	TX - R(3) - R(2) - RX	2.68	2
		TX - E(1) - E(2) - RX	TX - R(1) - R(3) - RX	3.06	2
	RX(2)	TX - R(2) - E(2) - R(2) - RX	TX - R(2) - R(3) - R(2) - RX	3.27	3
		TX - E(2) - R(2) - E(2) - RX	TX - R(3) - R(2) - R(3) - RX	3.65	3
		TX - E(2) - E(1) - RX	TX - R(3) - R(1) - RX	3.82	2
		TX - E(1) - E(2) - R(2) - RX	TX - R(1) - R(3) - R(2) - RX	4.40	3
		TX - R(2) - E(2) - E(1) - RX	TX - R(2) - R(3) - R(1) - RX	4.40	3
		TX - E(2) - E(1) - E(2) - RX	TX - R(3) - R(1) - R(3) - RX	4.78	3
		TX - E(1) - E(2) - E(1) - RX	TX - R(1) - R(3) - R(1) - RX	5.54	3



(a) (b)

Figure 12.17: Group delay dispersion curves of the bar (configuration 2.5 cm before the cut) superimposed to the spectrograms computed from the data obtained from the receivers placed (a) on the damaged side; (b) on the pristine side.



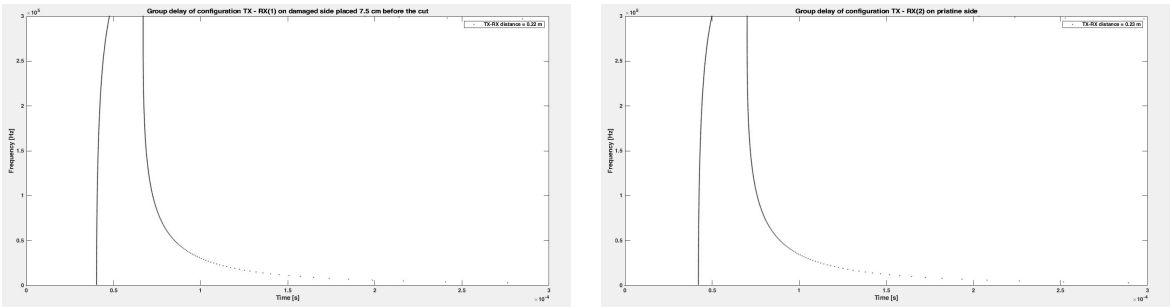
(a) (b)

Figure 12.18: Comparison between the dispersion curves of the steel bar and of the steel plate with the same thickness of the bar used in the qualitative experiments: (a) dreal wavenumber, (b) group velocity.

To make a first analysis, in order to begin to understand what happens in the waveguide, there is the need to simplify the representation. In fact, it would be useful to reduce the number of curves to plot for each reflection. In doing this, we can think of approximating the bar to an infinitely large plate, made of the same material and having the same thickness, and compute the dispersion curves for the plate Lamb waves. Obviously, the curves of a plate are not a faithful representation of the curves in a bar, but they allow to simplify the representation considerably, without loosing some fundamental information. The comparison between the dispersion curves of a bar (in black in the figure) and those of a plate (in blue) help explaining this concept (see Fig. 12.18). The representation is simplified, but the overall trends are kept.

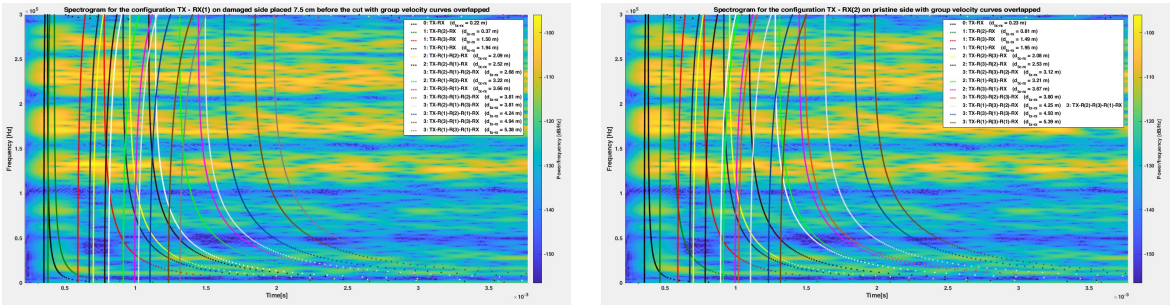
The curves in Fig. 12.18 were calculated by simulating a waveguide made of the same material: the change concerned only the geometry and, consequently, the mesh. Once understood (in broad terms) what happens using a more simplified representation, it is then possible to move on to a more complete and accurate representation, and observe every detail.

It is now possible to proceed with the analysis of the various configurations using the curves of a plate.



(a) (b)

Figure 12.19: Group delay dispersion curves (configuration 7.5 cm before the cut, transmitter-receiver distance = 22 cm) computed for the receivers placed (a) on the damaged side; (b) on the pristine side.



(a) (b)

Figure 12.20: Group delay dispersion curves of the bar (configuration 7.5 cm before the cut) superimposed to the spectrograms computed from the data obtained from the receivers placed (a) on damaged side; (b) on pristine side.

Regarding the first configuration (RX(1) placed 7.5 cm before the cut), the group delay curves for the direct path from TX to the two receivers, RX(1) and RX(2) are shown in Figs. 12.19(a) and 12.19(b), respectively.

Figure 12.20(a,b) shows the group delay curves for the two receivers; these are evaluated taking into account the length of all the propagation paths listed in Table 12.2. The figure is very useful to have a general picture but, despite the simplification of the representation, it is still difficult to interpret.

If each path corresponds to a set of group delay curves, these can be grouped and graphed in reduced sets, chosen so that the distance between one group and the other is maximized to avoid excessive overlaps. This has been done, and the result for this configuration is shown in Fig. 12.21.

A very similar analysis can be done, with the help of Table 12.2, both for the second receiver RX(2) (analyzing Fig. 12.21(b, d, f, h)), and for the other configurations

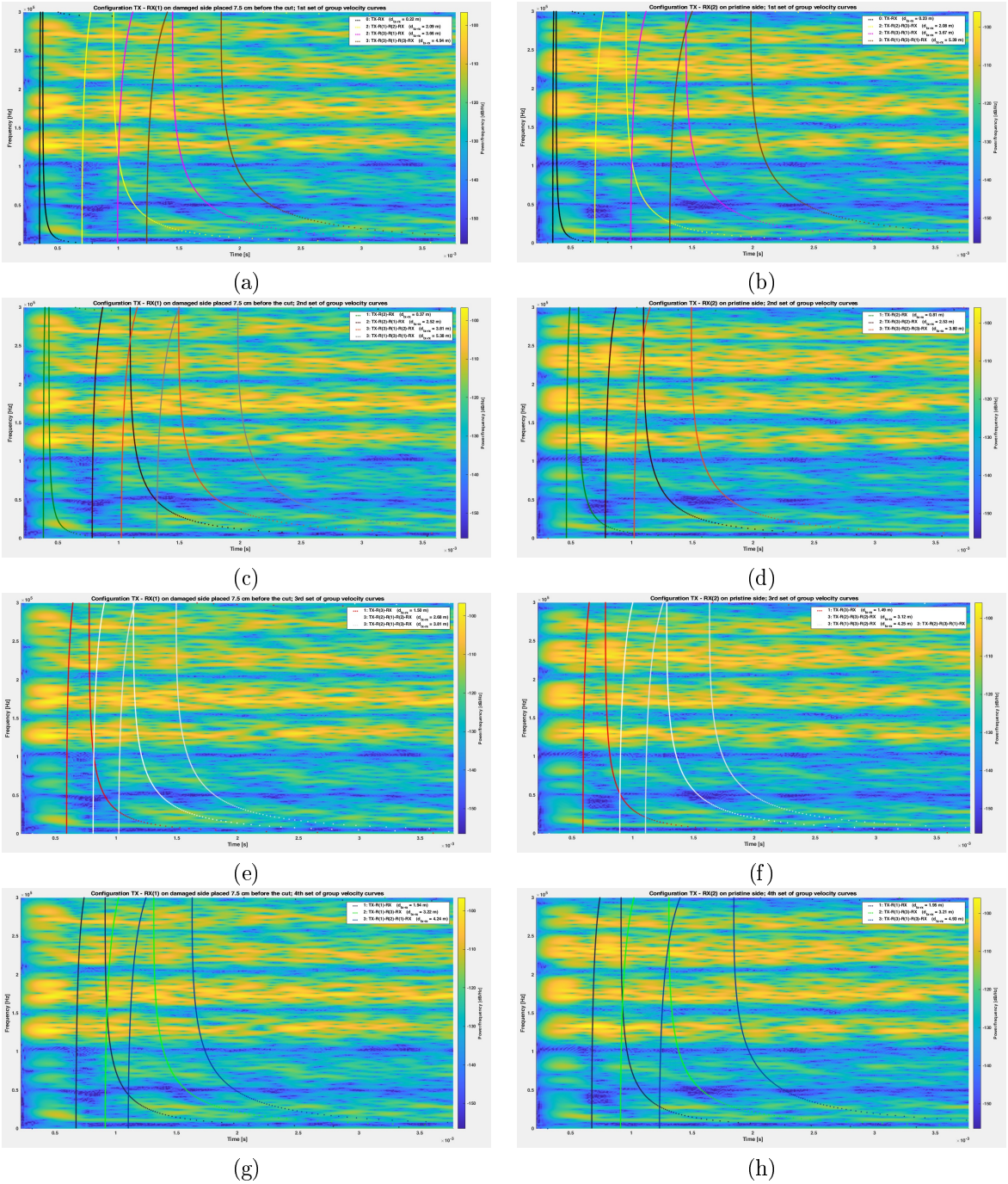
- Figures 12.22 for the second configuration (RX(1) 2.5 cm before the cut);
- Figures 12.23 for the third configuration (RX(1) 2.5 cm after the cut);
- Figures 12.24 for the fourth configuration (RX(1) 7.5 cm after the cut).

A detailed analysis can be omitted here so as not to overburden the discussion, but it is appropriate to focus on some details.

As far as the first configuration is concerned,we focus on the analysis of the data collected by the receiver RX(1) and shown in Fig. 12.21(a,c,e,g). We can to note that the first packet of waves to arrive traveled along the direct path. Immediately after, the second packet of waves is related to waves that, while propagating towards E(1), encounter the defect R(2), and are then reflected back towards the receiver RX(1). In fact, the additive contribution of the two wave packets is clearly highlighted by the spectrogram intensity, even at medium-low frequencies. The waves that are reflected only on the E(1) extremity arrive immediately after. Their contribution at frequencies below 100 kHz does not seem to be particularly intense, while it seems to be more significant at medium-high frequencies. A short time later, the contribution of the curves reflected from the opposite end of the waveguide with respect to the position of the transducer, E(2), is visible. The multiple reflections continue: we must keep in mind that the waves, once generated by the transmitter, because of reflection over the defects and on the extremities of the waveguide continue to propagate, passing multiple time under the receivers. The path legth gets longer and signal dispersion increases. Due to the dispersion and the different propagation speed of the modes, the contributions of the waves that have traveled paths with even rather different lengths, tend to overlap.

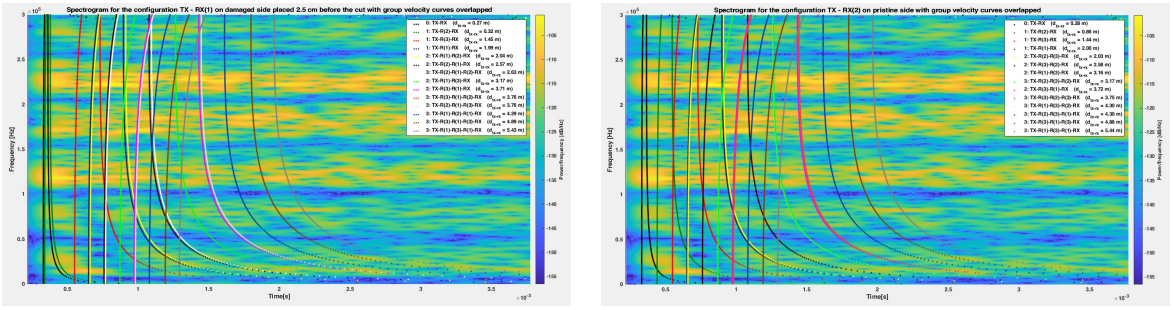
From the analysis of the data collected by the receiver RX(2) shown in Fig. 12.21(b,d,f,h), it can be inferred that even in this case the second wave packet arriving beyond the direct path is the one reflected by the defect. In this case, however, the waves traveled further the receiver RX(1). However, these two contributions appear to be quite relevant at both high and low frequencies. The same argument previously made on the waves reflected by the extremities is still valid: their energy seems to be more concentrated at medium and high frequencies (100-150 kHz and 200-250 kHz in particular). Note also that different contributions of waves reflected on different discontinuities arrive at the transducer simultaneously or almost simultaneously, and this could also cause effects of constructive interference.

Analyzing the data of the second configuration in Fig. 12.22 allows us to explain that the peaks seen in the signal received by RX(1) at relatively high frequency are probably due to the superposition and constructive interference of many paths, which especially involve multiple reflections of the waves on bar extremities.



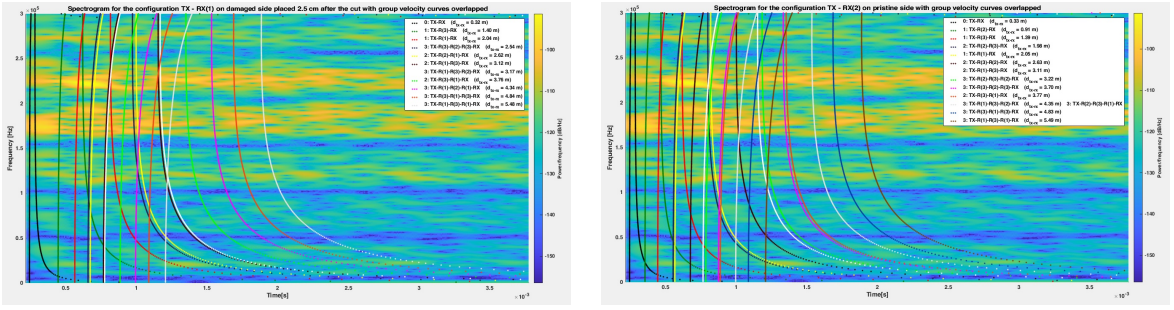
(a) (b) (c) (d) (e) (f) (g) (h)

Figure 12.21: Group delay dispersion curves of the bar (configuration 7.5 cm before the cut) superimposed to the spectrograms computed from the data obtained from the receivers placed (a), (c), (e) and (g) on damaged side; (b), (d), (f) and (h) on pristine side. (a) and (b) 1st set of curves; (c) and (d) 2nd set of curves; (e) and (f) 3rd set of curves; (g) and (h) 4th set of curves.



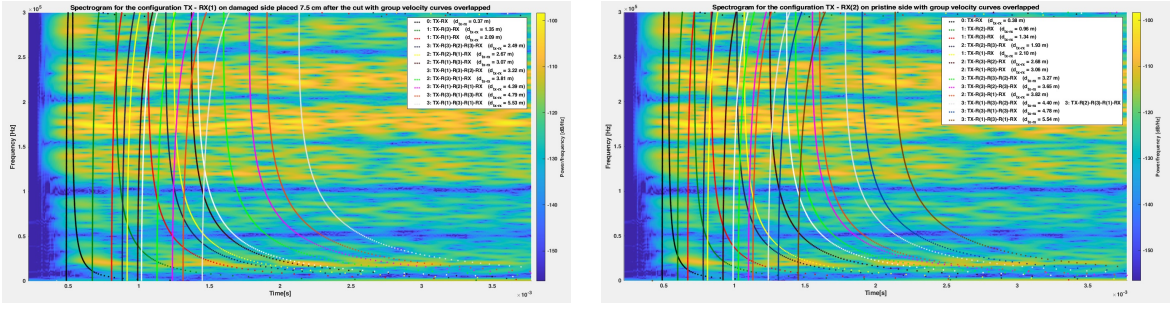
(a) (b)

Figure 12.22: Group delay dispersion curves of the bar (configuration 2.5 cm before the cut) superimposed to the spectrograms computed from the data obtained from the receivers placed (a) on damaged side; (b) on pristine side.



(a) (b)

Figure 12.23: Group delay dispersion curves of the bar (configuration 2.5 cm after the cut) superimposed to the spectrograms computed from the data obtained from the receivers placed (a) on damaged side; (b) on pristine side.



(a) (b)

Figure 12.24: Group delay dispersion curves of the bar (configuration 7.5 cm after the cut) superimposed to the spectrograms computed from the data obtained from the receivers placed (a) on damaged side; (b) on pristine side.

Even for the second configuration it is possible to make an interesting observation: some reflected waves seem to have a reduced energy content at low frequency (below 100 kHz) compared to others. By analyzing the graphs better, we realize that when the waves pass in correspondence of the defect, the wave front that is reflected has a good low-frequency spectral content. The low frequency waves that continue and are reflected on the extremity, when they return towards the transducer (passing once again through the defect), have a lower intensity. Instead, the waves that from the transmitter were directed towards the extremity far from the defect, when they reach the receiver have not yet encountered the defect and retain the low-frequency signal components.

This aspect is enhanced by analyzing the third and fourth configuration. Remember that, in this case, the receiver RX(1) is placed immediately after the defect, so any wave that reaches it (even the direct path) has passed through the defect. Consequently, a portion of the energy of those signals is reflected and this translates into a lower intensity of the measured signals, as already noted. It is also noted, however, that the signal intensity in the lower half of the spectrogram is lower for the signal measured by RX(1) than for the signal measured by RX(2). This can be explained by the fact that RX(2) not only collects waves that have not passed through the defect (for example the direct path or the paths involving the bar extremity E(2)), but also the components of the waves that were reflected by the defect itself even before reaching RX(1). The reflection of low-frequency waves by the defect is obviously not total, some of the energy still manages to pass and is visible in the spectrogram.

Previously, images of the spectrograms with superimposed the bar group delay curves were shown (see Figs. 12.17). It was stated that the analysis was difficult at first. With the experience gained by evaluating the behavior of the plate group delay curves, it is now possible to attempt a more in-depth analysis of those curves. It is immediately appreciable how these curves overlap even more faithfully with the underlying spectrogram, which in fact reproduces the behavior observable in higher-order modes. Obviously, the resolution obtainable from the graph does not allow to assign with absolute certainty each curve to each pattern in the graph, but the level of understanding is now much more complete. By refining the graphing of data in the joint domain of time and frequency, and by better defining the propagative characteristics of each mode, the analysis of a similar graph can help in identifying the most suitable modes for detecting a defect in the chosen wave. This graph allows to evaluate which discontinuities affect the propagation of certain modes, as each pattern can be linked to a specific mode (or to more than one).

Generally speaking, it should not be forgotten that it is always the signal radiated by the transmitter TX that propagates, initially in two directions:

- towards RX(1) and then towards the defect R(2) and the extremity E(1);
- towards RX(2) and then towards the extremity E(2).

It is this signal which, from reflection in reflection, passes several times under the receiver of interest. Each time the propagating signal is attenuated, modal conversion occurs, part of the energy is dispersed. In any case, the signal that the receiving transducer measures at any moment is given by the local contribution of the constructive or destructive interference of all the ultrasonic guided waves that cross the waveguide at that point.

12.2.6.3 Partial Conclusions

Spectrograms have shown that they are able to represent the group delay curves of the various reflected wave packets quite clearly, despite the compromise on time-frequency resolution. Consequently, the overlap between the simulated group delay curves and those obtained from the processing of the collected signals is also quite good. This allows us to say that the correctness of the simulation results has been tested in the field, given the good correspondence between the curves. This despite the numerous variable and unknown parameters. Analyzing the spectrogram with the help of the various sets of curves, a hypothesis that can be made is that the defect (the cut in the bar) reflects more low-frequency ultrasonic waves. This is an aspect that can be found by analyzing all the configurations. This hypothesis is worthy of further study.

12.2.7 Conclusions on the Study of the Effects Caused by the Presence of a Defect in a Waveguide

In this phase of the experimental activity, the focus was on the analysis of the effects of a defect in the UGW propagation along a waveguide. To carry out the experiments, a steel bar with a rectangular section was chosen. Despite the relative simplicity of the waveguide structure, the propagation environment is still relatively complex: it is expected that the rail environment is even more so.

In analyzing the results obtained one factor must be considered: the propagation speed of the UGW in the materials is particularly high, and the waveguide used in this experiment is not particularly long. Thus, the signal that propagated in the bar was reflected in a brief time interval over and over again on the ends of the waveguide itself, or on the defect. This only added complexity to complexity. In any case, the results obtained are still valid and interesting.

Analyzing the data collected, it is possible to realize how difficult it is to explain the trend of the signals in the time domain without resorting to more in-depth analysis. In particular, without knowing the signal (and reflections) behavior over time and frequency, it is difficult to associate a cause to each peak in the signal. Even analyzing the distribution of the spectral density of energy it is possible to somehow guess the presence of a defect, especially if it is located between transmitter and receiver. But it is not possible to establish, at least on the basis of the data collected so far, a reliable method to ascertain its presence.

Knowing the exact position of the discontinuities allowed to trace the origin of each peak in the measured signals or, on the contrary, to establish the effect of discontinuities on signals that can be measured. This was made possible by exploiting the representation of the data in the joint domain of time and frequency through the spectrogram.

Of course, the experiments conducted so far have had the task of suggesting a possible method to explain the consequences of a defect, not to certify the reliability of the method itself. It is still not possible to explain the causes of signal behavior if the position of the discontinuities is unknown. The method used up to now is not capable to identify defects in a signal collected on a waveguide of unknown conditions.

The purpose of this experimentation, however, was rather to begin to understand how the guided waves interact with each other and with the structural elements of the waveguide as they propagate. In these terms some positive results have been obtained, but it will still be necessary to carry out further research activities to investigate the aspects that emerged.

12.3 Experimental Observation of Rail Dispersion Curves

As anticipated at the beginning of this chapter, the second experiment concerned the actual in-field measurement of the dispersion curves of a rail, expressed in terms of frequency f and wavenumber k.

A representation of the dispersion curves of a waveguide expressed in these terms can be obtained starting from signals measured in the time domain (quite similar to those measured and analyzed in Section 12.2). To achieve this result, one method is to collect these signals at regular and close intervals along the waveguide, that is, by performing a *spatial scan* of the waveguide. The result is a bi-dimensional matrix in which each row contains the samples of the signal acquired in the time domain at a given position. This matrix, analyzed through a *bi-dimensional fast Fourier transform* (2DFFT) will result in a graphical representation of the dispersion curves, which can then be compared with those obtained from SAFE simulation.

This experimental activity also has the aim of verifying the possibility of obtaining directly the plot of the dispersion curves by using a relatively simple, robust and inexpensive instrumentation.

Figure 12.25: Measurement set-up for the experimental observation of the dispersion curves of a rail.

12.3.1 Measurement set-up

To perform a spatial scan consisting of a large number of acquisition points, with an uniform and small spacing between points as possible, it is not possible to resort to manual fixing of the transducers to the waveguide. In fact, this lacks the required precision, is laborious and takes too much time, preventing both the possibility of having a high number of points and, more importantly, their precise and regular positioning.

To solve the problem of manual fixing, the air probe described in Section 12.1 can be used. In fact, if a fixed transducer at one end of the waveguide is used as an ultrasonic signal generator, the air probe can then be used as a mobile receiver. To ensure positioning at short, even intervals, the air probe can be moved from point

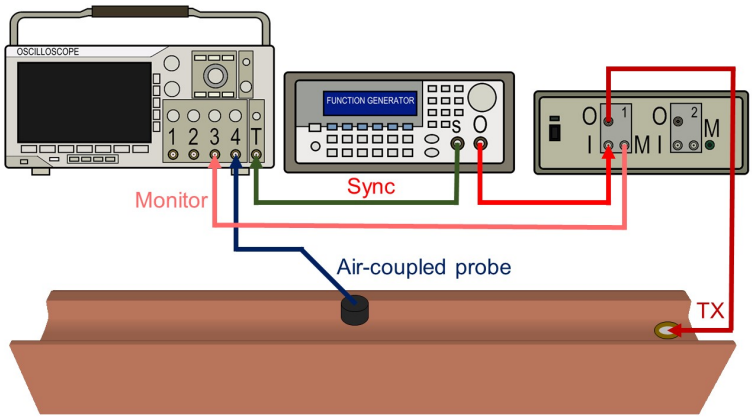


Figure 12.26: Schematic representation of the measurement set-up for the experimental observation of the dispersion curves of a rail.

Figure 12.27: Section of rail used as a waveguide.

to point by mechanical movement. To speed up the acquisition and ensure the best results, the mechanical movement can be controlled by computer and synchronized to the generation and acquisition of the signal.

The employed measurement set-up, shown in Fig. 12.25 and schematized in Fig. 12.26, made use of:

- a piezoelectric transducer (Murata 7BB-41-2L0 or DuraAct P-876.SP1) as transmitter;
- a Microacoustic BAT-1 ultrasonic air probe with Microacoustic Q-AMP pre-amplifier as receiver;
- a MyDIYCNC CNC mill: its movement system was used to move the air probe along the waveguide;
- an Agilent 33220A function generator to generate signals;
- a Tegam 2350 signal power amplifier to amplify the signals generated by the function generator before sending them to the transmission trnasducer;
- a Tektronix DPO-3014 oscilloscope to acquire the signals measured by the air probe.
- a PC to which the function generator, the oscilloscope and the cutter were connected; this device managed the signal acquisition process using a matlab algorithm based on a user-defined set of parameters.

A 1 m long section of 60E1 rail was used as a measurement sample (see Fig. 12.27). The exact type of steel alloy used to fabricate the given specimen is not known exactly, but analyzing the technical specifications of RFI [112] it is known that the most common type used on the network Italian is R260 steel[7]; it is characterized by the properties listed in Table 12.4.

Of the MyDIYCNC CNC mill, as anticipated, only the head movement system is used. The air probe is anchored to the CNC mill head by means of a support. In this way it is possible to control the displacement of the probe along the rail surface. The mill had a movement on 3 axes, x, y and z. To move the air probe, the movements of the y and z axes were therefore used: the y axis was used to displace the probe along the waveguide, the z axis to adjust the height of the probe with respect to the surface of the test specimen. The movement accuracy guaranteed by the mill manufacturer ensured precise management of the positioning of the air probe along the waveguide.

Given the complex shape of the rail, the results obtained from the analysis of the collected signals may vary depending on the part of the rail that is being analyzed. The UGW leaking out from two different parts of the

[7]Other steel alloys commonly used for the manufacture of rails used on the RFI network are: R320Cr chromium alloyed steel (1% Cr) not heat treated, and non-alloyed R350HT steel Carbon-Manganese (C-Mn) heat treated. Heat treated steels are generally used in tight curves to improve wear resistance.

Table 12.4: R260 rail steel characteristics.

Parameter	Value
Steel quality	R260 rail steel
HBW hardness (field of variation)	260 ÷ 300
Description	Non-alloy Carbon-Manganese (C-Mn) steel, not heat-treated
Density	7850 kg/m^3
Young modulus E	210 GPa
Poisson's ratio ν	0.25

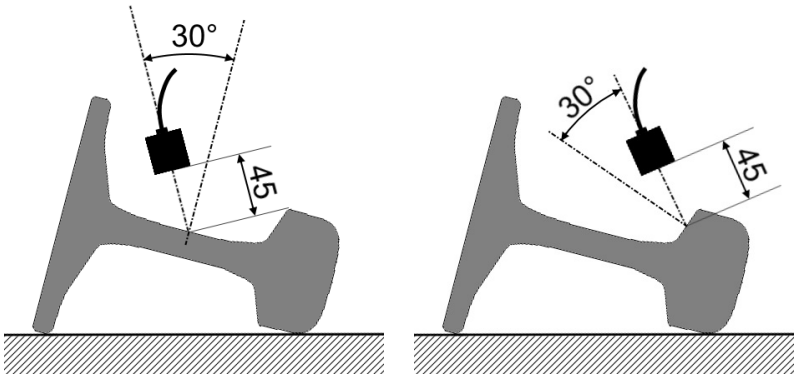


Figure 12.28: Position of air probe when measuring the UGW propagating in the web (a) or in the head (b).

rails were measured, the web and on the fishing surface under the head (see Chapter 2). The position of the air probe is depicted in Fig. 12.25 for the two cases.

The piezoelectric transducer was fixed to the surface of the rail by interposing *phenyl salicylate* ($C_{13}H_{10}O_3$) or *salol*, a crystalline substance characterized by a melting temperature of 41.5 °C. This crystalline material, once melted and re-crystallized, is able to fix the transducer to the surface of the sample being measured in a stable but easily reversible way. In addition, it guarantees a good acoustic coupling between the transducer and the waveguide.

Before the start of the acquisitions, it was ensured that the air probe was positioned in order to collect as much signal as it was dispersed by the waveguide into air. Using the movement along the z axis of the mill it was possible to focus the air probe. The optimal distance between the air probe and the waveguide surface was found to be around 4.5 cm. The inclination of the probe surface has also been set at 30° with respect to the surface of the waveguide to ensure optimal reception of signals. The inclination is useful to avoid that the sound waves, once they have reached the air probe diaphragm, are reflected back towards the surface of the specimen, returning once again towards the sensitive surface of the probe.

The instrumentation was connected as in the diagram in Fig. 12.26. Both the function generator and the oscilloscope are connected to and managed by a PC.

The generation and acquisition of signals take place in coordination with the movements of the air probe ensured by the CNC mill, managed in turn by the matlab script.

12.3.2 Excitation Signal

Once again, it was decided to use a broadband signal as the excitation signal. The underlying reasons remain those anticipated in Section 12.2: being interested in observing the waveguide dispersion curves, it is necessary to use a signal capable of exciting as many modes as possible in the widest spectrum available. The chosen signal is a sinusoidal chirp, characterized by a variable frequency and amplitude. The waveform is illustrated in Fig. 12.29 and its parameters are indicated in Table 12.5.The amplitude of the signal is variable because of the characteristics and performance limits of the function generator and power amplifier.

12.3.2.1 Acquired Signals

The result of the acquisition consists in a set of signals collected by the air probe. Each signal corresponds to a position along the waveguide, so that the total number of acquired signals equals the number of acquisition points.

Table 12.5: Chirp signal parameters.

Parameter	Value
Type of signal	Sinusoidal chirp
Initial frequency	15 kHz
Final frequency	500 kHz
Sweep time	20 ms
Chirp mode	logarithmic
Initial signal amplitude	dependent on the function generator settings
Final signal amplitude	10% of the initial signal amplitude
Signal offset	0 V
Number of pulses per burst cycle	1
Burst repetition interval	2 s

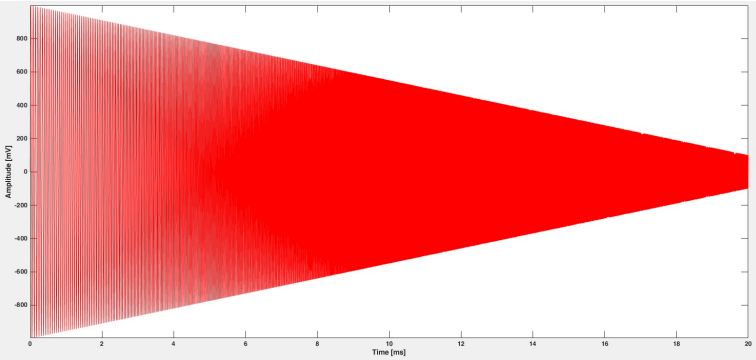


Figure 12.29: Chirp signal used as excitation signal. The frequency varies logarithmically in the range [15 kHz, 500 kHz], wile the amplitude decreases linearly towards the 10% of the initial amplitude. The function generator and the amplifier amplify the signal to the desired level.

In collecting the signals, the air probe starts its motion from a point close to the transmitter (which continuously generates the signal in bursts) and gradually moves away one step at a time. At each step, the stored signal is the result of a moving average calculated by the oscilloscope. The number of averaged signals (64) is the result of a compromise between the acquisition time and noise reduction.

The acquisition time window is 40 ms long, the sampling frequency is equal to 25 MS/s. An example of the collected data (and of the scaled copy of the signal actually transmitted) is shown in Fig. 12.30.

The probe's movement profile, or the characteristics of the spatial scan, are described by the so-called *mission profile*. In the mission profile the user specifies

- the number of steps;
- the direction of movement;
- the distance between one step and another.

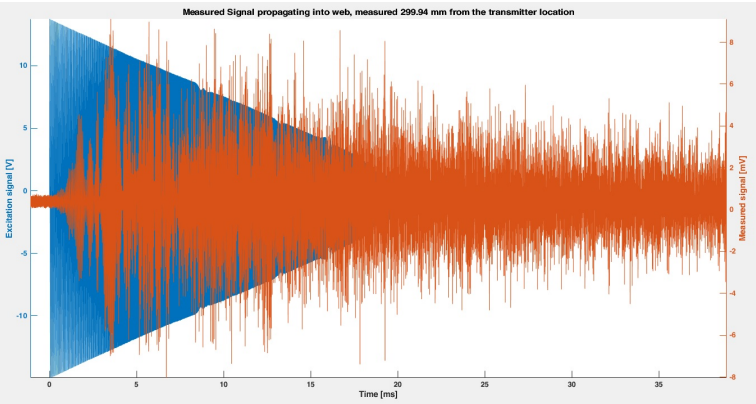


Figure 12.30: Example of received and transmitted signals in the time domain.

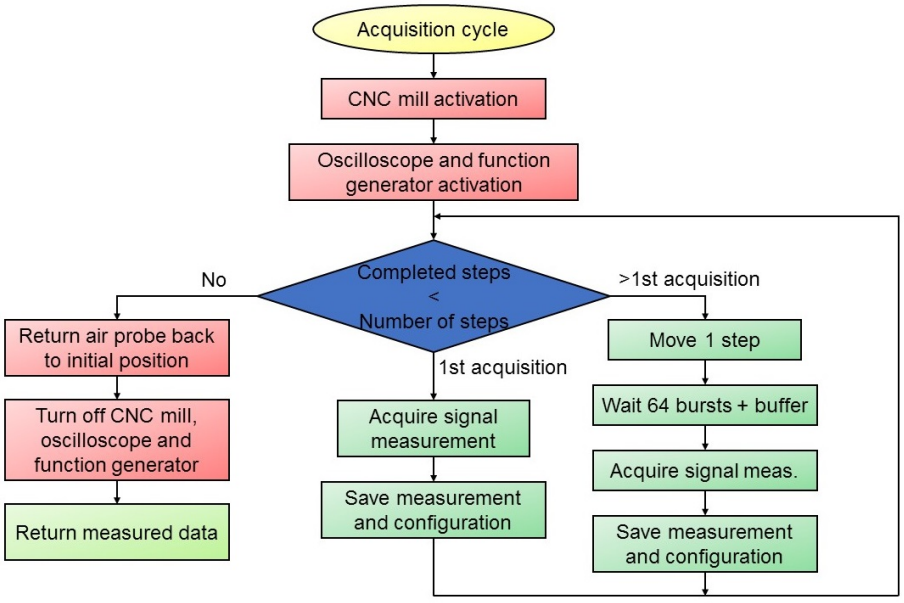


Figure 12.32: Flow chart of the algorithm implementing the spatial scan of the waveguide.

From these datathe overall distance covered by the spatial scan can be derived. The CNC mill used can offer a maximum head displacement along the y axis approximately equal to 36 cm. For each set of acquisitions, a file is stored containing the main settings used for carrying out the measurements.

12.3.3 Algorithms for the Acquisition of Signals and theirs Analysis

12.3.3.1 Algorithm for Data Collection

Data collection takes place following an algorithm sketched through the flow chart of Fig. 12.31. The script has a simplified (but useful) user interface for setting or loading the parameters necessary for the measurement. The data acquisition cycle is summarized in the flow chart of Fig. 12.32.

The acquired data are then stored (the time instances vector is saved only for the first acquisition, otherwise it is always the same) together with the updated distance of the acquisition point from the transmitter. Given that the acquisition cycle can have a considerable duration, it is preferred to save the measured signal from time to time, protecting against accidental interruptions of the acquisition and freeing up RAM memory in the PC.

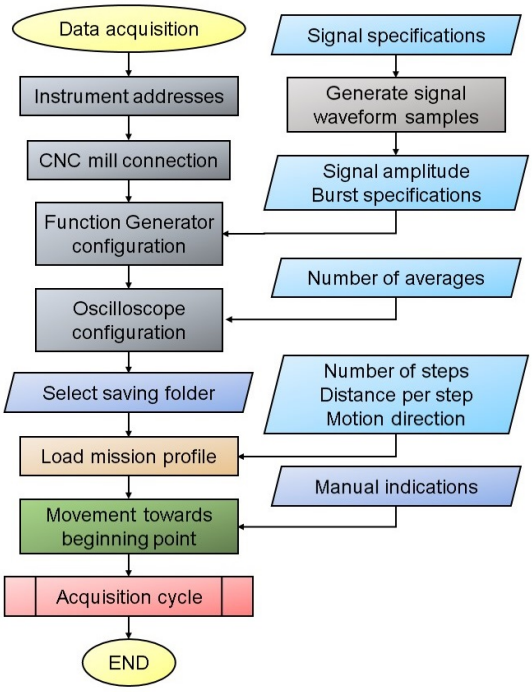
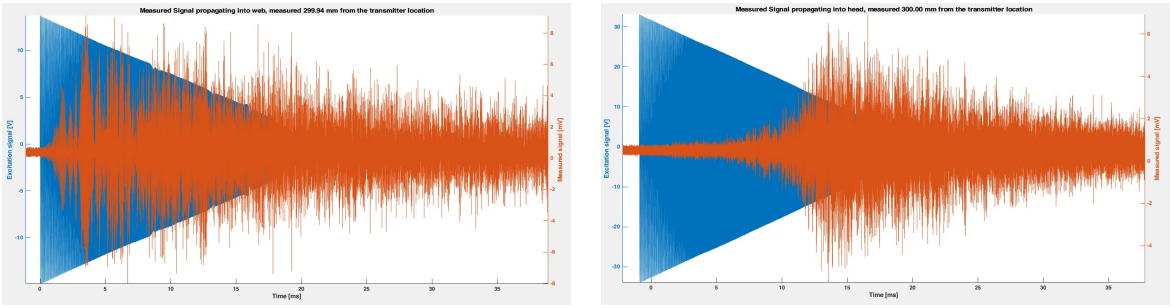


Figure 12.31: Flow chart of the signal acquisition algorithm.

12.3.3.2 Algorithm for Data Analysis

The collected data matrix undergoes a two-dimensional Fourier analysis in order to provide the representation of the waveguide dispersion curves. To carry out this analysis, a second algorithm is used, written to analyze the data collected in the previous phase. The algorithm for data analysis is summarized in the flow chart shown in Fig. 12.33.

First, a rough compensation of the delay due to the different velocity of the ultrasonic waves in air is carried out. In fact, while the waves group velocity in the rail steel reaches up to 5000 m/s, it does not exceed 343 m/s in air. The desire in this case is to measure the dispersion curves due only to the propagation in the waveguide and not through the trip into air. Knowing the approximate distance that the waves travel in air between the point where they exit the waveguide and the air probe, and the speed of sound, the



(a) (b)

Figure 12.34: Example of received and transmitted signals in the time domain acquired when scanning (a) the surface of the web; (b) the fishing surface of the rail under the head.

introduced delay t_{air} can be computed. This delay can be compensated for by neglecting the first t_{air} seconds in the measured signal. This is done by neglecting a given number of samples at the beginning of each signal; the number is calculated knowing the sampling period.

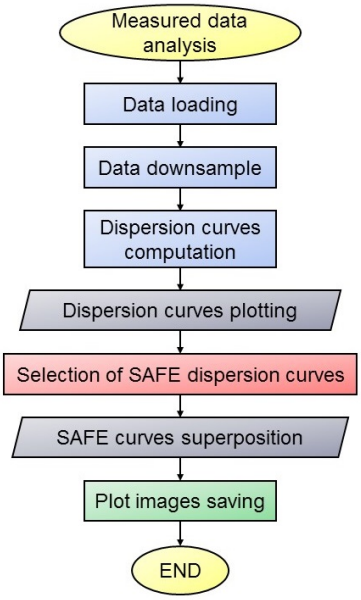


Figure 12.33: Flow chart of the algorithm used to analyze the acquired data.

Subsequently, the data is sub-sampled with a factor of 5 through the matlab `resample` instruction. This operation is necessary to simplify data processing by the computer; in any case, the representation of the data is in any case faithful, as Nyquist's law remains valid (the sampling frequency of the signal drops from 25 MS/s to 5 MS/s). It was preferred to store the signals collected with a very high level of detail so that, in case of future need, they can be analyzed without having to repeat the acquisition. At present, data analysis is also possible on a relatively high-performance computer by resorting to a downsample. During the downsample, from the collected signals is subtracted the value of their median, in order to eliminate any DC value. The product of this processing is still a matrix (this time of smaller dimensions).

The stored matrix can therefore undergo a *two-dimensional Fourier transform* (2DFFT) (instruction `fft2` of matlab). The result of the processing, i.e., the dispersion curves of the waveguide expressed in terms of wavenumber k and frequency f, is then plotted and displayed to the user, who can decide whether or not to save the resulting figure.

It is then possible to proceed with the superimposition on the plot of the dispersion curves of the waveguide simulated by SAFE (once again simulated using the GUIGUW software) and saving the new result. Again, a good match should be observed in the superposition between the simulated dispersion curves and those obtained from the measured data.

12.3.4 Analysis of the Processing Results

In this section we analyze the results produced by the analysis algorithm when processing the data collected on the web side and in the lower surface of the railhead (fishing surface). We analyze both the dispersion curves obtained from the collected data, and the results of the superposition with the simulated curves.

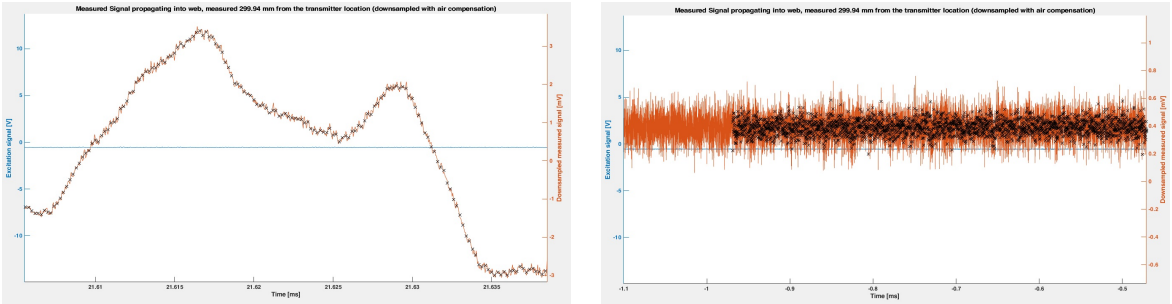
In Table 12.6 the parameters of the acquisition and of the analyzed signals are listed.

Fig. 12.34 shows two examples of data collected in the time domain for the two different parts of the analyzed rail. For both examples, the signals are acquired in a point placed at approximately 30 cm from the transmitter location. Fig. 12.35(a) shows an enlargement of the signal depicted in Fig. 12.34, to highlight the effect of downsampling. This operation also performs a filtering of the high-frequency noise signal that has remianed after the averaging of 64 acquisitions. The effect of the compensation of the air propagation is visible in Fig. 12.35(b): the very first milliseconds of the signal collected are neglected.

The signals collected have a particularly low intensity; this is due to the different cross-section thickness of the rail compared to the previous waveguide, and the use of the air probe. In the time domain signals the presence of a large number of peaks is also observed; this is due to the much more complex propagation environment constituted by the rail. The reduced length of the rail (1 m) also means that the effect of the

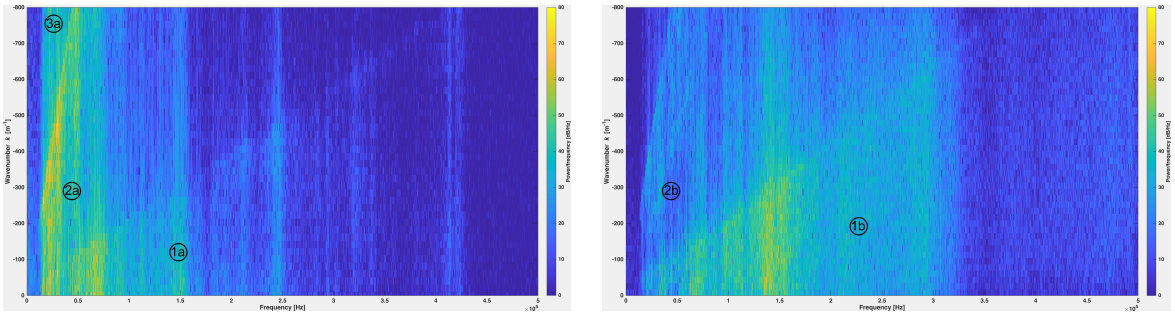
Table 12.6: Acquired dataset configuration data.

Parameter		Web	Head
Mission profile	*Number of acquisition points*	701	721
	Step interval	0.508 (0.02")	0.50 (0.0197")
	Total distance	356.11 mm	360.77 mm
	Initial distance from transmitter	145 mm	210 mm
Executed mission profile	*Number of acquisition points*	608	721
	Effective total distance	308.86 mm	360.77 mm
Chirp	*Initial frequency*	15 kHz	
	Final frequency	500 kHz	
	Mode	logarithmic	
	Duration	20 ms	
	Initial signal amplitude	30 V	70 V
	Final signal amplitude	3 V	7 V
Transmission transducer	*Type*	DuraAct P-876.SP1	Murata 7BB-41-2L0
	Placement	Head fishing surface	Web side
Burst	*Number of pulses per burst cycle*	1	
	Burst repetition interval	2 s	
Measured signal	*Duration*	40 ms	
	Number of samples	1,000,000	
	Sampling frequency	25 MS/s	
	Number of signal in average	64	
Analyzed signal	*Downsample rate*	1/5	
	Temporal sampling frequency	5 MS/s	
	Spatial sampling frequency	12368.47 rad/m	12556.83 rad/m



(a) (b)

Figure 12.35: Detail of the signal acquired when scanning the surface of the web. (a) Effect of downsample: in orange the signal sampled at 25 MS/s, the black crosses are the samples of the downsampled signal at 5 MS/s. (b) Effect of the compensation of the delay due to signal travel into air: in orange the non-compensated signal, the black crosses refer to the downsampled and compensated signal.



(a) (b)

Figure 12.36: Dispersion curves plot result of the bidimensional fourier transform of the set of signals acquired when scanning (a) the surface of the web; (b) the fishing surface of the rail under the head.

reflections on the extremities on the collected signals is soon felt. At this stage, however, our interest is not to analyze the trend of the signal in the time domain or to study the effect of reflections.

The result of the signal processing is shown in Fig. 12.36(a) for the wen and 12.36(b) for the head. Looking at the plots, some considerations can be easily inferred:

- the resolution, especially on the vertical axis, is not the best;
- two different energy concentrations are noted; these are identified by (1a) and (2a) in Fig. 12.36(a), and (1b) and (2b) in Fig. 12.36(b).

It is now possible to proceed with a point-by-point analysis of the results achieved.

The vertical resolution in the graph is directly proportional to the length of the spatial scan. In fact, to have a higher resolution in the graph it is not necessary to increase the resolution of the spatial scan, i.e., to collect more points at close intervals: it is necessary to cover a greater distance. The resolution of the spatial scan, that is, the number of collected points, only affects the limits of the vertical axis of the graph, thus the maximum wavenumber that can be represented. As the interval between one point and the next decreases, the resolution of the spatial scan increases and therefore the number of measurement points per acquisition campaign. Consequently, the maximum wavenumber value that can be plotted increases. Unfortunately, the CNC machine used to move the probe along the waveguide had a maximum distance of 36 cm; this sets a lower limit on the vertical resolution. The only parameter left to the user's choice is related to the maximum wavenumber that can be represented. The spatial resolution of the scan imposes a trade-off on the duration of the acquisition: as the number of points increases, the acquisition time increases in a directly proportional way, since it is necessary to wait at each step for the oscilloscope to average the signals. This is a point to be taken into account even if the distance covered by the spatial scan is increased, keeping the distance between the various points constant in order not to alter the maximum wavenumber value. Table 12.7 collects the effects on the resolution of the graph of the parameters used for the execution of the measurement campaign.

The two different concentrations of energy, which appear to reach asymptotes in the graphs of Fig. 12.36 consisting of oblique lines that unfold approximately from the origin, are in fact the representation of the dispersion curves expressed in terms of wavenumber vs frequency.

This will be become more apparent later with the superposition of the simulated dispersion curves; however, it it can be shown that the energy concentrations (1a) and (2a) are relative to the wave dispersion curves in the rail, while the concentrations (1b) and (2b) are relative to the propagation of sound into air. In fact, whenever a pulse is generated by the piezoelectric transmission transducer (a buzzer), part of the impressed energy travels as UGW in the rail, but part of the energy propagates as sound in the surrounding air (at the speed of sound). This energy is also measured by the air probe. Obviously, the energy that propagates into air and that can be measured is stronger as the intensity of the generated impulse becomes greater, and this is easily seen by comparing Figs. 12.36(a) and (b).

- In the case of the web the transducer (Murata 7BB-41-2L0) was excited with a pulse with an initial amplitude of 70 V;
- in the case of the measurement under the head, the transducer (DuraAct P-876.SP1) was excited with a pulse of initial amplitude 30 V.

In any case, the correctness of these intuitions can be easily confirmed by exploiting the definition of wavenumber. If

Table 12.7: Dispersion curves plot resolution and computation time.

Parameter	Web	Head	Resulting parameter	Web	Head
Signal duration t_s	40 ms				
Sampling frequency f_s	5 MS/s		*Horizontal resolution* $d_f = f_s/N_s$	20 Hz	20 Hz
Sampling period T_s	200 ns				
Number of samples N_s	250,000				
Horizontal scanning length l_y	308.86 mm	360.77 mm	*Vertical resolution* $d_k = 2\,\pi/l_y$	20.34 rad/m	17.43 rad/m
Distance between scanning points d_y	0.508 mm	0.50 mm	*Maximum wavenumber* $k_{max} = 1/d_y$	6163.9 rad/m	6265.8 rad/m
Number of points N_y	608	721			
Number of averages N_a	64		*Acquisition time* $t = N_y \times N_a \times t_b$	21h 47'	25h 50'
Burst repetition interval t_b	2 s				

Table 12.8: Observations from the plots of dispersion curves.

Frequency f [Hz]	Wavenumber k [rad/m]	Group velocity v_g [m/s]	Observations
41910	773	340.65	Sound speed into air
213300	427	3138.65	Group velocity of most UGW modes into rail steel (from SAFE simulation)

$$k = \frac{\omega}{v} = \frac{2\pi f}{v}, \tag{12.1}$$

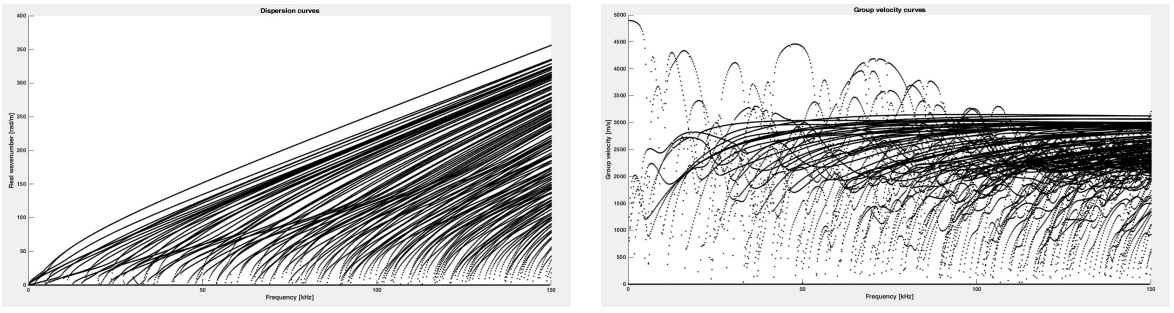
where f is the frequency of the wave oscillation and v is the speed of propagation, it is easily obtained that

$$v = \frac{\omega}{k} = \frac{2\pi f}{k}. \tag{12.2}$$

Therefore, taking two points on the asymptotes delimiting the energy concentrations in the figures, it is easily possible to refer to the maximum propagation speed of the waves that transported that energy. The results are visible in Table 12.8, for the results shown in Fig. 12.36(a).

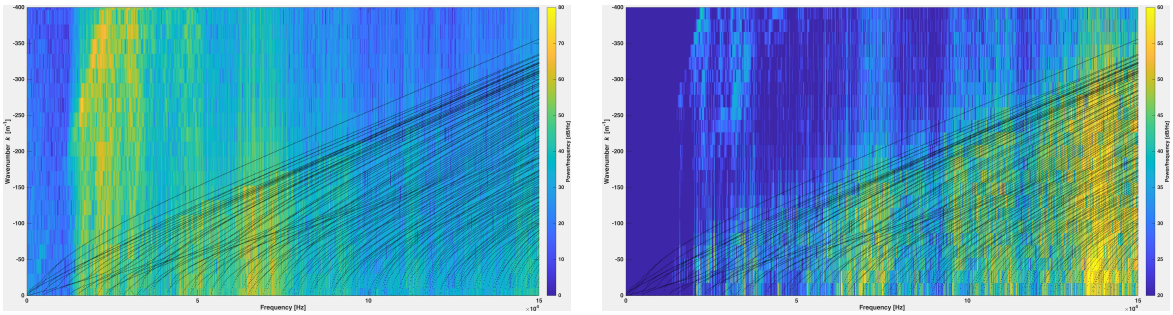
Given that the amplitude of the generated chirp varies, decreasing as the frequency increases, this also has an effect in the energy distribution in the measured signals. The part of the plots corresponding to the higher frequencies experiences a lower intensity than that at the lower frequencies. Likewise, since the chirp has an initial frequency of 15 kHz, a large spectral content cannot be expected at frequencies below that value.

Fig. 12.37 shows the dispersion and group velocity curves simulated by GUIGUW for a 60E1 rail (R260 steel alloy), identical to the one used for the measurements. In the frequency range considered in the graph, from



(a) (b)

Figure 12.37: Results of the SAFE simulation of dispersion characteristics of a 60E1 rail. (a) Dispersion curves wavenumber vs frequency; (b) Group velocity curves.



(a) (b)

Figure 12.38: Dispersion curves plot result of the bidimensional fourier transform of the set of signals acquired when scanning (a) the surface of the web; (b) the fishing surface of the rail under the head, with superimposed the SAFE simulated dispersion curves of a 60E1 rail.

0 to 150 kHz, GUIGUW simulated the behavior of 127 modes. It follows that the propagative environment is much more complex than that of the bar encountered in Section 12.2, as was to be expected.

Figures 12.38(a) and 12.38(b) show the superimposition on the calculated dispersion curves plots the dispersion curves simulated by GUIGUW for a 60E1 rail. It is noted that the intuition had in the previous paragraph regarding the concentration of energy (1a) and (2a) as due to the propagation of UGW in the rail is once again correct. The simulated curves, in both cases, largely overlap the energy concentration.

In the case of Fig. 12.38(a) (web) it can be observed that only some curves do not overlap the energy concentration, those having the highest wavenumber. One possible explanation may be that, while curves are simulated across the rail section, some modes are confined to the rail web, head or foot, as other researchers have already shown. Now, it may be that here modes that do not overlap the energy concentration are modes of propagation of the rail head or foot. This can be also due to the fact that these modes are not easily excited by the transducer, and therefore carry only a minimal part of the energy transmitted.

In the case of Fig. 12.38(b) (lower surface of the head), a better superposition of the modes with a high wavenumber and the energy concentration is observed. This could confirm, at least in part, the first of our previous hypotheses, namely that those slow modes propagate mainly in the head.

Unfortunately, however, due to the poor vertical resolution of the graph it is impossible to trace the trend of the various curves. This is a big limitation, difficult to solve given the available instrumentation. The only solution, in fact, would require increasing the length of the spatial scan, with the following three implications:

- the first is to create an ad hoc movement system to move the air probe;
- the second, it is necessary to use a longer waveguide;
- the third, an equally long acquisition time would be required.

In fact, if an acceptable vertical resolution value d_k is fixed, it is immediately clear that the length of the spatial scan must be given by

$$l_y = \frac{2\pi}{d_k}. \tag{12.3}$$

Instead, to obtain the resolution of the spatial scan, d_y, it is necessary to set the maximum wavenumber value that affects k_{max}, that is

$$d_y = \frac{1}{k_{max}}. \tag{12.4}$$

Given l_y and d_y, the number of steps N_s and therefore the number of acquisition points N_y is obtained very easily

$$N_y = \frac{l_y}{d_y} N_s = N_s + 1, \tag{12.5}$$

as we have seen that the first acquisition is made at the starting point.

Given the number of acquisition points, N_y, if we know (or hypothesize) the characteristics of the generated signal (burst interval, t_B) and the number of readings of the signal that must be averaged for each acquisition point (N_a), it is possible to estimate the lower limit of the time required to complete the acquisition campaign

$$t_{min,\,acq} = N_y \cdot N_a \cdot t_B; \tag{12.6}$$

this does not take into account the time required to move the receiver between one acquisition point and the next, and the impact of the various buffer times.

Another factor to be considered is the volume of collected data: as the number of acquisition points increases, the weight of the data that must be stored on disk increases proportionally. These data must also be loaded into a RAM for processing. The performance of the data analysis algorithm strongly depends on the performance and memory capacity of the machine on which it is run. The algorithm has relatively large margins for improvement in terms of performance, and it is possible to decide to use a lower sampling rate in data collection (if this is compatible with the desired level of detail). It is possible to lower the sampling frequency (using the oscilloscope settings) either by collecting a more limited number of samples per sequence, or by collecting shorter sequences (in time and/or in samples). To maintain a sufficient sampling rate, these two parameters must be balanced. However, it goes without saying that in the presence of a dataset with many acquisition points, the weight of the data begins to be significant, in any case.

12.3.5 Conclusions on the Experimental Observation of Rail Dispersion Curves

This experimental activity has allowed to show that a plot of a waveguide dispersion curves can be obtained using a fairly common equipment in research laboratories.

Although the approach is interesting, it suffers from several and heavy limitations, first of all the compromise between the time required and the length of the spatial scan to achieve a given vertical resolution in the plots. The air probe collects not only a less intense signal than a piezoelectric transducer attached to the waveguide, thus also worsening the signal-to-noise ratio, but also collects the acoustic waves emitted into air by the transmission transducer. This approach also requires the collection and processing of a massive amount of data.

12.4 Comparison Between the Proposed Methods

A comparison can be made between the methods proposed in the last two sections to evaluate the propagative conditions of a waveguide. The method described in Section 12.2 involves calculating the spectrogram of a signal in the time domain, while the method described in Section 12.3 calculates the two-dimensional Fourier transform of the matrix obtained from the spatial scanning of a waveguide.

The first method, despite its disadvantages, offers the important advantages of achieving a good balance between data acquisition and processing time, of requiring simple equipment adopted and of generating results of good quality (with room for improvement in this regard, for example by using a non-rectangular window in the spectrogram calculation). The only disadvantage of the first method, compared to the second one, is that it represents the dispersion curves of the waveguide directly in the form of group delay.

If the second method is adopted, generating results comparable to those of the first one requires performing a spatial scan of the waveguide with a large number of points and for a significant length; this would also take a long time to acquire and process data, and, consequently, more powerful computing resources.

Chapter 13

Validation of the Qualitative Results on the UGW Propagation in a Waveguide

As anticipated in Chapter 11, the third sub-goal of the experimental activity is related to the development of a method to validate and verify the qualitative results obtained in the previous phase and described in Chapter 12.

To validate the previously collected data, the first thing to do is to acquire a more practical knowledge on the propagation of UGW in metal waveguides. In the qualitative analysis of the results shown in Chapter 12, we have already had the opportunity to perceive the complexity of UGW propagation, even for simple waveguides such as a thin bar (not to mention the rail). If a better understanding of the phenomena that occur in the generation, transmission and reception of UGW is desired, it is necessary to divide the more complex problems into simpler ones. In this way, the results of each experiment will be influenced by a limited number of variables, that can be controlled more easily. Consequently, it will be possible to understand how each of these parameters acts in modeling the propagation mechanisms. In a subsequent phase it will be possible to put together the results and the knowledge acquired to understand how the various propagation mechanisms interact, forming the overall picture that could be observed in the experiments of Chapter 12.

Basically, we need to take a step back and start from the basics. In other words, we need to start from the simplest cases of UGW propagation and to study the results of the experiments in a more analytical and less intuitive way. The first experiment that can be performed relates to the study of the propagation of an ultrasonic signal in a pristine waveguide, analyzing the signals collected in the time domain, the spectrogram that can be obtained from these, and studying how the dispersion acts on the received modes.

13.1 Available Instrumentation

For the execution of the measures described in this Chapter, it was possible to rely on:

- an Agilent 33250A waveform generator;
- an Agilent 8114A Programmable Pulse Generator;
- a LeCroy Wavesurfer 44Xs oscilloscope.

To generate and sense the ultrasonic waves guided in the waveguides used during the experiments, the following transducers were used:

- Murata 7BB-41-2L0 piezoelectric transducers (hereinafter also referred to as "big" transducer);
- Murata 7BB-20-6L0 piezoelectric transducers (hereinafter also referred to as “small” transducer).

13.2 Study of the Propagation of an UGW Signal

As already mentioned in the introduction to this chapter, when studying ultrasonic propagation in a waveguide, it is necessary to prune the experimental campaigns of all those elements that complicate the understanding of the phenomena to be analyzed. In other words, it is required to face and analyze the factors with influence the propagation characteristics one at a time. In the following, two of the main aspects that influence the propagation of signals in waveguides will be addressed:

Table 13.1: Principal characteristics of waveguides and materials used in the experiments.

Parameters		6060 Aluminium	Steel
Geometry	*Length [mm]*	2000	1508
	Width [mm]	40.15	50.35
	Thickness [mm]	2.15	4.40
Elastic properties	*Density [kg/m^3]*	2710	7800
	Young modulus [GPa]	68.9	205.332
	Poisson's ratio	0.33	0.29345

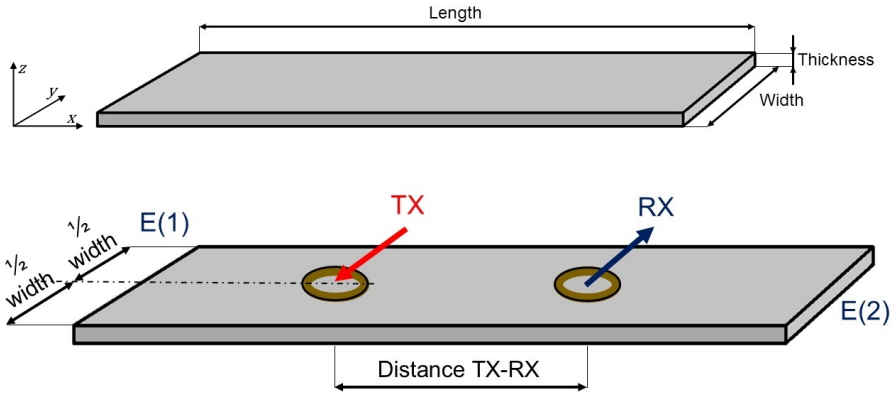


Figure 13.1: Schematic representation of the bar used for the experiments. TX is the transmitter placed in central position, RX is the receiver; transmitter and receivers are placed in laced in a paracentral position; E(1) and E(2) denote the extremities.

1. The first concerns the study of signal propagation in a waveguide. In this case, the goal is to analyze the propagation of UGW in a pristine waveguide. The considered signals are collected so that they are influenced as little as possible by the effects of reflections from the main discontinuities. In this context, the influences that different excitation signals types produce on the received signals, having different bandwidth and center frequencies, are also evaluated.

2. The first phase is completed with the study of the impact that a defect emulator (mass placed on the waveguide surface) has on guided waves propagation. At the same time, it is also analyzed how the weight distribution can affect the received signals.

During these experiments, the analysis tools developed and the experience gained during the execution of the analyzes described in Chapter 12 will be exploited.

13.3 Study of the UGW Signals Propagation in a Simple Waveguide

As far as the study of the propagation of signals in a waveguide is concerned, it was decided to analyze the simplest case. Two rectangular-section waveguides have been chosen: one made of aluminum and one made of steel. The experiments involved a pair of transducers in pitch-catch mode: one piezoelectric transducer is used as a transmitter and the other as a receiver. The distances between the transducers and the ends and the characteristics of the waveguides were chosen so that it was possible to carry out the measurements on signals free of influences from the reflections at the ends and with a number of propagation modes as small as possible. During the measurement campaign, the responses of the system to different types of excitation signals were tested.

13.3.1 Measurement Set-up

The two chosen waveguides have a rectangular cross-section, and a length that was the maximum possible to allow an easy transport. The materials of the two waveguides were 6060 aluminum and steel, unfortunately of an unknown alloy, respectively

The sizes and characteristics of the two waveguides are listed in Table 13.1; Fig. 13.1 shows the dimensional drawings of the wave guides.

Table 13.2: Relative distances between waveguide elements in Fig. 13.1.

Waveguide material	TX-RX distance [cm]	E(1) [cm]	TX [cm]	RX [cm]	E(2) [cm]
Aluminum	*50*	0	75	125	200
	70	0	65	135	200
Steel	*50*	0	50.4	100.4	150.8
	70	0	40.4	110.4	150.8

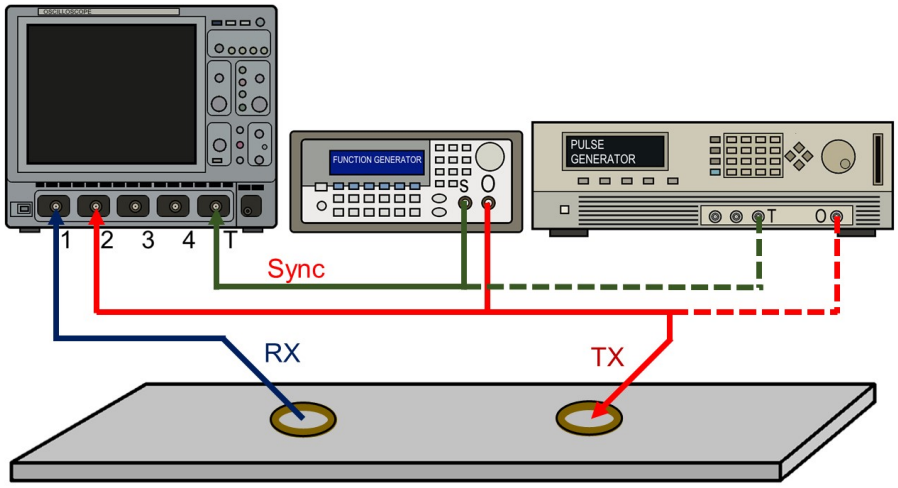


Figure 13.2: Schematic of the connections between instruments (function generator on the right, oscilloscope in center and impulse generator on the left) and transducers (TX is the transmitting transducer, RX is the receiving transducer). TX can be connected alternatively to both the function generator or the pulse generator.

- The thin aluminum waveguide was chosen to have a simple propagation environment as much as possible. The ultrasonic waves guided in aluminum are slower than in steel, 5,042.20 m/s compared to 5,130.70 m/s for the fastest mode. This allows to delay the arrival of the reflections on the extremities of the waveguide of the transmitted signal. The thin thickness, as can be inferred from the graphs of Fig. 5.5, has an influence in determining the frequency within which higher-order modes come into play. From those figures it is possible to see that, with the same frequency-thickness product and type of material, the curves have the same behavior. As a consequence, by reducing the thickness of the waveguide, the graph of the group velocity with respect to the frequency expands, thus delaying the entry into play of the various higher-order propagation modes as the frequency increases.

- The steel waveguide was chosen in consistency with the analysis carried out in the previous chapter and to evaluate a slightly more complex propagation environment.

As already mentioned above, two different types of transducers (*big*, Murata 7BB-41-2L0 and *small*, Murata 7BB-20-6L0) were used during the experiments. The transducers, used in uniform pairs as transmitters and receivers ("big" with "big" and "small" with "small"), were positioned on the waveguide in a paracentral position along the x axis (directed along the length of the waveguide), spaced 50 and 70 cm apart. Along the transverse y axis (directed along the width of the waveguide), the transducers were placed centered with respect to the edges of the waveguide. Depending on the waveguide, the relative distances between the various elements starting from one end are listed in Table 13.2, while a dimensional drawing of the position of the transducers is provided in Fig. 13.1.

To anchor the transducers to the waveguide, since it was not possible to find the phenyl salicylate, after some experiments with wax and bi-adhesive tape which introduced a low-pass filtering on the signal, it was decided to use cyanoacrylate glue (in the specific, methyl 2-cyanoacrylate). This adhesive, once crystallized, rigidly binds the transducer to the waveguide, maximizing the efficiency of the acoustic transmission. The only disadvantage related to the use of this glue is its permanent nature: to reposition the transducer it is necessary to bring the glue to complete crystallization by overheating, making it brittle. In this way, by forcing, it is possible to remove the transducer, which however is damaged both by mechanical action and heating.

The instruments and transducers were connected together according to the diagram of Fig. 13.2. Depending on the excitation signal to be generated, the Agilent 8114A pulse generator or the Agilent 33250A function

Figure 13.3: Picture of the bar and of the instrumentation used in the experiments.

Table 13.3: Excitation signal characteristics.

Parameter		**Rectangular pulse**	**Sinusoidal burst**	**Triangular wave burst**
Signal	*Amplitude*	20 V_{pp}		
Range of generated frequencies	*Minimum frequency*	1 µs (pulse width)	50 kHz	
	Maximum frequency	2 µs (pulse width)	500 kHz	
	Frequency step	-	25 kHz	
	Number of frequency steps	2	19	
Burst	*Number of cycles*	1	3	
	Burst repetition interval	500 ms		

generator were connected to the transmitter (and to the oscilloscope to monitor the signal generator output). A photo of the measurement set-up is shown in Fig. 13.3.

13.3.2 Excitation and Collected Signals

It was decided to evaluate the system response to the following three different types of excitation:

- rectangular pulse;
- 3-cycles sinusoidal signal burst;
- 3-cycles symmetric triangular wave burst.

The characteristics of the generated signals are listed in Table 13.3. The rectangular pulse is generated by the pulse generator, while the bursts, composed respectively of 3 cycles of sine wave and 3 cycles of symmetrical triangular wave, were generated by the function generator. It was also chosen to evaluate the response of the system to different oscillation frequency of the excitation signals; to do this, the frequencies of the signals spaced in a range from 50 kHz to 500 kHz in steps of 25 kHz. The amplitude of the excitation signals was set equal to 20 V_{pp} with zero offset. For the rectangular pulse signals, it was decided to follow a similar procedure, choosing to generate pulses whose duration is equal to 1 µs and 2 µs and an amplitude equal to 20 V_{pp}.

As for the collected data set is concerned, much has already been said above. In summary, for each of the two types of transducers ("big", Murata 7BB-41-2L0 and "small", Murata 7BB-20-6L0) measurements were

carried out taking into account two different distances between transmitter and receiver; for each of them, three types of different excitation signals at varying frequencies are considered. For each acquisition, the data set includes samples related to a copy of the transmitted signal, and to the received signal. For the acquisition of each data series, the LeCroy 44Xs oscilloscope was set to store a signal which was the result of an averaging window constituted by 64 readings for each channel. Each acquisition collects 500,000 samples, for a duration of 1 ms; this results in a sampling frequency of 500 MS/s.

13.3.3 Collected Data Analysis Algorithm

The algorithm used for the analysis of the collected data is implemented in matlab, and follows the flow chart shown in Fig. 13.4. First, the data and the parameters that useful in subsequent analyzes are loaded. The dataset to be analyzed is then selected, specifying to which type of waveguide (steel or aluminum), to what type of transducers ("big" or "small"), and to what type of excitation signal (rectangular pulse, sine burst or triangular wave burst) is referred to.

The data analysis takes place by examining in sequence the measures acquired as the excitation signal frequency was increased.

The first operation that is performed on the data is a low-pass filtering with a cut-off frequency of 750 kHz; this aims at removing the high frequency noise afflicting the collected signals. Immediately afterwards, a downsampling of the data with a factor of 1/10 is also performed. In fact, the number of samples per signal is reduced from 500,000 to 50,000: the sampling frequency goes from 500 MS/s to 50 MS/s without compromising the compliance with the Nyquist's law (the signals have been filtered).

Given:

- the maximum propagation speed of the modes in the selected waveguide,
- the length of the shortest path involving at least one reflection on the extremities,

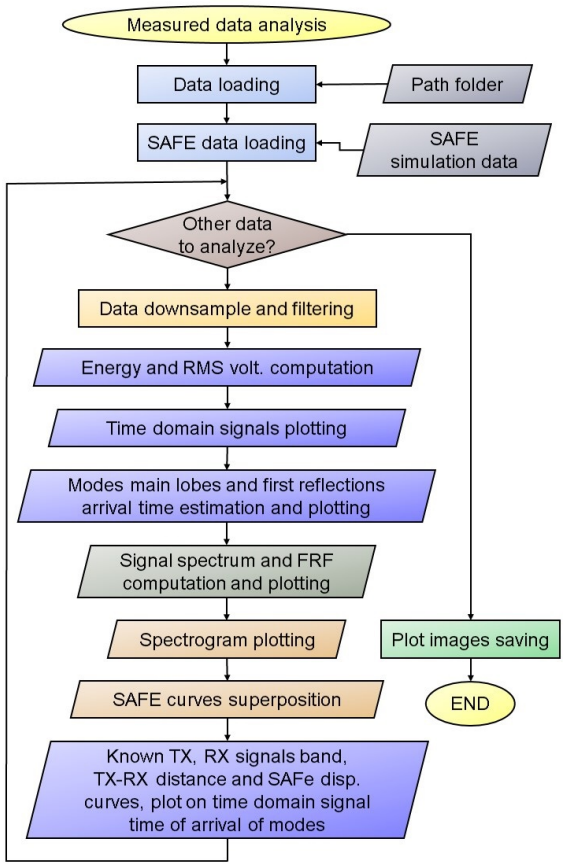


Figure 13.4: Flow diagram of the matlab algorithm used to analyze the data.

for each received signal, the time elapsing between the generation of the signal and the arrival of the first signal front related to a reflection from an extremity is calculated. In this way it is possible to know the time period in which the receiver measures the signal that has traveled the direct path only. Limiting to this portion of the signal, it is therefore possible to calculate the energy of the received signal, the energy of the excitation signal and the efficiency of the transmission. This for the given type of waveguide, for the type of transducers, for the distance between transducers, for the type of excitation signal at the given frequency. Parseval's theorem is used for the energy calculation

$$E = \sum_{n=1}^{N} \left| v\left(n\right)^2 \right|, \tag{13.1}$$

where N is the number of amplitude samples of the sampled signal v relative to the direct path only. At the same time, the maximum amplitude value of the received signal and its root mean square (RMS)

$$V_{RMS} = \sqrt{\frac{1}{N} \sum_{n=1}^{N} v\left(n\right)^2} \tag{13.2}$$

are also computed.

By comparing the values obtained for each combination it is possible to analyze their transmission efficiency.

It is then possible to proceed with the plotting of the collected signals in the time domain. To this graph will be added at a later time other information regarding the arrival time of signal useful to understand what happens in the course of UGW propagation.

Given

- the frequency of the excitation signal,
- the distance between the transducers,
- the trend of the curves of the group velocity for the waveguide,

it is possible to calculate the time taken by each mode to reach the receiver. These are the first information to be superimposed on the received signal time domain graph, together with the instant at which the first signal involving a reflection on the extremities begins to be measured.

The power spectrum of the data is then calculated and plotted, for both the transmitted and received signal. At the same tim,e the band occupied by the transmitted and received signals is also calculated through the matlab function `obw`. The latter returns the 99% occupied bandwidth of the input signal, specifying the frequency range. The frequency response of the system under the given conditions is the calculated and plotted, by performing the point-to-point ratio between the power spectra of the received signal and that of the transmitted signal.

The spectrogram of the received signal is then plotted. This operation is in no way different from the one already addressed in Chapter 12. In this way, we obtain a representation of the received signal behavior in the joint domain of time and frequency.

Once the distance between transmitter and receiver is known, it is possible once again to plot the curves of modes group delay, i.e., the time that the signal packets propagating at the group speed of the various modes take to reach the receiver once generated. These curves can then also be superimposed on the spectrogram, and it is once again possible to evaluate the actual correspondence between the curves predicted by the SAFE simulation and those obtained from the analysis of experimentally measured data.

The knowledge of the trend of the group delay curves and of the signal band also allows to evaluate the effects of dispersion. To highlight the effects of dispersion in the collected signals time domain graph, the knowledge of the signal band can be exploited. In fact, it is possible to identify the group velocity of the highest and lowest frequency components of both the excitation and received signals. Then it is easy to calculate the time taken by a signal traveling at the given speed to reach the receiver. This information can be superimposed not only on the signals in the time domain, but also on the spectrogram, and allows us to understand better how the contributions of the various modes occur and overlap in time and frequency.

13.3.4 Analysis of Elaboration Results

Given the considerable amount of collected data, in the following pages we will analyze the results obtained from the elaboration of a very limited subset of the data. However, these results are representative of what is happening on a general level and, for this reason, they were chosen.

First, the transmission efficiency can be analyzed by varying

- the type of waveguide,
- the excitation signal,
- the transducer and
- the distance between the transducer and the receiver.

We focus on the data collected in Table 13.4 and on the bar graphs of Figs. 13.5 (comparison between energies), 13.6 (comparison between maximum voltages) and 13.7 (comparison between RMS voltages).

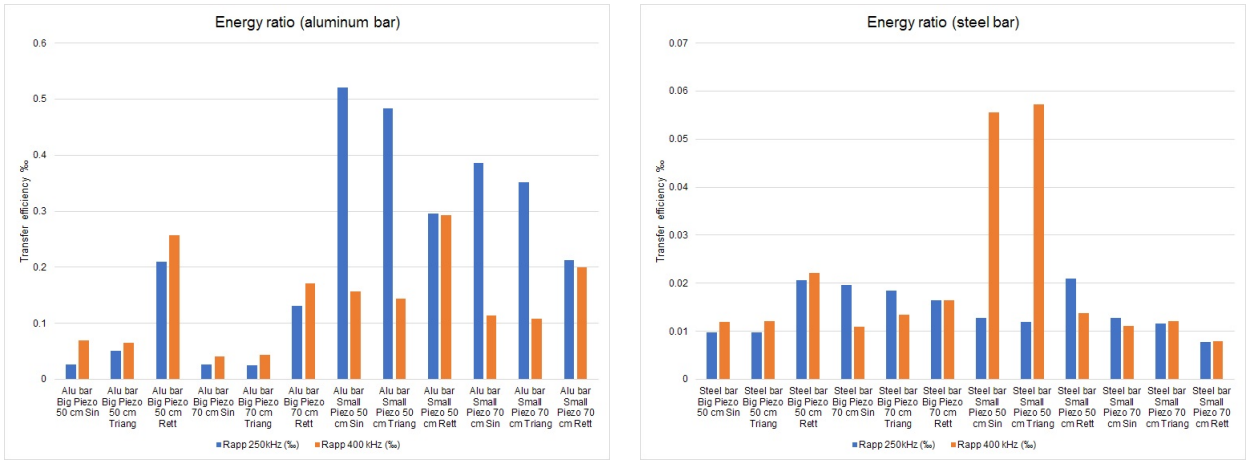
Various observations can be made from the analysis of the transmission efficiency graphs, that is the relationship between the energy of the measured signal (for the direct path) and the energy of the transmitted signal.

For the aluminum bar:

- the transmission efficiency is low in general;
- transmission efficiency drops as the distance between transmitter and receiver increases, as expected;
- on the average, the energy transmission is more efficient when the "small" transducer is used at a frequency around 250 kHz;

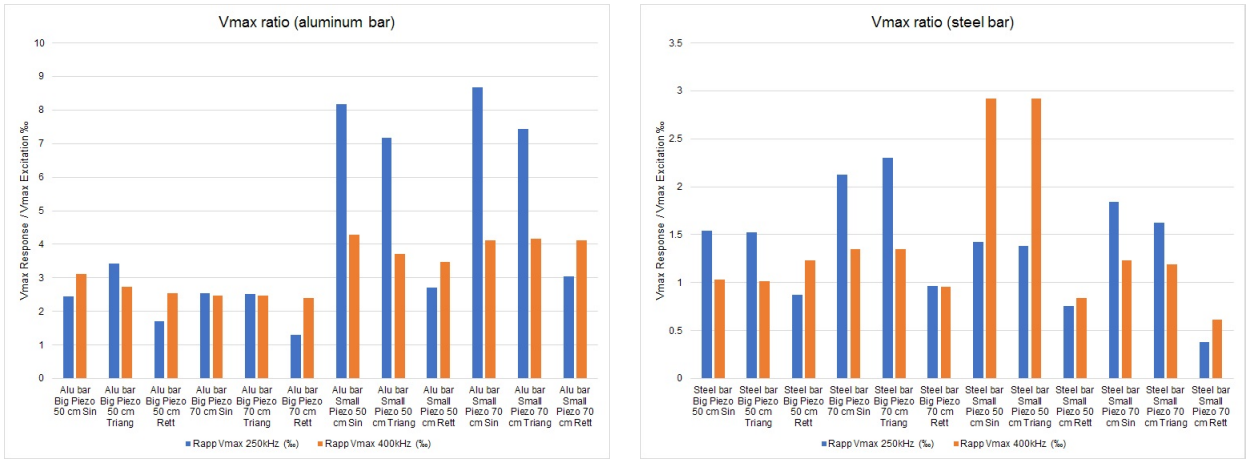
Table 13.4: Energy, Maximum voltage and RMS voltage measured for the different pulses injected in the different waveguides for the different transmitter-receiver distances.

Configuration				250 kHz (1 μs rectangular pulse)							400 kHz (2 μs rectangular pulse)						
Waveguide material	**Piezo size**	**TX-RX distance**	**Excitation signal**	**Energy Exc. [kJ]**	**Energy Resp. [J]**	**Energy ratio [ppm]**	**Vmax Exc. [V]**	**Vmax Resp. [mV]**	**Vrms Exc. [V]**	**Vrms Resp. [mV]**	**Energy Exc. [kJ]**	**Energy Resp. [J]**	**Energy ratio [ppm]**	**Vmax Exc. [V]**	**Vmax Resp. [mV]**	**Vrms Exc. [V]**	**Vrms Resp. [mV]**
Aluminum	*Big*	*50*	*Sin. Burst*	11.49	0.31	26.85	6.72	16.48	0.77	4.00	3.80	0.26	68.82	5.23	16.33	0.44	3.69
			Tri. Burst	7.48	0.38	51.18	5.52	18.86	0.62	4.46	2.45	0.16	64.46	4.22	11.54	0.36	2.87
			Rect. Pulse	8.19	1.71	209.53	13.93	23.73	0.65	9.44	24.25	6.24	257.26	18.07	46.04	1.12	18.01
		70	*Sin. Burst*	10.97	0.30	27.02	6.61	16.79	0.76	3.93	3.59	0.14	40.12	5.11	12.70	0.43	2.74
			Tri. Burst	7.14	0.18	25.08	5.42	13.59	0.61	3.05	2.34	0.10	43.93	4.15	10.24	0.35	2.31
			Rect. Pulse	7.76	1.01	130.49	18.91	24.45	0.83	7.26	23.16	3.96	171.00	19.97	48.04	1.30	14.34
	Small	*50*	*Sin. Burst*	23.02	11.98	520.31	8.89	72.73	1.09	24.95	11.26	1.77	156.96	7.99	34.17	0.76	9.58
			Tri. Burst	15.36	7.42	482.86	7.78	55.78	0.89	19.63	7.43	1.07	143.56	6.73	25.02	0.62	7.45
			Rect. Pulse	13.30	3.93	295.92	18.90	51.39	0.83	14.30	32.47	9.52	293.30	19.96	69.15	1.30	22.24
		70	*Sin. Burst*	23.01	8.86	385.32	8.88	77.01	1.09	21.46	11.24	1.27	113.28	8.00	32.89	0.76	8.13
			Tri. Burst	15.30	5.37	351.23	7.76	57.75	0.89	16.71	7.40	0.79	107.41	6.72	28.05	0.62	6.43
			Rect. Pulse	13.22	2.80	212.28	18.91	57.68	0.83	12.07	32.43	6.50	200.34	19.97	82.16	1.30	18.37
Steel	*Big*	*50*	*Sin. Burst*	11.46	0.11	9.71	6.73	10.38	0.78	2.44	3.79	0.05	12.01	5.25	5.39	0.45	1.56
			Tri. Burst	7.45	0.07	9.74	5.53	8.46	0.63	1.97	2.45	0.03	12.11	4.23	4.31	0.36	1.26
			Rect. Pulse	8.26	0.17	20.67	13.953	12.20	0.67	3.03	24.52	0.54	22.16	18.18	22.41	1.15	5.40
		70	*Sin. Burst*	11.20	0.22	19.64	6.70	14.28	0.78	3.44	3.76	0.04	10.94	5.22	7.04	0.45	1.49
			Tri. Burst	7.28	0.13	18.45	5.50	12.69	0.63	2.69	2.43	0.03	13.53	4.21	5.68	0.36	1.33
			Rect. Pulse	24.07	0.40	16.54	18.03	17.43	1.14	4.62	24.08	0.40	16.50	18.04	17.33	1.14	4.62
	Small	*50*	*Sin. Burst*	24.08	0.31	12.77	9.03	12.90	1.14	4.06	11.46	0.64	55.51	8.10	23.64	0.78	5.84
			Tri. Burst	23.87	0.28	11.92	9.01	12.43	1.13	3.91	11.36	0.65	57.17	8.05	23.54	0.78	5.90
			Rect. Pulse	13.58	0.28	20.91	19.03	14.38	0.85	3.90	33.02	0.45	13.70	20.03	16.80	1.33	4.93
		70	*Sin. Burst*	23.68	0.30	12.85	9.00	16.59	1.13	4.04	11.48	0.13	11.11	8.07	9.92	0.78	2.62
			Tri. Burst	15.58	0.18	11.59	7.85	12.73	0.91	3.11	7.54	0.09	12.04	6.82	8.13	0.64	2.21
			Rect. Pulse	13.27	0.10	7.85	19.00	7.17	0.84	2.36	32.60	0.26	7.96	20.01	12.37	1.32	3.73



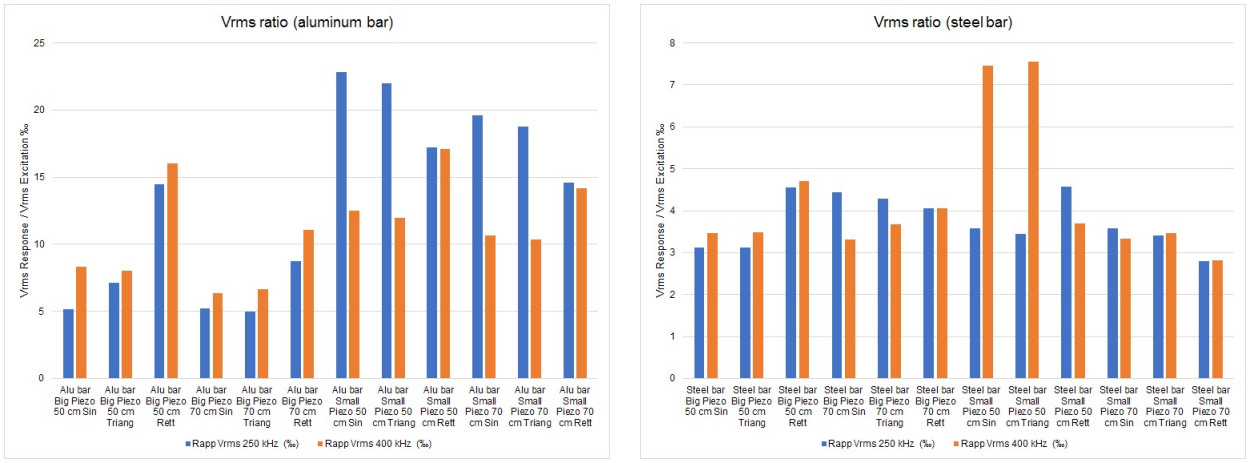
(a) (b)

Figure 13.5: Comparison between the transfer efficiency of the different pulses in the different situation (a) for the Aluminum bar; (b) for the steel bar.



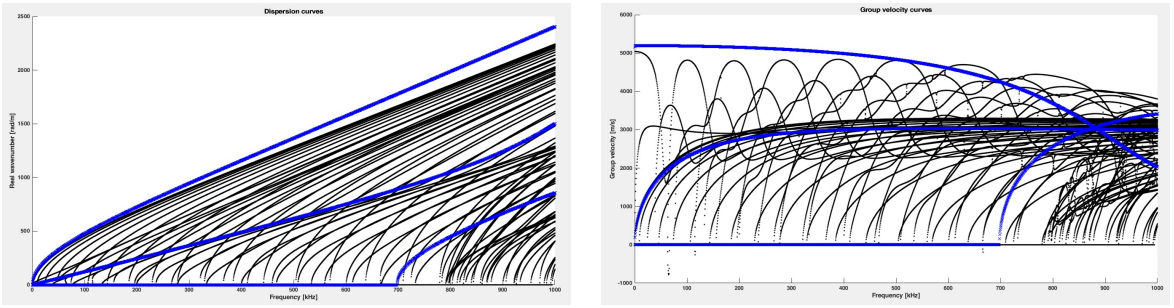
(a) (b)

Figure 13.6: Comparison between the measured maximum voltages Energy ratio of the different pulses in the different situation (a) for the Aluminum bar; (b) for the steel bar.



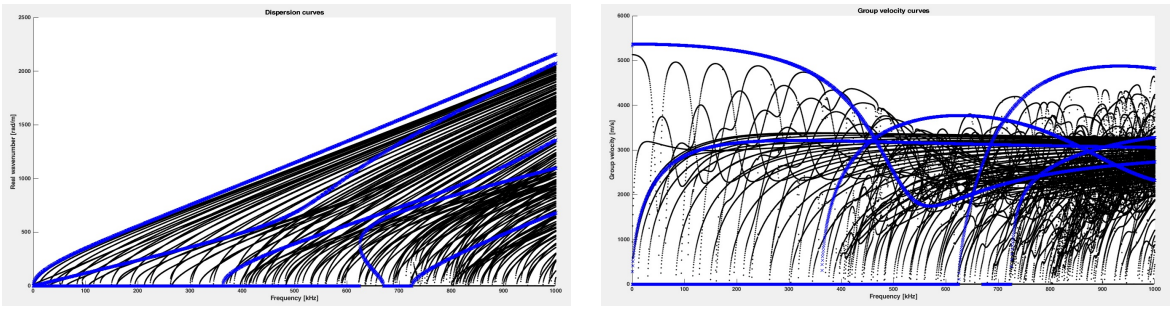
(a) (b)

Figure 13.7: Comparison between the measured RMS voltages of the different pulses in the different situation (a) for the Aluminum bar; (b) for the steel bar.



(a) (b)

Figure 13.8: Dispersion curves simulated for the aluminum waveguide. In black the dispersion curves for a bar and in blue those for a plate of the same thickness of the bar. (a) Wavenumber vs frequency; (b) Group velocity curves.



(a) (b)

Figure 13.9: Dispersion curves simulated for the steel waveguide. In black the dispersion curves for a bar and in blue those for a plate of the same thickness of the bar. (a) Wavenumber vs frequency; (b) Group velocity curves.

- as can be expected, the duration of the rectangular pulse has little influence on the efficiency of the transmission; efficiency is greater than other pulses due to its greater bandwidth;
- "small" transducers seem more efficient in energy transmission than "big" ones.

For the steel bar:

- the transmission efficiency is much lower than aluminum;
- transmission efficiency again drops as the distance between transmitter and receiver increases, as expected;
- there is not a big difference in transmission efficiency in using a "big" or "small" transducer, except in the case of a sinusoidal or triangular burst excitation at 400 kHz with two "small" transducers placed 50 cm away from each other.

The same observations hold when comparing the ratio between the maximum voltage value of the signal measured in reception and the maximum voltage value of the signal generated by the transducers, and also when comparing the RMS voltage values.

It is useful in the next analyses to keep in mind the behavior of the simulated dispersion curves for the two different waveguides, represented in Figs. 13.8 and 13.9. Once again, the overlap between the dispersion curves of the bar and those of an infinitely wide plate is illustrated. The former represent the modes that actually propagate; for convenience (given the high number of modes), the latter are used for the analyzes, as already done in Chapter 12. As shown in the diagrams, the curves of a plate once again summarize the main trends of the curves in a bar. Obviously, the representation is not faithful, but sufficient for our purposes. It is noted that in the steel bar the number of modes that propagate (141) is much greater than the number of modes capable of propagating in an aluminum bar (78) for the same frequency range (1 to 1 MHz). This is a significant aspect to take into account in the analyzes to come.

It is now possible to proceed with the analysis of time domain signals. As already mentioned in Chapter 12, analyzing these signals without knowing the trend of the dispersion curves can be quite pointless, but it is still possible to detect some interesting trends.

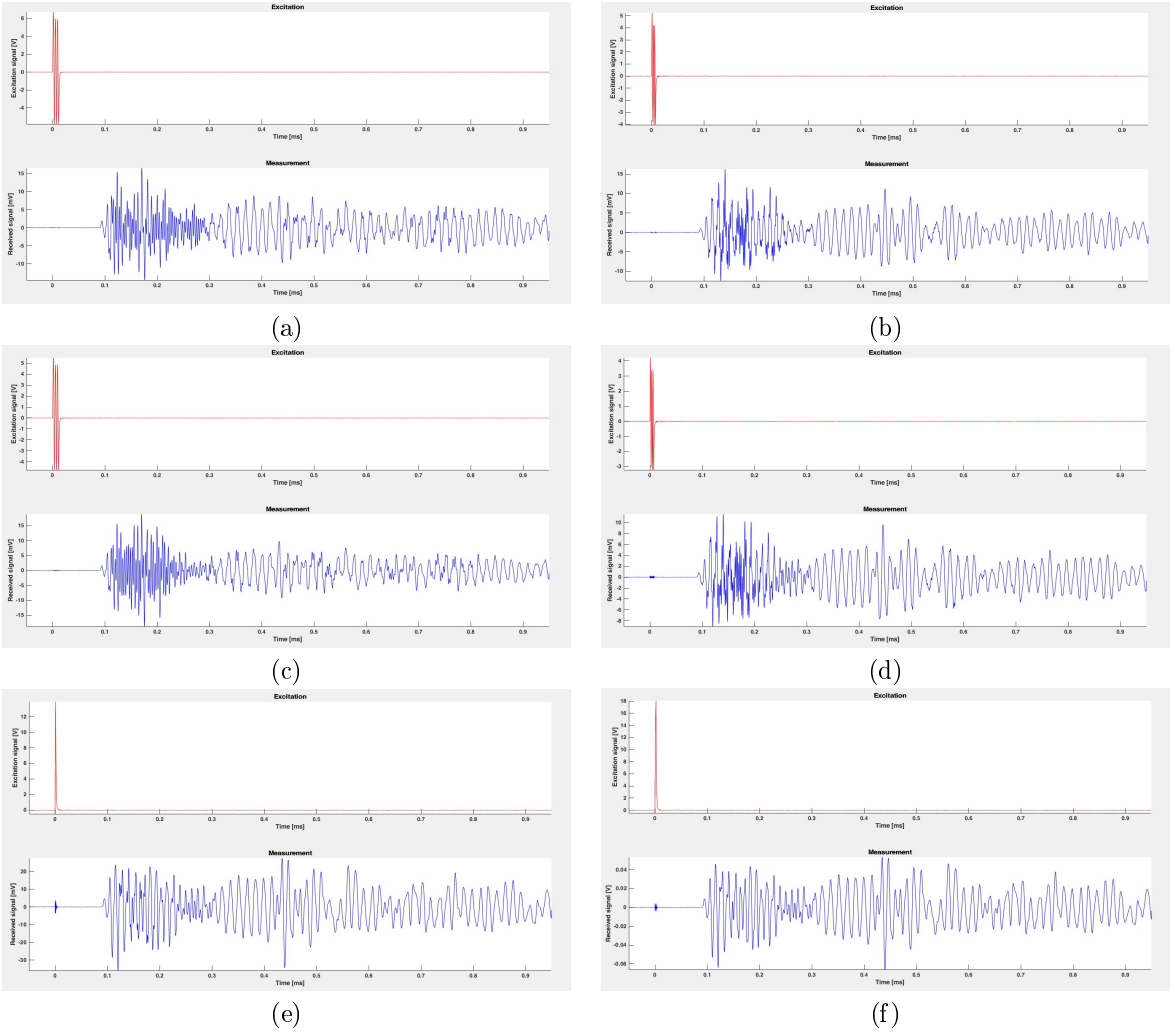


Figure 13.10: Time domain signals for the aluminum waveguide when the distance between transmitters and receivers ("big" size) is 50 cm. (a,b) sinusoidal burst; (c.d) triangular burst; (e,f) rectangular pulse; on the left column the excitation signal frequency was 250 kHz (the duration of the pulse was 1 μs); on the right column the excitation signal frequency was 400 kHz (the duration of the pulse was 2 μs).

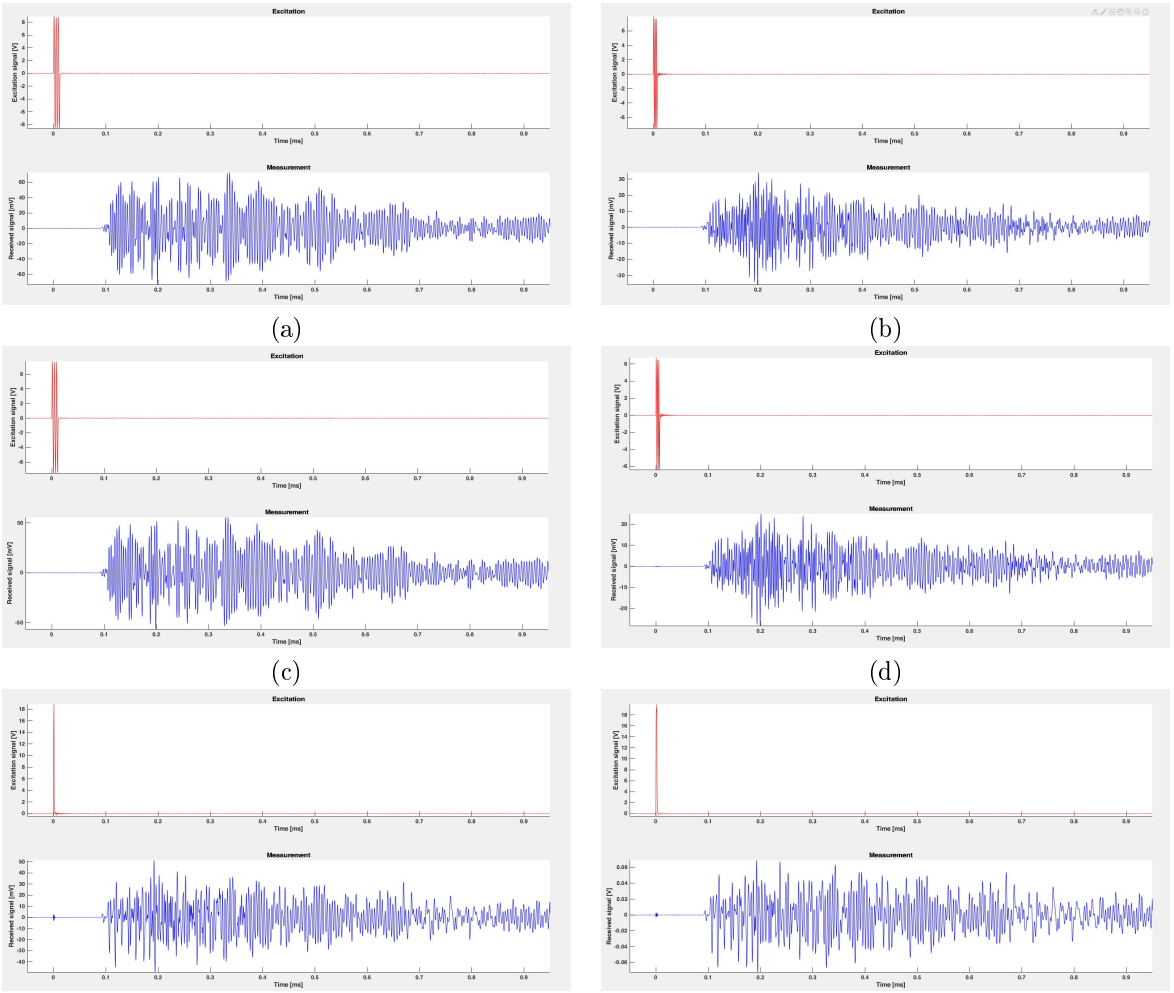


Figure 13.11: Time domain signals for the aluminum waveguide when the distance between transmitters and receivers ("small" size) is 50 cm. (a,b) sinusoidal burst; (c.d) triangular burst; (e,f) rectangular pulse; on the left column the excitation signal frequency was 250 kHz (the duration of the pulse was 1 μs); on the right column the excitation signal frequency was 400 kHz (the duration of the pulse was 2 μs).

For the aluminum bar and with the transducers placed at 50 cm from each other, we limit ourselves to observing in Figs. 13.10 and 13.11; as one might expect, the response to a triangular burst excitation is quite similar to that of the sinusoidal burst (and for this reason the responses to the triangular burst will be neglected in the following). Signals measured by "smaller" transducers (with higher resonant frequency) seem to have peaks and valleys sharper and closer in time. This result needs to be investigated further. A first hypothesis that can be made is that these transducers have a wider bandwidth and are able to measure higher frequency components than "bigger" transducers, which in fact have a lower resonant frequency. It also seems that with these transducers it is easier to discern the actual arrival of wave fronts that propagate. For the rest, it is not too surprising that signal trend is not very different as the type of pulse varies, given the excitation frequency: the waveguide is always the same, it has no defects and the position of the transducers is unchanged. Even when the excitation frequency varies, the differences are not substantial. The response changes only by changing the type of transducer.

This behavior is also visible when the distance between the transducers varies, as can be inferred from Fig. 13.12: in this case the transducers are placed at 70 cm from each other.

For the steel bar, in Figs. 13.13 and 13.14 are represented the measurements in the time domain taken with the transducers placed at 50 and 70 cm, respectively, from each other. The better definition of the excitation and measured signal achieved from the "small" transducers is evident once again. Many more modes propagate in the steel waveguide than in the aluminum one. The rectangular pulse, having a very wide band, is able to excite many of these modes, and this causes the interference events that can be guessed in the analysis of signals in the time domain.

It is now possible to evaluate the spectra of the signals and the frequency response of the system (i.e., the

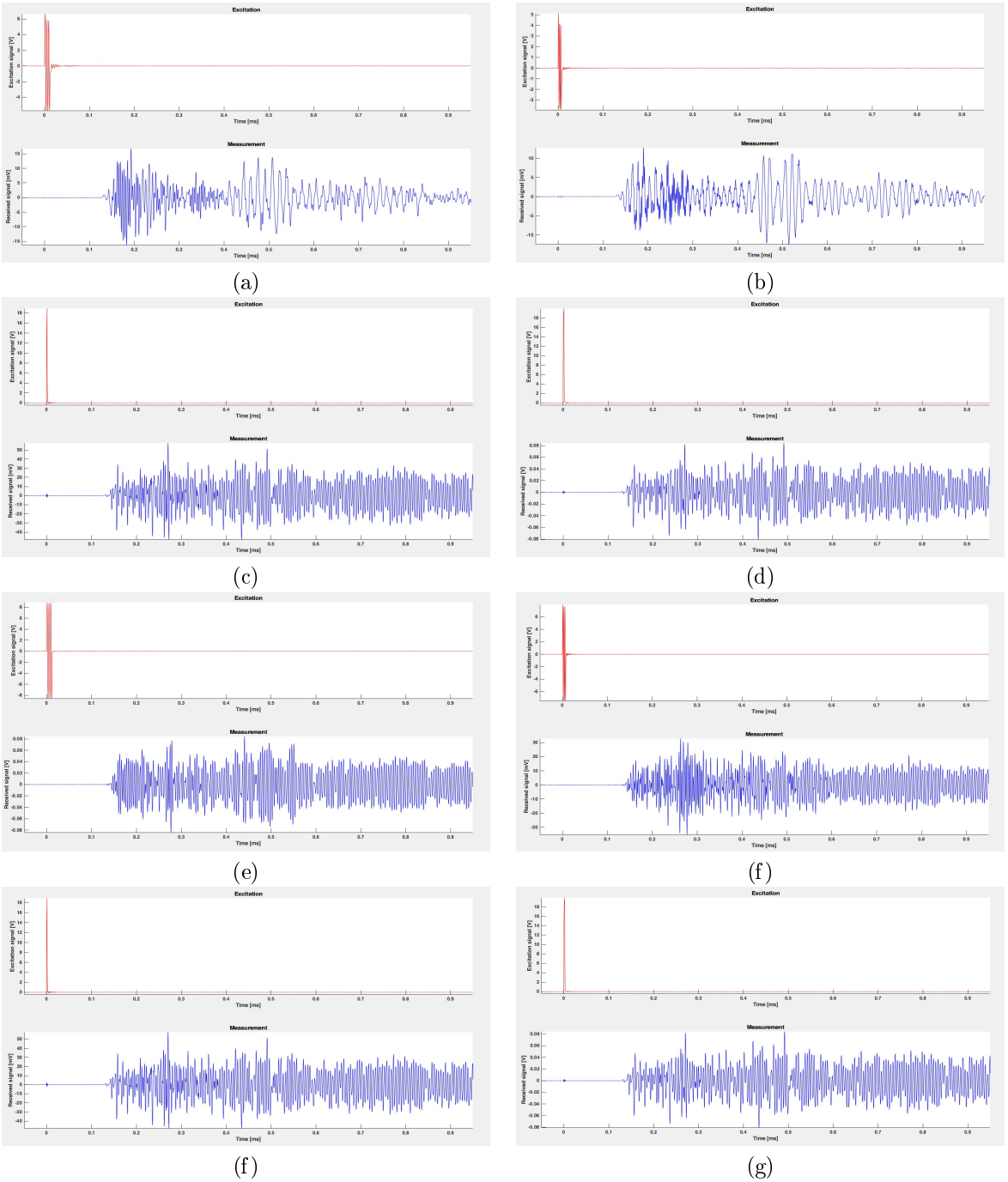


Figure 13.12: Time domain signals for the aluminum waveguide when the distance between transmitters and receivers is 70 cm. (a,b) sinusoidal burst injected by "big" transducers; (c.d) rectangular pulse injected by "big" transducers; (e,f) sinusoidal burst injected by "small" transducers; (g,h) rectangular pulse injected by "small" transducers; on the left column the excitation signal frequency was 250 kHz (the duration of the pulse was 1 μs); on the right column the excitation signal frequency was 400 kHz (the duration of the pulse was 2 μs).

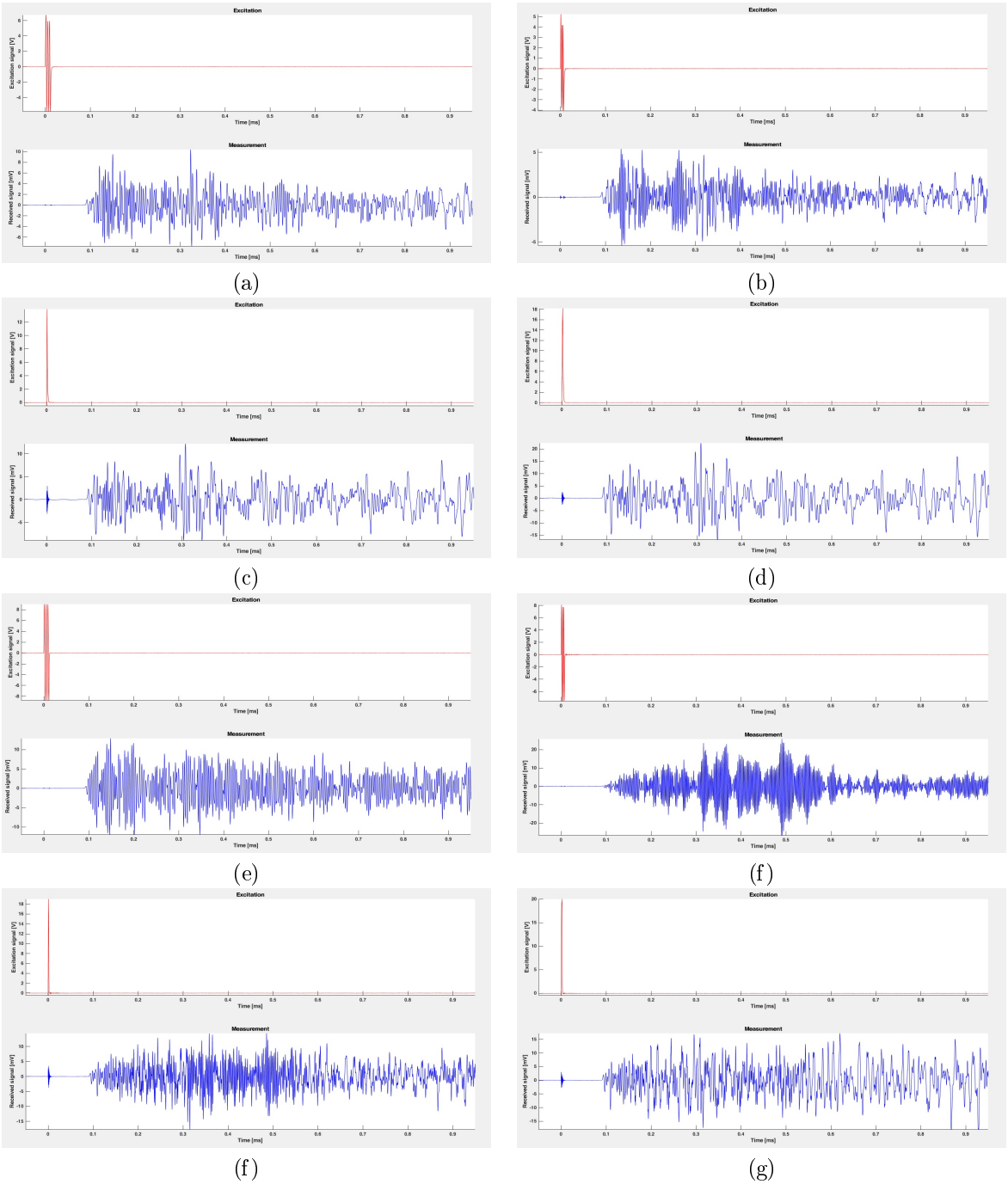


Figure 13.13: Time domain signals for the steel waveguide when the distance between transmitters and receivers is 50 cm. (a,b) sinusoidal burst injected by "big" transducers; (c.d) rectangular pulse injected by "big" transducers; (e,f) sinusoidal burst injected by "small" transducers; (g,h) rectangular pulse injected by "small" transducers; on the left column the excitation signal frequency was 250 kHz (the duration of the pulse was 1 μs); on the right column the excitation signal frequency was 400 kHz (the duration of the pulse was 2 μs).

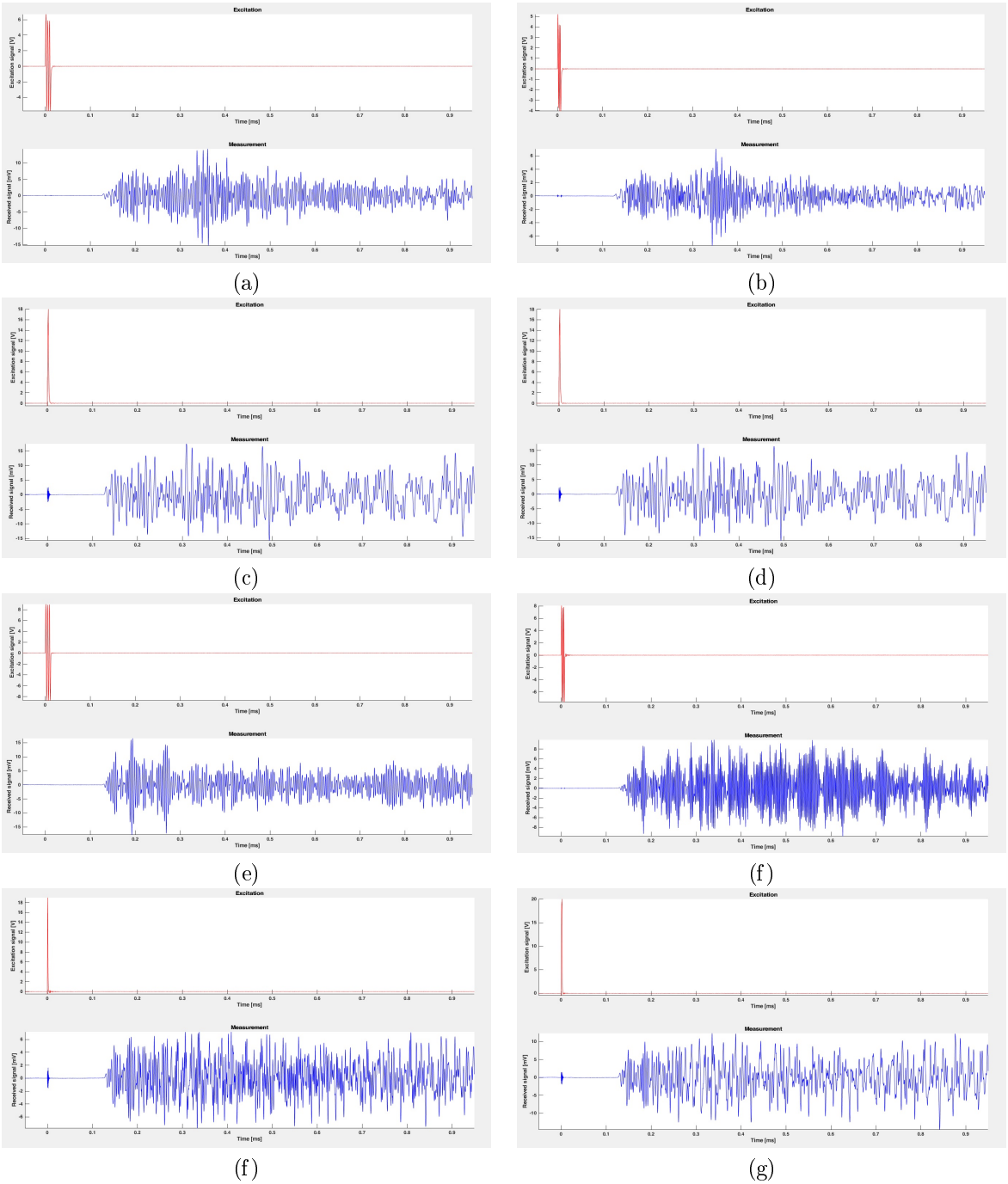


Figure 13.14: Time domain signals for the steel waveguide when the distance between transmitters and receivers is 70 cm. (a,b) sinusoidal burst injected by "big" transducers; (c,d) rectangular pulse injected by "big" transducers; (e,f) sinusoidal burst injected by "small" transducers; (g,h) rectangular pulse injected by "small" transducers; on the left column the excitation signal frequency was 250 kHz (the duration of the pulse was 1 µs); on the right column the excitation signal frequency was 400 kHz (the duration of the pulse was 2 µs).

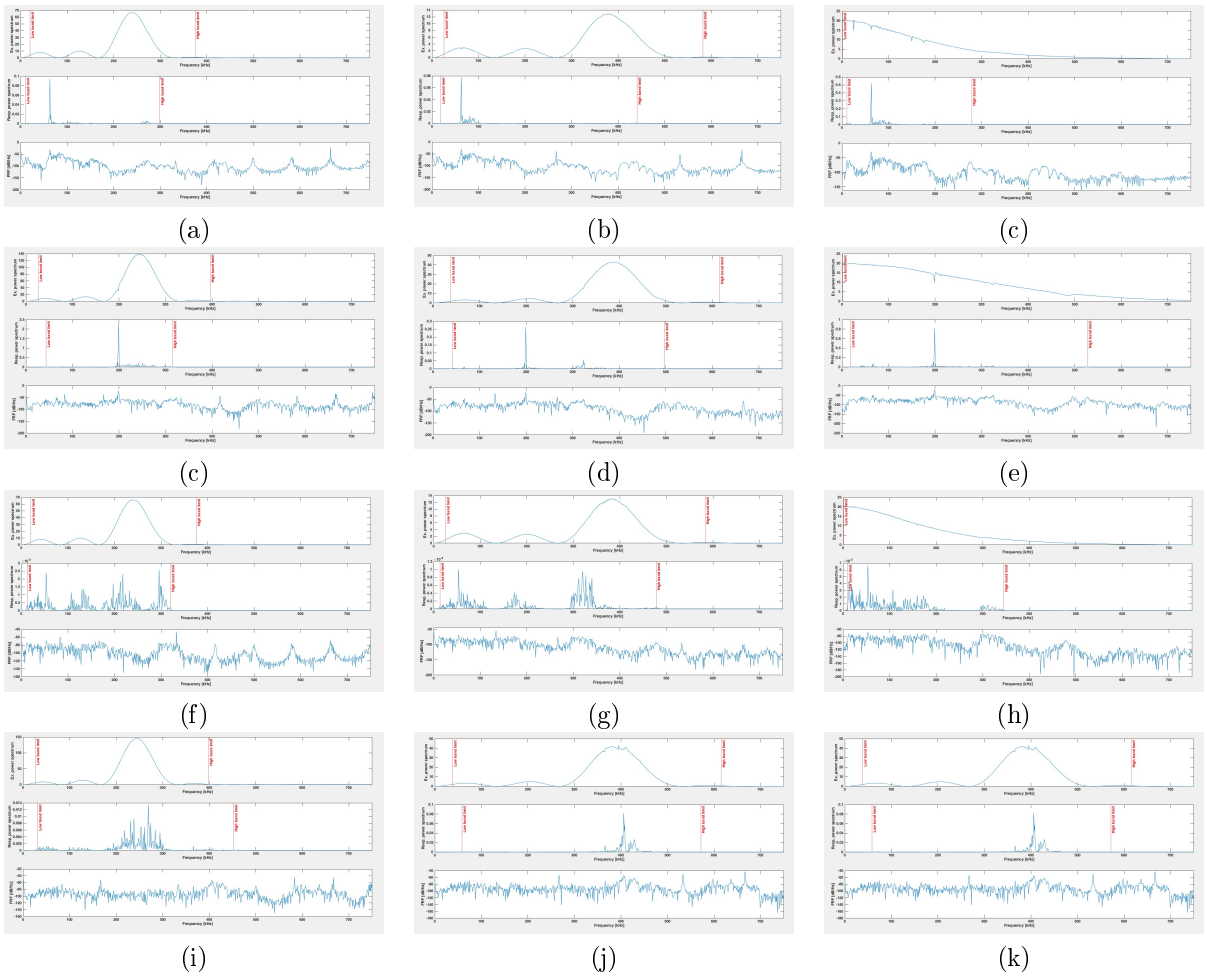


Figure 13.15: Power spectra of transmitted signal, received signal and frequency response function for different configurations. 1st row: aluminum, “big” piezo; 2nd row: aluminum, “small” piezo; 3rd row: steel, “big” piezo; 4th row: steel, “small” piezo. 1st column: sinusoidal burst excitation signal with frequency 250 kHz; 2nd column: sinusoidal burst excitation signal with frequency 400 kHz; 3rd column: rectangular pulse excitation signal (duration 1 μs).

point-to-point relationship between the power spectra of the received signal and the transmitted signal). To summarize, we evaluate only some of the spectra of the signals, represented in Fig. 13.15. From this figure, it can be seen how, in the aluminum waveguide and with the "big" transducers, a peak in the spectrum of the received signals is noted at around 63 kHz, whose intensity increases if the excitation signal is the rectangular burst. The band of the received signal increases as the band of the transmitted signal increases, as can be expected. The rectangular pulse, despite its very wide band, has a more intense spectrum mainly at low frequency. This causes the received signal to have a rather narrower band than the high frequency sinusoidal signals. It is assumed that the rectangular pulse is less capable of exciting component having high oscillation frequencies.

In the same waveguide, using the "small" transducers, the peak in the spectrum of the received signal moves around 200 kHz. This confirms the intuition we had in the analysis of signals in the time domain that the “small” transducers are more capable to excite high frequency components. This is especially evident by evaluating the spectrum of the rectangular impulse: the band of the excitation signal is now much wider, and this is also reflected in the response band.

The observations made so far are also valid for the steel waveguide, with one difference only: the spectrum of the received signal is much richer. In fact, there are many more peaks at different frequencies, much more extended in frequency. The position of the peaks changes as the type of excitation changes.

The next step is the analysis of the spectrograms. In those of Figs. 13.16 (for the aluminum waveguide) and 13.17 (for the steel waveguide), the group delay curves computed for a plate with the same thickness and the same material as the bar used to carry out the measurements have been superimposed. Each mode is represented by a triplet of curves to account for the duration of the excitation signal burst. In fact, the leftmost curve represents the arrival of the signal generated at the beginning of the burst; the rightmost curve represents the arrival of the signal generated at the end of the burst. In addition to these, the information relating to the

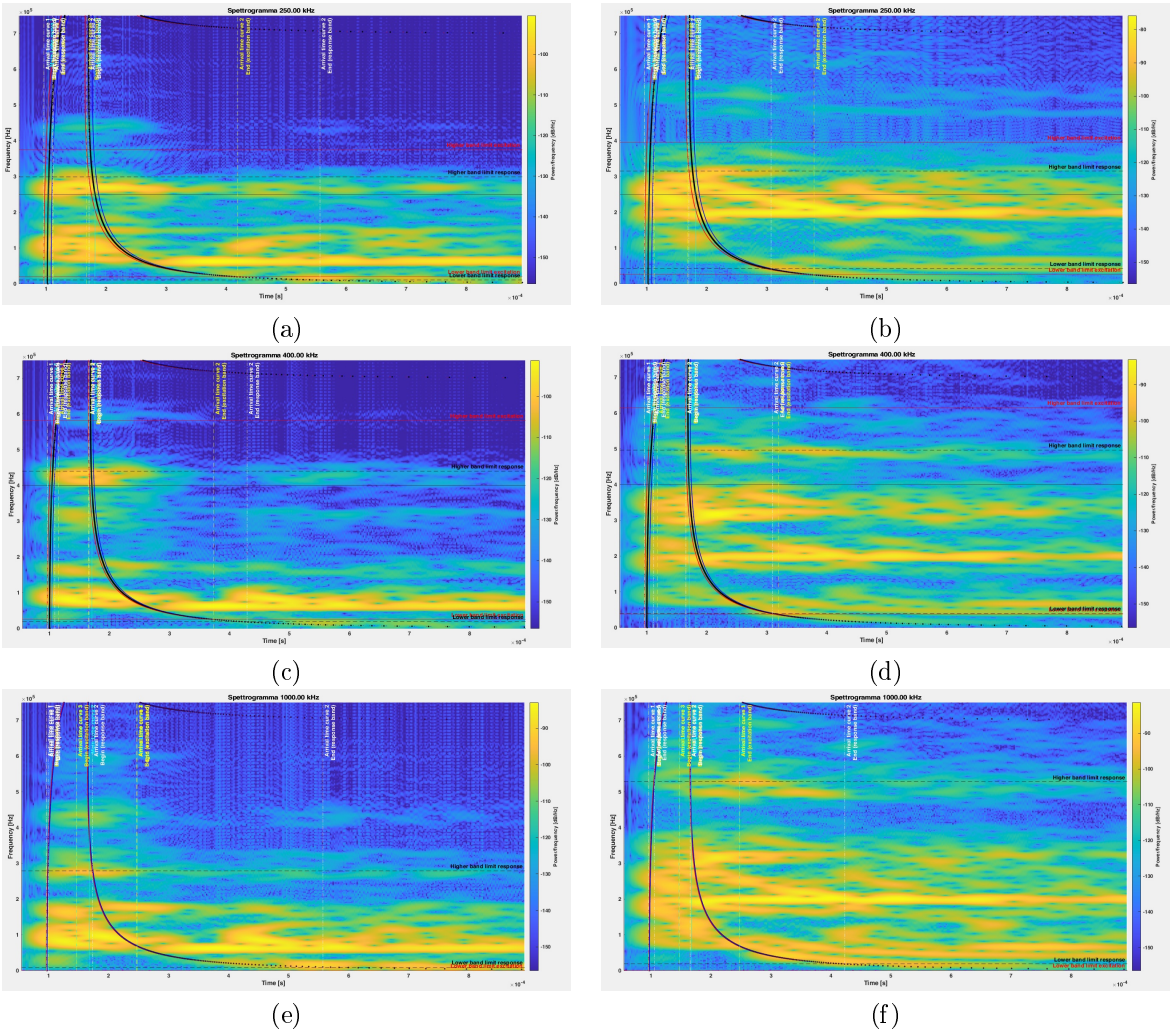


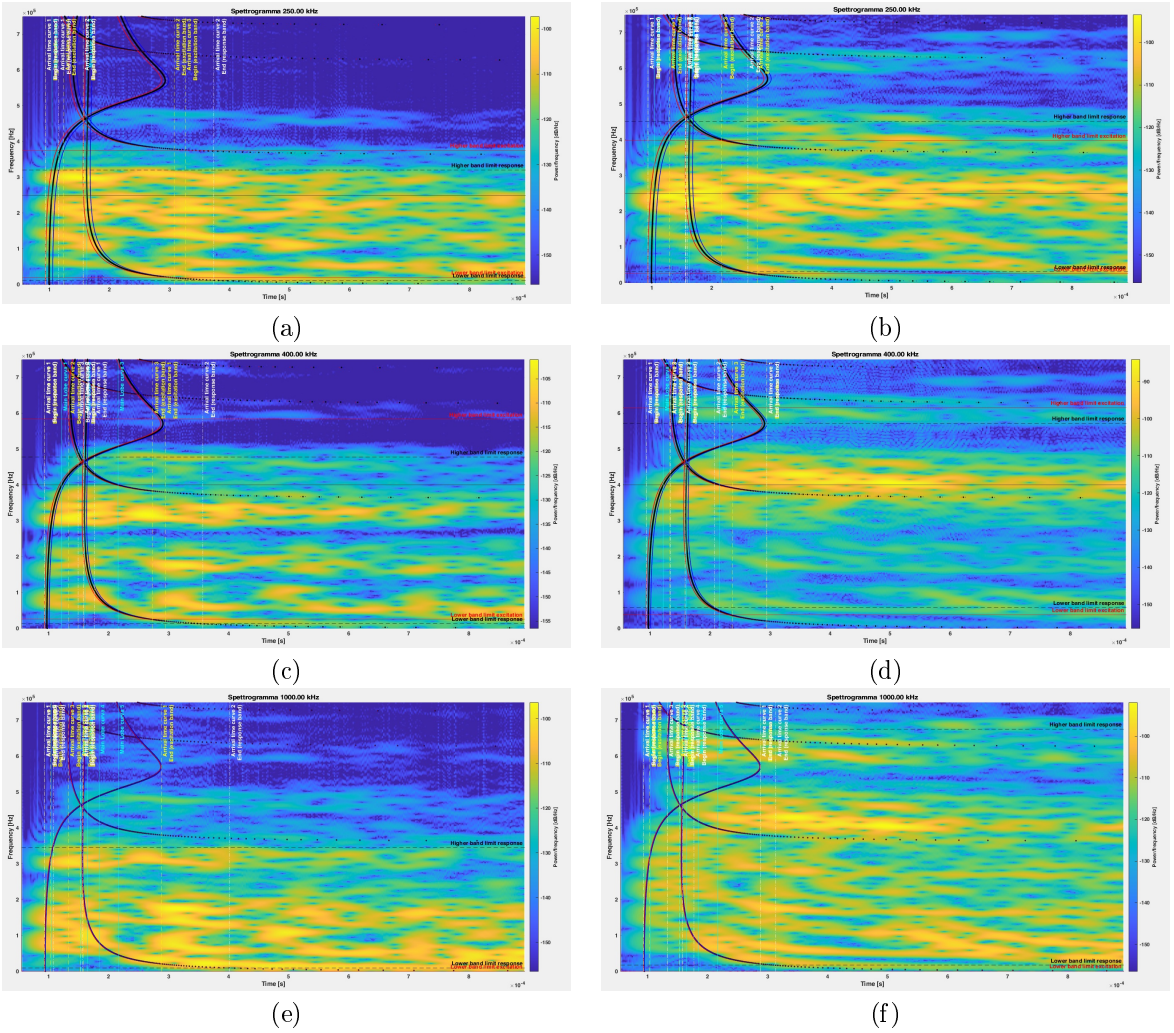
Figure 13.16: Spectrograms computed for the measured signals on the aluminum waveguide with transducers spaced 50 cm apart. On the left column the data transmitted and received from the "big" piezo, on the right the data transmitted and received from the "small" piezo. (a,b) sinusoidal excitation signal at 250 kHz; (c,d) sinusoidal excitation signal at 400 kHz; (a,b) rectangular pulse excitation signal (duration 1 μs).

band of the excitation signals and the measured signal is also superimposed to the spectrogram. We will return to the information obtainable from these curves in a while.

Quite the same observations made previously for the spectra can be made also for the spectrograms, as it naturally should be. What can be observed, once again, is how the spectrograms represent fairly faithfully the trend of the simulated dispersion curves. It should be noted that the graph of the dispersion curves of a bar is much richer than the graph of the dispersion curves of a plate (see Figs. 13.8 and 13.9). Consequently, many modes which are capable of propagating in the bar and whose presence appears in the spectrograms, are not represented by the superimposed curves, which are relative to the plate. In the current state, the same observations made in the case of qualitative analysis are valid: the resolution of the spectrogram is not yet sufficient to discriminate the trend of each mode that actually propagates in the wave guide. Further and more refined analysis are required.

The signals collected during these experiments have such a duration such that they refer, at least in part, only to propagation along the direct path. That is to say that, in the time interval from 0 to about 400 μs for the aluminum waveguide and 300 μs for the steel waveguide, the collected signal has not suffered from the interference of signal that have reflected on the extremities of the waveguide. Moreover, the waveguides are in pristine conditions, so there are no discontinuities over which the signals can be reflected. This has a significant advantage, as the graphs are much easier to read and faithful in representing the propagation of a single signal along the waveguide than those encountered in Chapter 12. One thing that can be noticed by looking at the graphs is a waving trend of the left edge of the areas of maximum signal intensity (close to $t = 0$). We will return to this observation later.

It was mentioned previously that the group delay curves of the modes and the horizontal lines representing



(a) (b) (c) (d) (e) (f)

Figure 13.17: Spectrograms computed for the measured signals on the steel waveguide with transducers spaced 50 cm apart. On the left column the data transmitted and received from the "big" piezo, on the right the data transmitted and received from the "small" piezo. (a,b) sinusoidal excitation signal at 250 kHz; (c,d) sinusoidal excitation signal at 400 kHz; (a,b) rectangular pulse excitation signal (duration 1 μs).

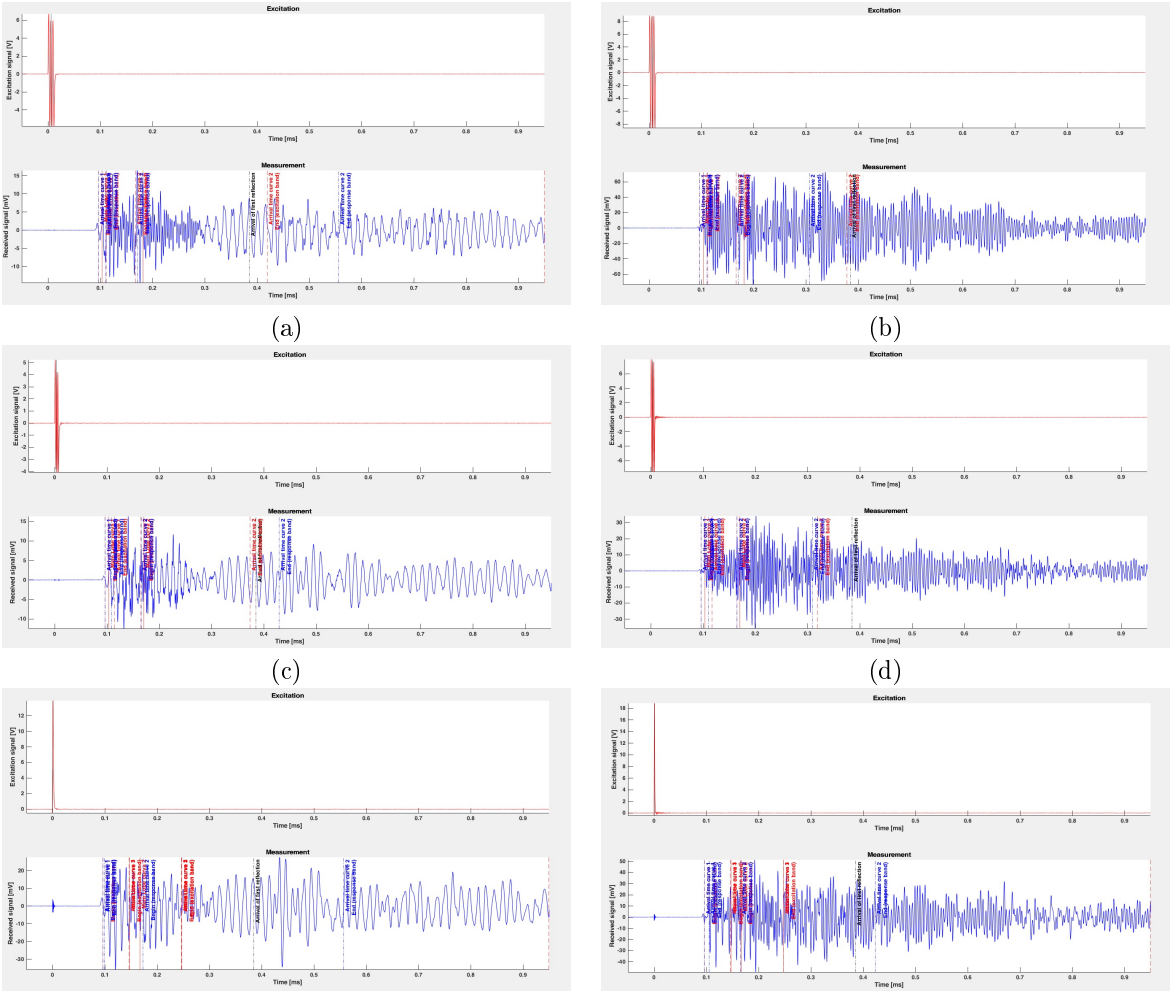


Figure 13.18: Instants of arrival of modes superimposed to the time-domain representation of the measured signals propagating along the aluminum waveguide with transducers spaced 50 cm apart. On the left column the data transmitted and received from the "big" piezo, on the right the data transmitted and received from the "small" piezo. (a,b) sinusoidal excitation signal at 250 kHz; (c,d) sinusoidal excitation signal at 400 kHz; (a,b) rectangular pulse excitation signal (duration 1 μs).

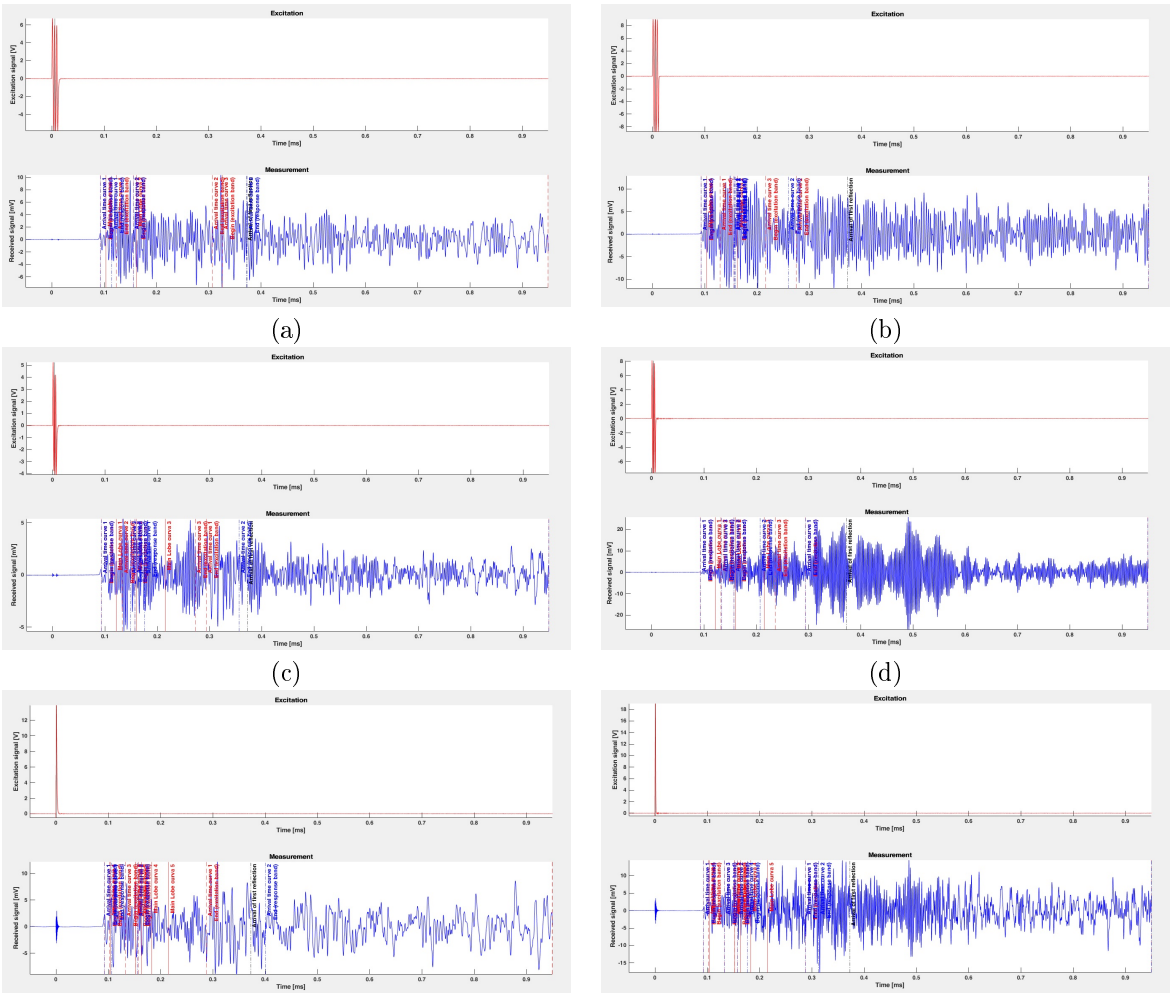


Figure 13.19: Instants of arrival of modes superimposed to the time-domain representation of the measured signals propagating along the steel waveguide with transducers spaced 50 cm apart. On the left column the data transmitted and received from the "big" piezo, on the right the data transmitted and received from the "small" piezo. (a,b) sinusoidal excitation signal at 250 kHz; (c,d) sinusoidal excitation signal at 400 kHz; (a,b) rectangular pulse excitation signal (duration 1 μs).

the transmitted and received signals band were superimposed on the spectrogram. By intersecting the dispersion curves with the band lines, it is possible to identify the interval of time in which each mode reaches the receiver. In other words, the intersecting points represent the begin and end instants of reception of each mode, called in short *instants of arrival* of a mode. These time instants will be different depending on whether we consider the band of the transmitted signal (so these are the theoretical arrival instants) and that of the signal actually received (actual arrival instants). By comparing the duration between the beginning and the end of the reception of a mode and the duration of the burst of the transmitted signal, one can immediately realize the distortion introduced by the variable group speed. Even better, it is possible to superimpose the arrival instants of the modes on the time domain signals. This makes it possible to evaluate even better the effects of dispersion on the signals received after propagation. In fact, it is possible to understand how the different modes interfere with each other and how their arrival instants overlap at the receiver. This is visible in Figs. 13.18 and 13.19.

The last observation that can be made is to analyze the superimposition the group delay curves simulated for a bar over the spectrograms. The result is visible in Fig. 13.20. It is possible to see how the overlap between the simulated curves and the curves obtained from the collected signals analysis is much more accurate. Even the rounded edges mentioned earlier now find an explanation in the behavior of the group delay curves. The same goes for the energy concentration bands that persisted over time: in reality they correspond to the instants of arrival of some modes for a frequency in which their group delay increases very quickly towards infinity. Basically, each concentration of energy, or spot, in the spectrogram corresponds to the trend of one or more modes expected for a bar.

13.4 Study of the Effects of a Defect Emulator on the Propagation of UGW in a Waveguide

In the previous section, the behavior of complex constituted by the transducers and the waveguide was evaluated in the basic propagation conditions. It has been shown in Chapter 12 how the presence of a defect (in that case a transverse cut in the waveguide) is reflected in the propagation of the signals. Since the defect was part of the same waveguide, and since it was not possible to recover an identical, undamaged waveguide, it was difficult to accurately determine the influx of the defect on signal propagation. It has been shown in ref. [102] that the adoption of a defect emulator (i.e., a mass) to be placed or glued to the waveguide is effective. In addition, the emulator allows to perform measurements with and without defects under the same conditions (position of the transducers, waveguide parameters, environmental conditions), allowing to carry out a direct comparison. In this section, the effectiveness of using a defect emulator will be evaluated first. That is to say that will be evaluated whether or not alterations in the measured signals actually occur in its presence; if so, what these variations consist of will also be assessed.

13.4.1 Measurement set-up

The measurement set-up is identical to that used to carry out the measurements analyzed in Section 13.3. In fact, the measurements in the presence and absence of the defect were taken at the same time. In this way it is possible to guarantee that

- the transducers are in the same position, and
- the environmental conditions are in fact very similar in the two cases.

In this way, from comparing the acquired signals it can inferred that the observable differences are largely due to the presence of the defect emulator.

As for the defect emulator is concerned, a common steel hammer was used, resting in the center of the waveguide, halfway between the transmitter and the receiver. Given the conformation of the hammer, two different positions were evaluated, visible in Fig. 13.21, which will be referred to in the rest of the Section:

- *tip position*, i.e., the hammer was resting on the tip over the upper surface of the waveguide;
- *side position*, i.e., the hammer was resting on one side of the head over the surface of the waveguide.

The two positions had the objective of evaluating the effect of the mass distribution in the point of support. In fact, the hammer resting on the tip unloads a large part of its mass (a part is discharged from the point of support of the handle) on a reduced and linear surface, placed in a transverse position with respect to the wave guide. The side of the hammer, on the other hand, unloads its mass over a much larger surface.

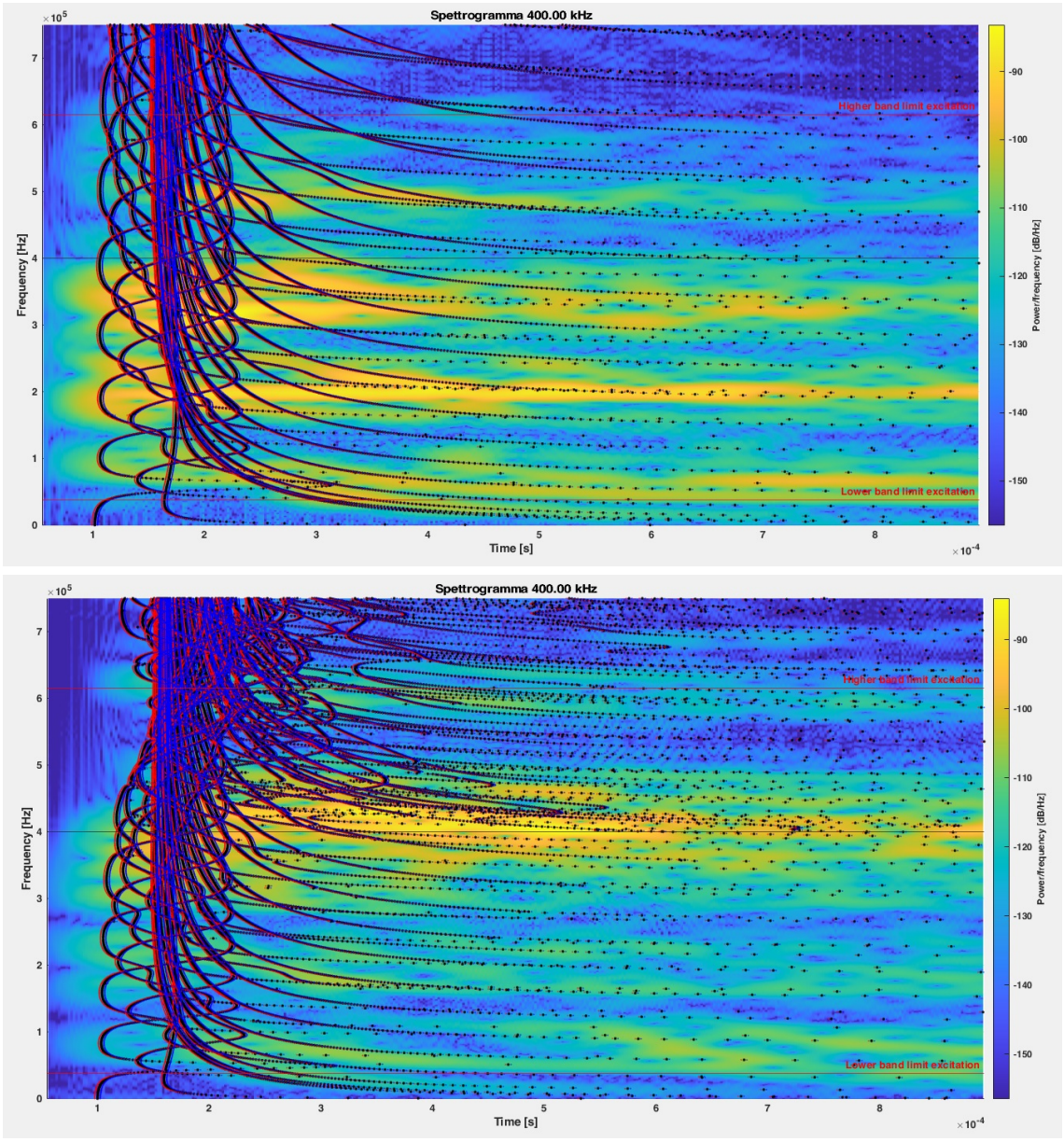


Figure 13.20: Superposition to the spectrogram of the group delay dispersion curves simulated for a bar; (a) Aluminum ("small" piezo spaced 50 cm apart, sinusoidal excitation at 400 kHz); (b) steel ("small" piezo spaced 50 cm apart, sinusoidal excitation at 400 kHz).

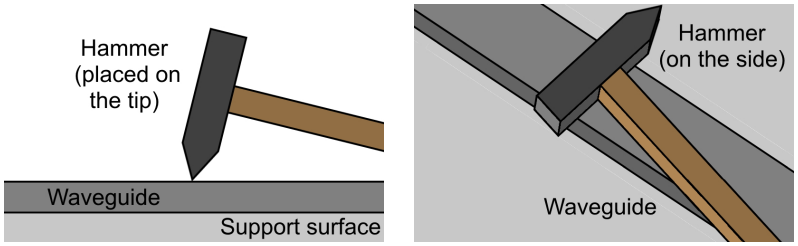


Figure 13.21: Different positions for the hammer used as defect emulator (a) hammer placed on the tip on the top of the waveguide; (b) hammer placed on the side on the top of the waveguide.

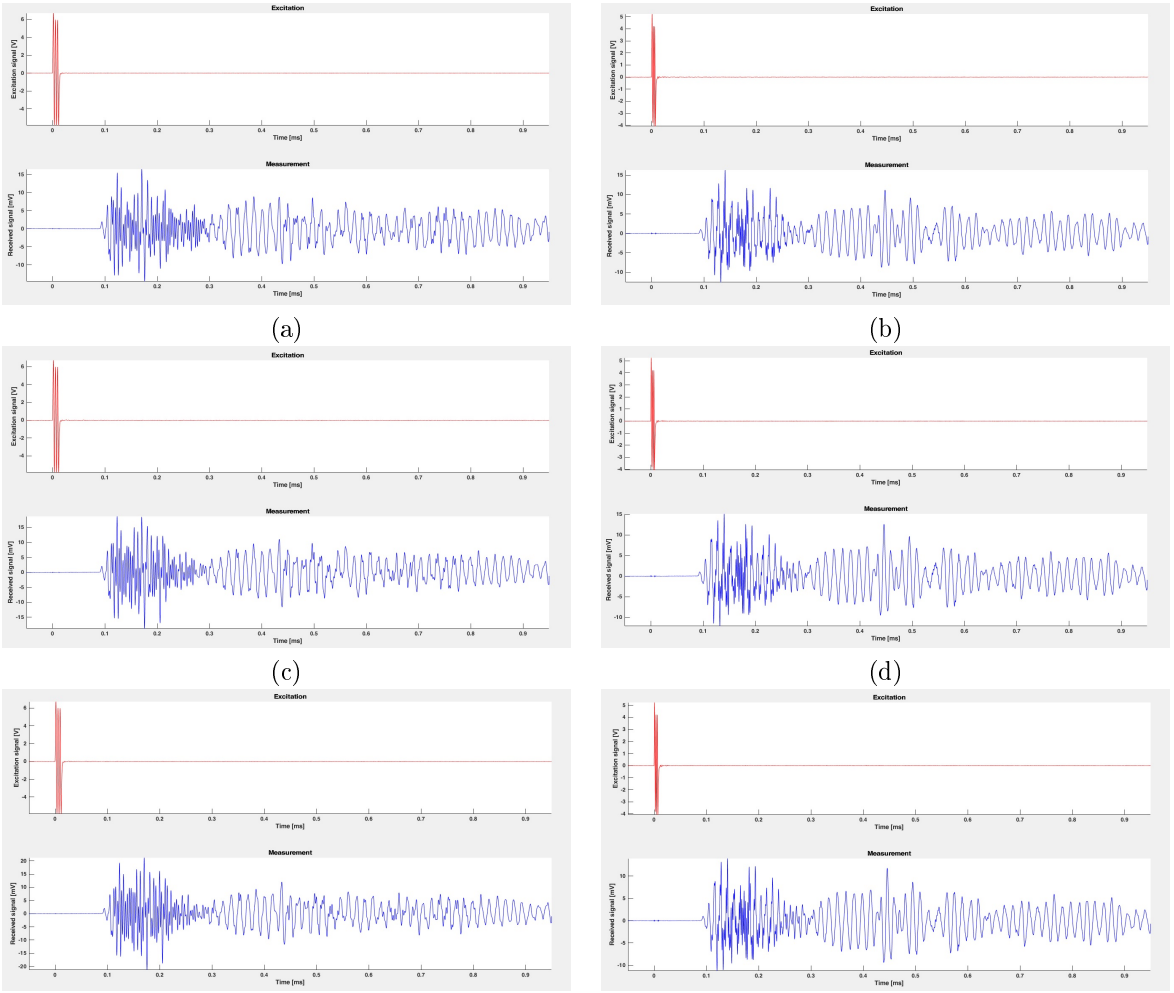


Figure 13.22: Time domain signals collected in presence or absence of defects for the aluminum waveguide by using the "big" piezos spaced 50 cm apart. (a,b) without hammer; (c,d) with hammer placed at the tip; (e,f) with hammer placed on the side. On the left column the measures taken with the excitation signal at 250 kHz; on the right column the measures taken with the excitation signal at 400 kHz.

13.4.2 Algorithm Employed in the Analysis of the Collected Data

The algorithm used for analyzing the collected data is identical to that described in Section 13.3.3. In the subsequent analysis of the results, only the study of the effects of dispersion on the collected signals is neglected.

13.4.3 Analysis of Processing Results

From the analysis of the waveforms collected in the aluminum bar with the "big" transducers placed at 50 cm from each other (see Fig. 13.22) it is possible to note how, effectively, there is a difference between the signals collected in the absence and in the presence of the defect, more concentrated in the first half of the signal collected. The trend of the signal is not drastically different, but there are some slight differences:

- more intense peaks in the area around time $t = 200$ μs;
- a more attenuated signal at about $t = 250$ μs;
- in the area from 450 μs onwards, the signal is slightly more intense.

These differences, as mentioned above, are small but encouraging: placing a mass on a waveguide actually produces measurable effects. There are no significant differences between the signal collected with the hammer resting on the tip and flat.

In Fig. 13.23 the waveforms acquired using "small" transducers are shown. In this case, the differences are much less visible to the naked eye, although they are present roughly in correspondence with the same points noted for Fig. 13.22.

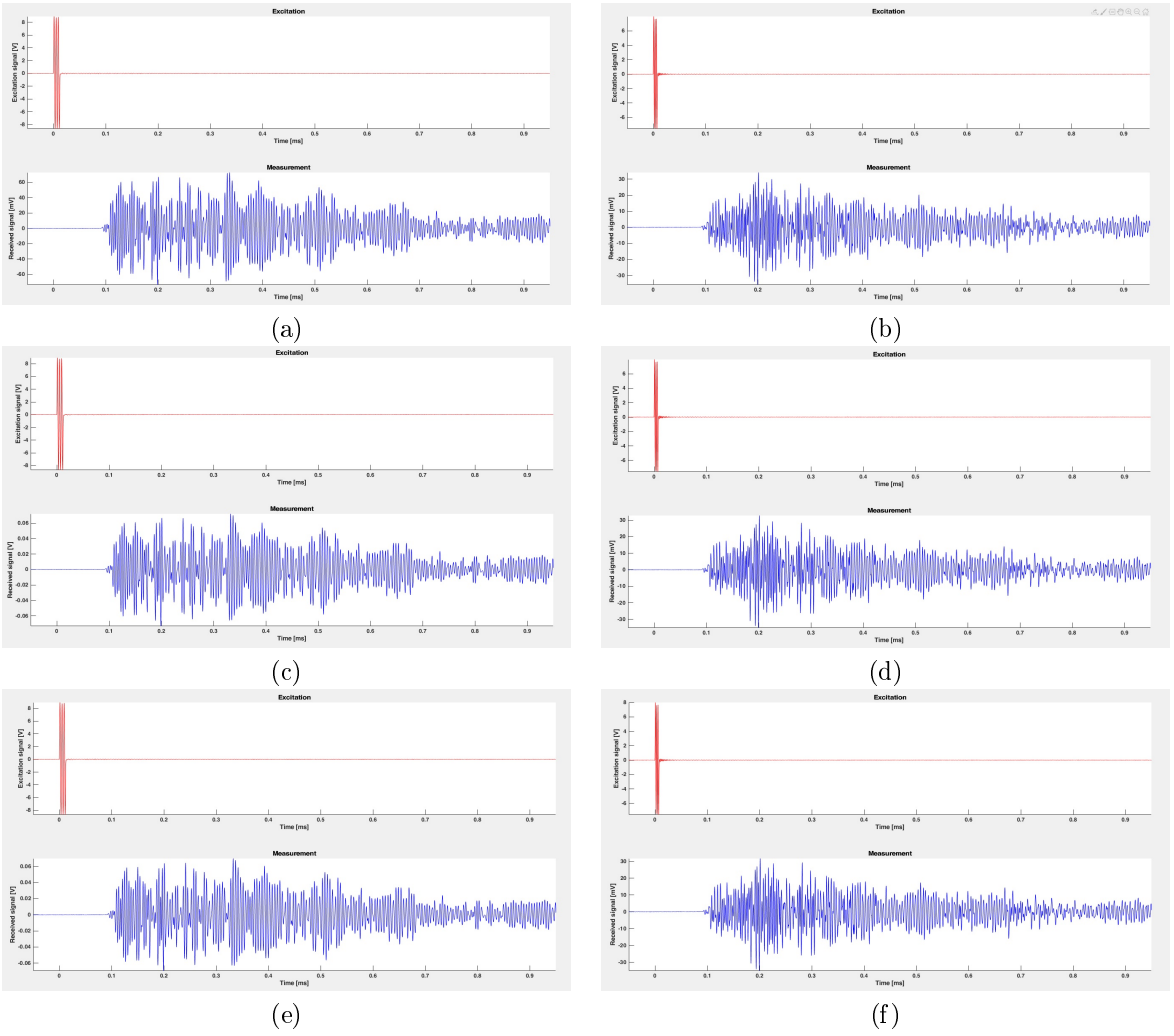


Figure 13.23: Time domain signals collected in presence or absence of defects for the aluminum waveguide by using the "small" piezos spaced 50 cm apart. (a,b) without hammer; (c,d) with hammer placed at the tip; (e,f) with hammer placed on the side. On the left column the measures taken with the excitation signal at 250 kHz; on the right column the measures taken with the excitation signal at 400 kHz.

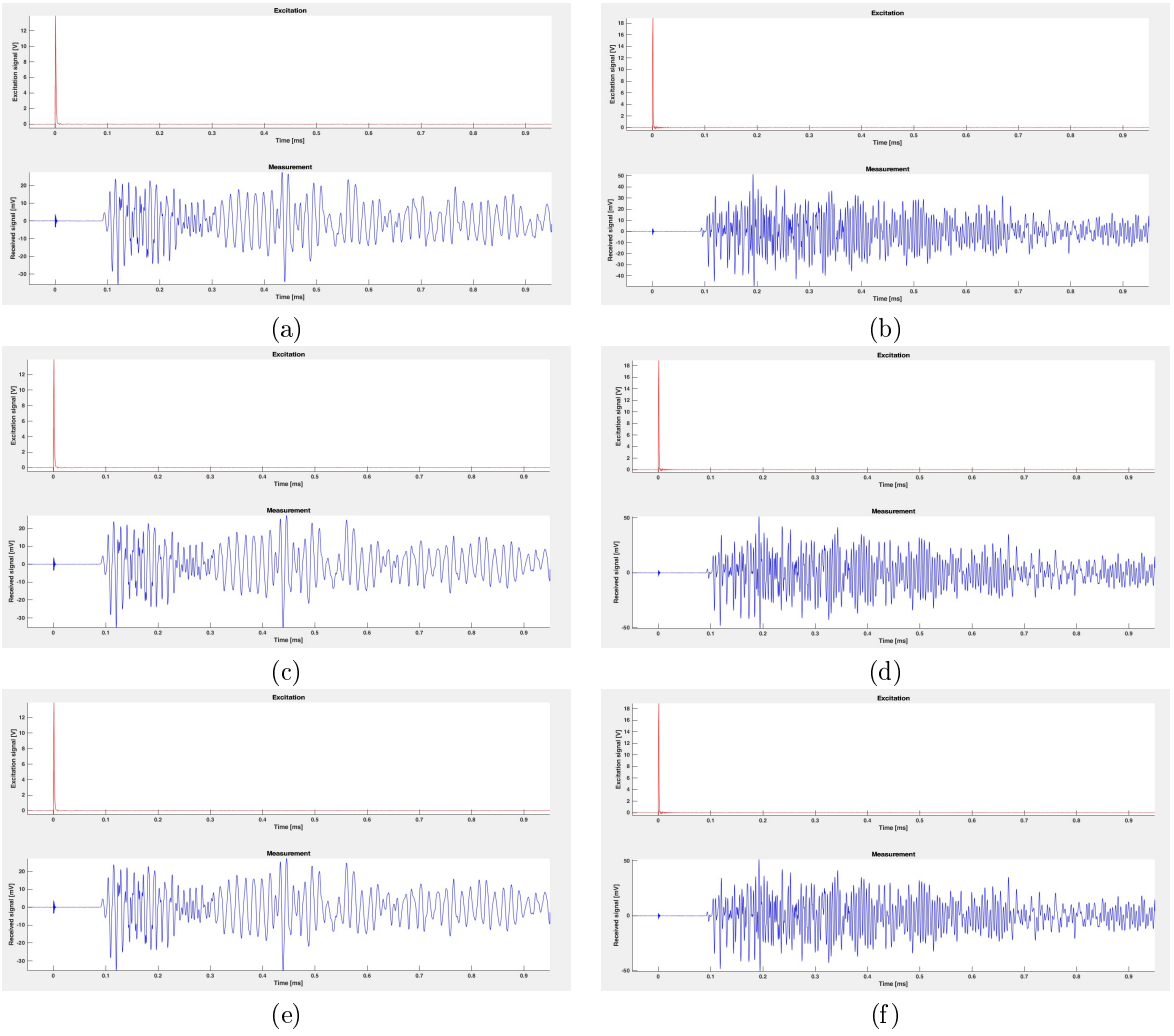


Figure 13.24: Time domain signals collected in presence or absence of defects for the aluminum waveguide by using the piezos spaced 50 cm apart with the excitation signal constituted by the rectangular pulse of duration 1 μs. (a,b) without hammer; (c,d) with hammer placed at the tip; (e,f) with hammer placed on the side. On the left column the measures taken with the "big" piezos; on the right column the measures taken with the "small" piezos.

The difficulty of detecting defects with the "smaller" transducers, which are able to excite the higher frequency components, could indicate that the presence of a defect mainly affects the propagation of the low frequency modes. We had made this observation also in the qualitative analysis, in which it was seen that the propagation paths that first encountered the defect had less intense components in the low frequency spectrogram. It will be interesting to evaluate this effect again.

In Fig. 13.24the waveforms acquired with the rectangular pulse excitation signal are shown. The measurements taken with "big" and "small" piezoelectric transducers are placed side by side. Again the differences between propagation with and without defect emulator are scarcely noticeable to the naked eye when analyzing the signal in the time domain. The major differences (albeit minimal), however can be seen better by analyzing the signals collected with the "smaller" piezos this time.

In the remaining part of the analyses, the acquisitions made with the hammer on the side will be neglected, as they have been seen to give less appreciable results, at least with the instrumentation and type of analysis in use. For the sake of brevity, we will consider only the analyses made with the transducers at a distance of 50 cm. In fact it is the position of the transducers that allows to collect as much signal as possible attributable only to the direct path before the influences of the reflections intervene. It was also chosen to excite the piezoelectric transducers with a sinusoidal burst signal with a frequency of 400 kHz.

In Fig. 13.25, the time domain signals collected for the steel bar are illustrated. Even in this case the differences between the signal collected in the time domain with or without hammer are minimal.

It is now possible to proceed with the analysis of the spectrograms.

In Fig. 13.26 the spectrograms calculated for the aluminum bar are illustrated. It can be seen that the

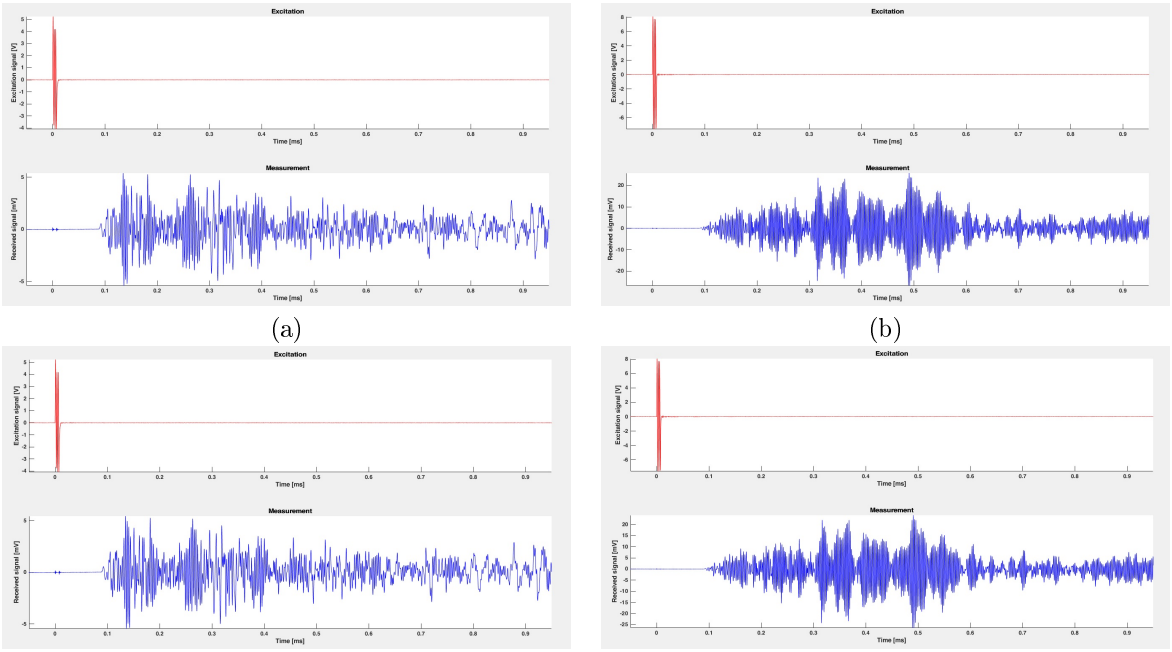


Figure 13.25: Time domain signals collected in presence or absence of defects for the steel waveguide by using the piezos spaced 50 cm apart with the excitation signal constituted by the sinusoidal burst with frequency 400 kHz. (a,b) without hammer; (c,d) with hammer placed at the tip. On the left column the measures taken with the "big" piezos; on the right column the measures taken with the "small" piezos.

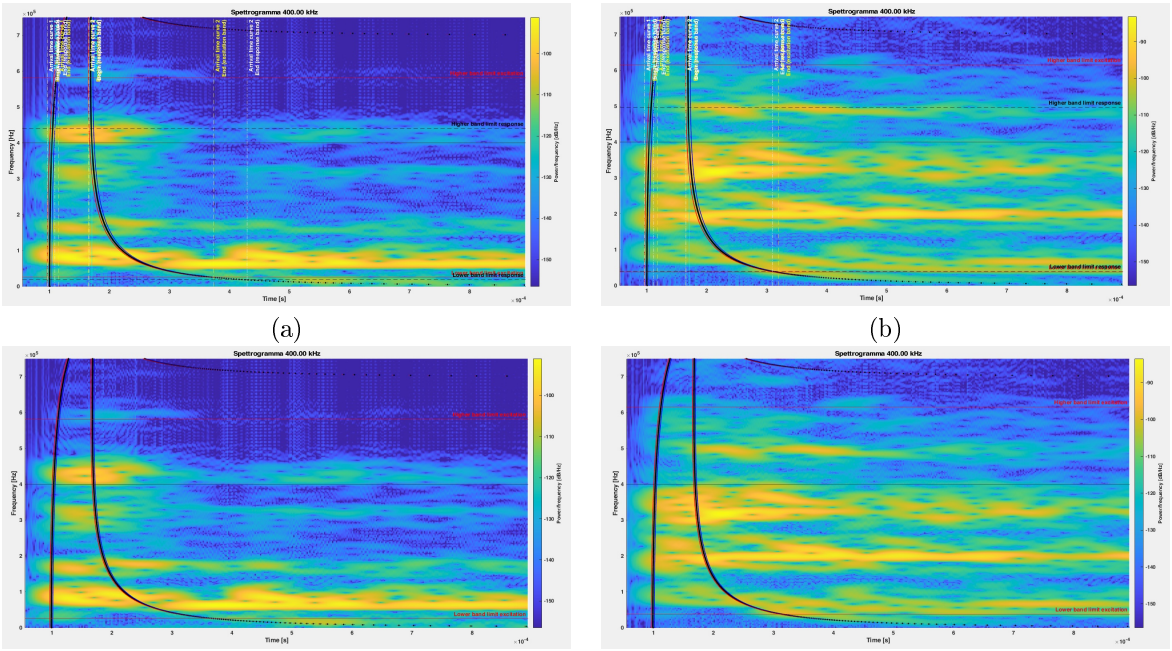


Figure 13.26: Spectrograms computed on the basics of the signals collected in presence or absence of defects for the aluminum waveguide by using the piezos spaced 50 cm apart with the excitation signal constituted by the sinusoidal burst with frequency 400 kHz. (a,b) without hammer; (c,d) with hammer placed at the tip. On the left column the measures taken with the "big" piezos; on the right column the measures taken with the "small" piezos.

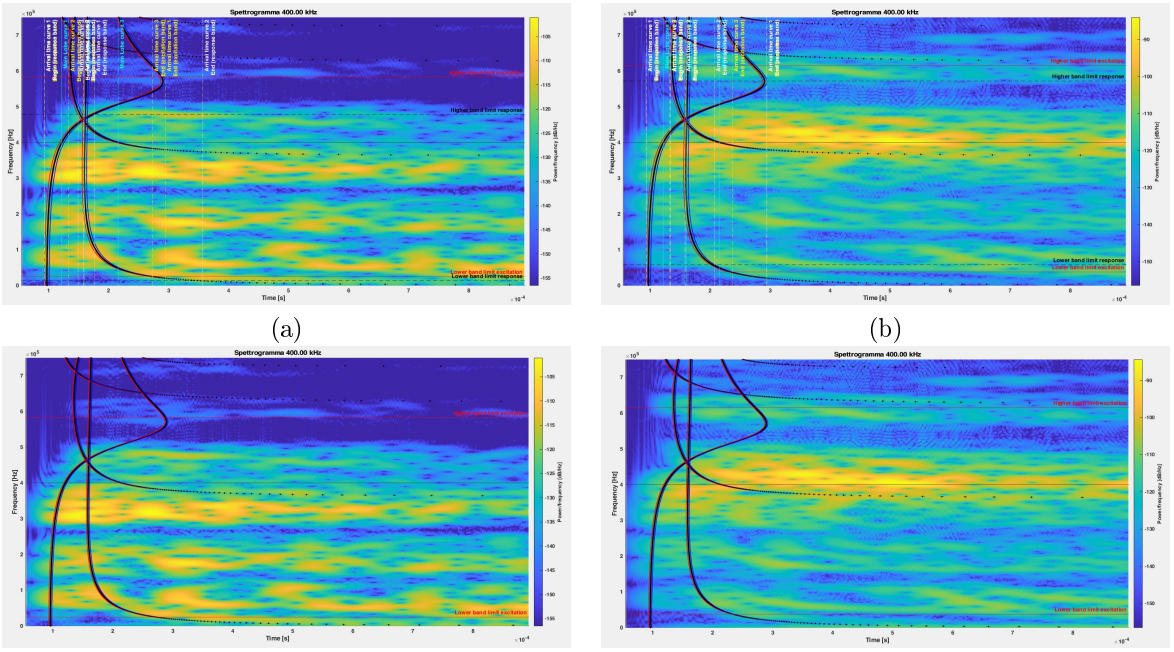


Figure 13.27: Spectrograms computed on the basics of the signals collected in presence or absence of defects for the steel waveguide by using the piezos spaced 50 cm apart with the excitation signal constituted by the sinusoidal burst with frequency 400 kHz. (a,b) without hammer; (c,d) with hammer placed at the tip. On the left column the measures taken with the "big" piezos; on the right column the measures taken with the "small" piezos.

differences between the two spectrograms calculated for the signals collected with "big" piezos (left column in Fig. 13.26) with and without defect emulator show some differences, although they are rather slight. For example, at $f = 250$ kHz it is noted that in the signal collected in the absence of a hammer (baseline) there is a concentration of energy which disappears when the hammer is added. In the baseline, moreover, some components at 450 kHz are excited, and they are not visible in the presence of the mass. As time passes, when the mass is applied, and therefore the reflections from the extremities and those produced by the defect emulator arrive, in the frequency ranges 300-325 kHz and 400-425 kHz, as well as below 50 kHz, a higher energy concentration of energy becomes possible. It could be the effect of the reflections caused by the presence of the mass. Even for the data collected with "small" piezos it is difficult to discern substantial differences.

In Fig. 13.27 the spectrograms calculated for the data collected in the steel waveguide are shown. Even in the case of steel, the differences between the spectrograms in the presence or absence of the defect are minimal. Generally speaking, it can be stated that the most evident differences are found at low frequency (below 20 kHz).

A final attempt can be made by comparing the frequency response functions calculated for the different combinations, represented in Fig. 13.28 for the aluminum waveguide and 13.29 for the steel waveguide. In general, for the aluminum waveguide, different heights can be noted in the peaks; this means that the energy is distributed differently in the signal which has propagated in the presence of the defect emulator. As for the steel waveguide, some differences can be observed; however, they are much more subtle than for the aluminum waveguide.

13.5 Conclusions for this Part of Experimental Activity

At the end of our analysis, some conclusions can be drawn.

The study of a simple waveguide such as a bar in the absence of defects has given good results. First of all, the combination of waveguide length, position of the transducers, absence of defects and acquisition time made it possible to understand the propagation behavior of UGW in the waveguide more clearly than the qualitative analysis in Chapter 12.

The correspondence between the dispersion curves simulated through SAFE method and what was obtained from the analysis of the experimental data was verified with even greater clarity, thanks to the substantial absence of reflections for the initial portion of the collected signals. In the case of qualitative experiments in fact, reflections sometimes overlapped to the arrival of the excitation signal along the direct path. In these

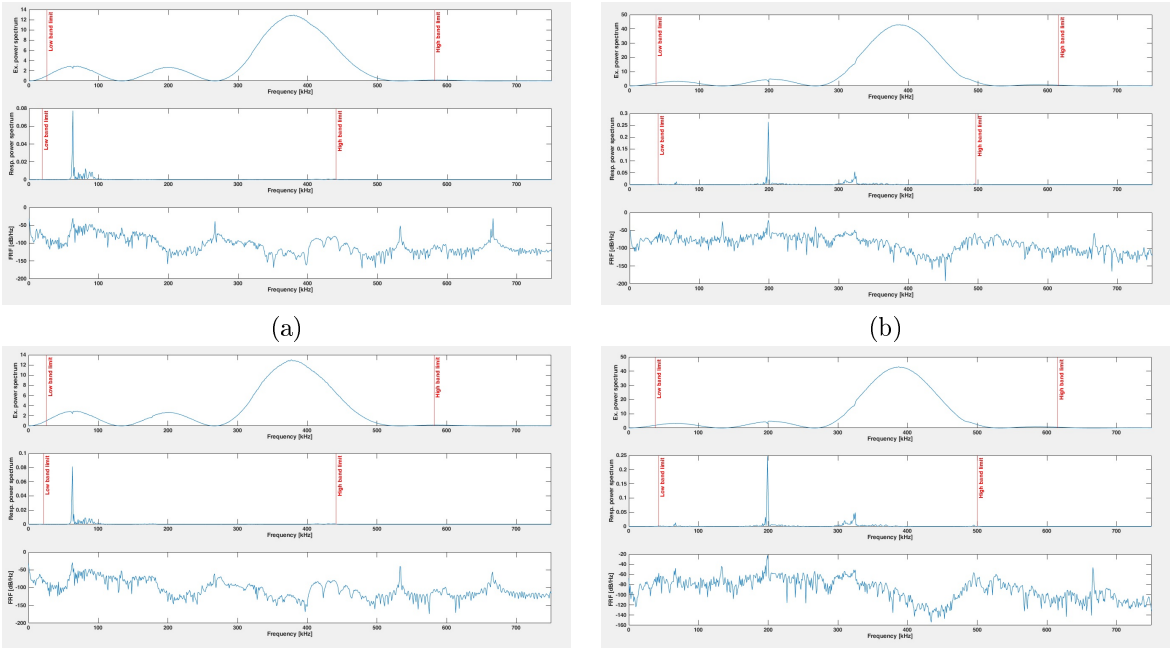


Figure 13.28: Spectra and frequency response function computed on the basics of the signals collected in presence or absence of defects for the aluminum waveguide by using the piezos spaced 50 cm apart with the excitation signal constituted by the sinusoidal burst with frequency 400 kHz. (a,b) without hammer; (c,d) with hammer placed at the tip. On the left column the measures taken with the "big" piezos; on the right column the measures taken with the "small" piezos.

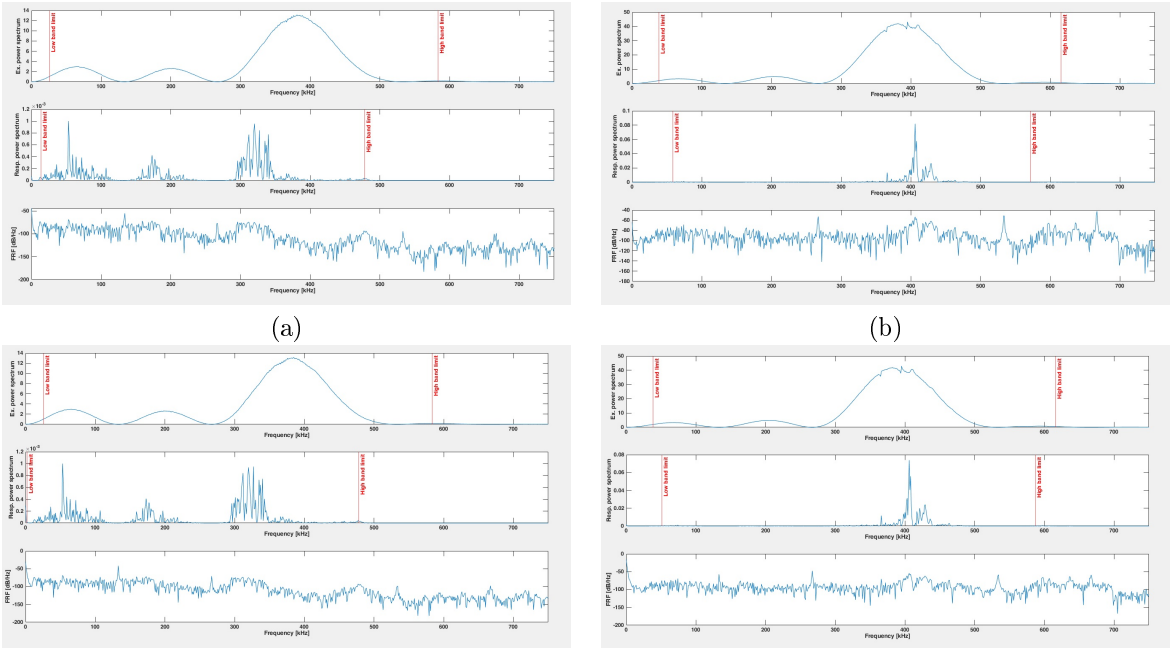


Figure 13.29: Spectra and frequency response function computed on the basics of the signals collected in presence or absence of defects for the steel waveguide by using the piezos spaced 50 cm apart with the excitation signal constituted by the sinusoidal burst with frequency 400 kHz. (a,b) without hammer; (c,d) with hammer placed at the tip. On the left column the measures taken with the "big" piezos; on the right column the measures taken with the "small" piezos.

experiments, it was possible to appreciate even more directly how the use of a different material or thickness for the waveguide leads to substantial differences in the propagation environment. Thanks to the absence of defects, it has been possible to better appreciate the richness of this propagation environment, with a large number of modes that propagate simultaneously. Unfortunately, once again, the available instrumentation did not allow us to achieve graphs and diagrams that would allow a full understanding of the phenomena at stake. To obtain even better results, it would be necessary not only to use more suitable transducers for the transmission and reception of signals in the given waveguide, but also to develop more efficient and sensitive software tools. There is wide room for improvement in both of these fields. For example, the use of a laser-vibrometer would allow to measure even more precisely the acoustic vibrations introduced in the waveguide. At the same time, the use of more refined analysis tools, such as the adoption of a non-rectangular window in the spectrogram plotting, or a wavelet analysis, could allow to better discriminate between the curves of the various modes capable to propagate in the waveguide. An interesting technique for the time-frequency analysis is the *warped frequency transform* (WFT); this is employed to produce a *warpogram* [117, 118], which should be interesting to test in a future development. Any analysis in time and frequency is affected by the time-frequency uncertainty principle, which limits the capacity of the technique to characterize closely spaced guided modes in a large frequency range. The WFT enables to overcome this limitation by implementing a more flexible tiling of the conjoint domain of time and frequency, mapped on the group velocity dispersion curves of the modes. The flexible tiling, as shown in ref. [117] allows to improve mode extraction. The WFT is fast, invertible and covariant to group delay shifts.

As for the defect emulator, the results obtained were less good. In fact, the influence of the mass on signal propagation is minimal, if not absent in some cases. The reasons for this need to be investigated, but some assumptions can be made. In fact, it can be thought that the hammer weight was not sufficient to induce appreciable variations in the way in which the signal propagate, or the hammer on the waveguide surface is not sufficient to produce effects. Perhaps, a more stable bond (through glue, for example) would have allowed to obtain more appreciable results. Further experiments would be needed to clarify this issue.

Chapter 14

Conclusions and Future Developments

The set of theoretical and experimental studies summarized in the first two parts of this manuscript only constitute a first step in the development of a system for the diagnostics of rail integrity. Once accomplished the study of the defects affecting the rails and of the main techniques in use today for their identification (Chapters 2 and 3, respectively), the discussion moved to the display of some of the basic notions on the propagation of ultrasonic guided waves in solid waveguides (Chapters from 4 to 8). The ongoing implementations of systems for the survey of rail making use of UGW were reviewed (Chapter 10), evaluating what has already been accomplished and what still remains to be done; describing what are the approaches followed and studied today and what are the their advantages and disadvantages. Based on this, guidelines were drawn for the possible development of a rail diagnostics system by means of UGW. In doing this, not only the achievable performances were taken into account, but also the requirements of each different type of implementation, both in terms of economic resources, knowledge, skills and testing possibilities.

14.1 Conclusions

The manuscript then focuses on the experimental phase which made possible to put the acquired theoretical knowledge into practice. As already known from the previous study, the field of UGW propagation is quite complex, and requires the development of adequate analysis tool and experience. In the experimental phase, therefore, have begun to lay the foundations for a practical knowledge of the propagation of UGW, with the development of the tools and methods of investigation. The results obtained, net of the available hardware and software tools limitations, and of the consequences due to external events, are encouraging. First of all, it was possible to affirm that the tools developed allow to produce results compatible with what is predicted by theory and simulation. In the results produced it is possible to follow the main characteristics of the signal propagation behavior and the variations due to the different type of excitation signal or transducers used. By analyzing these results it is possible to deduce the presence of a physical defect. Obviously the tools and methods developed up to now are not remotely sufficient for the development of a diagnostic system ready for implementation. However, this was not the purpose of this research activity: starting from zero or a little more, over time, with the knowledge and resources available, it was an unattainable goal. A much more realistic goal was the acquisition of practical knowledge specific to the field, the development of software tools useful for analyzing the signals that can be collected and useful for acquiring the aforementioned practical knowledge.

From a personal point of view, I would like to say that the objectives set for this activity have been achieved. The process to achieve these objectives had to clash with external and unpredictable causes. In the absence of these adverse events, the results obtained would probably have been even better, and the knowledge gained even greater. As I have already said, this is not an excuse for what has actually been achieved, but I believe that this is worth mentioning.

14.2 Future developments

In this manuscript, the first steps that need to be taken to develop an effective and reliable system have been described. The foreseeable future developments for this activity are manifold. By using more sensitive transducers and more refined analysis software tools, it will be possible to obtain a more accurate representation of the propagative environment of each waveguide. Once this has been accomplished, with the help of SAFE simulations it will be possible to proceed with the choice of the most suitable modes for propagation in the waveguide. These are the modes that propagate in the region of the waveguide of interest, with the least

possible distortion (i.e. with uniform speed as the frequency varies) and with the least attenuation. It will also be necessary to analyze how the environmental conditions and the slight variations in the geometry of the waveguide affect the propagation of the selected modes. Given the most suitable modes for propagation, i.e. those which, in addition to the characteristics stated above, are less affected by variations in environmental conditions and in the waveguide geometry, it will be possible to study the influence that the various types of defects will have on their propagation. In fact it will be interesting to see which defects reflect which modes and which ones will leave modes undisturbed. It will therefore be possible to choose a restricted subset of waveguide modes capable to identify and locate the defects of interest in the waveguide. The development of a rail structural integrity diagnostic system will then pass through the study of how to efficiently excite the chosen modes. That is, it will be necessary to choose the modes that are both capable of detecting defects and of being excited efficiently. Found a way to efficiently excite modes, it is necessary to develop a method to collect and analyze the signals, identifying the most appropriate techniques to clean up the collected signals from the inevitable alterations induced by the propagation and acquisition process. Once the signals have been cleaned up, it will be necessary to develop a method for the automatic detection of defects, such that this is accomplished reliably.

As seen, the procedure for obtaining a completed diagnostic system, given the starting point, it is still long and complex. It requires not only the acquisition of an even greater set of knowledge than what has already been obtained, but also the execution of much more articulated and complex experimental campaigns, even in the field. In fact, the field of ultrasonic guided waves propagation is a highly multidisciplinary research area, which blends advanced engineering knowledge in the civil, mechanical, electronic and IT fields. The development of a possible commercial system therefore requires the work of a multidisciplinary team and the availability of equipment and infrastructures that are not common to find in any research laboratory.

Part III

TCCSA - Track Circuit Contactless Signal Analyzer

Chapter 15

Introduction to the TCCSA

The *track circuit contactless signal analyzer* (TCCSA) is a portable device thought to be used on-field in the installation and maintenance of track circuits along railway lines. The use of conventional bench instrumentation generally poses several difficulties to the personnel: it is usually bulk, delicate and often requires grid supply (which can be difficult to find). The adoption of always more refined track circuit and train-track communication techniques, which need accurate tuning, calls for the use of a portable and low-cost albeit accurate device to quantify the real performance of the system. The signal acquisition hardware and software can also be used to extract signal features to be fed int a machine learning software in order to classify the signal behavior and thus assessing the health status of the track circuit components or of the rails. The main feature of TCCSA is that it can operate by leaving the track circuit working, without interrupting its normal operation and hence the circulation of trains.

A prototype of the system was already available, but it lacked of some functionalities. The performed activities regarded:

- the design of a compensation filter for a magnetic antenna that can be used to acquire track circuit voltages and current measures;
- the implementation of basic features in the user software, including the re-design of the GUI, the addition of the capability to save the input data stream and/or the statistics in .csv format, and the plots as images;
- the implementation of new functionalities in the software to plot the eye-diagram of the input data and to compute some related metrics, useful for following analysis.

The target of these activities was to enhance the capabilities of the TCCSA, allowing to be used not only as a device useful for installation and maintenance purposes for track circuits, but also to form the basis of a cost-effective data acquisition system that can be implemented to monitor the performance of the track circuit system as a whole, allowing to implement predictive maintenance strategies.

A brief background about track circuits will be given in Chapter 16; then, a literature review about existing projects involving track circuits used for the detection of damages or devoted to detect failures in the track circuit itself will be given in Chapter 17. Lastly, in Chapter 18 are described the performed activites to implement the new functionalities in the TCCSA.

Chapter 16

The Track Circuit

16.1 Introduction

To establish the presence of a train in a specific section of tracks, different solutions exist. One of the most common is to inject a carrier at one end of the section and detect its presence at the other end: the car axis, in fact, short-circuits the rails, preventing the transmitted carrier to reach the receiver. The carrier can be modulated by signals conveying useful information for train operation; the presence of cuts in rail can be avoided by using an audio-frequency carrier and electrical joints.

This system, performing safety critical operations, needs a careful and precise calibration, and to avoid track unavailability this should be accomplished under normal track operation.

This chapter is based principally on the manual [119] provided by the Italian state railways (FS) about signalling and safety systems, integrated with some contributions from Alstom documentation and other authors when specified.

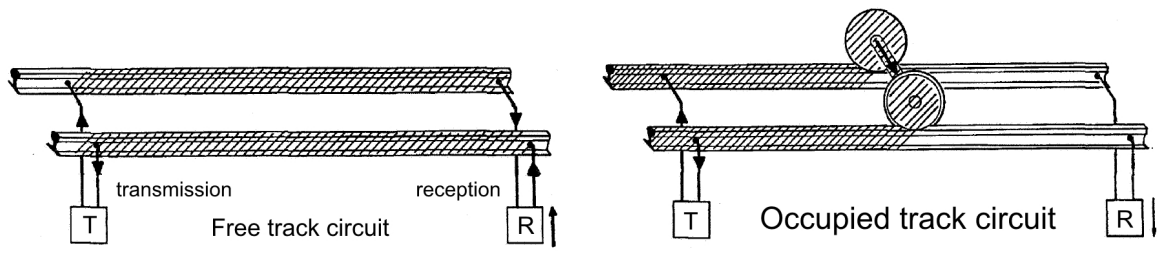
16.2 Principle of Operation

In railway signalling and interlocking applications, *track circuits* (hereinafter TC) are used to detect the presence of trains, or even only a single axle of a railway vehicle on a given track section. The name 'track circuit' originates from the fact that it is constituted by an electrical circuit whose conductors are the rails of the stretch of track. As it can be easily inferred from Fig. 16.1, at the extremities of the track circuit are placed, on one side the carrier *transmitter*, T; at the opposite side the *reception* device (usually a relay, named *track relay*), R.

- In Fig. 16.1(a) the circuit is represented under the status of *free section*: no trains are present along the track span overseen by the TC. As a consequence, the track relay is excited.
- In Fig. 16.1(b) the circuit is represented under the status of *occupied section*: at least one vehicle axle short-circuits the two rails, shunting the relay, which results de-energized

The device operates by *lack of current*: each and every supply shortage or circuit interruption provokes an occupation of the TC. This is since each TC must be fail safe, conceived according to the intrinsic safety dictates.

The smooth and safe operation of a TC assumes also that:



(a) (b)

Figure 16.1: Track circuit operating by lack of current. Pictures taken from [119].

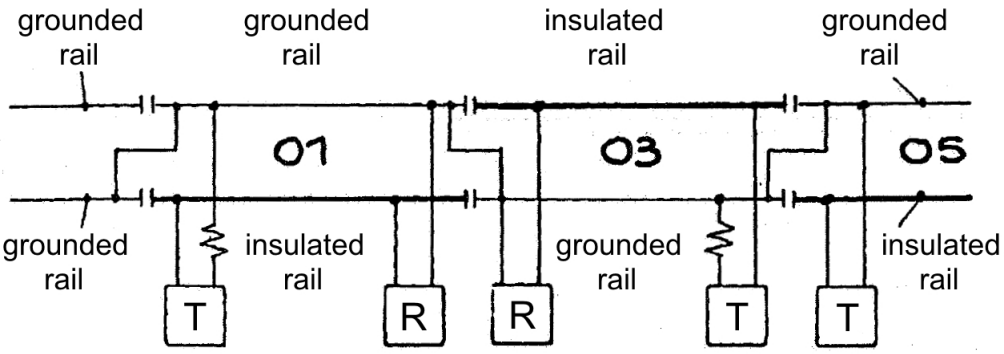


Figure 16.3: Contiguous track circuits. Picture taken from [119].

- the rails of the track section are sufficiently insulated both among themselves and with the other of the contiguous sections if they are separated by insulating joints;
- the length of the TC does not exceed a certain amount;
- no interferences are present, neither with close TCs, nor with the traction current return (if the line is electrified);
- the wheel-rail electric contact is ensured in each and every condition.

The track circuits can also be employed to communicate with the traveling train, by superimposing some coding currents to the TC current.

16.3 Traditional Track Circuits

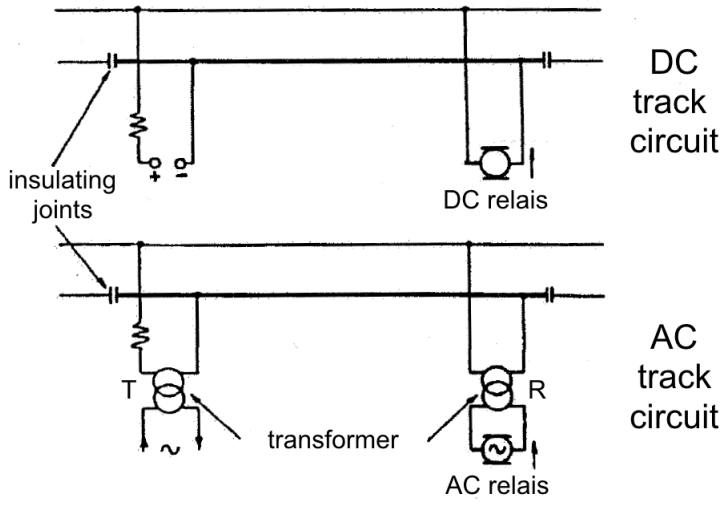


Figure 16.2: Operating principle of DC and AC track circuits. Picture taken from [119].

Traditional track circuits can work either with *direct current* (DC) or with *alternating current* (AC): the type of source and the type of relay must be changed accordingly, but the operating principle remains the same.

The operating principle of both DC and AC track circuit is sketched in Fig. 16.2. In both cases, a burden resistance is inserted to limit the injected current when the train axles short circuit the two rails. Traditionally both track circuits have *a single insulated rail*, since the other is grounded; in electrified lines, the return of traction current passes through the grounded rail. When multiple TCs succeed one another, insulated rails are alternated with grounded ones, as shown in Fig. 16.3. In such a case, insulating joints are needed at the extremities of each track section, and copper braid connections[1], named *Z-connections*. Each joints separates an insulated rail from one grounded, but the electric continuity of the traction current circuit is ensured by the Z-connection.

When the TC has a considerable length (between 700 and 2000 m), a system with *both insulated rails* is employed. In electrified lines, the return of traction current goes through both rails, passing from a TC to the subsequent by means of *inductive connections* (see Subsection 16.4.3.1) placed at each extremity of the TC in correspondence of the insulating joints.

The choice between the TC supply source, DC or AC, depends on the risk of interference with the return of traction current in electrified lines, which can be DC or AC too. Basically, if the line is electrified with DC

[1] Nowadays, to prevent thefts, they are usually made in aluminum.

current, DC track circuits cannot be used; if the line is electrified with an AC current of a given frequency, DC track circuits or AC ones operating at a frequency different from that of the traction current must be used[2].

16.4 Track Circuits Electrical Model

Rails act as conductors having different characteristics depending on which currents are taken into account: for traction currents, they act as a couple of ground-dispersive conductors in the close vicinity of trains; for the TC as a couple of imperfectly insulated conductors. In ref. [120], which forms a base for this section, some track models which account for the different mentioned situations are presented.

The knowledge of track admittance is important in the study of traction current return and TC operation. The track-to-earth connection model is complex and the rail-to-rail conductance, given their small distance, is not generally equal to the conductance towards ground of a single rail over long distances.

The representation of the terrain with a 2-level isotropic and homogeneous continuous function gives a reasonable correlation with experimental measures. The upper level is associated with the *inter-rail conductance*, and the lower one to the *rail-to-earth conductance*. The exact magnitude of the conductivity function depends on the environment, on the soil and on the substrate and cannot be exactly quantified.

16.4.1 Audio-Frequency Track Circuits Impedances

Audio-frequency track circuits are described in detail in Section 16.5. To increase the length of audio-frequency track circuits, it is necessary to control the permitted ballast conductance variation. The inherent track attenuation is a function of the rail-rail conductance: to reduce the attenuation, capacitors can be connected between the rails. The results in [120] show that the data transmission characteristics of tracks can be easily varied, although it should be remembered that connection of low impedances across the rails, such as track signalling apparatus, would itself modify the track transmission characteristics.

On electrified railways, if even some masts have a low impedance to ground, it is possible that the TC operation is affected: the TC relay could be energized also when the TC is occupied. To simulate this condition, the mast conductance to earth must be combined with the inter-rail conductance, but thus is not straightforward. In fact, the conductance to earth from the center of the equivalent inter-rail admittance is undefined.

16.4.2 Characteristic Elements of AC Track Circuit and Operating Parameters

This topic is analyzed by focusing on the following different topics:

- electric insulation between contiguous TCs;
- TC impedance;
- dispersion conductance between the rails.

16.4.2.1 Electric Insulation Between Contiguous Track Circuits

The delimitation of the stretch of track covered overseen by the TC is achieved by interrupting the rail continuity by means of insulating joints. The rails are physically separated by a piece of insulating material. Environmental conditions imply a great variability of the *joint electric resistance*, spanning

- from some hundreds of ohms with dry weather and good joint conditions,
- to few tens of ohms with wet weather and bad joint conditions.

At each TC extremity, the joints on both rails should be aligned for safety reasons: if not, an axle in-between the two joints cannot shunt the TC, resulting in a wrong free status.

[2] In Italy, on the conventional lines and independently of the electrification, AC track circuits operating at industrial frequency, 50 Hz, are employed. The track circuit can be used also to communicate the signal aspects in advance to trains by modulating the carrier; to achieve a richer amount of information, a carrier at 178 Hz can be used in addition to the one at 50 Hz. In lines electrified with 2x25 kV 50 Hz AC current, the track circuits operate at 83.3 Hz, when they are not substituted by the audio-frequency track circuits which have an electric joint (see Section 16.5).

16.4.2.2 AC Track Circuits Impedance

The TC comprises the two rails, transformers, transmission and reception resistances and various connections. *The impedance of each TC should be the lowest and the most constant in time possible.* It is given by:

$$Z = \sqrt{R^2 + (2\pi f L)^2} \tag{16.1}$$

where R is the ohmic resistance of the circuit, f is the operating frequency and L is the self-induction coefficient. To achieve a low impedance, being impossible to operate on the frequency f (which is fixed), or on the inductance L (depending essentially on the track geometric elements), it is necessary to keep low the resistance R. While rails are in themselves good conductors (having a great section per intensity of AC current), the problem sits in joints, if the rails are not welded together to form a long stretch of welded rail. Joints can rust or come loose, and this increases their resistance. To overcome this inconvenient, the extremities of rails are electrically connected one another by means of conductive connections. In case of welded rails the resistance of welds is negligible.

In practice, the impedance, is *directly proportional* to the length of the TC[3].

16.4.2.3 Dispersion Conductance in AC Track Circuits

While the TC impedance depends mainly on rails resistance, *the dispersion conductance towards ground* depends

- on the type of ballast,
- on the health status and type of sleepers,
- on weather conditions.

The conductance is *directly proportional* to the TC length, if the dispersion is homogeneous along its span. Sometimes, in place of the dispersion conductance G the *insulation resistance* R between the rails is considered $\left(G = \frac{1}{R}\right)$; therefore R is *inversely proportional* to the TC length.

In practice, the electric insulation between the two rails is low and variable with environmental conditions. The conductance variations can be relavant particularly in tunnels, also in function of the weather and humidity conditions. It is appropriate that the dispersion conductance has minimum variations in the different environmental conditions, and therefore a minimum limit value is established[4].

16.4.3 Drawbacks of AC Track Circuits Operating on Lines Electrified in DC

In TC having a single insulated rail, it is practically impossible to achieve a complete separation between the two currents, the alternating of the TC in the insulated rail and the continuous due to traction in the grounded rail. The alternating current of the TC has to go through both rails and the traction return current can partially circulate also in the insulated rail, for two reasons:

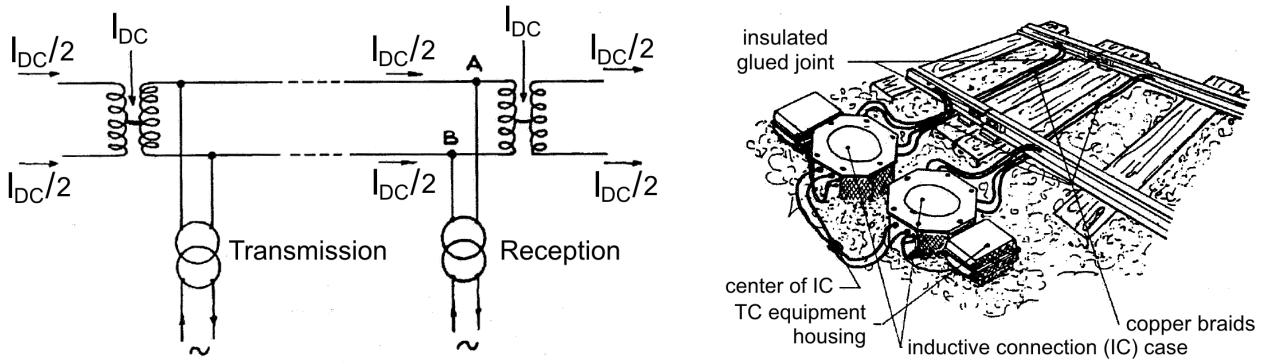
- because of the dispersion of traction current from a rail to the another through sleepers and ballast;
- because of the voltage drop caused in the grounded rail by the traction current.

The direct current, circulating inside the transformer windings, tends to saturate their magnetic nuclei and, consequently, to reduce their impedance: the result is an irregular operation of the TC. In fact:

- At the transmission end, the absorbed AC current increases significantly: at the limit, the supply protection devices could intervene and on the secondary of the supply transformer a sensible voltage drop is observed.
- At the reception end, the voltage on the secondary of the reception transformer can decrease to the point that the TC relay could be de-energized, even if the TC is free.

[3] In Italy, the track impedance at 50 Hz should have an impedance value around 0.6-1 Ω·km for an open line TC and of 1.3 Ω·km for a station TC comprising switches.

[4] In Italy, the nominal resistance and conductance values are with wet weather and good ballast conditions: resistance 2.5-5 Ω·km, conductance 0.4-0.2 S/km; with dry weather, the maximum admitted values are: resistance 10-20 Ω·km, conductance 0.1-0.005 S/km. In tunnels and in particular pollution conditions, the conductance can also reach values above 1 S/km. The minimum limit value adopted in Italy is of 2 Ω·km for the insulation resistance (0.5 S/km conductance) for open line TCs, and of 1.4 Ω·km (0.71 S/km) for station TCs.



(a) (b)

Figure 16.4: TC with inductive connections: between points A and B there is no voltage drop: (a) schematic view; (b) TC elements installed at one extremity of the TC. Pictures taken from [119].

16.4.3.1 Measures to Cope With the Traction Current Circulation Effects - Inductive Connections

Among other measures, like:

- limitation of the length of the TC;
- regulation of the limiting resistances at both ends;
- use of transformers capable to sustain higher powers at transmission end;
- use of a bridge made of resisitive and inductive components at the reception end;
- adoption of a band-pass filter at both ends;

a commonly used countermeasure is the adoption of *inductive connections* (IC), especially when the TC has a considerable length, greater than 700 m. In such a case, the transformer saturation cannot be avoided: the solution is to allow the return of traction current in both rails.

As shown in Fig. 16.4(a), the return traction current circulating in both rail is halved and the instantaneous voltage drop in the two rails are equal. Therefore, the points A and B shown in Fig. 16.4(a) are equipotential for DC current but not for the AC one. If the TC is balanced, in the TC transformer windings circulates only AC current.

Inductive connections are installed at both ends of the TC, as shown in Fig. 16.4(b). The inductive connection consists in a reactor having a laminated ferromagnetic nucleus wrapped with a center-tap winding. The winding wire has a considerable section since traction current runs through it, and it is made of copper or aluminum of flat section.

The operating principle of inductive connections is quite simple. The traction current, which cannot go through the insulating joint, circulates on the two semi-windings of the IC placed before the joint until it reaches the center. It then passes from the center of the first IC to that of the next one placed after the joint; it is brought back to the two rails of the next track section through its two semi-windings. The DC traction current is countered only by the very small ohmic resistance of the big winding (its size is in the order of some thousandths of ohms), and then can easily pass from one track section to the another towards the substation.

The TC alternating current instead finds an obstacle in the IC winding because of its impedance (1 Ω at 50 Hz): the AC current does not find in the IC a short-circuit. After injection of AC current at transmission side, it is then possible to obtain at the reception side a voltage sufficient to excite the TC relay if the track is free.

If the traction current is divided into two equal parts between the two rails, the IC impedance value is stable; the nucleus is not magnetized and therefore there is no danger of saturation. If the traction current in the two rails is unbalanced, a non-zero magnetizing flux is verified, provoking a sensible decrease in IC impedance. This can result in a irregular occupation of the TC. To avoid saturation, the IC nuclei are realized by means of a magnetic circuit having a given air gap. With the adoption of welded rails unbalances of currents in rails could happen only in case of faulty connections of rails to IC winding or if a rail is broken (limit case).

Generally IC must work correctly in a large temperature range[5] and be able to sustain high currents[6] and high overloads[7].

16.4.4 Safety Checks and TC Operation Regularity - Tuning

Before the entry into service of the TC, and after during periodic maintenance, a series of controls and calibrations must be performed to ensure safety and regularity of operation:

- effective TC occupation checks (certain de-excitation of the TC relay);
- regular excitation of the TC relay checks;
- calibration and tuning.

16.4.4.1 Effective TC Occupation Checks

The TC relays must be capable to de-excite themselves even in the most unfavorable conditions:

- bad rail-wheel contact conditions (rusted surfaces);
- light vehicles;
- minimum dispersion conductance values (maximum resistance between rails) in dry weather.

The check consists in verifying that the TC relay de-energizes itself when a given resistance is applied between the rails at the receiving end instead of a short circuit.

16.4.4.2 TC Relay Regular Excitation Checks

The TC relays must be able to excite themselves and stay excited even in the subsequent most unfavorable conditions:

- maximum dispersion conductance values (minimum resistance between rails) in wet weather;
- fluctuation inside given limits of supply characteristics (voltage, phase and frequency).

Once supply voltage and maximum conceivable dispersion conductance are fixed, the calibration is performed by varying the transmission and reception resistances, taking into account also the effect that such variation has on other TC operating characteristics.

16.4.4.3 TC Calibration and Tuning

Given the sensitivity of the exposed issues, before the TC entry into service, some calibration and tuning procedures are required to allow an optimal operation. Through application of measurement devices and shunt resistances, the variable elements of the circuits (transformers winding ratios, transmission and reception resistances) are tuned. As said, the tuning involves the TC occupation and liberation conditions checks and the measures of impedance and conductance to ensure that these stay inside the established limits.

16.5 Audio-Frequency Track Circuits

The *audio-frequency track circuits* (AFTC) works, instead of 50 Hz frequency, with frequencies lying in the audible band. They offer a great number of advantages:

- *elimination of traditional insulating joints* which delimit the TC length: the riding comfort is enhanced and also the maintenance costs are lowered: the track mechanical continuity is ensured;
- *possibility to eliminate interference among close* TCs by using different frequencies;
- *possibility of superimposition to other* TCs operating at different frequencies (both audio or traditional): as a consequence, it is possible to inject in the track a higher quantity of information;
- *possibility to increase security* by modulating the carrier frequency;

[5] Between -30° C and +60° C.
[6] Rail currents from 300 A to 1000 A and unbalance current of 120 A to 200 A depending on the type and use of the IC.
[7] 160% for 20 minutes and over 260% for 5 minutes.

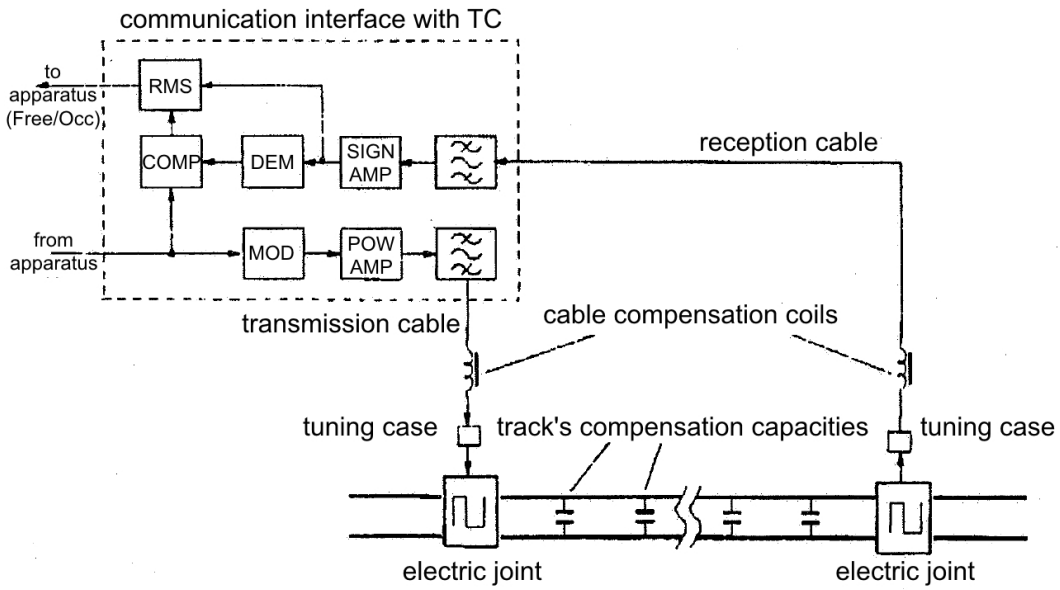


Figure 16.5: Audio-frequency track circuit block scheme, with transmission and reception apparatus for train-ground communication. Picture taken from [119].

- *possibility of remote power supply*;
- *insensibility to disturbs* due to the return of traction current or to harmonics generated by electric drives of locomotives;
- *probable lower energy consumption* respect to traditional TCs.

16.5.1 Description

Alstom Bologna (previously SASIB) developed a digital audio-frequency track circuit system (DIGICODE), which is widely employed. The block diagram of such a track circuit is shown in Fig. 16.5. This system uses the track to both detect the presence of a train and to transmit and receive data to and from the train. The mechanical insulating joint is substituted by an electric one, constituted by an S-shaped cable (named *S-bond*, see Fig. 16.6) properly placed between the rails of the track; the installation is completed by some passive elements housed in a case placed on the trackside (named *tuning case*). The S-bond offers a high level of directionality (defined as the ratio between the useful voltage transmitted in the track section of interest and the lost voltage, transmitted in the adjacent track section) and allows a correct two-way data transmission between ground and train.

Figure 16.6: S-bond electric joint installation on test site.

The separation between two adjacent TCs is realized through a different operating frequency bands (see Subsection 16.5.2). These frequencies depend on the chosen system; depending on the frequency varies the electrical joint length and the number of available channels and the channel width. The channels are cyclically repeated along the track (by ensuring that there are no near TCs also in parallel tracks operating on the same channel). As already stated above, this type of TC is also able to transmit and receive information to and from the train by modulating the carrier frequency by means of a *minimum displacement frequency modulation* (MSK) with two amplitude levels.

The features of audio-frequency track circuits vary among type, technology, system and country. Those of the Alstom's DIGICODE system are: operating range from -30° C to 75° C. In open line, by adopting

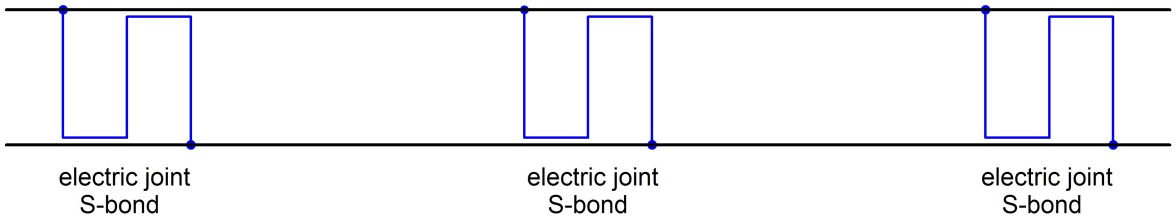


Figure 16.7: Stretch of track subdivided in 2 track sections by means of 3 electric joints.

some trimming capacitors between the rails[8], the TC length can vary from 100 m to 2000 m. The operating frequencies are chosen between 1900 Hz and 4300 Hz, thus allowing for an electrical joint length between 17 and 26 m; in the above range up to 8 channels are allocated. The breakage of a rail in the electrical joint area is detected by the DIGICODE audio-frequency track joint; this is a valuable feature of this technology respect to other systems.

16.5.2 Operating Principle

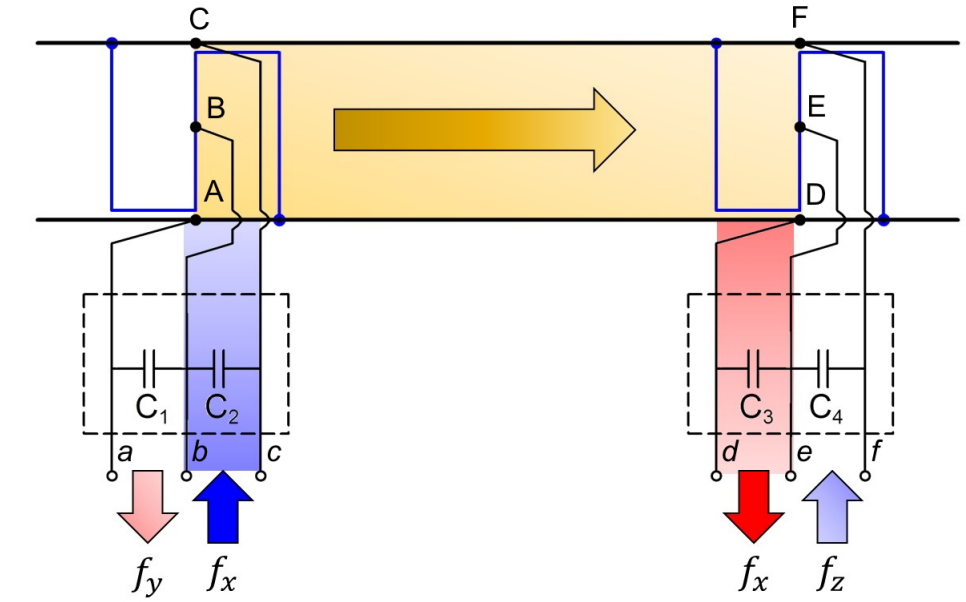
As already stated above, the electric joint is constituted by a S-bond and a tuning case; a generic section of track is delimited by two S-bonds, see Fig. 16.7. The effect of resonance is used at the S-bond to confine the given audio-frequency carrier inside the track section of interest.

As shown in Fig. 16.8(a), the tuning case contains two capacitors, and its three trackside terminals are connected to three specific points: the two external terminals (e.g., points *a* and *c*) are connected to the two rails of the track, in correspondence of the transverse section of the S-bond (points A and C); the common terminal (e.g., point *b*) is connected to the center of the transverse section of the S-bond (point B). The longitudinal branches of the S-bond, being parallel to the rails, constitute two inductors. Therefore, thanks to the connections to the capacitor in the tuning case, each S-bond can be schematized as two L-C resonators connected in series between the two rails of the track, as visible in Fig. 16.8(b). By changing the capacitances, it is possible to tune each of these two resonators to resonate at a different frequency. It is known that a RC resonator has a given impedance at the frequency of resonance, zero otherwise. Each TC is delimited by two resonators having the same resonance frequency, one used for transmission (②) and one for reception (③); in Fig. 16.8(b) this frequency, chosen between those that the system made available, is indicated with f_x. The adjacent resonators (① and ④), and hence the adjacent TCs are tuned to operate at a frequency different from f_x, and usually different one from another: in Fig. 16.8(b) are indicated with f_y and f_z, respectively.

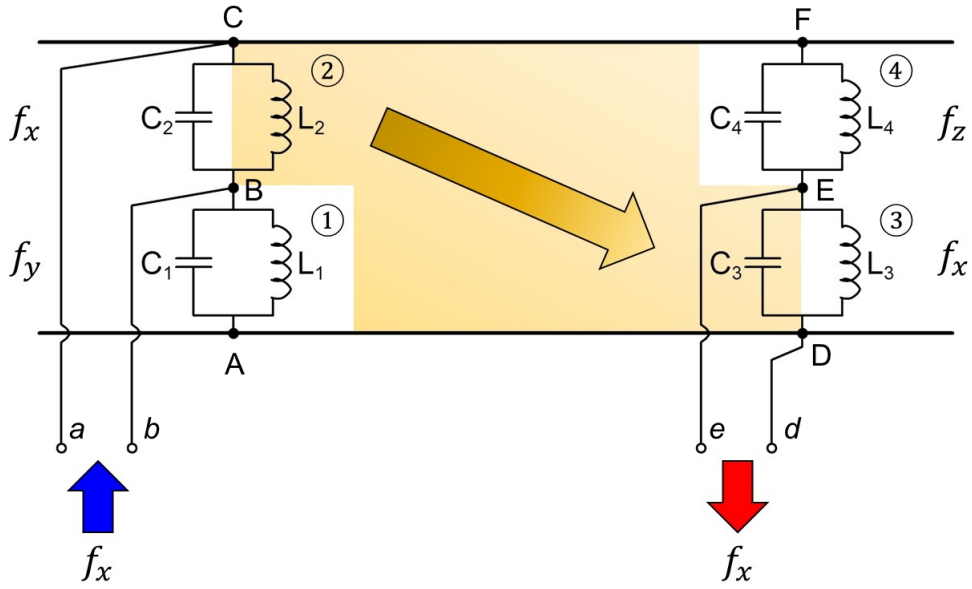
At the transmit side, a signal having frequency f_x is injected between the terminals *a* and *b* of the tuning case: since the pair ②, L_2-C_2, is tuned in order to have resonance frequency f_x, the impedance seen from *a* and *b* is finite. The pair ①, L_1-C_1, has oscillating frequency f_y, different from f_x; therefore, the impedance of the couple for the input signal at frequency f_x is ideally zero, and the pair acts as a short circuit. Electrically, the central point B is coincident with A: therefore, the input signal is transmitted on the right of the leftmost S-bond shown in Fig. 16.8(b). In practical applications, the ideal case scenario does not apply: hence, a little part of the injected signal is transmitted also on the right, on the adjacent track circuit. Care must be taken in order to avoid that this interference affects the operation of the track circuit, by ensuring safety and regularity. This is achieved by selecting an operating frequency (and hence inductance and capacitance values) such that the resonant circuit has a very low impedance for signals oscillating at the adjacent TC's resonant frequency. At the receive side, a symmetric process happens: the couple ③, L_3-C_3, is tuned to work at frequency f_x, while the couple ④, L_4-C_4, is tuned to work at yet another frequency, f_z. Therefore the resonant parallel ③, connected to the terminals *e* and *d*, is used to receive the signal at frequency f_x transmitted through the couple ②, since at such frequency the pair ④ is seen as a short-circuit (E is electrically coincident with F). Each electric joint is therefore transmitter for a frequency and receiver for another frequency.

Compared to an insulating joint, an electric joint is longer: this means that when a train passes over the joint its position is uncertain. This concept is illustrated in Fig. 16.9. This is not an issue from the perspective of safety: both the adjacent track sections are occupied, but there is no more a sharp separation between the two like in traditional TCs using insulating joints.

[8] Used to correct the phase of TC signals.



(a)



(b)

Figure 16.8: Audio-frequency track circuit (a) operating principle, (b) schematic diagram.

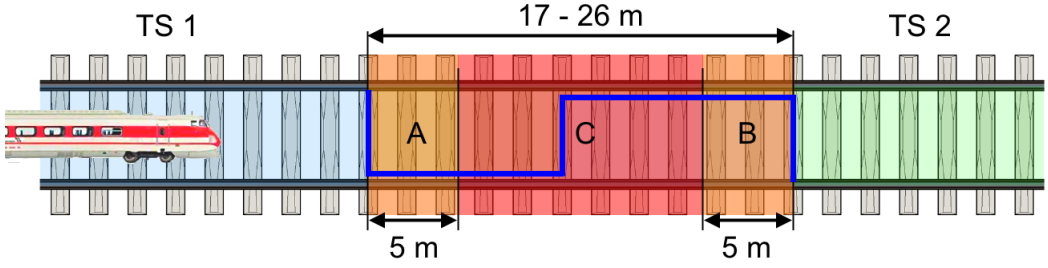


Figure 16.9: Uncertain position of a train over an electric joint (open line). In A, the track section (TS) 1 is certainly shunted (i.e., occupied by a train), TS 2 may be shunted; in B, TS 2 is certainly shunted, TS 1 may be shunted; in C, both track section are shunted and the exact position of the train is unknown (it is in C).

Chapter 17

Literature Review on Projects Regarding Track Circuits

In the technical literature, different projects have been published about new techniques either to detect faults in track circuits or that use currents and voltages in tracks to detect defects in track circuit elements and in rails. In the following, a brief literature review is presented, distinguishing between the two abovementioned approaches and focusing on those articles of potential interest.

17.1 Track Circuit Fault Detection

Different authors proposed methods to diagnose the health status of the track circuit using its currents and voltages. In general, some features are extracted, processed and then fed to a machine learning algorithm, able to infer the presence of faults in track circuit elements, and also to classify the type of fault.

17.1.1 Use of Recurrent Neural Networks

A first example that can be considered is the method proposed by de Bruin et al. in ref. [121]. They described a method to identify eventual faults along a TC by means of a *recurrent neural network* (RNN) with a *long short term memory* (LSTM) architecture. The analysis was based on the analysis of the TC currents measured in 5 TCs in the same area for a long period of time. This enables not only to diagnose the presence of a fault, but also to discriminate its nature.

The five types of fault they considered were:

- insulating joint fault;
- conductive object across the joint;
- defective rail;
- electric disturbance;
- ballast degradation.

A generative model, based on a qualitative comprehension of the system and of the effects of faults based on a limited set of real data has been developed to allow the development, testing and comparison with other methods of condition monitoring. It generates the amplitude of the TC current $I(t)$ in the receiver when a train passes over the section of track covered by the TC. Four phases have been recognized during the train passing:

1. Between t_0 and t_1 the train is still not arrived in the track section $\Rightarrow I(t) = I_h$;
2. In $t = t_1$the first axle enters in the TC $\Rightarrow I(t) \Rightarrow I_h \rightarrow I_l$;
3. Between t_2 and t_3 the train leaves the TC $\Rightarrow I(t) \Rightarrow I_l \rightarrow I_h$;
4. After t_3 the train has leaved the TC, but in some cases an overshoot of current occurs $\Rightarrow I(t) \rightarrow I_h$.

A simplified model of the current in the TC has been defined; this is characterized by four different variable parameters (named I_h, I_l, R, and ΔI_3) that depend on weather, hour of the day, type of train, etc.

The faults afflicting the TC performance change their influence very slowly over time. It is therefore important to sample the current in a manner which is informative enough, while keeping the minimum number of values possible, to ensure a high density of information. The values of measured current are observed for the 5 TCs and sent to the RNN at every time step T (train passage).

The RNN is a natural choice if the task is to detect temporal dependencies based on the training data since its structure allows to memorize past events. The LSTM enhances the performance of plain RNN by increasing the time period in which events are remembered, so solving the vanishing gradient problem. The current values obtained at each train passage over each of the 5 TCs are given as input to the first hidden layer of the RNN. The output layer is constituted by 6 softmax classifiers, one for each TC status: 5 fault conditions and healthy status. Two datasets have been employed, one for training the RNN and one for its test; in the sequences, the TC and fault properties were determined stocastically. At every train passing event T, the RNN was trained to classify the corresponding sequence. The objective of the classification $t\,(T)$ was to determine if the TC was in healthy status (normal case) or if it contained a fault of severity greater than 0.15 in a scale from 0 to 1. The threshold was chosen in order to detect a fault as soon as possible without declaring a false positive. A comparison between a RNN and a convolutional neural network (CNN) evidenced that the two types of neural networks are most suited to detect different types of fault. A RNN-LSTM is most suited to intermittent faults, but achieved an overall performance better than that of a CNN, since it is able to remember events that have took place far in time. The two techniques have complementary strengths, and their use can be successfully combined to achieve a better accuracy.

Four temporal dependencies have been considered in the classification of fault types: linear, exponential, abrupt, or intermittent. The difference between faults is influenced also by the velocity in which these manifest and develop, independently on their linear of exponential dependence on time. It is also possible to distinguish among spatial differences, if a fault involves only a TC, multiple TCs on the same track, or more TC on the same geographic area.

The results of the test, performed using a dataset containing 1500 sequences, led to the concsuion that the RNN was able to correctly classify the 99.7% of the input sequences (1495), without generating false positives.

17.1.2 Use of Neuro-Fuzzy Systems

Chen et.al. in ref. [122] described an hybrid quantitative/qualitative approach to diagnose faults: the *neuro-fuzzy* system. It combines the advantages of fuzzy logic and neural network, the capacity to cope with system imprecision and learning through training. The proposed method is able to detect and diagnose correctly the most common TC faults, and makes use of a test bench made of real TC components to emulate the operation of an AFTC. This is the most interesting aspect of the article, which dates back to 2008: the performance of the system is worse than that shown in the previous article, but the parameters used to classify the data and the types of considered fault are of interest.

The test bench was used to verify the correctness of the model and to gather data about faults. The TC has been modeled through a combination of lumped components representing concentrated and distributed parameters. The concentrated parameter components represented the terminations of the TC, while the distributed parameter components were used, repeated over 7 sections, to represent the section of track as a transmission line (each one of the 7 sub-sections represented a 50 m span of track, so the total TC length considered was of 350 m). The test bench was realized keeping in mind the practical limits about the installations of sensors on a real track under operation and the difficulties faced in the direct measure of the current flowing in rails or the difference of potential between rails. The sensors had not to interfere with the correct operation of the TC in any case.

Eight voltages and currents were measured and sampled[1] on the test bench:

- currents on electric joint branches (4 currents);
- transmitter and receiver currents (2 currents);
- output transmitter voltage;
- internal receiver control voltage (proportional to receiver current at the specified frequency).

Sensor data were gathered when different fault types were simulated on the test bench: transmitter and receiver faults, cable disconnections, tuning case or electric joint faults. In this case 10 types of fault were considered

[1] At 60 kHz.

along with the healthy condition; the fault were much closely related (respect to [121]) to TC components voltages and currents[2].

17.1.3 Detection of Fault by Means of On-Board Analysis of Currents and Voltages

Various manuscripts, like those by Oukhellu et.al. (see ref. [123]), or by Lin-Hai et.al. (see ref. [124]) analyze the effects of fault in the trimming capacitors placed at regular intervals between the track rails. These capacitors are used to compensate for the inductive behavior of the track and, hence, are useful to re-phase the currents traveling from transmitter to receiver in order to limit the attenuation and to improve the transmission level. If one of such capacities is faulty, the current changes its behavior in the remaining path towards the receiver, and this can cause the malfunction of the entire TC which the capacitor belongs. Both the authors suggest the use of an equipped train capable to measure the behavior of the current along the section of track and to use these data to assess the presence of faults on the trimming capacitors. This is done by the two articles by using different methods.

In ref. [123] the final decision on the detection and localization of a fault in the system is taken on the basis of the outcomes of a neural network and of a decision tree classifier fused together. The Dempster–Shafer theory is used to fuse data; this provides a framework to handle imprecision and uncertainty in the decision problem. The suggested method is based on a statistical pattern recognition approach. Simulation results on shown that a correct detection of fault can be obtained in more than 99% of cases, while a correct location in more than 92% of cases.

In ref. [124] a collaborative system of diagnosis was used to detect the presence of faulty trimming capacitors. It employs the *adaptive optimal kernel time-frequency representations* (AOK-TFRs) technique to diagnose the fault of a single trimming capacitor and the *adaptive generic algorithm* (AGA) to deal with multiple faults. Experiments show that, by using a collaborative approach, the system has a good diagnosis efficiency.

Sun and Zhao in ref. [125], instead, make use of the data gathered by a moving train measuring the TC current in the track section to detect the faults of the insulating joints between two adjacent track sections. It makes use of a *support vector machine* (SVM) applied to the polynomial quadratic coefficients used to describe the first three arches visible on the graph representing the behavior of the current when the trains enters a new TC. The system diagnosed correctly 96% of cases in the performed experiments with simulated data.

17.2 Rail Defect Detection Using Track Circuits

In refs. [126] and in [127] Espinosa et al. proposed a method to detect and locate qualitatively rail breakages in one or more rails of a double-track or a single-track railway line. The described approach is based on the monitoring of the electric discontinuity manifested by one or two rails of a double-track line; it ia assumed that the rails are electrically insulated from the remaining part of the railway network at the operating frequency of 800 Hz. The system makes use of 7 km-long detection sections delimited by emission and reception nodes. Emitters supply the section with a coded signal which provides a great noise immunity and enables to detect the emitter in case of simultaneous emission thanks to the great auto-correlation value of the code used. Analyzing the transmitted signals enables to detect rail breakages and to locate a breakage (i.e., not only which rail of which track is broken, but also its approximative location). This analysis process happens in a period of time lower than a minute in which the track section must be free of trains.

The system includes insulation interfaces between TCs and electronic modules used to receive and analyze data, is compatible with railway supply systems and can be remotely controlled. The system has a claimed advantage over conventional TC in the length increase of the track sections used to detect the rail breakages, in the better immunity to noise and also in the capacity to approximately locate the breakage.

The system is able to detect only broken rails which interrupt the electric continuity of the track section.

[2]Such as increase or decrease of voltages, currents, resistance or capability to de-energize the TC relay.

Chapter 18

The Track Circuit Contactless Signal Analyzer - TCCSA

The TCCSA is a portable commissioning and maintenance tool for audio-frequency track circuits. The idea is to implement a single and low-cost instrument for all the track circuit needs for commissioning, maintenance and technical appraisals, guaranteeing fast, safe, non intrusive, and accurate operation. The system is able to acquire all the signal carriers available on the TC and then demodulate the interested carrier to measure (in voltage and optionally current) its level or decode the transmitted information. After having done that, these information are analyzed to quantify the real performance of the system, by classifying the signal behavior in order to assess the health status of the TC or of the rails. This is done by leaving the track circuit working, without interrupting its normal operation and hence the circulation of trains.

The TCCSA could also be used to detect the presence of rail defects of significative magnitude which could affect the performance of the TC.

Alstom AFTC5 - Audio Frequency Track Circuit 5th Generation

AFTC5 is the 5th generation of Alstom audio frequency track circuit for heavy rail and light rail transit applications; AFTC5 is a highly configurable, jointless audio frequency track circuit. It operates between the interlocking system (i.e. Alstom's VPI® vital processor interlocking system) and trackside *audio frequency track circuit* (AFTC) equipment to generate and receive coded track signals for multiple track circuits. Additionally, AFTC5 can generate and receive bidirectional non-vital train-wayside communication signals. It makes use of Ethernet communications with other AFTC5 units and interlocking processors.
AFTC5 implements the following main functions:

- Vital communications with the signaling computer (i.e. VPI® system);
- Non-vital communications with other AFTC5;
- Vital generation and reception of rate-coded carrier for train detection;
- Vital generation of rate-coded carrier for speed commands to the train. During vital generation the rate coded carrier is checked to verify vitally the correct transmission;
- Non-vital generation and reception of TWC, *train-to-wayside communication*, and wayside-to-train communication signals;
- Berthing function, which ensures that the train is properly berthed before giving permission to the wayside equipment to transmit the door open command to the train.

Therefore, the general role of AFTC5 is to function as a vital, intelligent link between the higher-level vital interlocking controller and the track circuit equipment. The interlocking controller monitors the overall traffic conditions and either determines the track circuit speed commands directly, or provides vital information to AFTC5 enabling it to compute speed commands and other associated control data, as well as optional non-vital TWC data.
The primary operational mission of AFTC5 is to generate the correct vital track waveforms and power levels to the track circuit equipment based on current conditions (local inputs and network inputs from interlocking and any relevant adjacent AFTC5) and report the vital status of the track circuit and any required non-vital information (e.g. TWC) as required. Secondary missions include non-vital control and reporting between interlocking and AFTC5, including maintenance data and fault reporting.

AFTC5 Key Technical Features

Parameters, such as frequencies, code rates, transmit/receive levels and IP addresses, are software configurable, allowing identical track circuit hardware to be used.

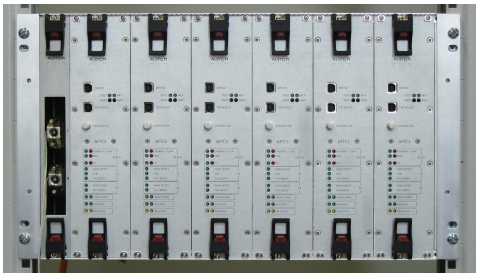


Figure 18.1: AFTC5 subrack panel.

Vital Train Detection

- Up to eight train detection carrier frequencies, in the range of 1 kHz–10 kHz;
- Up to three train detection code rates, in the range of 1.25 Hz–27.5 Hz;
- Single code (same train detection and cab signal code rates) or dual code (different train detection and cab signal code rates) operation;
- Train detection transmit/receive, transmit only, receive only or receive/receive;
- Digital signal processing filters to ensure proper rejection of substation harmonics and other undesired signals;
- Broken rail detection.

Vital Cab Signals

- Up to two cab signal carrier frequencies in the range of 990 Hz–10 kHz;
- Up to 15 cab signal code rates, in the range of 1.25 Hz–27.5 Hz;
- Code rate phase synchronization between adjacent track circuits;
- Temporary speed restriction settable to any available code rate.

Non-Vital Train-Wayside Communication Signals

- Frequency Shift Keying with 9650 Hz ±150 Hz or 9800 Hz ±150 Hz carrier frequencies;
- Gross bit rate of 100 or 110 bits/s;
- Manchester encoding, Return to Carrier encoding or no encoding.

Digital Signal Processing

- All signals are transmitted and received using digital signal processing techniques, allowing for future upgrades of filters and modulation types.

18.1 TCCSA: Purpose and Role of the System

The TCCSA has been initially thought as an internal Alstom device used for commissioning and maintenance of audio-frequency track circuit technologies. The idea was born during the AFTC5[1] qualification tests on some North American metro lines (Chicago red line and Toronto Metrolinx railway).

In the AFTC5 environment, different types of signals are injected on the tracks and travel along them:

- rate-coded *carrier for train detection* (TD);

[1] A track circuit technology mainly used in metros, see box at pages 179 and 180.

- signals used for *track to train communication* (CAB[2]);
- signals used for *train to wayside communication* (TWC).

All these signals have different characteristics and necessities, and their joint presence makes difficult to tune the audio-frequency system. To make a practical example, a speed command (CAB) needs 200 mA on a specific part of the track, and tuning this transmission requires to switch off all the transmitted signals different from the interested CAB one. In this way the commissioner is capable to measure the flowing current using a multimeter on the track. However, the exclusive presence of only CAB signal on the track is uncertain; for some reason (faults, errors, etc.), some other signals can come from other TCs. Switching off of the various signals meant that the TC, during the tuning, could not work in nominal mode; therefore, train operations are impelled, also because any technician has to work in close proximity to tracks, and this could be risky from a safety perspective if trains continue normal operations.

The system has been designed to allow two kind of track circuit activities:

- Support to commissioning an maintenance operations, by performing:
  - track circuit voltage measurements,
  - track circuit code-rate measurements,
  - track circuit current measurements;
- Perform deep engineering field analysis:
  - track circuit spectrum analysis;
  - track circuit channel demodulation;
  - field device analysis (bandwidth measurement and tuning of resonating frequencies);
  - track circuit integrity check;
  - unbalancing current measurements;
  - field insulation equipment check.

Therefore, the TCCSA makes possible to:

- *acquire all the signals carriers* present on the track circuit;
- *demodulate the interested carrier* to measure the level or to decode the information;
- *analyze the real performance of the system*, measuring the *signal-to-noise ratio* (SNR), analyzing the eye and constellation diagrams;
- *classify the signal behavior* through neural networks to give more statistics and prevent accidents.

All this is accomplished *leaving the transmitters to work in nominal mode* and by means of *contactless probing*. This enables to keep the trains running as usual, without interrupting the service.

The main advantages of such a solution are:

- It is an *all-in-one instrument*: the devices comprises accurate voltmeters, current meters, oscilloscope and spectrum analyzer in a single case.
- it allows an *easy and safe commissioning*: the signal components such as TD, CAB, TWC, berthing, noise and spurious signals can be separated, demodulated and analyzed singularly. This allows to discover each TC configuration adopted, allowing for a faster and accurate commissioning in case of equipment renewal or maintenance.
- It allows *real-time processing* of signals, so enabling to verify in real-time the performance of a commissioning or maintenance operation, during normal train activities (for CAB or TD signals).
- It is *compatible* with AFTC nominal operation: it does not affect in any way the operating and safety characteristics of Alstom's track circuit systems, independently of their technology.
- *Track circuits are available during the commissioning*; moreover, the measurements does not affect the signals traveling on the track thanks to the electromagnetic coupling between track and antenna. This enables to avoid to stop train circulation on the track and, even better, real-time processing allows dynamic measurements during the train passage under normal operation.

[2]CAB stands for train cab.

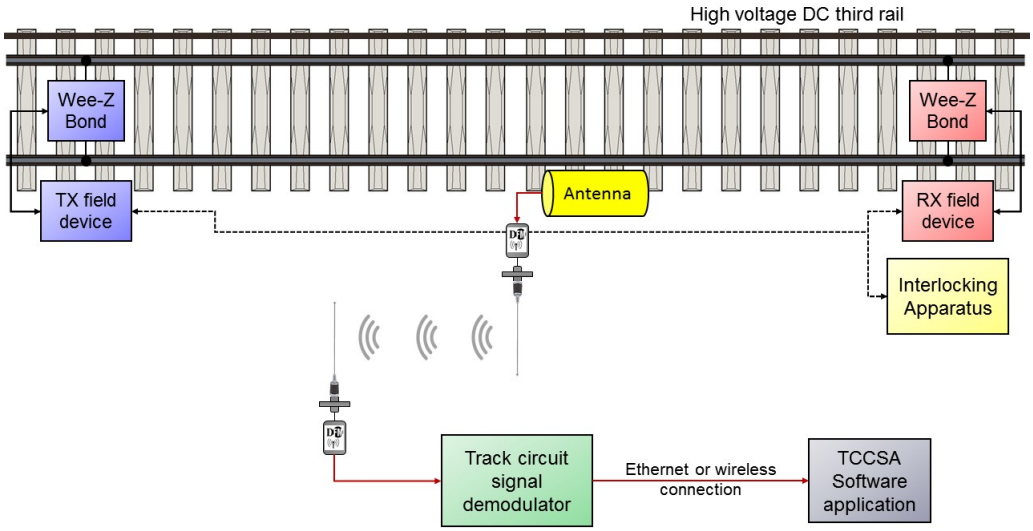


Figure 18.2: TCCSA system overview. The TC elements connected with the interlocking apparatus are also included.

- The *instrumentation is portable*: being an all-in-one instrument different, delicate and heavy specialized instrumentation does not have to be carried on field; the supply with a standard 5 V USB or a battery power pack allows to avoid mains power supply which can be difficult to find on field.
- It is a *cheap solution*: being an all-in-one instrument allows to avoid to carry expensive and delicate instrumentation on field, but highly accurate measurements can still be obtained in real-time.
- The system is designed to allow a *DSP easy customization*, making it possible to easily implement all kinds of measurement needed (e.g. different types of demodulation for different TC signals) thanks to a user-friendly DSP framework.
- The *safety of the operator is guaranteed*: the commissioner does not need to touch the track in which the return of traction current circulates (in case of faulty grounding it can be dangerous, given the high currents involved); it is sufficient to keep the antenna close to the track using a special support. Moreover, once the antenna is properly placed, it is not required to stay close to the tracks, with the risk of being hurted by a passing train.
- It allows *data sharing among operators*: the information acquired from the track and processed by the DSP are available on the Ethernet interface, therefore potentially available to every operator along the track circuit. In fact, the system allows up to four point monitoring shared on LAN, a special windows based software capable to represent up to four signals in the same time. The signal analysis therefore can be done on site or from remote; the signals can be recorded and the analysis done off-line.
- It enables to develop a predictive *maintenance strategy*: by analyzing the behavior of signals during time it is possible, by means of machine learning software, to assess the status of track circuit and track components, determining the presence of a deteriorating element, which can be kept monitored.

18.2 TCCSA: General Overview

A schematic overview of the TCCSA system is provided in Fig. 18.2. The system is composed by the following three main components:

1. *Antenna*: these devices enable the contactless probing; antennas are designed to pick-up the *extreme low frequency* (ELF) signals traveling on the track circuit (TD, CAB and TWC) without affecting their behavior. Two types of antennas are provided:
    - *Electric (capacitive) extremely low sidelobe antenna*, suitable for voltage measurements, of which a prototype is visible in Fig. 18.3(a);

(a) (b)

Figure 18.3: TCCSA antennas: (a) electric (capacitive) ELS antenna prototype; (b) magnetic (inductive) ELS antenna prototype.

Figure 18.4: TCCSA track circuit signal demodulator.

- *Magnetic (inductive) extremely low sidelobe antenna*, suitable for current and power measurements, of which a prototype is visible in Fig. 18.3(b);

2. *Track circuit signals demodulator device*: a device designed to demodulate the composite TC signal gathered by the antennas. The device is shown in Fig. 18.4, and it is composed by two main boards:

   - *Demodulator board*: this is used to acquire the signal coming from the antenna. It is able to measure, for each of its four channels, current and voltage of the composite signal running on the rails. Through a DSP it is then able to demodulate the received signals, and distinguish between TD, CAB, TWC and harmonics.
     In fact, the employed DSP (AnalogDevices ADAU1701 DSP) has 2 audio inputs and 4 DAC outputs; it can operate at sampling rates of 44.1 and 48 kHz. It can be programmed through an Arduino-uno board to configure the demodulator and the input signal compensation filter; the configuration files are generated through the software SigmaStudio. The pre-installed ASK demodulator is suited to the AFTC5 system, and it is able to demodulate up to 4 channels at the same time.
   - *Acquisition board*: it is used to acquire directly TC signals from the antenna or the demodulated signals from the DSP board; the acquired signal samples are made available on the Ethernet port to allow their analysis on a remote device running a specific software application.

3. *Software analyzer application*: it is used to show and measure the signals collected and demodulated by the antenna and the TC signals demodulator device. A screenshot of the main window is shown in Fig. 18.14. The various properties of each observed signal are used to gather information about different TC operating characteristics and settings; e.g.:

   - *Voltage*: liberation/occupation settings;
   - *Current*: speed command settings;
   - *Frequency*: disturbance analysis;
   - *SNR*: disturbance analysis;

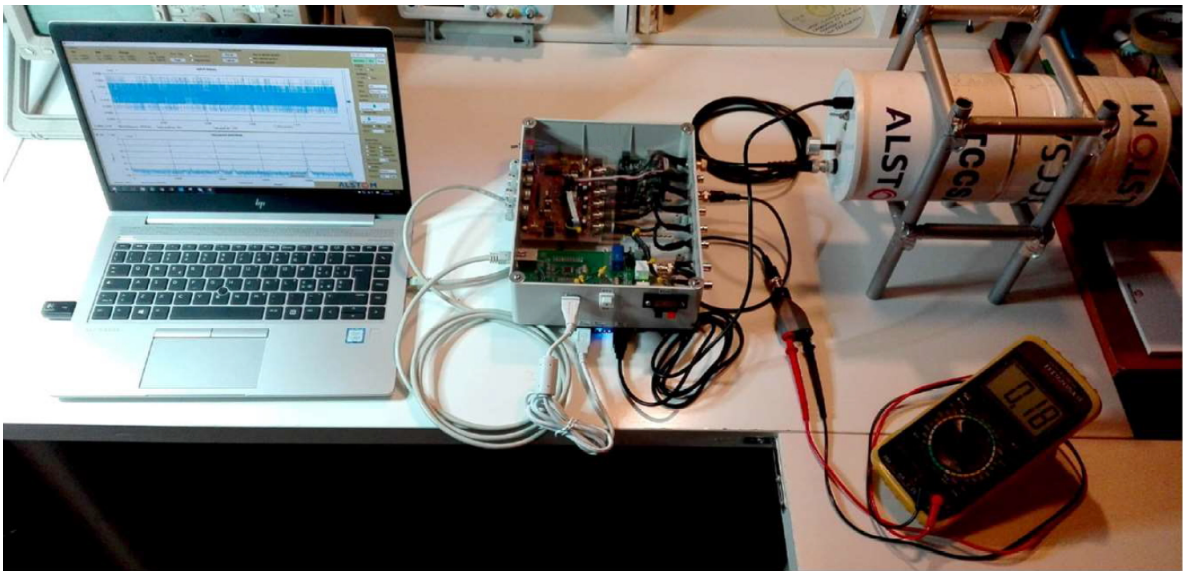


Figure 18.5: TCCSA prototype on the workbench; from right to left it is possible to recognize the capacitive antenna, the TCCSA device containing the demodulator and acquisition boards, which is connected to a portable PC which is running a version of the TCCSA software.

- *Eye Diagram*: performance;

The overall representation of the TCCSA is given in Fig. 18.5, where a prototype is being tested on the workbench.

18.3 Performed Activities

18.3.1 Design of a Compensation Filter for a Magnetic Antenna

The magnetic antenna has a non-flat frequency response, while for the following processing a flat response is desired. The antenna frequency response therefore needs to be compensated for. This is accomplished in the acquisition-board by means of a 150th order finite impulse response all-pass filter (FIR) implemented on the DSP. The filter coefficients are tailored to compensate exactly for the frequency response of the specific antenna; this is done by means an application written in matlab, whose front panel is illustrated in Fig. 18.6.

Given the antenna's frequency response samples as input (expressed in different formats), it is then possible to execute an interpolation and data extrapolation of data (by using different interpolating functions) to obtain the curve of the frequency response over the given range of frequencies. From the curve, the FIR filter coefficients will be computed, as well as the *inverse FFT* (IFFT) of the antenna and filter responses. The FIR and IFFT coefficients will be provided as outputs in a text file.

An example of the software operation can be given as follows. The antenna frequency response is constituted by few samples of voltage measured at the antenna output terminations when it is subjected to a periodic signal constant in amplitude oscillating at different discrete frequencies. The conversion in decibel it is therefore executed, using as reference the antenna measured voltage at 50 Hz. Then the frequency response samples are represented on a Cartesian plane, as shown in Fig. 18.7.

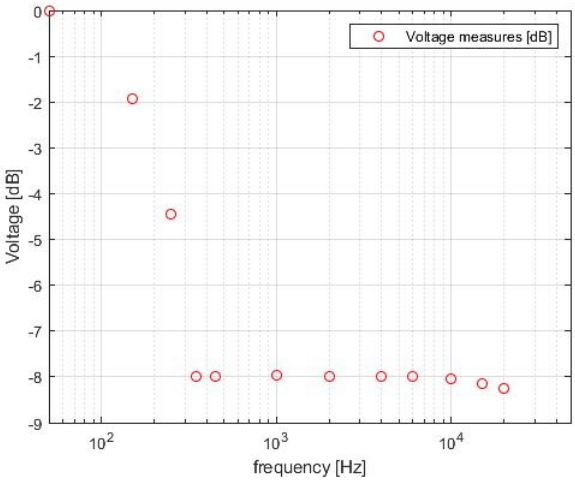
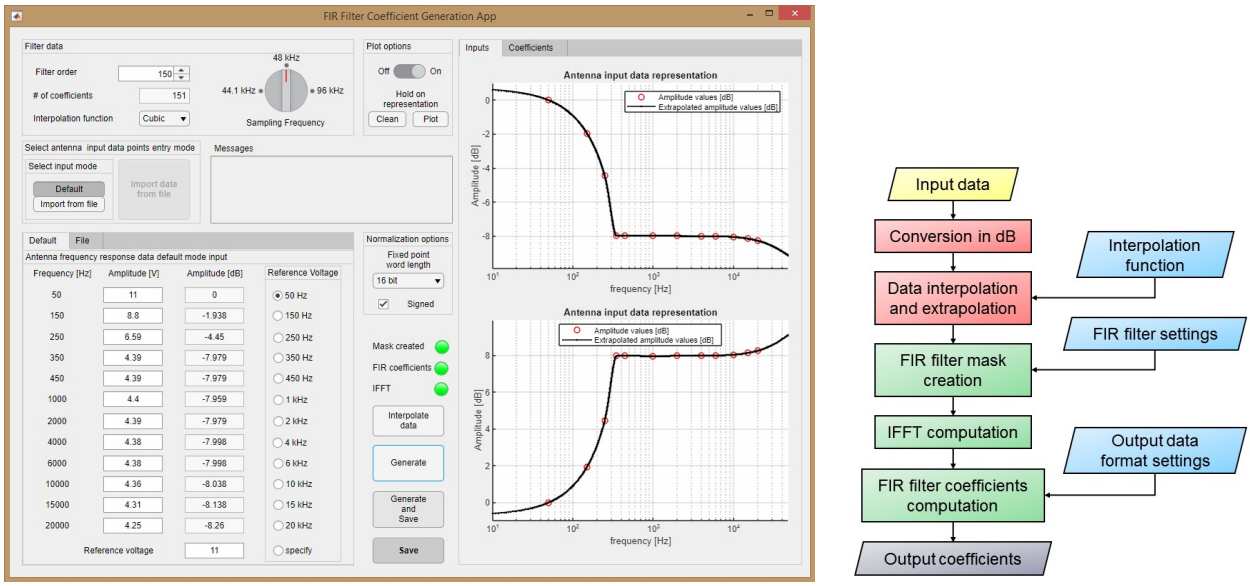


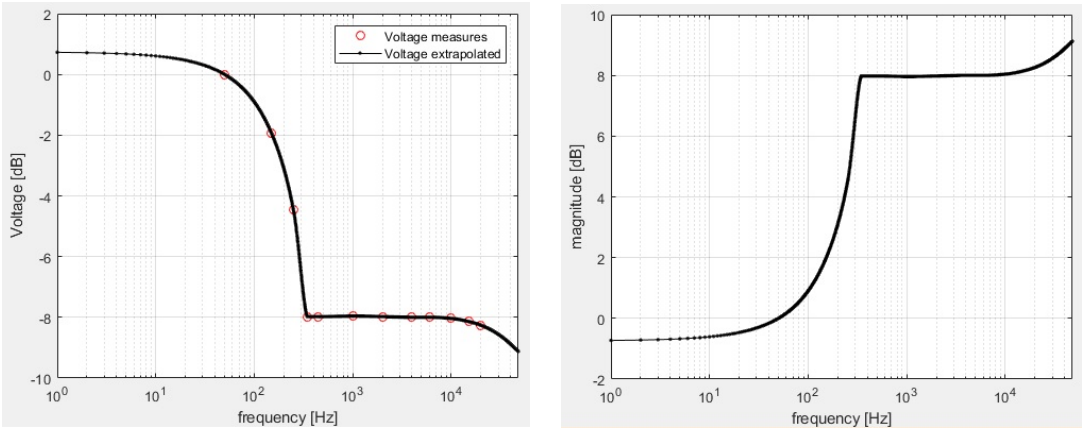
Figure 18.7: Antenna frequency response samples plot. The frequency axis is in logarithmic scale here.

The data are then interpolated and extrapolated by means of the appropriate matlab function `interp1`. Different interpolation functions can be chosen (cubic or spline); the result obtained after that the cubic extrapolation is performed on the initial samples is shown in Fig. 18.8(a).



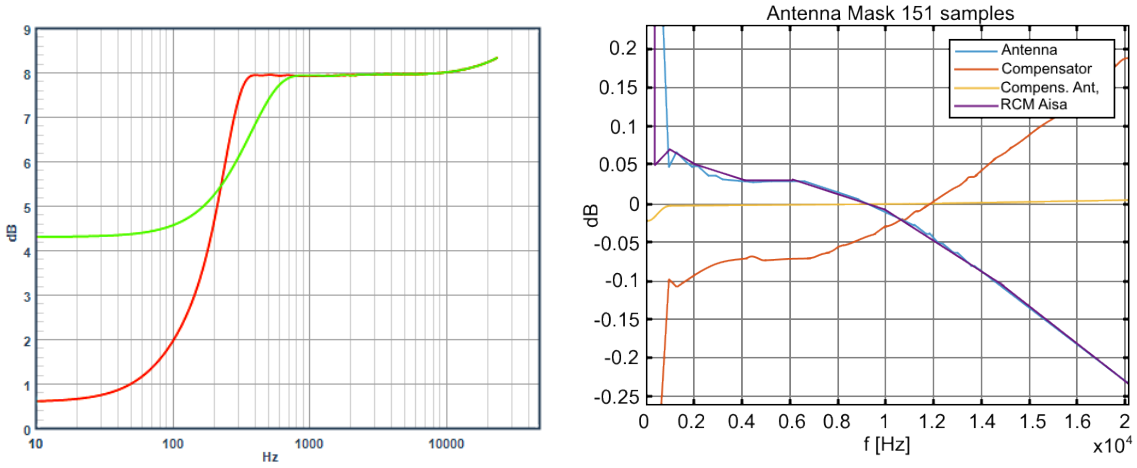
(a) (b)

Figure 18.6: FIR compensation filter coefficient generation application (a) front panel; (b) algorithm.



(a) (b)

Figure 18.8: (a) Antenna frequency response samples interpolation and extrapolation using a cubic function. (b) Filter frequency response computed as the reciprocal respect to 0 dB of the antenna frequency response.



(a) (b)

Figure 18.10: SigmaStudio comparison between different compensation filter responses. (a) The order of the filter is varied: red line, filter of order 700; yellow line, filter of order 150. (b) Different compensation filters are used, and the outcome of the antenna+filter system is represented.

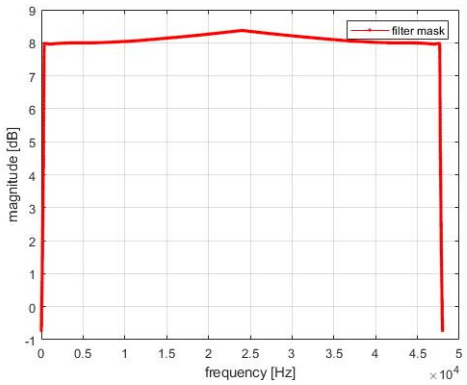


Figure 18.9: Filter mask sampled at 48 kHz.

As already stated above, the overall amplitude response of the system antenna and compensation filter must be flat; it was established that a 0 dB level would be suitable to the scope. Therefore, the compensation filter response is computed as the reciprocal respect to 0 dB to the antenna's filter response. This, obviously, brings to an ideal transfer function; however, a filter having the required characteristics cannot be implemented using the limited resources of the available DSP. The achieved filter frequency response is shown in Fig. 18.8(b).

After evaluating the antenna and filter frequency responses, the masks for filter (see Fig. 18.9) and antenna are computed, with a sampling frequency f_s of 48 kHz (other sampling frequencies can be adopted, 44.1 kHz or 96 kHz). The mask is obtained by using a method that, albeit being brute force, is quite effective. A couple of vectors is generated: in the first the frequency values from 0 to f_s are contained, in steps of 1 Hz; in the second the corresponding mask amplitude values are memorized. As a consequence, both the vectors have the same length.

The response of the compensation filter and antenna must be windowed at the half of sampling frequency to withstand the Shannon's Theorem; therefore, the first amplitude samples of the extrapolated signal, from 0 to $\frac{f_s}{2}$ are copied in the mask amplitude vector, in correspondence to the frequencies spanning from 0 to $\frac{f_s}{2}$. This is possible since the extrapolated values to be copied are also sampled at a frequency f_s and samples are spaced 1 Hz one from another, so their number is equal to $\frac{f_s}{2}$ in Hz. The same selected samples are then mirrored with respect to the frequency axis and appended to their un-mirrored counterparts in the mask amplitude vector, until the sample corresponding to frequency f_s is reached. The final result, the mask, is achieved by fitting the two vectors (that of the frequencies and that of the corresponding amplitudes) by means of the Matlab instruction `fit` (depending on the user choice, a cubic or spline function can be adopted).

In the meantime, the IFFT coefficients of the filter and antenna frequency response are computed by means of the Matlab function `ifft`.

The coefficients required by the FIR filter block of the software used to configure the DSP are obtained from the filter mask; the `fir2` Matlab function and Sigmastudio by Analog Devicesis are employed to this end. Being the maximum implementable filter order in the given configuration equal to 150, 151 coefficients are thus needed. For the implementation in the DSP board, a 16 bits fixed point format is required for each coefficient, so the computed coefficients are converted by means of the Matlab instruction `fi`; the application is able to convert and output the coefficients in formats other than 16 bit fixed point, such as 8 bit fixed point and 23 bit double (as they are internally computed).

A comparison between two filters of different order (one of order 700, the other one of order 150) is shown in Fig. 18.10(a).

As it can be inferred from Fig. 18.10(a), even with a filter of order 150, it is not possible to retrace the

behavior of the ideal curve: this brings to a partial compensation of the antenna response. The simulated compensated signal (available at the output of the antenna and compensation filter described in this Section) is shown in Fig. 18.10(b). The compensation is not perfect, but still well acceptable, given the limitations of the available hardware, in which the compensation filter constitutes only a small portion of the tasks to be accomplished.

The impact of the difference between the almost ideal behavior and what it is effectively obtained needs to be verified through measures on the real hardware. The results illustrated in Fig. 18.11(a) show that the compensation filter, fed by a signal having comparable characteristics of that received from the antenna (sinusoidal signal with constant amplitude of 2 V), performs satisfying compensation. The amplitude of the output signal acquired by the oscilloscope shows little variation as the frequency of the input signal varies, as expected from the simulation results; the graph shown in Fig. 18.11(b and c-n) represents the real situation. Only at 50 Hz a minimum attenuation still remains, but it is not judged detrimental, and was expected as well.

It is thus possible to state that the implementation of the 150-th order all-pass compensation FIR filter on the DSP installed on the DSP board, given the hardware limitations, performs in a satisfying way.

18.3.2 TCCSA Oscilloscope and Signal Analyzer Board and PC Software

18.3.2.1 Premise

The TCCSA oscilloscope and signal analyzer board constitutes the hardware extension of the PC software.

Both the board and software where developed by Alstom engineers on the basis of an open source project (the *network connected signal analyzer*, NCSA) published on the elektor magazine [128]. This project was chosen initially by Alstom engineers to speed-up the development time, which was very strict.

18.3.2.2 The Initial Project

The employed project was the base for the development of the TCCSA oscilloscope and signal analyzer that consists in a basic oscilloscope, spectrum analyzer and signal generation integrated system of purposely designed hardware and software. The oscilloscope sampling rate is of up to 1 MHz, it supports sub-sampling; the maximum input signal level is of 0 dBm (0.225 V_{RMS}), while it is able of a sensitivity of down to -80 dBm (22.5 μV_{RMS}). The electronic board is able to connect via Ethernet to a PC running the software, which controls the board, adjusts the operating parameters and elaborates the signal. The analysis capabilities, in addition to those of a basic digital oscilloscope, include also those of a spectrum analyzer, enabling to produce the frequency-domain power spectrum of the input signal; the NCSA system is able also to generate analog or digital signals.

The NCSA board is where the signal is conditioned and amplified before analog-to-digital conversion, operation that allows to send the digitized data via Ethernet to the application software. The remaining signal processing tasks are performed in a PC before displaying the final results; through the user interface embedded in the software it is also possible to configure the operation of the hardware module. As already mentioned, the NCSA board contains also basic analogue signal generation capabilities that can be used externally or fed back to the board input to be analyzed. A second software digital signal generator can be used to run the software without using the hardware module.

The NCSA board is based on the Microchip dsPIC33EP512MC502 microcontroller, connected to the WIZnet W5500 Ethernet controller IC and the Analog Devices ADP151 low dropout low noise voltage regulator. The dsPIC33 microcontroller is suited to applications requiring high-speed ADC (it features an embedded 10/12 bit ADC), it has a built-in amplifier and has a large RAM (50 kB) enabling to memorize large data blocks on-chip. The high-speed PWM is used for signal and noise generation. The W5500 chip is used to manage the communication to and from the PC through Ethernet, and is controlled by the dsPIC33 via SPI port. To isolate the analogue ADC supply and to minimize the digital noise crossover, two separate 3.3 V DC regulators are used. The ADP151 voltage regulator has a very low noise factor (9 μV_{RMS}) and a very low dropout voltage (140 mV), respectively ensuring a pure ADC supply and avoiding the need for an heatsink. The supply voltage of 5 V comes through a standard mini USB connector. The user selects how to route the input and output signals by means of jumpers (it is possible to generate and receive distinct signals or to loop back the generated signal towards the input). Also four preset IP addresses configured for the Ethernet connection can be selected by the user by acting on a jumper.

The firmware included in the dsPIC33 was written in C, while the GUI application software is written in C#.

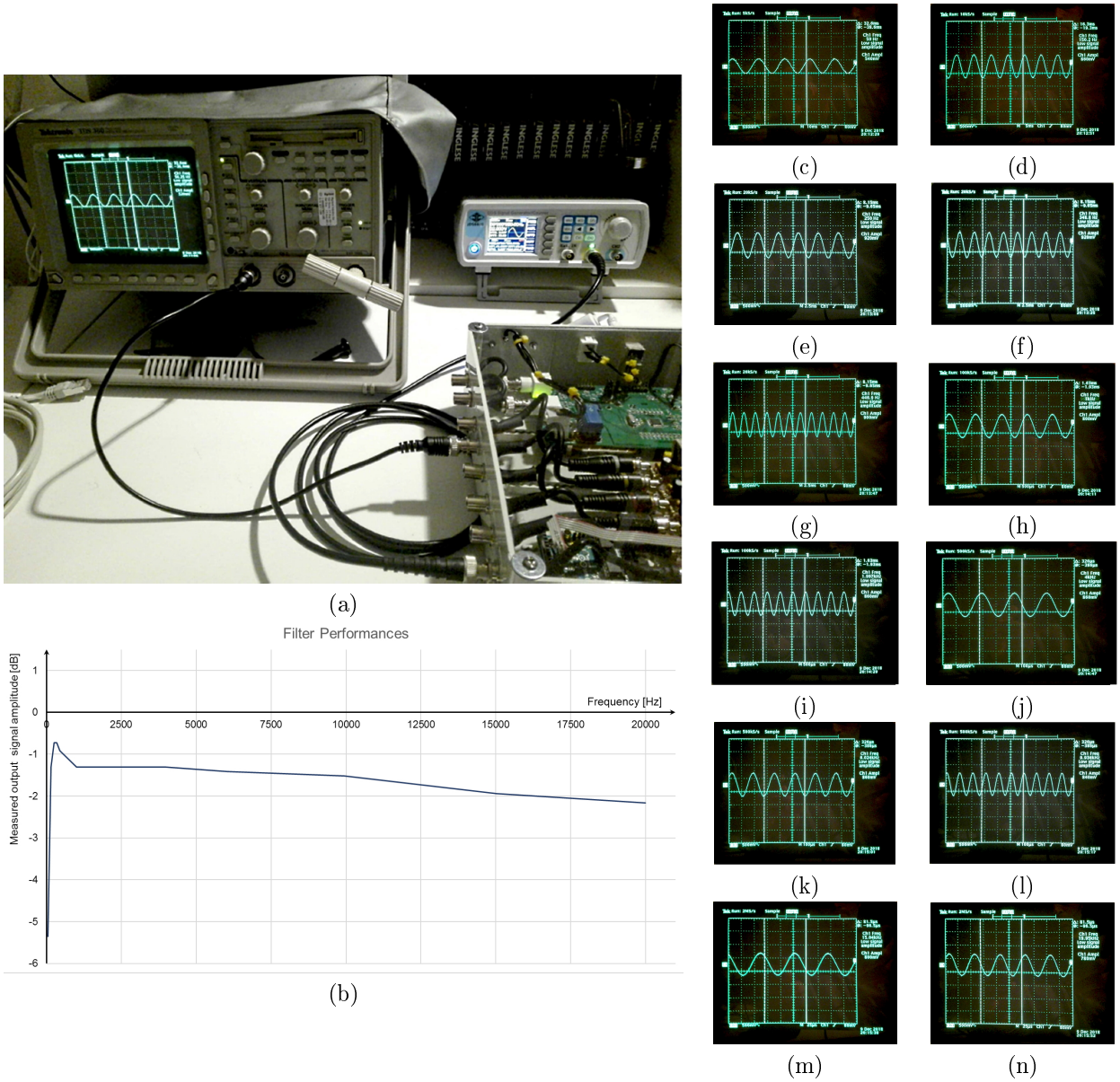


Figure 18.11: Experimental measures of the compensation filter. The DSP was fed by an oscillating signal having constant peak-to-peak amplitude of 2 V and received by a magnetic antenna having the same frequency response as in Table **??**. (a) Measurement set-up; (b) Plot of the performance of the compensation filter: received signal amplitude (in dB) versus frequency (in Hz); (c-n) Oscilloscope screen snapshots of some measured waveforms at different frequencies.
(c) 50 Hz, 540 mV, -5,35 dB; (d) 150.2 Hz, 860 mV, -1,31 dB; (e) 250 Hz, 920 mV, -0,72 dB; (f) 348.8 Hz 920 mV, -0,72 dB; (g) 446.9 Hz, 900 mV, -0,91 dB; (h) 1.000 kHz, 860 mV, -1,31 dB; (i) 1.997 kHz, 860 mV, -1,31 dB; (j) 4.000 kHz, 860 mV, -1,31 dB; (k) 6.024 KHz, 850 mV, -1,41 dB; (l) 9.934 kHz, 840 mV, -1,51 dB; (m) 15.040 kHz, 800 mV, -1,94 dB; (n) 20.000 kHz, 780 mV, -2.16 dB.
The measures were performed with the TCCSA prototype by Giuseppe Casulli in Alstom (Bari).

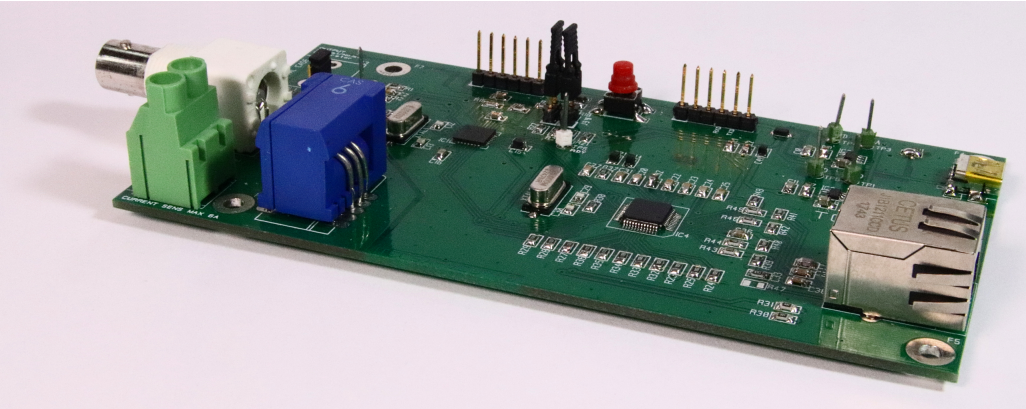


Figure 18.12: TCCSA acquisition board prototype accounting for the modifications of the original project made by Alstom.

18.3.2.3 Initial Project Defects and Modifications by Alstom

The original project presented various defects; for this reason, both the electronic board and the software were modified and enhanced, by adding some functionalities useful to the scopes of the TCCSA project.

- Despite the low noise components installed and the PCB arrangement used, the original hardware produced signals having a very low SNR, which required a complete redesign of mass plans and the use of different voltage regulators (Analog devices ADP150, having a greater output voltage stability respect to temperature variations).

- The board was unable to acquire current signals; therefore an on-board current sensor had to be implemented, and the firmware upgraded (by also adding an on-board FFT capability with meaningful frequency resolution limit). The current transducer implemented is the LEM CAS6-NP, a closed-loop compensated multi-range current transducer used for the current-to-voltage conversion. When the current sensing is used, the jumpers can be configured such that the output voltage of the sensor can be fed into the input of the board for digital conversion.

- The original GUI was not working, and so an external company had worked on the software and firmware to make the system operate correctly.

- In addition, the software was improved by adding some lacking fundamental features:

  - it was implemented the triggering capability to the oscilloscope;
  - it was improved the capability to measure some characteristic parameters of the input signal, like frequency, peak to peak value, or about the FFT.

The situation of the project when I stepped in was a stable hardware (shown in Fig. 18.12); this produced signals with a good (and stable) SNR and a working current acquisition chain. The software was working reliably and had triggering capabilities that in the original software were missing.

I had the task of improving the capabilities, by adding the required functionalities in order to create a finished product, able to evolve also in an instrument useful to detect the presence of faults in the TC environment.

18.3.2.4 Implementation of New Functionalities: Preliminary Work

The software, even after the modifications introduced by Alstom engineers, was lacking of some basic features, which needed to be implemented.

- The GUI after all the modifications and the additions was messy, so it was re-arranged and rationalized.

- The capability to store the input data stream and/or the statistics in .csv format, and the plots as image was added.

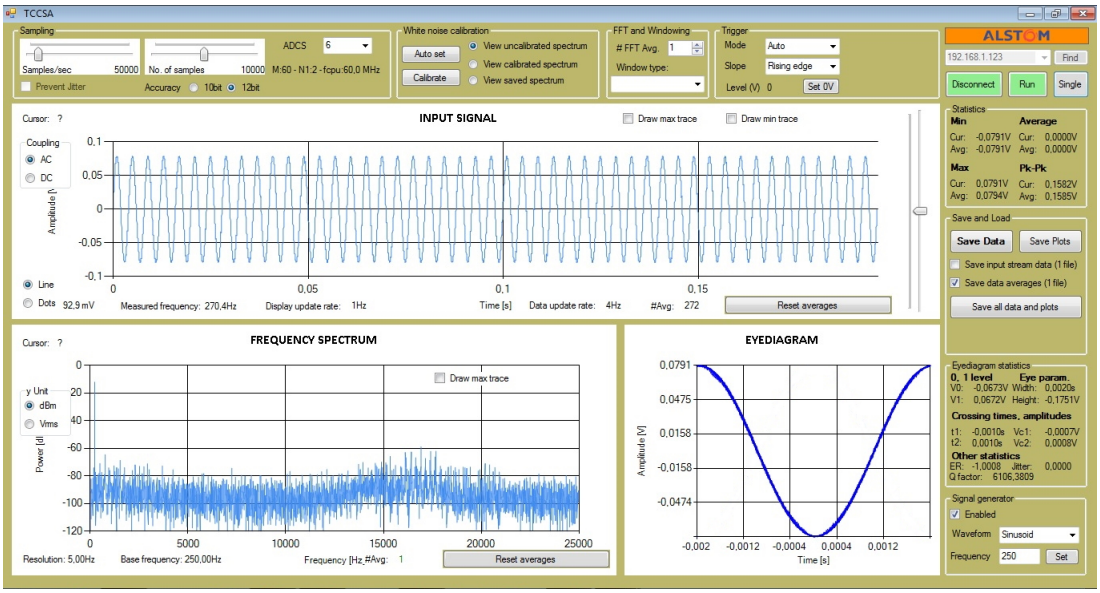


Figure 18.14: TCCSA oscilloscope and signal analyzer software.

18.3.2.5 Implementation of New Functionalities: Eye-Diagram Plotting

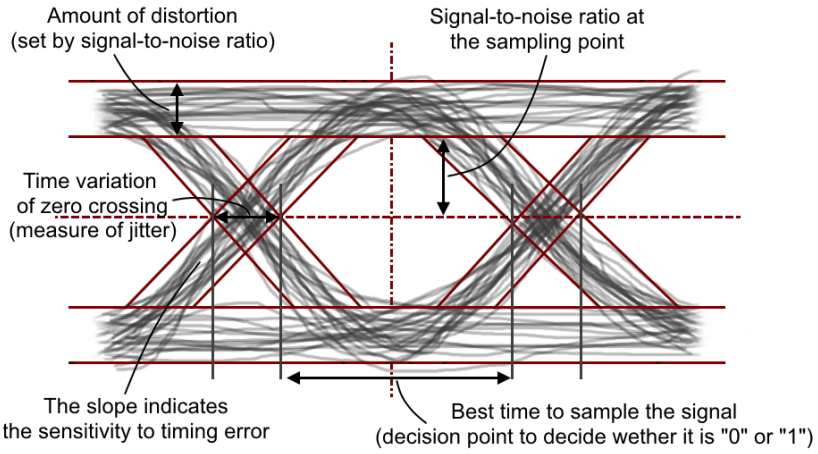


Figure 18.13: Eye-diagram indications about the influence of noise and intersymbolic interference.

The eye-diagram is a type of plot used to evaluate the quality of a digital signal by putting in evidence some of its key features, related to the modulation characteristics. It enables to evaluate the combined effects of channel noise and inter-symbol interference affecting a signal which travels in a transmission mean by analyzing the characteristics of the eye. Some of the key features of the signal that can be deduced from the analysis of the eye diagram are shown in Fig. 18.13. If no distortion due to noise or intersymbol interference is present, the eye is wide open, but when these sources of non-idealities, the eye begins to close.

The eye diagram is generated through an oscilloscope by overlapping all the possible one-zero combinations on the instrument display, and by leaving persistent each sweep. The triggering information, the data sampling frequency and the oscilloscope clock frequency must be taken into account and kept synchronized in order to correctly represent the eye [129]. A similar process is performed in the TCCSA application.

To represent the eye-diagram, the signal is subdivided into *chunks* on the basis of triggering information; each chunk can contain one or more periods of the signal, and this quantity is user-defined. Each point in the signal is defined by 2 coordinates: time and amplitude. The chunks forms the basis of the eye-diagram: in fact to obtain the plot, the chunks are overlapped one above the other. The amplitude information is kept unaltered, while at the end of each chunk the time information is reset to 0, thus enabling the samples to overlap; each overlapping chunk constitutes a *sweep* of the eye-diagram. In this case, for correct representation, the number of signal samples contained in an eye-diagram sweep is of paramount importance. The number of samples of each period is computed through the ratio between the sampling frequency and signal frequency, i.e. the fundamental frequency inferred from the FFT analysis.

In the TCCSA application, the eye diagram is plotted inside one of what in C# UI applications are known as chart areas. The library functions require that the coordinates of the points to be plotted in the chart must be formatted properly. The TCCSA software with the resulting eye-diagram plotting is represented in Fig. 18.14.

The number of samples per each sweep, `n`, is computed by multiplying the number of samples per period by the number of periods contained in each sweep. The parameter `n` is an integer; therefore, in order to have a correct representation of the eye-diagram without artifacts, it is required that the ratio between fundamental and sampling frequency is an integer. If this is not true, since the sampling frequency is user-defined and also the signal frequency can vary, the ratio could generate a fractional number, that must be therefore rounded to have an integer `n`. Hence, the practical (integer) number of samples per period, `n`, does not coincide anymore with the (fractional) theoretical one `n2`[3]. A mechanism has been implemented in order to compensate for this effect by skipping the sample in excess when necessary. Moreover, the initial samples before the first triggering point can alter the representation, and -in a first phase- must be skipped.

Another problem that can be encountered is that the triggering information can be imprecise: even if a test signal having constant frequency is used, the trigger point spacing is not constant. In this case, in fact, a trigger pulse is expected every `n` samples; however, in the list of trigger point indexes, sometimes, the trigger point is given at `n`/2 intervals, or after very few samples from the previous[4]. A spacing slightly different from `n` could be due to jitter, and this is an information that must be kept unaltered. Conversely, when the sampling is half of the theoretical value `n`, or is very little, this is an evident non-ideality in the triggering process, probably caused by noise in the signal, and produces results not utilizable for a correct representation of the signal in the eye-diagram. This effect is compensated for by means of finite state machine used to group the data contained in the original sequence in a rational manner, so that the spacing between every trigger point is close to the theoretical one, by keeping the jitter information. [5].

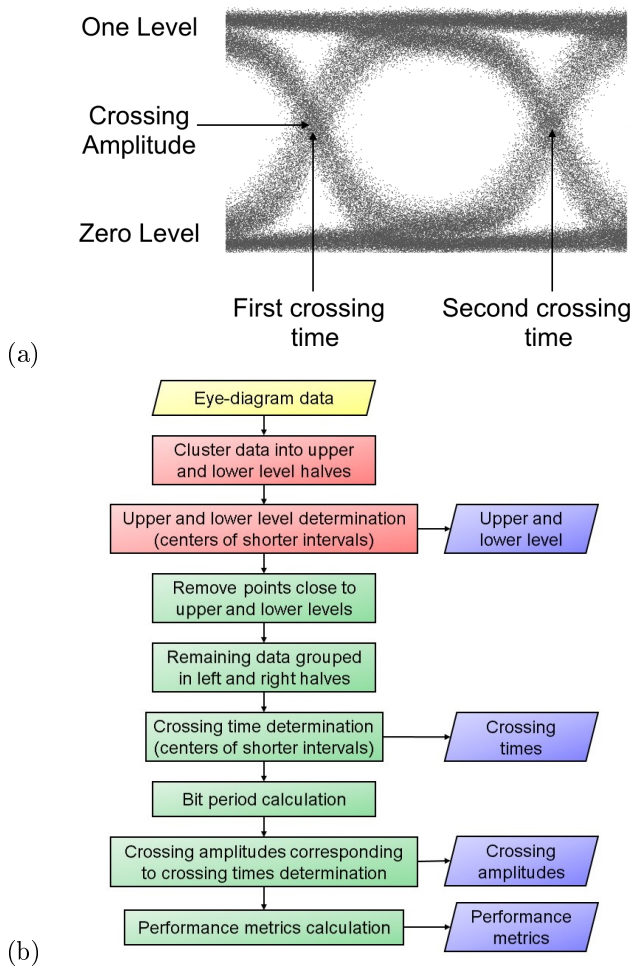


Figure 18.15: (a) Fundamental parameters of an eye-diagram. (b) Flowchart representing the method used to compute the eye-diagram metrics. Pictures taken from [129].

In the plot representation of the eye-diagram is always considered the theoretical integer value of samples per sweep `n`. The target is to keep the eye always in the center of the plot area, so

- if the number of samples per each sweep is greater than `n`, the difference `diff` is computed, and then the first `diff`/2 and the last `diff`/2 samples of the sweep sequence are discarded[6];
- if the number of samples per each sweep is lower than `n`, the difference `diff` is computed, and then the last `diff`/2 samples of the previous sequences are repeated, and the first `diff`/2 samples of the following sequences are anticipated (and then plotted again in the following sweep)[7].

In this way the jitter is taken into account in a correct way, and the eye diagram is always in the center of the plot area.

[3] For example, if the ratio between the fundamental frequency and the sampling frequency gives `n2` = 193.26, it is rounded as `n` = 193. If each sweep contains 4 periods of the signal, by using `n`, 772 samples are gathered instead of the theoretical number of 773.04: every sweep a sample that should have been included in the previous is included in the new one, and this builds up each and every sweep. The result is a drift of the represented curves that alter the eye-diagram representation.

[4] If the theoretical triggering point for a test signal having constant amplitude is every `n` = 500 samples, it could happen that the system gives triggering points spaced -for example- every 499, 2, 250, 251, 502, 499, 251, 248, 2, etc. samples.

[5] The original triggering sequence of Note 4 can be reconstructed as 500, 502, 502, 499, 500, etc.

[6] If `n` = 500 and the sweep contains for example 502 samples, the difference `diff` = 2, so the first and last samples of the sequence are discarded.

[7] If `n` = 500 and the sweep contains for example 498 samples, the difference `diff` = -2, so the last sample of the previous sequence is repeated and the first sample of the sequence is anticipated.

18.3.2.6 Implementation of New Functionalities: Eye-Diagram Metrics

The computation of eye-diagram parameters is carried on by following the methods illustrated by Jargon et.al. in [129], with some adaptations to cope with the characteristics of the available signals.

From the eye diagram plot, different parameters are extrapolated:

- *Fundamental parameters* of the eye-diagram (see Fig. 18.15(a)):
  - one level and zero level;
  - crossing times and amplitudes.
- *Extinction ratio*, the ratio between the average power used to transmit a logic "1" level and the average power used to transmit a logic "0" level [136].
- *RMS jitter*, the standard deviation of data in time, computed in a little time window surrounding the crossing amplitude [137, 138].
- *Performance metrics*, such as:
  - eye height and width;
  - Q factor, which measures the quality of the signal in terms of SNR.

The theoretical method to compute these parameters is outlined in Fig. 18.15(b) and illustrated in the following paragraphs.

Fundamental Parameters: One and Zero Level Determination The eye-diagram points are divided in 2 halves by using a *K-means algorithm* on the basis of the signal amplitude.

The K-means algorithm is a non-supervised learning algorithm used to divide a dataset into a finite number of clusters, groups of points similar among themselves. The k-means is an iterative algorithm, formed by three phases.

1. First of all, the available data are arbitrarily divided into a number of groups equal to the number of clusters (in this case, two), and then the mean point, or centroid, is computed.
2. The whole set of data points is now analyzed, and each data-point is assigned to the closest centroid (determined by using the Euclidean distance).
3. The clusters are now formed by a different set of points, so their centroid is computed again.

The last two steps are repeated over and over again until when a stop condition is reached. Possible stop conditions are:

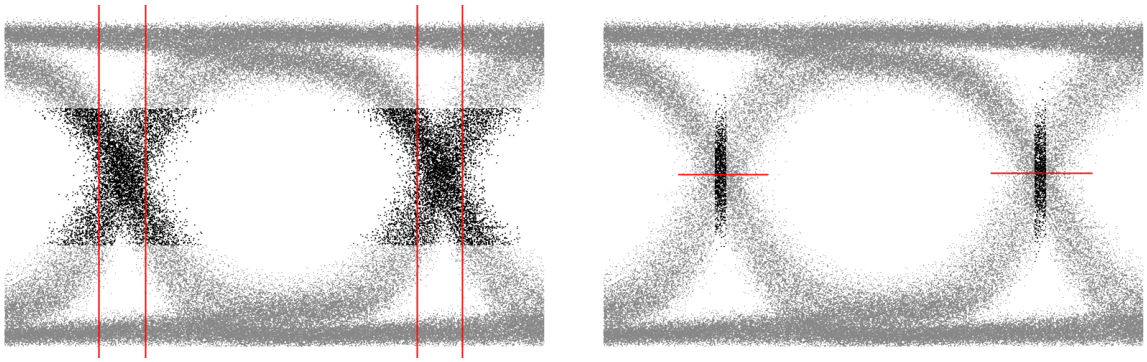
- No data point changes cluster anymore;
- the sum of the distances of points to their centroid is reduced to the minimum (convergence);
- a maximum number of iterations is reached.

The 2 halves are used to determine the upper and lower level of the eye-diagram.

The method of the *shortest interval* (i.e., of the interval which contains a given amount of the data of the given dataset, in this case, 50%) is used: the target is to find a mode in the data. The midpoint of the shortest interval is taken as a mode estimator and is used to estimate the one or zero level. In ref. [130] it has been shown that the estimated mode through the shortest interval is also a *least median of squares* (LMS) estimator for the data, which is a very robust estimator compared to the mean or the median, tolerating up to 50% of contaminated data.

The shortest interval is the interval that, in a ordered data sequence of length n, such that $y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n)}$, produces the smallest of the following differences

$$y_{(h)} - y_{(1)}, \quad y_{(h+1)} - y_{(2)}, \ldots, y_{(n)} - y_{(n-h+1)} \tag{18.1}$$



(a) (b)

Figure 18.17: (a) Crossing points determination. The set of points considered in the crossing times computation is evidenced in black. The red vertical lines delimit the shortest interval containing 50% of data in the left and right clusters. (b) Evaluation of crossing amplitudes; the data used to compute the crossing amplitude (evidenced by the red horizontal lines) are represented in Fig.. Pictures taken from ref. [129].

in which $h = \lfloor \frac{n}{2} \rfloor + 1$. The midpoint of this shortest interval, i.e. the mean of the values delimiting the shortest interval, is the LMS estimate for the data sequence.

By applying the shortest interval midpoint calculation to both the upper and lower halves two of the fundamental parameters, namely the one and zero level, (denoted $\overline{v}_1$ and $\overline{v}_0$, respectively) are evaluated A graphical representation is given in Fig. 18.16.

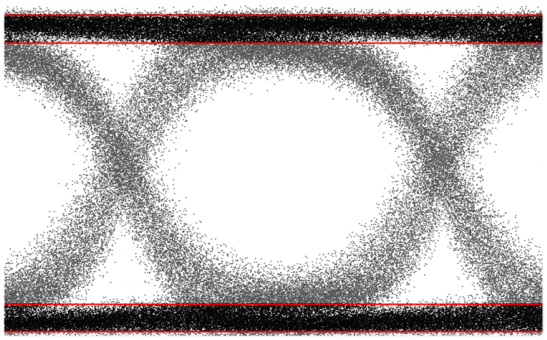


Figure 18.16: One and zero level determination. With the red lines are delimited the shortest intervals that contain the 50% of the upper and lower halves of the original dataset. Picture taken from [129].

Fundamental Parameters: Crossing Times Determination Once the upper and lower levels have been determined, we can focus on the crossing times. To do so, of the entire dataset, it is useful to keep only the points belonging to the central fascia of the eye-diagram in which the eye crossings are located, as illustrated in Fig. 18.17(a). This is achieved by creating a new dataset in which the points close to the upper and lower levels are excluded. For this reason, two level thresholds, denoted $v_{t,low}$ and $v_{t,high}$ are computed on the basis of the one and zero level

$$v_{t,low} = \overline{v}_0 + p\,(\overline{v}_1 - \overline{v}_0)\,, \qquad v_{t,high} = \overline{v}_1 - p\,(\overline{v}_1 - \overline{v}_0) \tag{18.2}$$

here, p is a percentage value (in ref. [129] the suggested value is 25%, but this can be tuned). In generating the new subset of data, the points whose amplitude is lower than $v_{t,low}$ or higher than $v_{t,high}$ are neglected.

The new subset of data is split in two clusters by using again the k-means algorithm. The target is to identify the clusters containing the two crossing points, so the clustering is now based on the horizontal axis. The first and second crossing points, $\overline{t}_1$ and $\overline{t}_2$, are found through the shortest interval algorithm, set to operate on the x axis.

Fundamental Parameters: Crossing Amplitudes Determination The crossing amplitudes $\overline{v}_{c1}$ and $\overline{v}_{c2}$ are determined by considering only the samples in each of the left and right cluster that are close to the crossing times $\overline{t}_1$ and $\overline{t}_2$, as evidenced in Fig. 18.17(b). The bit period T_B is computed as:

$$T_B = \overline{t}_2 - \overline{t}_1. \tag{18.3}$$

Then, only the data whose x coordinate falls between $\overline{t}_1 \pm q\,T_B$ for the left cluster and $\overline{t}_2 \pm q\,T_B$ for the right cluster are considered; here, q is a percentage value (in ref. [129] the suggested value is 1%).

Once the points are computed, the LMS mode of the y data coordinate is computed by means of the shorter interval method for each cluster. The results are the crossing amplitudes $\overline{v}_{c1}$ and $\overline{v}_{c2}$.

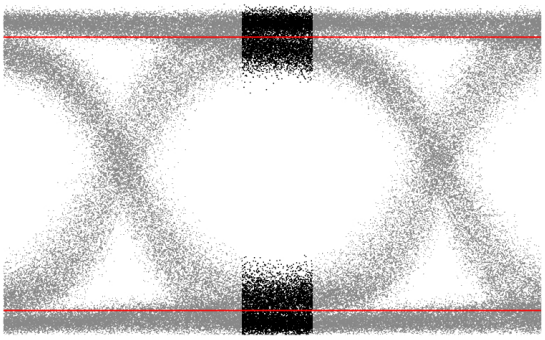


Figure 18.18: Evaluation of the extinction ratio. The points falling around the central point are evidenced in black (the band width corresponds to 20% of the bit period). The red lines correspond to the high and low level at the central point. Picture taken from ref. [129].

Metrics: Extinction Ratio (ER) Determination The extinction ratio is computed by considering the ratio between the high and low amplitude levels, $\overline{v}_{high}$ and $\overline{v}_{low}$, around the central point of the eye-diagram t_c, in a span that, according to various international standards [131, 132, 133, 134, 135], must be large as 20% of the bit period (see Fig. 18.18).

The central point of the eye diagram is computed as the mean value between the crossing points, $\overline{t}_1$ and $\overline{t}_2$:

$$t_c = \frac{\overline{t}_1 + \overline{t}_2}{2}. \tag{18.4}$$

The mean value between one and zero level to be used as horizontal separator is also computed as

$$v_c = \frac{\overline{v}_0 + \overline{v}_1}{2}. \tag{18.5}$$

To delimit the 20% time span around the central point, the bounds are placed at $t_c \pm 0.1\,T_B$. The horizontal separator is used to divide the data into two sets of points, those having an amplitude greater than v_c and those having an amplitude lower than v_c. In the implementation, to save computation effort, are chosen the points falling between the horizontal bounds around $t_c$in the higher and lower clusters already selected.

For each of the achieved clusters, the mean and standard deviation are computed: the results are the low and high amplitude levels $\overline{v}_{low}$ and $\overline{v}_{high}$ in correspondence of the central points and the relative standard deviations σ_{low} and σ_{high}.

The extinction ratio ER is then computed as:

$$ER = \frac{\overline{v}_{high}}{\overline{v}_{low}}. \tag{18.6}$$

Metrics: Evaluation of RMS Jitter To compute the *root-mean square* (RMS) jitter σ_t ,the crossing amplitudes, $\overline{v}_{c1}$ and $\overline{v}_{c2}$, the one and zero level, $\overline{v}_1$ and $\overline{v}_0$, and the clusters of data (left and right) used to find the crossing times are considered.

The points of each cluster of data are selected: only the points whose amplitude falls between $\overline{v}_{c1} \pm r\ (\overline{v}_1 - \overline{v}_0)$ for the left cluster or $\overline{v}_{c2} \pm r\ (\overline{v}_1 - \overline{v}_0)$ for the right cluster are considered; here, r is a percentage value equal to 0.1%. The selected points are shown in Fig. 18.19. Then, the standard deviations of each of the new sets of data, σ_{t_1} and σ_{t_2}, are evalauted; these are the RMS jitter in correspondence of the first and second crossing point, respectively.

The overall RMS jitter σ_t is given by the average of the RMS jitters at each crossing point, i.e.

$$\sigma_t = \frac{\sigma_{t_1} + \sigma_{t_2}}{2}. \tag{18.7}$$

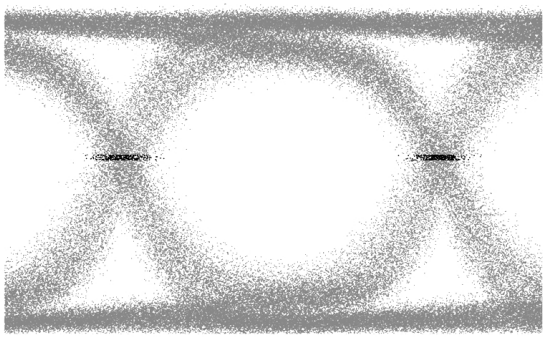


Figure 18.19: RMS jitter determination. In black are evidenced the data used to compute the left and right RMS jitters and then the overall RMS jitter. Picture taken from [129].

The use of both crossing point RMS jitters instead of only one is explained in ref. [129]. It is taken into account the possibility that not all combinations are present on both sides of the eye: so, unlike in the ideal case, the two crossing areas are not one the duplicate of the other, and neither the resulting jitters are equal. The suggested overall RMS jitter provides a more robust estimation.

Performance Metrics: Evaluation of Eye Height and Width The eye height and eye width are defined in ref. [137, 138] by using respectively the high and low level around the center point and the crossing times, along with their standard deviations:

$$\text{eyeheight} = (\overline{v}_{high} - 3\,\sigma_{high}) - (\overline{v}_{low} - 3\,\sigma_{low}) \tag{18.8}$$

$$\text{eyewidth} = \left(\overline{t}_2 - 3\,\sigma_{t_2}\right) - \left(\overline{t}_1 - 3\,\sigma_{t_1}\right). \tag{18.9}$$

Performance Metrics:Evaluation of Q Factor The signal quality in terms of signal-to-noise ratio is given by the Q factor, which is again given by the amplitude levels around the center point of the eye-diagram and the corresponding standard deviations:

$$Q\,\text{factor} = \frac{\overline{v}_{high} - \overline{v}_{low}}{\sigma_{high} + \sigma_{low}}. \tag{18.10}$$

18.3.2.7 Implementation of New Functionalities: Future Plans Outline

Workflow - Advanced and Normal User As already mentioned above, the TCCSA software suite is composed by two components:

- TCCSA Signal Analyzer;
- TCCSA Signal Generator.

Alstom envisions two types of users:

- a *normal* user, which uses the TCCSA for maintenance operations;
- an *advanced* user, which uses the TCCSA for commissioning operations.

The two types of users have different levels of skills, knowledge of the system, and different needs; in particular, during commissioning, operations that should not be allowed during normal maintenance are permitted, or it is not required to the maintainer to know the internal operation of the software. Therefore, depending on the type of user, it should be possible to made available or not some capabilities of the software, after an authentication procedure.

The *signal analyzer in normal mode* should be able to visualize and manage:

- Device connection settings options;
- Simplified sampling frequency settings[8];
- Simplified sampling length settings[9];
- Data in (voltage and current), spectrum and eye-diagram plots, and data mode of visualization (lines or points);
- Signal measurements;
- Trigger options;
- DC or AC coupling settings;
- Calibration settings;
- Simplified analog-to-digital converter and clock settings[10];
- Simplified number of averages and windowing settings[11];
- Simple signal generator capability[12].

The *signal analyzer in advanced mode* should be able to visualize and manage:

- All the capabilities of the normal mode;
- Settings that are not enabled during the normal user operation (sampling frequency, sampling length, analog-to-digital converter, clock and windowing, etc.);
- Individual time-domain, FFT or eye plots;
- A window in which are plotted the demodulated (ASK and PSK) signals (functionality to be implemented in the software);

[8] By allowing the user to select among a restricted set of sampling frequency values.
[9] By allowing the user to select among a restricted set of sampling length values, compatible with other settings.
[10] By allowing the user to select only among "low accuracy" and "high accuracy", the parameters being too strictly related to the internal operation of the hardware and firmware.
[11] By allowing the user to select only among a restricted set of values.
[12] Capable to generate only predefined and unchangeable standard test signals.

- Full signal generator capabilities.

As stated above, the signal generator in normal mode should allow only simple capabilities, giving more freedom only to the advanced user.

A possible way to differentiate between normal and advanced mode of operation is to use a Ribbon-type menu bar. This type of menu gives to the user in a single screen a vast and well-organized choice of icons, controls and fields to show results. By enabling or disabling some tabs on the basis of the type of user that authenticated, the distinction between the different capabilities is readily made, without the need to design (and maintain) different forms for the different types of users.

The structure could include:

- a *home* tab, in which all the principal settings of each feature, and the file management are contained;
  - this should be the only tab available in normal mode, containing the simplified controls;
  - in advanced mode the tab should include also some sub-menus, to contain the full controls of each feature. This will enable to keep a simple and neat interface for the routine operations, but allowing also to finely tune the settings if needed.
- some *advanced* tabs for advanced mode capabilities, like optional graphs and controls.
  - In this way, when an additional feature is added to the software, it is simply needed to add a tab in the UI, and it is no more necessary to modify existing tabs.

Machine Learning - Predictive Maintenance The TCCSA is thought also to introduce a predictive maintenance capability. To do so, some signal parameters must be selected to feed the machine learning algorithm. These parameters should concisely represent the signal under analysis, allowing to establish the presence of a defect in the track circuit apparatus, and possibly to predict the type of defect.

A possible features vector could comprise:

- The signal characteristics:
  - Average current and voltage values;
  - Peak-to-peak current and voltage values;
  - Principal signal spectral components, by indicating frequency and amplitude of each one;
  - Eye-diagram metrics: extinction ratio, RMS jitter, eye width, eye height, Q factor;
- Environmental conditions:
  - Temperature;
  - Humidity;
  - Type of environment.

In any case, if a supervised learning algorithm is adopted, it is of paramount importance to build a training dataset which is representative of the real situation, comprising all the possible operating and environmental conditions. To do so it is also thinkable to use track circuit models.

18.4 Conclusions of the Activity

The goal of implementing new features in the TCCSA device was successful. Basic functionalities missing in the original implementation have been added; moreover, other important functionalities have been added to determine the operating parameters of a track circuit, both for installation and maintenance purposes, and to use the system as a device for data acquisition with a view to implementing a continuous diagnosis of track circuits.

Some possible developments have been planned, both regarding the interface towards a commissioner/maintainer user (which has the role of installing or maintaining track circuits) and the interface towards a machine learning software for data analysis.

Part IV

Indices and Bibliography

List of Figures

List of Tables

Bibliography

[1] European Union Agency for Railways, ERA. Report on Railway Safety and Interoperability in the EU 2020. Luxembourg: Publications Office of the European Union, 2020. doi:10.2821/30980 1.1

[2] Marco Gallini. Dalla manutenzione ciclica alla manutenzione predittiva, L'ingegnerizzazione dei processi di manutenzione dell'infrastruttura, la diagnostica mobile ed i relativi progetti innovativi in RFI. Webinar CIFI 23 November 2020. http://www.cifi.it/UplDocumenti/WebinarCIFI23112020-Dalla%20manutenzione%20ciclica%20alla%20predittiva.pdf 1.1

[3] Ferreira, L. and Murray, M. Modelling rail track deterioration and maintenance: current practices and future needs. Transp. Rev., 1997, 17(3), 207–221. 1.2.1

[4] European Union Agency for Railways, ERA. Railway Safety in the European Union. Safety overview 2017. https://www.era.europa.eu/sites/default/files/library/docs/safety_interoperability_progress_reports/railway_safety 1.1, 1.2.1

[5] G. Zumpano, M. Meo, A new damage detection technique based on wave propagation for rails, International Journal of Solids and Structures, Volume 43, Issue 5, 2006, Pages 1023-1046, ISSN 0020-7683, https://doi.org/10.1016/j.ijsolstr.2005.05.006. 1.2.1, 2.2.1, 2.2.2, 2.2.3, 2.2.3, 2.2.3, 1, 2, 3

[6] Ghofrani F, Pathak A, Mohammadi R, Aref A, He Q. Predicting rail defect frequency: An integrated approach using fatigue modeling and data analytics. Comput Aided Civ Inf. 2019;1–15. https://doi.org/10.1111/mice.12453. 2, 2.2, 2.2.6

[7] Sadeghi, J. M., & Askarinejad, H. (2011). Development of track condition assessment model based on visual inspection. Structure and Infrastructure Engineering, 7(12), 895–905. 2.2

[8] Ph Papaelias, M, C Roberts, and C L Davis. "A Review on Non-Destructive Evaluation of Rails: State-of-the-Art and Future Development." Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit 222, no. 4 (July 2008): 367–84. doi:10.1243/09544097JRRT209. 1.2.1, 2, 2.2, 2.2.1, 2.2.3, 2.2.6, 6, 3.2.1, 3.2.3, 1, 3.2.5, 3.2.6, 3.3.1, 1

[9] Muravev, V.V., Boyarkin, E.V., 2003. Nondestructive testing of the structural-mechanical state of currently produced rails on the basis of the ultrasonic wave velocity. Russian Journal of Nondestructive Testing 39 (3), 24–33. 2.2.1

[10] Abbaszadeh, K., Rahimian, M., Toliyat, H.A., Olson, L.E., 2003. Rails defect diagnosis using wavelet packet decomposition. IEEE Transactions on Industry Applications 39 (5). 2.2.1

[11] Jeffrey, B.D., Peterson, M.L., 1999, Assessment of rail flaw inspection data. Mountain–Plains Consortium, Report No. MPC-99-106. 2.2.1, 2

[12] Vadillo, E.G., Tarrago, J.A., Zubiaurre, G.G., Duque, C.A., 1998. Effect of sleeper distance on rail corrugation. Wear 217, 140–146. 2.2.3

[13] Bohmer, A., Klimpel, T., 2002. Plastic deformation of corrugated rails—a numerical approach using material data of rail steel. Wear 253, 150–161. 2.2.3

[14] Cannon, D. F., Edel, K. O., Grassie, S. L., & Sawley, K. (2003). Rail defects: An overview. Fatigue and Fracture of EngineeringMaterials and Structures, 26(10), 865–886. 1.2.1, 2.1.2, 2.2.1, 2.2.3, 2.2.3, 2.2.4, 2.2.5, 2.2.6, 1, 2

[15] Cannon, D.F., Pradier, H., 1996. Rail rolling contact fatigue research by the European Rail Research Institute. Wear 191, 1–13. 2.2.3

[16] Grassie, S., Nilsson, P., Bjurstrom, K., Frick, A., Hansson, L.G., 2002. Alleviation of rolling contact fatigue on Swedens heavy haul railway. Wear 253, 42–53. 2.2.3

[17] Grassie S. L., Railhead Damage, Stuart Grassie Engineering Ltd, March 2018, http://railmeasurement.com/railhead-damage/intro/. 2.2.3, 2.2.3

[18] L. Mayer, 1970. Impianti ferroviari: tecnica ed esercizio. Collegio Ingegneri Ferroviari Italiani. 2.1.1, 2.1.2, 1, 2.1.2

[19] M. Guerrieri, 2017. Infrastrutture ferroviarie, metropolitane, tranviarie e per ferrovie spaciali: elementi di pianificazione e di progettazione. Maggioli editore. 2.2.4

[20] Rail Defects Handbook: Some Rail Defects, their Characteristics, Causes and Control RC2400, Issue A, Revision 0, March 2006, Australian Rail Track Corporation Ltd. 1.2.1, 2.1.2, 2.1, 2.1.3, 2.1.4, 2.2.1, 2.2.3, 2.2.3, IV

[21] Rose, J.L.. (2014). Ultrasonic Guided Waves in Solid Media. Ultrasonic Guided Waves in Solid Media. 1-512. 10.1017/CBO9781107273610. 3.1, 4.15, 5, 5.1, 5.1.1, 5.2, 5.2, 5.3.2.1, 5.5, 6.1, IV, IV

[22] Cawley, Peter & Lowe, M.J.S. & Alleyne, D.N. & Pavlakovic, B & Wilcox, Paul. (2003). Practical long range guided wave testing: Applications to pipes and rail. Materials Evaluation. 61. 66-74. 3.3.3, 3.3.3, 7, 7, 7, 7.1, 7.1

[23] P. Rizzo, 16 - Sensing solutions for assessing and monitoring railroad tracks, Editor(s): M.L. Wang, J.P. Lynch, H. Sohn, In Woodhead Publishing Series in Electronic and Optical Materials, Sensor Technologies for Civil Infrastructures, Woodhead Publishing, Volume 56, 2014, Pages 497-524, ISBN 9781782422426, https://doi.org/10.1533/9781782422433.2.497. 1.2.1, 2.2.4, 3.2, 3.2.1, 3.2.3, 3.2.4, 3.2.5, 3.3, 3.3.1, 3.3.2

[24] Clark, R., 2004. Rail flaw detection: overview and need for future developments. NDT&E International 37, 111–118. 3.2.1, 3.3.1

[25] Marais, J.J., Mistry, K.C., 2003. Rail integrity management by means of ultrasonic testing. Fatigue and Fracture of Engineering Materials and Structures 26, 931–938. 3.3.1

[26] Coccia, S., Bartoli, I., Salamone, S., Phillips, R., di Scalea, F. L., Fateh, M., & Carr, G. (2009). Noncontact Ultrasonic Guided Wave Detection of Rail Defects. Transportation Research Record, 2117(1), 77–84. https://doi.org/10.3141/2117-10. 3.3.3

[27] Ryue, Jungsoo & Thompson, David & White, Paul & R. Thompson, D. (2008). Wave Propagation in Railway Tracks at High Frequencies. 10.1007/978-3-540-74893-9_62. 3.3.3

[28] Olympus NDT (2011), Advances in Phased Array Ultrasonic Technology Applications, http://www.olympus-ims.com/data/File/advances_book/Applications_Cha1. en.pdf. 3.3.2

[29] Shull P. J., Nondestructive Evaluation: Theory, Techniques, and Applications, (2002) New York, Marcel Dekker. 3.2.1

[30] Bray D. E. and Stanley R. K. (1997), Nondestructive Evaluation. A Tool in Design, Manufacturing, and Service , Boca Raton, CRC press. 3.2.1

[31] Mix P (2005), Introduction to Nondestructive Testing, A Training Guide , 2nd Edn., Hoboken, New Jersey, John Wiley and Sons. 3.2.1

[32] Deutschl E, Gasser C, Niel A and Werschonig J (2004), 'Defect detection on rail surfaces by a vision based system', Intelligent Vehicles Symposium IEEE, 14–17 , 507–511. 3.2.3

[33] Singh M, Singh S, Jaiswal J and Hempshall J (2006), 'Autonomous rail track inspection using vision based system', IEEE Intl. Conf. on Computational Intelligence for Homeland Security and Personal Safety, 16–17, 56–59. 3.2.3

[34] Molina L F, Resendiz E, Edwards J R, Hart J M, Barkan C P L and Ahuja N (2011a), 'Condition monitoring of railway turnouts and other track components using machine vision', TRB 11–1442. 3.2.3

[35] Molina L F, Edwards J R and Barkan C P L (2011b) 'Emerging condition monitoring technologies for railway track components and special trackwork', 2011 ASME/ ASCE/IEEE Joint Rail Conference, JRC2011, Pueblo, Colorado, USA. 3.2.3

[36] Oukhellou L, Côme E, Bouillaut L and Aknin P (2008), 'Combined use of sensor data and structural knowledge processed by Bayesian network: application to a railway diagnosis aid scheme', Transportation Research Part C , 16 , 755–767. 3.2.3

[37] Magnus, D. L. Non-contact technology for track speed rail measurement (ORIAN). SPIE Proc., 1995, 2458, 45– 51. 3.2.3

[38] Backinsky, G. S. The electronic BAR gauge (a customised optical rail profile measurement system for rail grinding application. SPIE Proc., 1995, 2458, 52–63. 3.2.3

[39] Sasama, H., Ukai, M., and Okimura, Y. The development of rail-gap inspection system. Q. Rep. RTRI, 1991, 32(1), 21–29. 3.2.3

[40] Marino, F. et al. A real time visual inspection system for railway maintenance: automatic rail detection and tracking. InternalReportDEE–Politecnico di Bari, 2005. 3.2.3

[41] The MER MEC track surface defects detection system, available from www.mermec.com. 3.2.3

[42] Stella, E., Mazzeo, P. L., Nitti, M., Cicirelli, G., Distante, A., and D'Orazio,T.Visual recognition of missing fastening elements for railroad maintenance. In Proceedings of the IEEE-ITSC International Conference on IntelligentTransportation systems, Singapore, 2002, pp. 94–99. 3.2.3

[43] Mazzeo,P. L.,Nitti, M.,Stella,E.,and Distante, A.Visual recognition of fastening bolts for railroad maintenance. Pattern Recognit. Lett., 2004, 25(6), 669–677. 3.2.3

[44] Cawser, S. J., Hardy, A. E. J., and Wright, C. E. Acoustic track monitoring. Phase 1: Initial data gathering analysis, AEA Report for RSSB, December 2002. 3.2.6

[45] Weise,V. andWaite,G.Developments of railflawinspection techniques within the UK rail industry. In the Proceedings of the 7th International Conference on Maintenance and Renewal of Permanent Way; Power and Signalling; Structures and Earthworks, London, UK, 6–7 July 2004. 3.2.6

[46] Clark, R. Rail flaw detection: overview and needs for future developments, NDT&E Int., 2004, 37, 111–118. 3.3.1

[47] Krautkramer, J. and Kraukramer, H. Ultrasonic testing of materials, 4th ed., 1990 (Springer, Berlin, Germany). 3.3.1

[48] Vandone A, Rizzo P and Vanali M (2012), 'Two-stage algorithm for the analysis of infrared images', Research in Nondestructive Evaluation , 23 (2), 69–88. 3.2.4

[49] Wirtu, L. Bayissa, M. Dhanasekar (2011). High speed detection of broken rails, rail cracks and surface faults. Advanced project management training needs, CRC for Rail Innovation, Brisbane (Australia). 3.2.7

[50] Ihara, Ikuo. (2008). Ultrasonic Sensing: Fundamentals and its Applications to Nondestructive Evaluation. 10.1007/978-3-540-69033-7_14. 5.1, 5.2

[51] L. Rose, Joseph. (2002). A Baseline and Vision of Ultrasonic Guided Wave Inspection Potential. Journal of Pressure Vessel Technology. 124. 273. 10.1115/1.1491272. 5.2

[52] Loveday, Philip & Long, Craig & Ramatlo, Dineo. (2017). Mode Repulsion of Ultrasonic Guided Waves in Rails. Ultrasonics. 84. 10.1016/j.ultras.2017.11.014. 7

[53] Ryue, Jungsoo & Thompson, David & White, Paul & R. Thompson, D. (2009). Decay rates of propagating waves in railway tracks at high frequencies. Journal of Sound and Vibration - J SOUND VIB. 320. 955-976. 10.1016/j.jsv.2008.09.025. 7

[54] Ryue, Jungsoo & Thompson, David & White, Paul & R. Thompson, D. (2008). Investigations of propagating wave types in railway tracks at high frequencies. Journal of Sound and Vibration - J SOUND VIB. 315. 157-175. 10.1016/j.jsv.2008.01.054. 7, 7.1, 7.2, 7, IV

[55] Long, Craig & Loveday, Philip. (2014). Prediction of Guided Wave Scattering by Defects in Rails Using Numerical Modelling. AIP Conference Proceedings. 1581. 10.1063/1.4864826. 3.4, 7.2, 7.4, 7.2, IV

[56] Wilcox, Paul & Pavlakovic, Brian & Evans, Mark & Vine, Keith & Cawley, Peter & Lowe, Michael & Alleyne, David. (2003). Long Range Inspection of Rail Using Guided Waves. 657. 10.1063/1.1570142. 3.3.3, 7.1, 7.2

[57] Fraden, Jacob. (2010). Handbook of Modern Sensors. 10.1007/978-1-4419-6466-3. 8.1, 8, 8.2, IV

[58] Meissner, A. (1927). Über piezoelectrische Krystalle bei Hochfrequenz. Zeitschrift für Technische Physik, 8(74). 8

[59] Alan V. Oppenheim and Ronald W. Schafer. 1999. Discrete-Time Signal Processing (2nd. ed.). Prentice Hall Press, USA. 9, 9.1, 9.2, 9.2.1, 9.3, 9.4.1, 9.4.2, 9.4.2, 9.4.2

[60] Smith, Julius O. Mathematics of the Discrete Fourier Transform (DFT) with Audio Applications, Second Edition, W3K Publishing, http://books.w3k.org/, 2007, ISBN 978-0-9745607-4-8. 9, 9.4.1, 9.4.2, 9.4.2, 12.2.4.1

[61] Ruye, Wang. Two-Dimensional Fourier Transform in Fourier Analysis and Image Processing lectures. 2007-11-15 http://fourier.eng.hmc.edu/e101/lectures/Image_Processing/node6.html visited on 2021-02-13. 9.5

[62] Mariani, S., Nguyen, T., Phillips, R. R., Kijanka, P., Lanza di Scalea, F., Staszewski, W. J., ... Carr, G. (2013). Noncontact ultrasonic guided wave inspection of rails. Structural Health Monitoring, 12(5–6), 539–548. https://doi.org/10.1177/1475921713498533. 3.3.3, 1, 10.2.1, 10.3.1.1, 10.3.1.3, 10.7.1

[63] Mazzoldi P., Nigro M., Voci C. Fisica Vol. 1 - Meccanica-Termodinamica, chapter 8. EdiSES. 4.1, 4.2

[64] Mazzoldi P., Nigro M., Voci C (1991). Fisica Vol. 2 - Elettromagnetismo - Onde, 2nd edition, chapter 12. EdiSES. 4.2

[65] Landau L. D., Lifshitz E. M., "Fisica Teorica VII - Teoria dell'elasticità", Editori Riuniti university press (2010), chapter III. 4.2.4, 4.2.12.2

[66] Calvini P., "Elementi di acustica", INFN Genova lectures, March 2, 2015. https://www.ge.infn.it/~calvini/acustica.pdf 4.2.12.2

[67] Coccia, Stefano & Bartoli, Ivan & Salamone, Salvatore & Phillips, Robert & Scalea, Francesco & Fateh, Mahmood & Carr, Gary. (2009). Noncontact Ultrasonic Guided Wave Detection of Rail Defects. Transportation Research Record. 2117. 77-84. 10.3141/2117-10. 6.1, 6.2, 6.2

[68] Loveday, Philip. (2008). Simulation of piezoelectric excitation of guided waves using waveguide finite elements. IEEE transactions on ultrasonics, ferroelectrics, and frequency control. 55. 2038-45. 10.1109/TUFFC.895. 6.1, 6.2

[69] L. Gavrić, "Computation of propagative waves in free rail using a finite element technique," J. Sound Vibrat., vol. 185, no. 3, pp. 531–543, 1995. 6.2

[70] T. Hayashi, W.-J. Song, and J. L. Rose, "Guided wave dispersion curves for a bar with an arbitrary cross-section, a rod and rail example," Ultrasonics, vol. 41, pp. 175–183, 2003. 6.2

[71] I. Bartoli, A. Marzani, F. L. di Scalea, and E. Viola, "Modeling wave propagation in damped waveguides of arbitrary cross-section," J. Sound Vibrat., vol. 295, pp. 685–707, 2006. 6.4

[72] Graff, K. F. (1991). Wave Motion in Elastic Solids. Dover Publications. 5.3.2.1, 7

[73] Bocchini, Paolo & Marzani, Alessandro & Viola, Erasmo. (2010). Graphical User Interface for Guided Acoustic Waves. Journal of Computing in Civil Engineering - J COMPUT CIVIL ENG. 25. 10.1061/(ASCE)CP.1943-5487.0000081. 6.1, 6.2, IV

[74] Piervincenzo Rizzo, Marcello Cammarata, Ivan Bartoli, et al., "Ultrasonic Guided Waves-Based Monitoring of Rail Head: Laboratory and Field Tests," Advances in Civil Engineering, vol. 2010, Article ID 291293, 13 pages, 2010. https://doi.org/10.1155/2010/291293. 1, 10.3.1.1, 10.3.1.3

[75] Mariani, S., & di Scalea, F. L. e(2018). Predictions of defect detection performance of air-coupled ultrasonic rail inspection system. Structural Health Monitoring, 17(3), 684–705. https://doi.org/10.1177/1475921717715429 1, 10.2, 10.1, 10.3.1.2, 10.3.1.3, 10.3.1.3, IV, IV

[76] Mariani, Stefano & Nguyen, Thompson & Zhu, Xuan & Lanza di Scalea, Francesco. (2017). Field Test Performance of Noncontact Ultrasonic Rail Inspection System. Journal of Transportation Engineering, Part A: Systems. 143. 04017007. 10.1061/JTEPBS.0000026. 1, 10.3.1.1, 10.3.1.4, 10.7, 10.7.1, 10.7.1, 10.7.1, 10.2, IV

[77] Benzeroual, H.; Khamlichi, A.; Zakriti, A. Detection of Transverse Defects in Rails Using Noncontact Laser Ultrasound. Proceedings 2020, 42, 43. 10.2.1

[78] Teidj, S. Defect Indicators in a Rail Based on Ultrasound Generated by Laser Radiation. Procedia Manuf. 2020, 46, 863–870. 10.2.1

[79] Lanza Di Scalea, Francesco & Zhu, Xuan & Capriotti, Margherita & Liang, Albert & Mariani, Stefano & Sternini, Simone. (2018). Passive Extraction of Dynamic Transfer Function From Arbitrary Ambient Excitations: Application to High-Speed Rail Inspection From Wheel-Generated Waves. ASME Journal of Nondestructive Evaluation, Diagnostics and Prognostics of Engineering Systems. 1. 10.1115/1.4037517. 1, 10.2.1, 10.3.2.1, 10.4, 10.3.2.3, 10.3.2.4, 10.7.2, IV

[80] Liang A., Sternini S., Capriotti M., and Lanza di Scalea F. High-Speed Ultrasonic Rail Inspection by Passive Noncontact Technique. Materials Evaluation, Volume 77, Issue 7. 1 July 2019. https://ndtlibrary.asnt.org/2019/HighSpeedUltrasonicRailInspectionbyPassiveNoncontactTechnique. 1, 10.2.1, 10.3.2.1, 10.3, 10.3.2.2, 10.3.2.3, 10.7.2, 10.7.2, IV

[81] Albert Liang, Simone Sternini, Margherita Capriotti, Peter (Xuan) Zhu, Francesco Lanza di Scalea, Robert Wilson, "Passive extraction of Green's function of solids and application to high-speed rail inspection," Proc. SPIE 10970, Sensors and Smart Structures Technologies for Civil, Mechanical, and Aerospace Systems 2019, 109700R (27 March 2019). 10.3.2.1, 10.7.2, 10.7.2

[82] Ultrasonic Broken Rail Detector Overview, RailSonic Inc., Institute for Maritime Technology. 2, 10.2.2, 10.6, 10.4.1, 10.4.2, IV

[83] Philip W. Loveday, Francois A. Burger and Craig S. Long. (2018). Rail Track Monitoring in SA using Guided Wave Ultrasound. http://www.saint.org.za/wp-content/uploads/2018/03/Rail-Track-Monitoring-in-SA-using-Guided-Wave-Ultrasound.pdf 2, 10.2.2, 10.1, 10.4.3, 10.5.1, IV

[84] B. M. Steyn, J. F. W. Pretorius & F. Burger, "Development, testing and operation of an Ultrasonic Broken Rail Detector System," Computers in Railways IX, 2004, WIT Press, www.witpress.com, ISBN 1-85312-715-9. 2, 10.2.2, 10.5, 10.4.1, 10.4.2, IV

[85] Yuan, Lei & Yang, Yuan & Hernández, Álvaro & Shi, Lin. (2018). Feature Extraction for Track Section Status Classification Based on UGW Signals. Sensors. 18. 1225. 10.3390/s18041225. 2, 10.4.3

[86] Setshedi, Isaac & Long, Craig & Loveday, Philip & Wilke, Daniel. (2018). Adaptive SAFE model of a rail for parameter estimation. 2, 10.2.2, 10.5.5.1, 10.5.5.2

[87] Loveday, Philip & M. C. Taylor, Rebecca & Long, Craig & Ramatlo, Dineo. (2018). Monitoring the reflection from an artificial defect in rail track using guided wave ultrasound. AIP Conference Proceedings. 1949. 090003. 10.1063/1.5031566. 2, 2, 10.2.2, 10.5.1, 10.5.2.2, 10.5.2.3, 10.5.3.5, 10.5.4, 10.5.4.1

[88] Hyvarinen, A.; Oja, E. Independent component analysis: Algorithms and applications. Neural Netw. 2000, 13, 411–430. 10.5.4.1

[89] Loveday, Philip & Long, Craig. (2016). Influence of resonant transducer variations on long range guided wave monitoring of rail track. AIP Conference Proceedings. 1706. 150004. 10.1063/1.4940616. 2, 10.2.2, 10.5.2.2

[90] Setshedi, Isaac & Loveday, Philip & Long, Craig & Wilke, Daniel. (2019). Estimation of rail properties using semi-analytical finite element models and guided wave ultrasound measurements. Ultrasonics. 10.1016/j.ultras.2018.12.015. 2, 10.2.2, 10.10, 10.5.5.2, 10.5.5.2, IV

[91] Yuan, Lei & Yang, Yuan & Hernandez Alonso, lvaro & Li, Shuyu. (2018). Application of VMD Algorithm in UGW-based Rail Breakage Detection System. 1-6. 10.1109/ICVES.2018.8519587. 2, 10.4.3

[92] Loveday, P.W., Dhuness, K. and Long, C.S. 2018. Experimental development of electromagnetic acoustic transducers for measuring ultraguided waves. Eleventh South African Conference on Computational and Applied Mechanics (SACAM 2018), Vanderbijlpark, South Africa, 17-19 September 2018, pp. 693-702 2, 10.2.2

[93] Enekom RCFSS UsersManual V4. 2, 10.6.1

[94] Xing, Bo & Yu, Zujun & Xu, Xining & Zhu, Liqiang & Shi, Hongmei. (2019). Research on a Rail Defect Location Method Based on a Single Mode Extraction Algorithm. Applied Sciences. 9. 1107. 10.3390/app9061107. 2, 10.11, 10.6.2, 10.6.2.1, IV

[95] Liu C, Dobson J, Cawley P (2017). Efficient generation of receiver operating characteristics for the evaluation of damage detection in practical structural health monitoring applications. Proceedings Royal Society, 437: 1-26. 10.5.4, 10.7.4

[96] A. Burger, Francois & Loveday, Philip & Long, Craig. (2015). Large scale implementation of guided wave based broken rail monitoring. AIP Conference Proceedings. 1650. 771-776. 10.1063/1.4914679. 10.7.3

[97] Xiaoyuan Wei, Yuan Yang, Wenqing Yao, Lei Zhang, An automatic optimal excitation frequency tracking method based on digital tracking filters for sandwiched piezoelectric transducers used in broken rail detection, Measurement, Volume 135, 2019, Pages 294-305, ISSN 0263-2241, https://doi.org/10.1016/j.measurement.2018.11.033. 2, 10.6.3

[98] Xiaoyuan Wei, Yuan Yang, Wenqing Yao, Lei Zhang, Design of full bridge high voltage pulser for sandwiched piezoelectric ultrasonic transducers used in long rail detection, Applied Acoustics, Volume 149, 2019, Pages 15-24, ISSN 0003-682X, https://doi.org/10.1016/j.apacoust.2019.01.012. 2, 10.6.3

[99] X. Xining, Z. Lu, X. Bo, Y. Zujun and Z. Liqiang, "An Ultrasonic Guided Wave Mode Excitation Method in Rails," in IEEE Access, vol. 6, pp. 60414-60428, 2018. doi: 10.1109/ACCESS.2018.2875123 2, 10.6.2

[100] Shi, Hongmei & Zhuang, Lu & Xu, Xining & Yu, Zujun & Zhu, Liqiang. (2019). An Ultrasonic Guided Wave Mode Selection and Excitation Method in Rail Defect Detection. Applied Sciences. 9. 1170. 10.3390/app9061170. 2, 10.6.2

[101] Xining, & Bo, & Lu, & Hongmei, & Liqiang,. (2019). A Graphical Analysis Method of Guided Wave Modes in Rails. Applied Sciences. 9. 1529. 10.3390/app9081529. 2, 10.6.2

[102] Loveday, P. W., Long, C. S., & Ramatlo, D. A. (2019). Ultrasonic guided wave monitoring of an operational rail track. Structural Health Monitoring. https://doi.org/10.1177/1475921719893887 2, 10.2.2, 10.5, 10.5.2, 10.5.2.1, 10.5.2.2, 10.7, 10.5.3.4, 10.5.3.5, 10.5.4, 10.7.4, 10.7.4.2, 10.7.4.3, 13.4, IV

[103] Long CS and Loveday PW. Prediction of guided wave scattering by defects in rails using numerical modelling. AIP Conf Proc 2014; 1581: 240–247. 10.5.2.3

[104] Wilcox PD. A rapid signal processing technique to remove the effect of dispersion from guided wave signals. IEEE T Ultrason Ferr 2003; 50: 419–427. 10.5.3.3

[105] Moustakidis S, Kappatos V, Karlsson P, et al. An intelligent methodology for railways monitoring using ultrasonic guided waves. J Nondestruct Eval 2014; 33: 694–710. 10.5.3.4

[106] Liu C, Harley JB, Berge´s M, et al. Robust ultrasonic damage detection under complex environmental conditions using singular value decomposition. Ultrasonics 2015; 58: 75–86. 10.5.4, 10.5.4.1

[107] Dobson J and Cawley P. Independent component analysis for improved defect detection in guided wave monitoring. Proc IEEE 2016; 104: 1620–1631. 10.5.4

[108] Wei, X.; Yang, Y.; Ureña, J.; Yan, J.; Wang, H. An Adaptive Peak Detection Method for Inspection of Breakages in Long Rails by Using Barker Coded UGW. *IEEE Access* **2020**, *8*, 48529–48542. 10.6.3

[109] Yuan, L.; Yang, Y.; Hernández, Á.; Shi, L. Novel adaptive peak detection method for track circuits based on encoded transmissions. IEEE Sens. J. 2018, 18, 6224–6234. 10.6.3

[110] Wilcox PD. Guided-wave array methods. In: Boller C, Chang FK and Fujino Y (eds) Encyclopedia of structural health monitoring. Chichester: John Wiley & Sons, 2008, p. 1305. 10.5.3.2, 10.5.3.2

[111] Bocchini, Paolo & Marzani, Alessandro & Viola, Erasmo. (2010). Graphical User Interface for Guided Acoustic Waves. Journal of Computing in Civil Engineering, 25(3), DOI: 10.1061/(ASCE)CP.1943-5487.0000081. 1, 2

[112] RFI, Specifications for rails and bars for needles, document RFI TCAR SF AR 02 001 D. 12.3.1

[113] Keysight Technologies, Keysight 33250A Function/Arbitrary Waveform Generator data sheet.

[114] Agilent Technologies, 8114A 100 V/2 A Programmable Pulse Generator.

[115] LeCroy Corporation, Wavesurfer XS series oscilloscopes operators manual, 2008.

[116] Murata Manufacturing Co., Ltd., 7BB-20-6L0 Piezoelectric Diaphragm data sheet.

[117] L. De Marchi, A. Marzani, S. Caporale and N. Speciale, "Ultrasonic guided-waves characterization with warped frequency transforms," in IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control, vol. 56, no. 10, pp. 2232-2240, October 2009, doi: 10.1109/TUFFC.2009.1305. 13.5

[118] L. De Marchi, M. Ruzzene, B. Xu, E. Baravelli and N. Speciale, "Warped basis pursuit for damage detection using lamb waves," in IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control, vol. 57, no. 12, pp. 2734-2741, December 2010, doi: 10.1109/TUFFC.2010.1747. 13.5

[119] Finzi V. (1996), "Apparecchiature e impianti di sicurezza. Manuale del ferelettrico - Vol IV - Parte I," Collegio Ingegneri Ferroviari Italiani. 1.2.2, 16.1, 16.3, 16.2, 16.4, 16.5, IV

[120] R. J. Hill, D. C. Carpenter and T. Tasar, "Railway track admittance, earth-leakage effects and track circuit operation," Proceedings., Technical Papers Presented at the IEEE/ASME Joint Railroad Conference, Philadelphia, PA, USA, 1989, pp. 55-62, doi: 10.1109/RRCON.1989.77281. 16.4, 16.4.1

[121] T. de Bruin, K. Verbert and R. Babuška, "Railway Track Circuit Fault Diagnosis Using Recurrent Neural Networks," in IEEE Transactions on Neural Networks and Learning Systems, vol. 28, no. 3, pp. 523-533, March 2017, doi: 10.1109/TNNLS.2016.2551940. 17.1.1, 17.1.2

[122] J. Chen, C. Roberts, P. Weston, Fault detection and diagnosis for railway track circuits using neuro-fuzzy systems, Control Engineering Practice, Volume 16, Issue 5, 2008, Pages 585-596, ISSN 0967-0661, https://doi.org/10.1016/j.conengprac.2007.06.007. 17.1.2

[123] Latifa Oukhellou, Alexandra Debiolles, Thierry Denoeux, Patrice Aknin. Fault diagnosis in railway track circuits using Dempster–Shafer classifier fusion. Engineering Applications of Artificial Intelligence, Elsevier, 2010, 23 (1), pp.117-128. (10.1016/j.engappai.2009.06.005). (hal-00467941) 17.1.3

[124] Lin-Hai, Zhao & Jian-Ping, Wu & Yi-Kui, Ran. (2012). Fault diagnosis for track circuit using AOK-TFRs and AGA. Control Engineering Practice. 20. 1270–1280. 10.1016/j.conengprac.2012.07.002. 17.1.3

[125] S. Sun and H. Zhao, "Fault Diagnosis in Railway Track Circuits Using Support Vector Machines," 2013 12th International Conference on Machine Learning and Applications, Miami, FL, 2013, pp. 345-350, doi: 10.1109/ICMLA.2013.146. 17.1.3

[126] F. Espinosa et al., "Detector of Electrical Discontinuity of Rails in Double-Track Railway Lines: Electronic System and Measurement Methodology," in IEEE Transactions on Intelligent Transportation Systems, vol. 18, no. 4, pp. 743-755, April 2017, doi: 10.1109/TITS.2016.2586538. 17.2

[127] F. Espinosa, J.J.Garcia, A.Hernandez, M.Mazo, J.Ureaa, J.A.Jimenez, I.Fernandez, C.Perez, J.C.Garcia, Advanced Monitoring of Rail Breakage in Double-Track Railway Lines by means of PCA Techniques, Applied Soft Computing Journal https://doi.org/10.1016/j.asoc.2017.11.009 17.2

[128] N. Martini, Network Connected Signal Analyzer (1), Elektor Magazine (March & April 2016) www.elektormagazine.com. 18.3.2.1

[129] Jargon, Jeffrey & Wang, Chih-Ming Jack & Hale, Paul. (2008). A Robust Algorithm for Eye-Diagram Analysis. Lightwave Technology, Journal of. 26. 3592-3600. 10.1109/JLT.2008.917313. 18.3.2.5, 18.15, 18.3.2.6, 18.17, 18.16, 18.3.2.6, 18.3.2.6, 18.18, 18.19, 18.3.2.6, IV

[130] P. Rousseeuw and A. Leroy, Robust Regression and Outlier Detection. New York: Wiley, 2004. 18.3.2.6

[131] OFSTP-4A: Optical Eye Pattern Measurement Procedure, Telecommunications Industries of America/Electronics Industries of America (TIA/EIA) Standard, Nov. 1997. 18.3.2.6

[132] IEC 61280-2-2: Fibre Optic Communication Subsystem Test Procedures— Part 2-2: Digital Systems— Optical Eye Pattern, Waveform, and Extinction Ratio Measurement, International Electrotechnical Commission, Apr. 2005. 18.3.2.6

[133] GR-253-CORE: Synchronous Optical Network (SONET) Transport Systems: Common Generic Criteria, Telcordia Technologies, Jan. 1999. 18.3.2.6

[134] CSMA/CD Access Method and Physical Layer Specifications—Amendment: Media Access Parameters, Physical Layers, and Management Parameters for 10 Gb/s Operation, IEEE Std. 802.3ae, Institute of Electrical and Electronics Engineers, Aug. 2002. 18.3.2.6

[135] ITU-T Recommendation G.957: Optical Interfaces for Equipments and Systems Relating to the Synchronous Digital Hierarchy, Telecommunication Standardization Sector of International Telecommunications Union (ITU-T), Mar. 2006. 18.3.2.6

[136] J. Gowar, Optical Communication Systems. Englewood Cliffs, NJ: Prentice-Hall, 1984. 18.3.2.6

[137] D. Derickson, Fiber Optic Test and Measurement. Englewood Cliffs, NJ: Prentice-Hall, 1998. 18.3.2.6, 18.3.2.6

[138] M. Hart, C. Duff, and S. Hinch, "Firmware measurement algorithms for the HP83480 digital communications analyzer," Hewlett-Packard J., Dec. 1996.

18.3.2.6, 18.3.2.6