DOCTORATE SCHOOL IN
INDUSTRIAL AND ENVIRONMENTAL ENGINEERING

CURRICULUM INDUSTRY 4.0

XXXIII CYCLE

UNIVERSITY OF MODENA AND REGGIO EMILIA

“ENZO FERRARI” ENGINEERING DEPARTMENT

Ph.D. DISSERTATION

Deep Learning Methods for Audio-Visual Speech Processing in Noisy Environments

Candidate:
Giovanni MORRONE
Advisor:
Prof. Sonia BERGAMASCHI
Co-Advisor:
Dr. Leonardo BADINO
Director of the School:
Prof. Alberto MUSCIO

SCUOLA DI DOTTORATO IN
INGEGNERIA INDUSTRIALE E DEL TERRITORIO

CURRICULUM INDUSTRIA 4.0

XXXIII CICLO

UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

DIPARTIMENTO DI INGEGNERIA "ENZO FERRARI"

TESI PER IL CONSEGUIMENTO DEL TITOLO DI DOTTORE DI RICERCA

Metodologie di Apprendimento Profondo per l'Elaborazione Audio-Video del Parlato in Ambienti Rumorosi

Candidato:
Giovanni MORRONE
Relatore:
Prof. Sonia BERGAMASCHI
Correlatore:
Dr. Leonardo BADINO
Il Direttore della Scuola:
Prof. Alberto MUSCIO

Keywords:

Audio-Visual
Deep Learning
Speech Enhancement
Speech Recognition
Speech Inpainting

Abstract

Human communication is often an audio-visual experience. Indeed, while people hear words uttered by speakers, they can see facial movements and other gestures which convey speech information. However, speech communication can be negatively affected by background noises and artifacts, which are very common in real environments. Restoring clean speech from degraded audio sources is crucial for many applications, e.g., automatic speech recognition and hearing aids. Neuroscience research proved that looking at a talking face enhances the human capability to focus auditory attention on a particular stimulus while muting external noisy sources. This dissertation is an attempt to exploit the bi-modal, i.e., audio-visual, nature of speech for speech enhancement, automatic speech recognition and speech inpainting.

We start by presenting a novel approach to solve the problem of extracting the speech of a speaker of interest in a cocktail party scenario. Contrary to most previous work, we exploit a pre-trained face landmark detector and use facial landmarks motion as visual features in a deep learning model. In that way, we relieve our models from the task of learning useful visual features from raw pixels. We train and test our models on two widely used limited size datasets and we achieve speaker-independent speech enhancement in a multi-talker setting.

Motivated by these results, we study how audio-visual speech enhancement can help to perform automatic speech recognition exploiting a multi-task learning framework. Then, we design a strategy where speech enhancement training phase is alternated with speech recognition phase. We observe that, in general, the joint optimization of the two phases shows a remarkable improvement of speech recognition compared to audio-visual baseline models trained only to perform speech recognition.

Finally, we explore if visual information can be useful for speech inpainting, i.e., the task of restoring missing parts of an acoustic speech signal from uncorrupted audio context. We design a system that is able to inpaint variable-length missing time gaps in a speech signal. We test our system with time gaps ranging from 100 ms to 1600 ms to investigate the contribution that vision can provide for time gaps of different duration. Experiments show that the performance of audio-only baseline models degrades rapidly when time gaps get large, while the proposed audio-visual approach is still able to plausibly restore missing information.

Sommario

Spesso la comunicazione tra persone è un'esperienza audio-visiva. Infatti, una persona ascolta le parole pronunciate da un interlocutore e contemporaneamente può anche vedere i movimenti faciali ed altri segni che possono trasmettere informazioni sul parlato. Tuttavia, la comunicazione attraverso la lingua parlata può essere influenzata negativamente da rumori di sottofondo ed artefatti, i quali sono molto comuni in ambienti reali. Recuperare il parlato ripulito a partire da sorgenti sonore degradate è fondamentale per molte applicazioni, ad esempio per il riconoscimento vocale automatico oppure per gli apparecchi acustici. La ricerca nell'ambito delle neuroscienze ha dimostrato che guardare il volto di una persona mentre sta parlando migliora la capacità umana di focalizzare l'attenzione su uno stimolo sonoro specifico, silenziando sorgenti rumorose esterne. Questa tesi ha l'obiettivo di provare a sfruttare la natura bi-modale, ovvero audio-visiva, del parlato per eseguire lo speech enhancement, il riconoscimento vocale automatico e lo speech inpainting.

Iniziamo presentando un nuovo approccio per risolvere il problema di estrazione della voce di un interlocutore di interesse in uno scenario cocktail party. A differenza della grande maggioranza dei lavori precedenti, noi sfruttiamo un rilevatore pre-allenato di punti salienti facciali ed usiamo il movimento di tali punti come input video in un modello di apprendimento profondo. In questo modo, sollevi-

amo i nostri modelli dal compito di imparare le caratteristiche visive direttamente dai pixel contenuti nei fotogrammi dei video. I nostri modelli sono allenati e testati su due dataset largamente utilizzati e di dimensione limitata, e sono in grado di eseguire lo speech enhancement in presenza di più interlocutori che parlano contemporaneamente, ed anche per persone che non sono osservate durante l'addestramento.

Motivati da questi risultati, analizziamo in che modo lo speech enhancement audio-visivo può aiutare il riconoscimento vocale automatico, sfruttando un'architettura di apprendimento multi-task. Quindi, abbiamo ideato una strategia in cui la fase di addestramento dello speech enhancement è alternata con la fase di riconoscimento vocale. Osserviamo che, in generale, l'ottimizzazione congiunta delle due fasi fornisce un notevole miglioramento dell'accuratezza del riconoscimento vocale rispetto ai modelli baseline audio-visivi addestrati solamente per eseguire il riconoscimento vocale.

Infine, indaghiamo se l'informazione visiva può essere utile per lo speech inpainting, ovvero il ripristino di parti mancanti di un segnale acustico a partire dalle parti integre del segnale. Progettiamo un sistema in grado di ripristinare intervalli multipli mancanti e di lunghezza variabile all'interno di un segnale contenente il parlato. Il nostro sistema è testato con intervalli da 100 ms fino a 1600 ms per analizzare il contributo che la visione artificiale può fornire per intervalli mancanti di durate differenti. Gli esperimenti mostrano che le prestazioni dei modelli baseline basati solo sull'audio peggiorano rapidamente con l'aumentare della durata degli intervalli, mentre l'approccio audio-visivo proposto è comunque in grado di ripristinare l'informazione mancante con segnali plausibili.

"Two things fill the mind with ever new and increasing admiration and awe, the more often and steadily we reflect upon them: the starry heavens above me and the moral law within me."

I. Kant, *Critique of Practical Reason*

"A pair of wings, a different mode of breathing, which would enable us to traverse infinite space, would in no way help us, for, if we visited Mars or Venus keeping the same senses, they would clothe in the same aspect as the things of the Earth everything that we should be capable of seeing. The only true voyage of discovery, the only fountain of Eternal Youth, would be not to visit strange lands but to possess other eyes, to behold the universe through the eyes of another, of a hundred others, to behold the hundred universes that each of them beholds, that each of them is; and this we can contrive with an Elstir, with a Vinteuil; with men like these we do really fly from star to star."

M. Proust, *Remembrance of Things Past*, *Vol. 5: The Prisoner*

Acknowledgments

I owe my deepest gratitude to both my advisor Professor Sonia Bergamaschi, and co-advisor Leonardo Badino, for their precious support and guide during my Ph.D. I would like to thank Professors Jesper Jensen and Zheng-Hua Tan, for their hospitality and successful collaboration during my stay at Aalborg University. Special thanks go to my colleagues, my friends, my family, and all other people, who have accompanied me in this long journey.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AM	Acoustic Model
AO	Audio-Only
AO-SE	Audio-Only Speech Enhancement
AO-SI	Audio-Only Speech Inpainting
AO-SS	Audio-Only Speech Separation
ASR	Automatic Speech Recognition
AV	Audio-Visual
AV-ASR	Audio-Visual Automatic Speech Recognition
AV-SE	Audio-Visual Speech Enhancement
AV-SI	Audio-Visual Speech Inpainting
AV-SS	Audio-Visual Speech Separation
BLSTM	Bi-directional Long-Short Term Memory
BPTT	Back-Propagation Through Time
CNN	Convolutional Neural Network
CTC	Connectionist Temporal Classification
DNN	Deep Neural Network
EDC	Event-Driven Camera
FC	Fully Connected
FFT	Fast Fourier Transform

GAN	Generative Adversarial Network
HMM	Hidden Markov Model
IAM	Ideal Amplitude Mask
LSTM	Long-Short Term Memory
LTASS	Long-Term Average Speech Spectrum
MLP	Multi-Layer Perceptron
MSE	Mean Squared Error
MTL	Multi-Task Learning
NMF	Non-negative Matrix Factorization
PER	Phone Error Rate
PESQ	Perceptual Evaluation of Speech Quality
PIT	Permutation Invariant Training
RNN	Recurrent Neural Network
SDR	Source-to-Distortion Ratio
SE	Speech Enhancement
SI	Speech Inpainting
SNR	Signal-to-Noise Ratio
SS	Speech Separation
STFT	Short-Time Fourier Transform
STOI	Short-Time Objective Intelligibility
TBM	Target Binary Mask
TF	Time-Frequency
VAE	Variational Auto-Encoder
ViSQOL	Virtual Speech Quality Objective Listener
VL2M	Video-Landmark to Mask

List of Publications

[A] **Giovanni Morrone**, Luca Pasa, Vadim Tikhanoff, Sonia Bergamaschi, Luciano Fadiga, and Leonardo Badino, "Face Landmark-based Speaker-Independent Audio-Visual Speech Enhancement in Multi-Talker Environments", in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6900-6904, 2019.

[B] Luca Pasa, **Giovanni Morrone**, and Leonardo Badino, "An Analysis of Speech Enhancement and Recognition Losses in Limited Resources Multi-Talker Single Channel Audio-Visual ASR", in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 7309-7313, 2020.

[C] Ander Arriandiaga, **Giovanni Morrone**, Luca Pasa, Leonardo Badino, and Chiara Bartolozzi, "Audio-Visual Target Speaker Enhancement on Multi-Talker Environment using Event-Driven Cameras", in *Proceedings of IEEE International Symposium on Circuits and Systems (to appear)*, 2021.

[D] **Giovanni Morrone**, Daniel Michelsanti, Zheng-Hua Tan, and Jesper Jensen, "Audio-Visual Speech Inpainting with Deep Learning", in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (to appear)*, 2021.

Chapter 1

Introduction

1.1 Motivation

When I was writing this dissertation, most of us had to stay at home due to the outbreak of the COVID-19 pandemic. During this time, our private and working relationships were mainly depending on voice or video conference platforms, where speech was the major tool for communication. Indeed, speech is the most natural way for human interaction. Without it, people are not able to share ideas, thoughts and feelings efficiently.

However, in real world environments, speech is often corrupted by background noises originated by various sound sources like other speakers' speech, TV, wind, street noises, and so on. Despite the presence of disturbing sources, humans are very good at focusing to words uttered by a speaker they are interested in, while muting other concurrent sounds [12, 113]. This phenomenon is well-known and is referred to as *cocktail party effect* [21, 82] (Fig. 1.1).

On the other hand, hearing impaired listeners can experiment huge degradations of speech intelligibility and quality in such challenging conditions, especially when the Signal-to-Noise Ratio (SNR) is less than or equal to +10 dB [51, 53].

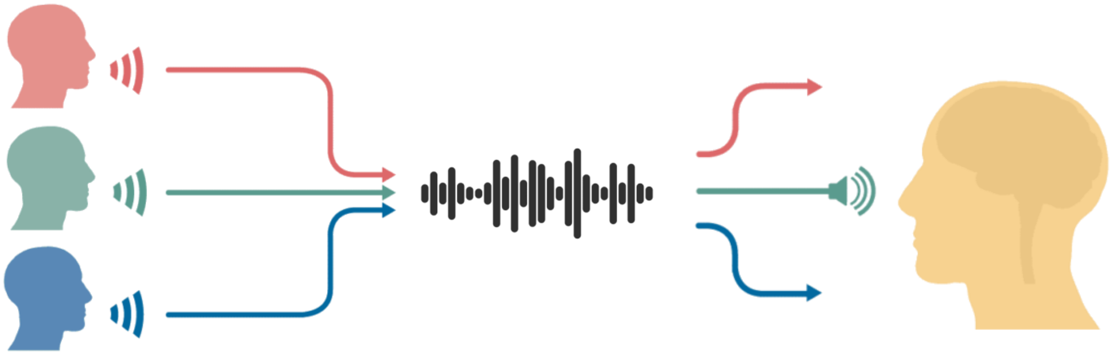


Figure 1.1: Cocktail party effect.

Additionally, many speech processing tasks, like Automatic Speech Recognition (ASR) and speaker recognition, performed poorly under these adverse noisy conditions [6, 72]. For these reasons, designing a computer system capable to extract the speech of interest in a cocktail party scenario is of great importance.

Humans can exploit additional cues to enhance a specific sound source. For instance, having two ears is beneficial, since it enables the detection of the direction of arrival of a sound. Other cues can be facial movements of the target speaker. Facial movements include the motion of lips, jaw, tongue and eyes, and also the entire head. Listeners often use these visual cues in addition to the audio to improve speech understanding. Neuroscience research proved that looking at a talking face enhances the human capability to focus auditory attention on a particular stimulus [43]. The visual stream becomes more important when the SNR is low because the audio stream is very susceptible to acoustic noise, while the visual stream is not affected by background noises. Sumby and Pollack [116] demonstrated that by carrying out experiments on several subjects. They presented to subjects Audio-Only (AO) and Audio-Visual (AV) signals of words contaminated by noise. The subjects had to select the words they perceived from a given vocabulary list. The work showed that the relative contribution of the visual stream, over the AO stream, to the word recognition performance was independent of the SNR and the absolute contribution was higher at lower SNR. McGurk and MacDonald [83] showed that vision can affect speech perception even when disturbing noises are absent. In particular, when a listener was presented with a mismatch between the audio and visual information, he tended to perceive an intermediate sound between the audio and video modalities. This phenomenon is known as *McGurk effect*.

These findings demonstrated that speech perception is essentially an AV process. Since vision plays an important role, in this dissertation we attempt to develop deep learning models that are able to exploit the AV nature of speech for several tasks, like Speech Enhancement (SE), Automatic Speech Recognition (ASR), and Speech Inpainting (SI).

1.2 Objectives

In this dissertation, we aim at developing speech processing systems that use vision to improve performance over AO systems in very adverse environments. Specifically, we address the following tasks and aspects:

- Background noises removal is crucial for development of successful speech processing applications. In particular, multi-talker environments are a very challenging scenario. The target and noisy sources consist of speech signals, generating mixtures of both acoustic and linguistic information. Moreover, if we deal with mixed-speech single channel recordings, the problem of extracting the speech of a target speaker is ill-posed. Indeed, many different hypotheses about what the target speaker says are consistent with the mixture signal. We try to address this issue by exploiting face movements of the target speaker to condition our Audio-Visual Speech Enhancement (AV-SE) system. The majority of previous works relied on Deep Neural

Networks (DNNs) to extract visual features. Otherwise, we propose the use of *face landmarks* [63] motion extracted with a pre-trained system. In this way, our models do not require to learn useful visual features from raw pixels. Based on these observations, we aim at achieving speaker-independent single-channel AV-SE using limited size datasets, which is a very common scenario in real-world applications.

- Some robust ASR systems process the audio signal through a speech enhancement or separation stage before passing it to a speech recognizer. We propose to add a preliminary AV-SE pre-processing step to analyze whether it helps in performing phone recognition in multi-talker scenarios. We experiment with a Multi-Task Learning (MTL) [15] approach, where SE and phone recognition tasks are trained together in a deep learning framework. Our main aim is to study the interaction between the two tasks, and understand how it is advantageous to train them jointly.

- In addition to disturbing sound sources, audio signals are often corrupted by accidental distortions. Impulsive noises, clicks and even transmission errors might wipe out audio intervals. The task of restoring the lost speech segments from context information is known as Speech Inpainting (SI). Previous work only exploited audio signals as context information. We aim at addressing the problem of Audio-Visual Speech Inpainting (AV-SI), where visual features are used with acoustic features to reconstruct missing speech segments. We study the contribution of vision for various lengths of missing information. Besides, we explore whether it is beneficial for AV-SI to add a phone recognition task in a MTL learning framework.

Our main work regards both the speech enhancement (SE) and speech separation (SS) tasks. Since SE and SS terms are used in previous work with different meanings, let us clarify these terms to avoid potential confusion. From now on, with SE we refer to as the process of estimating the speech of a target speaker from noisy input, regardless of the types of the disturbing sources. When all sound sources in the mixture have to be estimated, the task is denoted with SS.

1.3 Organization

The rest of this dissertation is organized as follows.

The next chapter reviews related works about AV-SE, robust ASR and SI. In particular we focus on deep learning methods.

Chapter 3 presents several AV-SE models which exploit face landmarks motion to extract the target speaker audio from mixed-speech recordings.

In order to investigate how AV-SE can help to decrease phone recognition error, Chapter 4 presents an end-to-end joint SE and phone recognition system. We analyze several training strategies based on MTL that reveal some interesting and unexpected behaviours.

In Chapter 5 we present our AV-SI system and we discuss how performances are affected by duration of missing segments.

Chapter 6 summarizes the contribution of this dissertation and presents future research directions.

Chapter 2

Background

In this chapter, we present related previous work about speech enhancement and separation, robust automatic speech recognition and speech inpainting. We review both single-channel AO and AV methods, especially the ones based on deep learning which are the most related to this dissertation.

2.1 Speech Enhancement and Separation

Several signal processing techniques have been developed to perform *speech enhancement* (SE), which is the task of recovering the clean speech of a target speaker in a noisy environment. If the noisy environment consists of speech from concurrent speakers the task is also denoted as *target speaker extraction.* On the other hand, some applications require the estimation of multiple target signals. In that case, the task is known as *source separation* or *speech separation* (SS), when the target signals are all speech signals.

Traditional SE and SS algorithms leveraged on the knowledge of statistical characteristics of the signals involved to attempt to estimate the target speech signals (cf. [76, 125] and references therein). Spectral subtraction [11, 75], Wiener filtering [81], and short-term spectral amplitude estimation [29] were some of the earliest algorithms to perform noise reduction. Since these approaches depended on strong statistical assumptions, the enhancement became problematic when such assumptions did not hold under more complex noisy conditions. The generalization problem was recently addressed by several methods based on *deep learning* which reached impressive results compared to knowledge-based models, especially for SS [19, 60, 68, 77]. SE and SS were transformed in supervised learning problems [126], allowing the use of a plethora of model architectures and training strategies that were proven to be very successful for other contexts, such as computer vision and natural language processing. Hershey et al. [54, 60] were the first to address speaker-independent multi-talker SS in the DNN framework. They proposed the deep clustering method, which combined DNN-based feature learning and spectral clustering. The DNN was trained to assign similar embeddings to Time-Frequency (TF) bins that belonged to the same speaker. During inference, the K-means algorithm was applied to cluster the TF bins into speaker clusters, according to the distance between embeddings. Finally, the speaker clusters were converted to binary TF masks, which were employed to extract the clean spectrograms of the speakers involved in the mixture. Chen et al. [19] proposed the deep attractor network, which was an improved version of the deep clustering technique. Similarly to deep clustering, high-dimensional embeddings for TF bins were learned by using DNNs. The system created attractor points for each speaker in order to pull TF bins dominated by different speakers to their corresponding attractors. Then, binary TF masks were generated by comparing embedded points and each attractor. These approaches were not able to estimate the target sources directly with DNNs, since a clustering step was required. This problem was solved by Yu et al. [68, 134], who proposed the Permutation-Invariant Training (PIT) criterion. A DNN was trained to output S TF masks, each of which was applied to mixed-speech to produce a source estimate. The Mean Squared Error (MSE) cost function was dynamically computed during training for each of the $S!$ possible assignments between the target and estimated sources. The assignment with the lower MSE was chosen and the neural network was trained to minimize the corresponding MSE. This strategy matched the SS results obtained with deep clustering, although it was much simpler.

We refer to the techniques mentioned above as Audio-Only Speech Enhancement (AO-SE) and Audio-Only Speech Separation (AO-SS), since they only considered acoustic information. However, speech perception is multimodal. Indeed, having two ears is beneficial to detect the direction of arrival of target speech and we can also see locations and movements of articulatory organs (e.g., lips, jaw, tongue) which convey speech information. Neuroscience [43] and speech perception [116] research showed that watching speaker's talking face could dramatically improve human ability to focus auditory attention on a particular acoustic stimulus. These findings motivated the development of the first Audio-Visual SE (AV-SE) [40] and Audio-Visual SS (AV-SS) [26] systems. Similarly to AO approaches, classical statistical systems were outperformed by deep learning methods. Extensive overviews on traditional AV-SS and deep-learning-based AV-SE/AV-SS were provided in [85] and [105], respectively. Here we focus on the deep learning methods that are the most related to this dissertation.

The first works that investigated to solve the AV-SE task with deep learning were presented in [56] and [133]. Wu et al. [133] proposed to use a Convolutional Neural Network (CNN) and a Multi-Layer Perceptron (MLP) to extract visual and acoustic features, respectively. These features were concatenated and fed into a Bi-directional Long-Short Term Memory (BLSTM) which outputted the estimated log power spectrum of the clean speech signal. Hou et al. [56] showed that using information about lip positions could help to improve SE. A video feature vector was obtained computing pair-wise distances between 18 landmark points of the speaker's mouth. Then, the visual features were concatenated with noisy acoustic features and used as input to a MLP to obtain the mel-scaled spectrogram of the denoised speech.

However, mere concatenation of AV feature vectors did not allow to control the multi-modal fusion process. AV-SE models might not be able to fully exploit visual information and might be dominated by audio, or vice versa. Several solutions were proposed to mitigate this problem. For example, in [57] the system was trained to output visual frames of speaker's mouth, in addition to acoustic output. In this way, they forced the network to exploit visual information, since it was much easier to reconstruct visual frames by using video input rather than audio. They also used the *multi-style training* strategy [22, 91], where one of the input modalities was removed to ensure that an input did not dominate over other ones. Another strategy tried to use a mixture of speech signals from the same speaker as input [36]. It was effective because the only way to extract the target speech was by exploiting vision. Nowadays, state-of-the-art approaches to fuse AV features are attention-based methods [23, 49]. Each input speech frame can contain silence, target speech only, interfering speech only, and overlapped speech. Attention mechanisms updated the importance of audio and visual modalities according to the characteristics of each frame. Weights referred to audio and video were dynamically computed and multiplied with plain audio and visual features to obtain their corresponding weighted features. Finally, the AV fusion was performed by concatenating the weighted features.

The availability of large-scale AV datasets enabled a clear breakthrough for AV-SE over the previous systems in the last few years [4, 31, 93, 94]. These AV corpora consisted of recordings from thousands of speakers in real-world scenarios which were exploited to achieve speaker-independent AV-SE. Indeed, these systems were able to extract speech of target speakers unseen at training time. Afouras et al. [4] proposed a CNN-based architecture that consisted of two modules: a magnitude sub-network and a phase sub-network. The first sub-network was fed with the noisy spectrogram and the speaker video, and estimated the enhanced spectrogram. Then, the enhanced spectrogram and the noisy phase were passed to the phase sub-network to estimate the enhanced phase. Ephrat et al. [31] exploited a network pre-trained on face recognition to extract face embeddings. The face embeddings of each speaker involved in the mixture and the Short-Time Fourier Transform (STFT) of the noisy audio were processed by a deep learning model that outputted complex-valued TF masks. Ochiai et al. [93] also used similar face embedding features and, additionally, an enrollment utterance of the target speaker. They proposed an additive attention framework inspired by [9] to emphasize the most informative speaker cue at each frame.

All the mentioned deep learning-based approaches were generally trained to learn a mapping between synthetic noisy speech signals and target speech signals in a supervised learning framework. On the other hand, Sadeghi et al. [108] proposed an unsupervised learning approach based on Variational Auto-Encoders (VAEs). VAEs were exploited to learn a prior probability distribution over clean speech signals. In particular, they developed a conditional VAE where the speech generative process was conditioned by visual information. At inference time, the AV speech generation model was used with a noise variance model based on Non-negative Matrix Factorization (NMF) to perform SE. Contrary to supervised methods, the training process did not require neither clean and noisy speech pairs, nor multiple noise types and levels to ensure good generalization.

2.2 Joint Speech Enhancement and Recognition

Although state-of-the-art speech recognition systems have reached very high accuracy, their performance drops significantly when the signal is recorded in challenging conditions (e.g., mismatched noises, low SNR, reverberation, multiple voices).

Classical robust ASR systems processed the audio signal through a noise removal stage before passing it to the recognizer. Since SE (or SS) and ASR modules were trained independently, a channel mismatch between denoised features and ASR inputs was introduced. Narayanan et al. [89] addressed this problem introducing a learnable non-linear function that mapped the denoised features to their clean counterparts. Alternative approaches [90, 130] proposed DNN-based frameworks that unified denoising and acoustic modeling via joint adaptive training. The unification of the two modules was done by inserting between them appropriate layers with fixed weigths and by tuning all the trainable weights for a few additional epochs. The fixed layers included logarithmic compression, feature normalization, delta calculation, and feature splicing. Such pre-processing steps generated a better input representation for acoustic models. All these methods exploited some kind of TF masking technique to perform SE.

Multi-Task Learning (MTL) is another way to improve noise-robustness of ASR systems. Indeed, it is well known that simultaneously learning multiple related tasks from data can be more advantageous rather than learning these tasks independently [15, 137]. Several speech processing applications are tightly related, then MTL methods can improve performance and reduce generalization error. In particular, robust ASR models showed better accuracy when they were trained with other tasks [20, 119]. Chen et al. [20] developed a LSTM-based model trained with a weighted sum of ASR and SE losses. It consisted of several shared hidden layers and performed ASR and SE in two different output layers. Tang et al. [119] proposed a multi-task recurrent network to perform speech and speaker recognition simultaneously. In particular, the output of the ASR sub-model at the current frame was used as auxiliary information by the speaker recognition sub-model when processing the next frame, and vice-versa.

Several recent studies showed significant advancements in target speaker extraction from mixed-speech [128, 141], and SS [19, 60, 68, 77]. Many of these SS methods were exploited to perform target speaker ASR [27, 129] and end-to-end multi-speaker ASR, which aimed at recognizing all the utterances from a mixture of multiple speakers' speech [16, 103, 111]. Qian et al. [103] presented a variant of the PIT [134] technique to estimate senone posterior probabilities with a joint optimization scheme. The architecture consisted of two PIT-based modules. Firstly, a front-end SS module was optimized with a MSE loss, then a back-end recognition module was trained with cross entropy loss while the weights of the SS module were frozen. Finally the parameters of both modules were jointly tuned to minimize the cross entropy ASR loss. However, this method still required a single-speaker ASR system to obtain the senone alignment labels. Chang et al. [16], and Seki et al. [111] solved this drawback by unifying SS and ASR models in an end-to-end architecture that combined an attention-based encoder-decoder network with the Connectionist Temporal Classification (CTC) [46] objective function. The attention mechanism and the CTC were both used to align input speech features to reference character labels. Additionally, the CTC loss was used to find the correct output-label permutation. Therefore, the models only required the speech mixture and corresponding transcriptions of each speaker at training time. Note that in all PIT-based systems the targets of all speakers involved in the scene were needed at training time. Moreover, the use of PIT imposed an upper bound on the maximum number of speakers in the mixture. Delcroix et al. [27] addressed this issue by using a speaker extraction and recognition neural network, "SpeakerBeam", that was independent of the number of the sources in the mixture. They employed a speaker adaptation layer that was used to inform the extraction network about which speaker to extract. This approach required enrollment utterances of the target speaker which were exploited to adjust the parameters of the speaker adaptation layer.

In Section 2.1 we have reviewed many AV-SE and AV-SS deep learning-based approaches which outperformed their AO counterparts in extremely challenging scenarios. Vision can also be used to boost ASR performances [88, 92]. Chung et at. [22] were the first to propose an end-to-end AV-ASR system based on an encoder-decoder architecture. Petridis et al. [98] used two different streams for feature extraction from raw images and waveforms. The streams were concatenated and fed to a bi-directional Recurrent Neural Network (RNN) that processed the fused representation and emitted word labels. Afouras et al. [3] proposed a combination of a sequence-to-sequence (seq2seq) [117] and a transformer self-attention [122] model. All these systems were trained and tested with general environmental noises. The first attempt to deal with overlapping speech was reported in [18]. They experimented with several combinations of audio, visual, and speaker identity information as input to a DNN-Hidden Markov Model (DNN-HMM). The AV model showed a large improvement over the AO baseline when it was tested on 2-speakers mixtures. A very recent work [135] outperformed [3] on clean and overlapped speech using a hybrid architecture. The input features were processed by a Time-Delay Neural Network (TDNN) which emitted the aligned grapheme-state units. The entire model was trained to optimize the Lattice Free-Maximum Mutual Information (LF-MMI) [100] discriminative criterion.

2.3 Speech Inpainting

In real life applications, audio signals frequently suffer from undesired localized corruptions. These corruptions can be originated by packet-loss in transmission, clicks, issues during the recording, scratched CDs, and so on. The process of restoring the lost information from the audio context is known as *audio inpainting* [2], regardless of the causes of the loss. In the literature, this restoration is also denoted as *audio interpolation* [34], *audio extrapolation* [61], or *waveform substitution* [44]. When applied to speech signals, we refer to it as *Speech Inpainting* (SI).

Generally, previous work assumed that the distorted data was missing and their location was known. Moreover, the most studied problem regarded the restoration of time gaps in audio signals. The system presented in Chapter 5 also follows this setting. The number and the duration of time gaps can vary a lot. For instance, corruptions can be frequent in presence of clicks, but their duration is usually very short, e.g., a few milliseconds or less. This case is referred as inpainting of *short gaps*. On the other hand, we can experiment very long missing gaps, e.g., hundreds of milliseconds, or seconds. Lost connections in audio transmission, damaged physical media, unwanted high noises might last for seconds. We refer to this situation as inpainting of *long gaps*.

The first audio inpainting works aimed at restoring short gaps by exploiting traditional signal processing techniques. Adler et al. [2] proposed an algorithm based on Orthogonal Matching Pursuit (OMP), which exploited sparse representations to efficiently model audio signals. This work inspired a lot of research on sparsity-based audio inpainting [1, 74, 86, 120]. Other works used TF representations in a regression model [132], or in a NMF model [70, 114]. Although these methods obtained good performance for short gaps, they did not extend well for longer gaps.

For inpainting long gaps several methods have been proposed. In general, early solutions leveraged repetitions and stationarity of signals to find the most suited missing segment extracted from uncorrupted audio context. In that case, the objective was to output a plausible solution rather than a perfect reconstruction. For example, Bahat et al. [8] tried to fill missing gaps using pre-recorded speech examples from the same speaker. Perraudin et al. [97] exploited self-similarity graphs within audio signals. However, the first approach required a different model for each speaker, and the second one was less suitable for speech, since it could only inpaint stationary signals. Prablanc et al. [101] proposed a text-informed solution to inpaint missing speech combining speech synthesis and voice conversion models.

Recently, several researchers attempted to solve audio inpainting using deep learning. In [79] a CNN model was used to inpaint missing audio from adjacent context. Other works exploited Generative Adversarial Networks (GANs) to generate sharper TF representations [28, 78]. Zhou et al. [138] demonstrated that exploiting visual cues improved inpainting performance. However, these approaches only restored music signals, which usually have longer time dependencies than speech. Chang et al. [17], and Kegler et al. [64] both tried to generate speech from masked signals with convolutional encoder-decoder architectures. A very recent work proposed a two-stage enhancement network where binary masking of a noisy speech spectrogram was followed by inpainting of TF bins affected by severe noise [50].

At the best of our knowledge, there are no works that address the problem *Audio-Visual Speech Inpainting* (AV-SI), i.e., the task of restoring the missing parts of an acoustic speech signal using audio context and visual information. However, when missing time gaps are very long, i.e., one second or more, the contribution of vision to the reconstruction should be predominant. Therefore, the task of *speech reconstruction (or synthesis) from silent videos* is very related to ours. Indeed, we can see this task as a particular case of AV-SI when the acoustic context is missing. Le Cornu and Milner [69] developed the first speech synthesis system only using the silent video of a speaker's frontal face. They based their system on STRAIGHT vocoder [62], which was able to generate speech from several parameters. Only the spectral envelope parameter was estimated with a neural network model, while the other vocoder parameters were artificially produced without considering visual information. Ephrat and Peleg [32] proposed to solve the video-to-speech task with a DNN-based regressor. Their model mapped raw pixels to linear predictive coding (LPC) coefficients computed from the target audio signal. They improved their system in a follow-up work [30], where a CNN model estimated the mel-scale spectrogram from video frames and optical flow. Another work [5] employed an autoencoder to extract bottleneck features from the auditory spectrogram. Subsequently, the bottleneck features were used as targets in the main video-to-speech synthesis neural network.

The main limitation of the previous systems was that they were speaker-dependent, meaning that they needed different models for each speaker. Prajwal et al. [102] developed a sequence-to-sequence system inspired by Tacotron 2 [112], a state-of-the-art text-to-speech model. They mainly aimed at reconstructing speech of specific speakers in unconstrained settings. However, they also proposed a multi-speaker approach. Speaker embeddings were extracted from reference signals of the speakers and were used to condition the system. We can not consider this method as speaker-independent as it needs prior information about the speaker. The first pure speaker-independent system was proposed in [124], where a GAN was exploited to estimate directly time-domain speech from video frames. Although this model was able to reconstruct intelligible speech of speakers unseen at training time, speech quality was lower than speaker-dependent methods mentioned above. Michelsanti et al. [84] tried to improve speech quality by using vocoder features as training targets in place of raw waveforms. They used the WORLD vocoder [87] to synthesize waveforms from parameters estimated by a deep learning model. In addition, they proposed a MTL approach where a visual speech recognition task was learned together with the estimation of the vocoder parameters. The results showed that the MTL approach was beneficial both for speech intelligibility and quality.

Chapter 3

Audio-Visual Speech Enhancement at the Cocktail Party

This chapter addresses the problem of enhancing the speech of a speaker of interest in a cocktail party scenario when visual information, i.e., face landmarks motion, is available. The work described in this chapter has been presented at the *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* [A] [1].

3.1 Introduction

SE aims at extracting the voice of a target speaker. When two or more speakers are involved in the input mixture, all the utterances spoken by each speaker can be treated as valid outputs. This problem can be solved by exploiting some additional

[1] Code and demos of our AV-SE models are available at `https://dr-pato.github.io/audio_visual_speech_enhancement/`.

information associated to the speaker of interest to guide the SE model to output the utterance of the target speaker. We attempt to use video of the target speaker's talking face as additional information.

The majority of AV-SE systems exploit DNN architectures to extract visual features from raw pixels [85]. In general, in most previous work visual features are learned directly during SE training through backpropagation [36, 42, 57]. Other systems extract visual features from a hidden layer of DNN models trained for other tasks like face recognition [31, 93] and visual speech recognition [4, 108].

Instead, Hou et al. [56] show that using information about facial landmarks positions can help to improve SE. Landmarks are salient points on the face such as the corners of the mouth, the eyebrows, the eyes, the nose, and the jaw (see Fig. 3.1). The video feature vector is obtained computing pair-wise distances between any mouth landmarks. In this chapter, we also use face landmarks as video inputs. Contrary to our approach, they use position-based features while we use motion features of the whole face that in our experiments turn out to be much more effective than positional features. Finally, in [56] the system is only evaluated in a speaker-dependent setting.

We extract the face landmarks employing an efficient pre-trained landmark extractor available from Dlib [65]. By using face landmarks as input, we relieve our AV-SE models from learning useful visual features from raw pixels. That aspect is particularly relevant when the training AV datasets are small.

In this chapter we aim at exploring different ways of mapping such visual features into TF masks, which are exploited to clean the acoustic noisy spectrogram.

Additionally, we propose to substitute the visual pipeline implemented with traditional frame-based sensors with an equivalent pipeline based on the Event-Driven Cameras (EDCs) [96]. EDCs are a novel type of vision sensors that asynchronously measure the brightness change for each pixel. These features enable

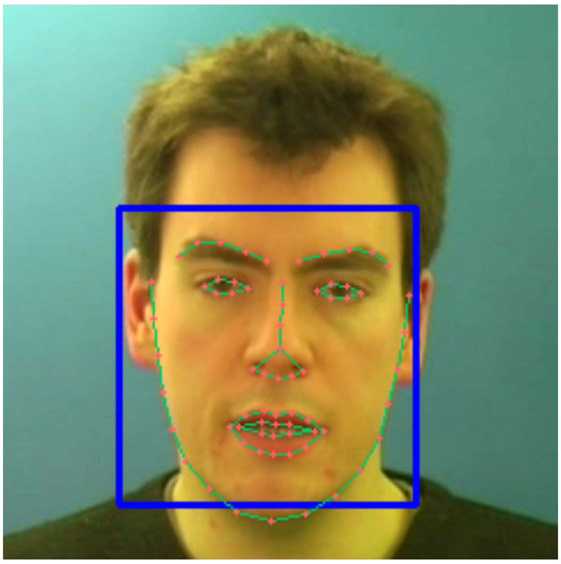


Figure 3.1: Example of a face bounding box (blue square) and face landmarks (red crosses) extracted with Dlib.

the extraction of motion at lower computational cost and latency.

The chapter is organized as follows. The problem formulation is presented in Section 3.2. The description of the overall system and of the AV-SE models is provided in Section 3.3. Experimental setup and evaluations are given in Section 3.4 and Section 3.5, respectively. We provide a short overview of the EDC-based approach in Section 3.6. Finally, we conclude this chapter in Section 3.7.

3.2 Problem Formulation

Let $x[n]$ and $d[n]$ denote the target clean speech signal and the noise signal, respectively, where n indicates a discrete-time index. We can model the observed noisy speech signal, $y[n]$, as:

$$y[n] = x[n] + d[n]. \tag{3.1}$$

We define the problem of monaural, or single-channel, AO-SE as the task of finding a function, $\mathcal{F}_{a-se}$, that estimates $x[n]$ from $y[n]$. The estimated clean signal is denoted by $\hat{x}[n]$. In Eq. 3.1 we only consider additive noise signals, although SE generally deals also with non-additive noises and distortions, e.g., reverberation. In the case of AV-SE, the function $\mathcal{F}_{av-se}$ has to estimate $x[n]$ from $y[n]$ and an additional visual signal, $v[m]$. m is a discrete-time index typically different from n, since visual and acoustic signals are generally sampled at different rates.

The Eq. 3.1 can also be expressed in the TF domain as:

$$Y(k,l) = X(k,l) + D(k,l), \tag{3.2}$$

where k and l are a frequency bin index and a time frame index, respectively, while $Y(k,l)$, $X(k,l)$ and $D(k,l)$ represent the Short-Time Fourier Transform (STFT) coefficients of the noisy speech, target clean speech and noise signals, respectively. As above, the problem of SE is finding a function that provides an estimate, $\hat{X}(k,l)$, of $X(k,l)$. The STFT coefficients are complex-valued, then both real and imaginary parts have to be estimated or, equivalently, the magnitude, $|X(k,l)|$, and the phase, $\angle X(k,l)$. Logarithmic or power-law compression is usually applied to STFT magnitude to obtain a spectrogram, $X_s(k,l)$, which is often used as acoustic input to speech processing algorithms. The majority of SE systems only estimate the STFT magnitude or the spectrogram, since the target STFT phase is considered less important [38, 127]. In that case, noisy phase, $\angle Y(k,l)$, is exploited to apply inverse STFT, which reconstructs the time-domain signal. In this dissertation we follow this approach. However, recent AV-SE work propose alternative solutions that attempt to estimate the target phase [4], $\angle X(k,l)$, the STFT complex-valued coefficients [31], $X(k,l)$, or the time-domain signal [59], $x[n]$, directly.

In general, SE aims at extracting the voice of a single target speaker from signals affected by background noise and other artifacts. In some situations the noisy signal may consist of two or more speech signals. In the two-speakers mixture case, the observed acoustic signal can be modelled as:

$$y[n] = x_1[n] + x_2[n]. \tag{3.3}$$

In that case, the task of estimating the speech of the target speaker, $x_1[n]$ or $x_2[n]$, is also denoted as *target speaker extraction*, and it is the scenario we address in this chapter and in Chapter 4. On the other hand, the task of estimating all the target speech signals, $x_1[n]$ and $x_2[n]$, in the mixture is known as AO-SS. If visual signals are provided then the task is denoted as AV-SS.

3.3 System Overview

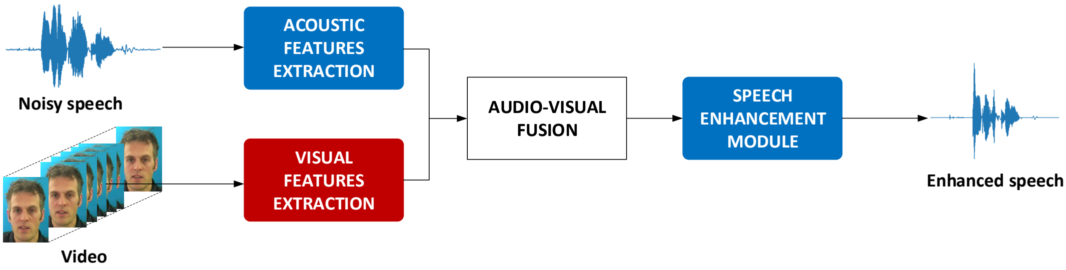

Noisy speech
Video
ACOUSTIC FEATURES EXTRACTION
VISUAL FEATURES EXTRACTION
AUDIO-VISUAL FUSION
SPEECH ENHANCEMENT MODULE
Enhanced speech

Figure 3.2: Schematic diagram of the general architecture of our AV-SE systems.

The general architecture of our AV-SE algorithms is depicted in Fig. 3.2. The architecture consists of four building blocks: *acoustic features extraction*, *visual features extraction*, *audio-visual fusion* and *speech enhancement module*. In the following subsection we describe the *audio-visual fusion* and the *speech enhancement module*. The details about the *acoustic features extraction* and the *visual features extraction* are provided in Subsections 3.4.3 and 3.4.4, respectively.

3.3.1 Model Architectures

We experiment with four models. All models receive in input the target speaker's landmark motion vectors and the power-law compressed spectrogram of the single-channel mixed-speech signal. All of them perform some kind of TF masking operation. Generally, the TF mask is element-wise multiplied with the noisy spectrogram to obtain the estimated clean spectrogram. The analysis of landmark-dependent masking strategies is motivated by the fact that SE mediated by explicit masking is often more effective than mask-free enhancement [136].

Video-Landmark to Mask

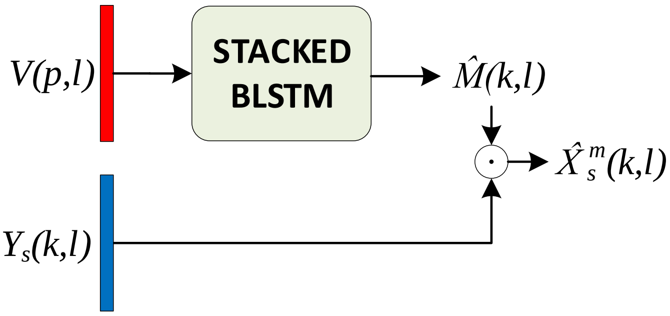


Figure 3.3: VL2M model.

At each time frame, the Video-Landmark to Mask (VL2M) model (Fig. 3.3) estimates a TF mask, $M(k,l)$, from visual features of the target speaker, $V(p,l)$, without considering acoustic information. More formally, VL2M performs a function $\mathcal{F}_{vl2m}(V(p,l)) = \hat{M}(k,l)$, where $\hat{M}(k,l)$ is the estimated mask, and p, k and l are a visual feature index, a frequency bin index and a time frame index, respectively.

The training objective for VL2M is a Target Binary Mask (TBM) [7, 67], computed using the spectrogram of the target speaker only. This is motivated by our goal of extracting the speech of a target speaker as much as possible independently of the concurrent speakers. An additional motivation is that the model takes as only input the visual features of the target speaker, and a target TBM that only depends on the target speaker allows VL2M to learn an one-to-one mapping function. This condition is not met by masks dependent on background noisy sources.

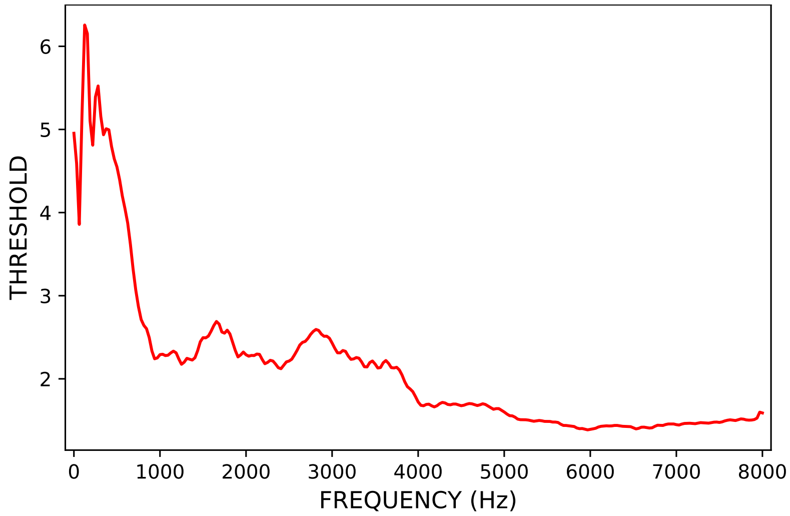
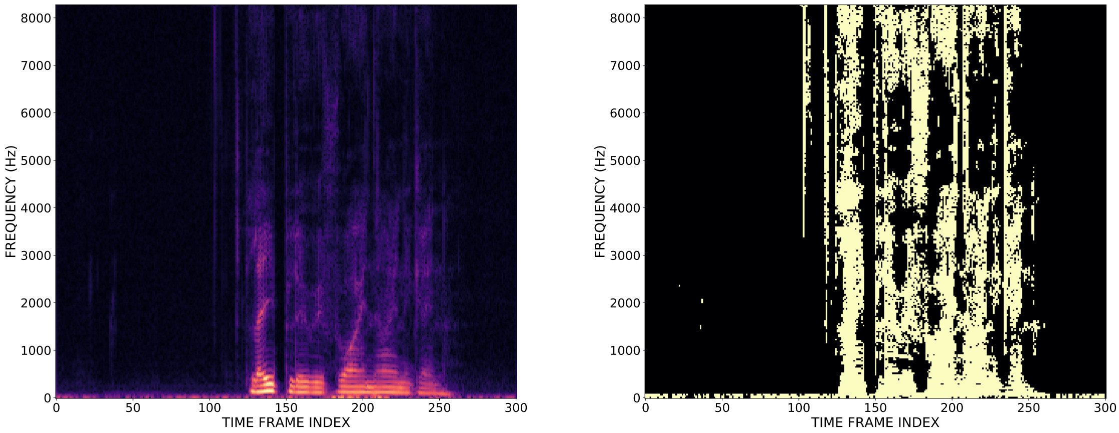


Figure 3.4: Example of a thresholding function based on the LTASS of a male speaker.

Given a clean speech spectrogram of a target speaker i, $X_s(k,l)$, the TBM is defined by comparing, for each frequency bin index $k \in [1, K]$, the target speaker value $X_s(k,l)$ vs. a reference threshold $\tau^i(k)$. This threshold indicates if a TF unit is generated by the speaker or refers to silence or noise. As in [35], we use a function of Long-Term Average Speech Spectrum (LTASS) as reference threshold (see Fig. 3.4). The LTASS provides a means of viewing the energy distribution across frequencies in continuous speech. In our case, the LTASS is computed separately for each speaker.

The process to compute the LTASS and TBMs from the spectrograms of the speaker i is as follows:

1. The mean, $\mu_s^i(k)$, and the standard deviation, $\sigma_s^i(k)$, along the time axis, are



(a) Spectrogram.

(b) Target Binary Mask (TBM).

Figure 3.5: Example of a spectrogram and the corresponding Target Binary Mask (TBM).

computed for all frequency bin indices of spectrograms included in training data and belonging to the speaker i.

2. The threshold, $\tau^i(k)$, is defined as $\tau^i(k) = \mu_s^i(k) + 0.6 \cdot \sigma_s^i(k)$, where 0.6 is a value selected by manual inspection of several spectrogram-TBM pairs.

3. The threshold is applied to every speaker's speech spectrogram, $X_s(k,l)$.

$$M(k,l) = \begin{cases} 1, & \text{if } X_s(k,l) \geq \tau^i(k), \\ 0, & \text{otherwise.} \end{cases}$$

The mapping $\mathcal{F}_{vl2m}(\cdot)$ is carried out by a stacked BLSTM network [47]. The outputs of the last BLSTM are then forced to lay within the $[0,1]$ range using the sigmoid activation function. Finally, the estimated TBM, $\hat{M}(k,l)$, and the noisy spectrogram, $Y_s(k,l)$, are element-wise multiplied to obtain the estimated clean spectrogram $\hat{X}_s^m(k,l) = \hat{M}(k,l) \odot Y_s(k,l)$.

The model parameters are estimated to minimize the binary cross-entropy loss:

$$J_{vl2m} = \sum_{k=1}^{K} \sum_{l=1}^{L} -M(k,l) \cdot \log(\hat{M}(k,l)) - (1 - M(k,l)) \cdot \log(1 - \hat{M}(k,l)). \quad (3.4)$$

VL2M-ref model

VL2M generates TF masks that are independent of the acoustic context. We may want to refine the masking by including such context.

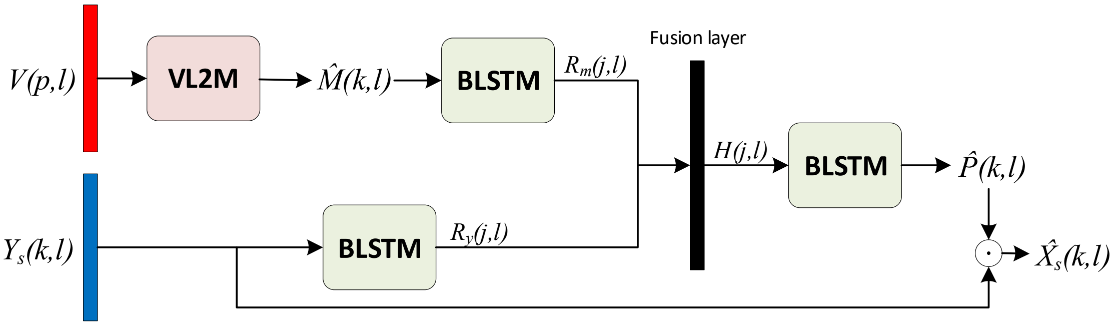


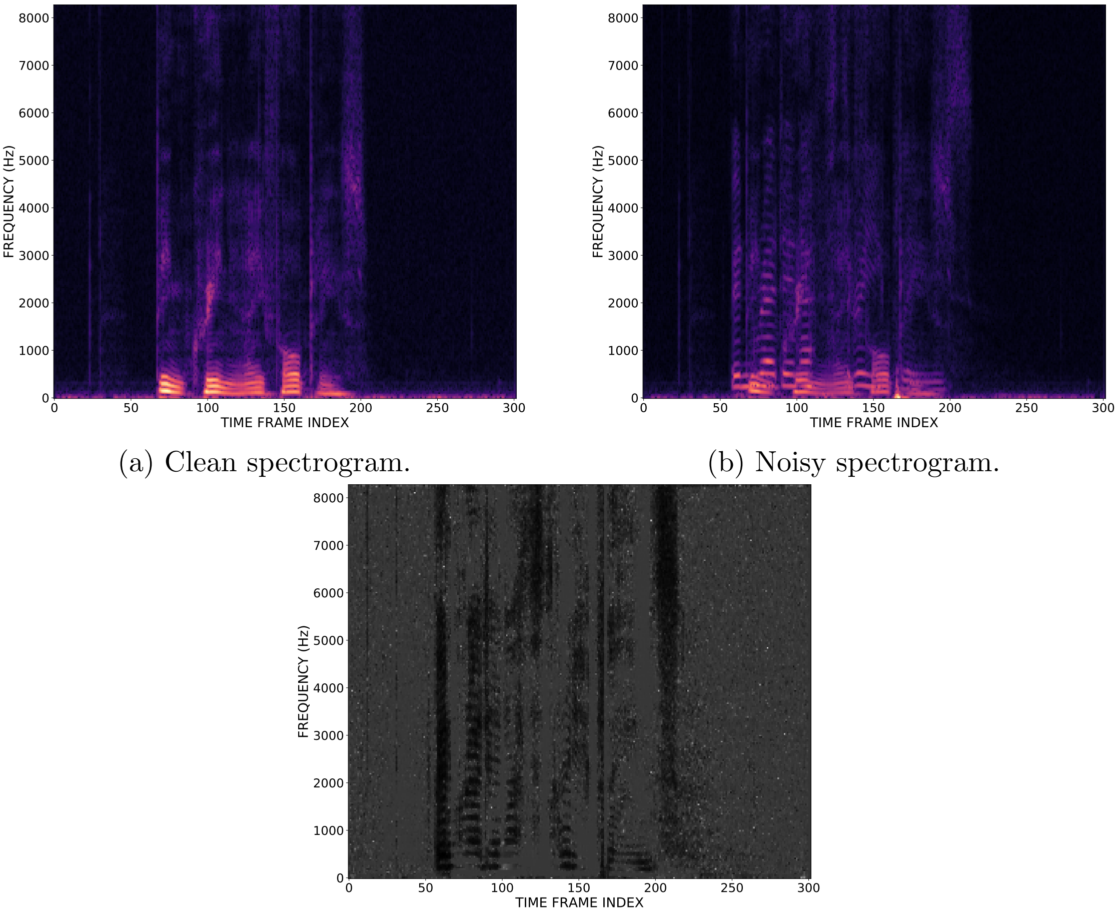
Figure 3.6: Vl2M-ref model.

This is what the novel VL2M-ref does (Fig. 3.6). The estimated TBM, $\hat{M}(k,l)$, and the noisy spectrogram, $Y_s(k,l)$, are the input to a function that outputs an Ideal Amplitude Mask (IAM), $P(k,l)$ (known as FFT-MASK in [136]). Unlike TBM, IAM allows perfect recovery of the target spectrogram. Given the target clean spectrogram, $X_s(k,l)$, and the noisy spectrogram, $Y_s(k,l)$, the IAM is defined as:

$$P(k,l) = \frac{X_s(k,l)}{Y_s(k,l)}. \quad (3.5)$$

Note that IAM computation only uses the target and mixed-speech spectrograms, unlike several widely used masks, e.g., Ideal Binary Mask (IBM), and Ideal Ratio Mask (IRM), which also require the spectrogram of interfering sources.

The target speaker's spectrogram, $X_s(k,l)$, is reconstructed by multiplying the input spectrogram with the estimated IAM. Since the IAM is not upper bounded, values greater than 10 are clipped to 10 in order to obtain better numerical stability as suggested in [136].



(a) Clean spectrogram.

(b) Noisy spectrogram.

(c) IAM.

Figure 3.7: Example of a clean spectrogram, a noisy spectrogram and the corresponding Ideal Amplitude Mask (IAM).

The model performs a function $\mathcal{F}_{mr}(V(p,l), Y_s(k,l)) = \hat{P}(k,l)$ that consists of a VL2M component plus three different BLSTMs $\mathcal{G}_m$, $\mathcal{G}_y$, and $\mathcal{H}$.

$\mathcal{G}_m(\mathcal{F}_{vl2m}(V(k,l)) = R_m(j,l)$ receives the VL2M mask as input, and $\mathcal{G}_y(Y_s(k,l)) = R_y(j,l)$ is fed with the noisy spectrogram. Their outputs, $R_m(j,l)$ and $R_y(j,l)$, can be treated as matrices, where the l^{th} column represents an intermediate vector representation at the time frame l, and j indicates an index of a value in the vector. We apply a linear combination of $R_m(j, l)$ and $R_y(j, l)$ to obtain $H(j, l)$, which can be regarded as a joint AV representation:

$$H(j, l) = \mathbf{W}_{hm} \cdot R_m(j, l) + \mathbf{W}_{hy} \cdot R_y(j, l) + \mathbf{b}_h, \tag{3.6}$$

where $\mathbf{W}_{hm}$, $\mathbf{W}_{hy}$ and $\mathbf{b}_h$ are the first weight matrix, the second weight matrix and the bias of the linear layer, respectively. $H(j, l)$ is the input of the third BLSTM $\mathcal{H}(H(j, l)) = \hat{P}(k, l)$, where $\hat{P}(k, l)$ lays in the [0,10] range.

As training objective, we use an indirect mapping scheme [131], which minimizes the MSE loss between the target and estimated spectrogram with an implicit masking:

$$J_{mr} = \frac{1}{KL} \sum_{k=1}^{K} \sum_{l=1}^{L} (\hat{P}(k, l) \odot Y_s(k, l) - X_s(k, l))^2. \tag{3.7}$$

Audio-Visual Concat

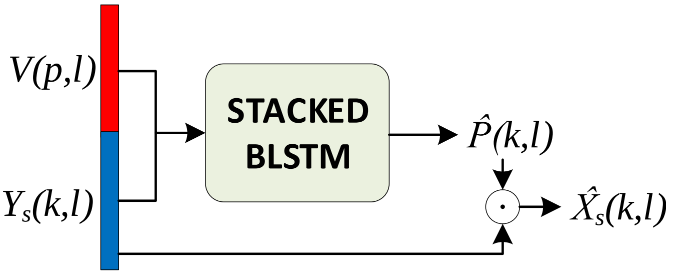


Figure 3.8: AV Concat model.

The third model (Fig. 3.8) performs early fusion of AV features. The fusion is made by concatenating the audio and video along the feature axis. The model consists of a single stacked BLSTM that computes the IAM, $\hat{P}(k, l)$, from the fused representation. The training loss is the same J_{mr} (Eq. 3.7) used to train

VL2M-ref. This model can be regarded as a simplification of VL2M-ref, where the VL2M operation is not performed.

Audio-Visual Concat-ref

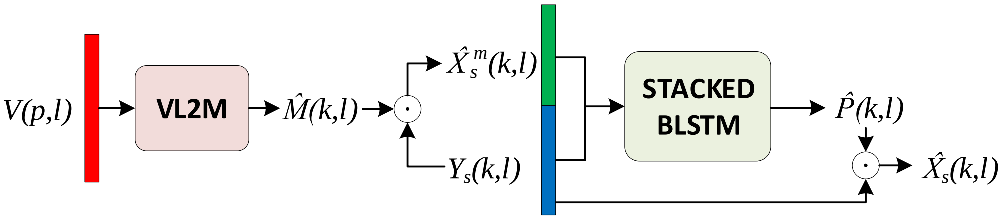


Figure 3.9: AV Concat-ref model.

The fourth model (Fig. 3.9) is an improved version of the AV Concat model. The only difference is the input of the stacked BLSTM that is replaced by the concatenation of the noisy spectrogram, $Y_s(k,l)$, and the denoised spectrogram returned by VL2M operation, $\hat{X}_s^m(k,l)$. This architecture introduces a second-stage enhancement network to refine the target speech estimation, similarly to [60].

3.4 Experimental Setup

3.4.1 Datasets

All experiments are carried out using the GRID [24] and TCD-TIMIT [52] AV datasets, which consist of frontal face recordings. For each of them, we create a mixed-speech version.

The GRID corpus consists of AV recordings from 33 speakers (one has to be discarded), each of them uttering 1000 sentences. The sentences are drawn from the following simple syntax: $command^{(4)} + color^{(4)} + preposition^{(4)} + letter^{(25)} + digit^{(10)} + adverb^{(4)}$, where the number denotes how many word choices there are for each of the 6 word categories. Each recording is 3 s long with an audio sample rate of 50 kHz and a video frame rate of 25 fps. For each speaker, who is denoted as target speaker, we first randomly select 200 utterances (out of 1000). Then, for each utterance, we create 3 different audio-mixed samples. Each audio-mixed sample is created by mixing at the same volume level the chosen utterance with one utterance from a different speaker. That results in 600 audio-mixed samples per target speaker. Audio-mixed samples and videos of target speakers are used as inputs of our models. Our systems are evaluated in a speaker-independent setting, with 25 speakers (s1-20, s22-25, s28) used for training, 4 speakers (s26-27, s29, s31) for validation, and 4 speakers (s30, s32-34) for testing. Each set consists of the same number of male and female speakers, except for the training set which contains 13 males and 12 females.

In addition, we create a speaker-dependent version of the GRID corpus to compare our systems with other studies which are only evaluated in a speaker-dependent setting [35, 36]. We split the utterances of speaker 2 and 3 into disjoint sets of 600/200/200 utterances for training/test/validation, respectively. Similarly to the speaker-independent case, each utterance of one speaker is mixed with another utterance from the other speaker in the same set.

The TCD-TIMIT corpus consists of 59 speakers (we exclude 3 professionally-trained lipspeakers) and 98 utterances per speaker. The speakers are recorded saying various sentences from the TIMIT dataset [37]. The audio sample rate is of 48 kHz and the video frame rate is of 29.97 fps. The mixed-speech version is created following the same procedure as for GRID, with one difference. Contrary to GRID, TCD-TIMIT utterances have different duration. Thus two utterances are mixed only if their duration difference does not exceed 2 seconds. For each utterance pair, we force the non-target speaker's utterance to match the duration of the target speaker utterance. If it is longer, the utterance is cut at its end, whereas if it is shorter, silence samples are equally added at its start and end. The resulting dataset is split into disjoint sets of 51 (s1-50), 4 (s51, s53-55), and 4 (s56-59) speakers for training, validation and testing, respectively.

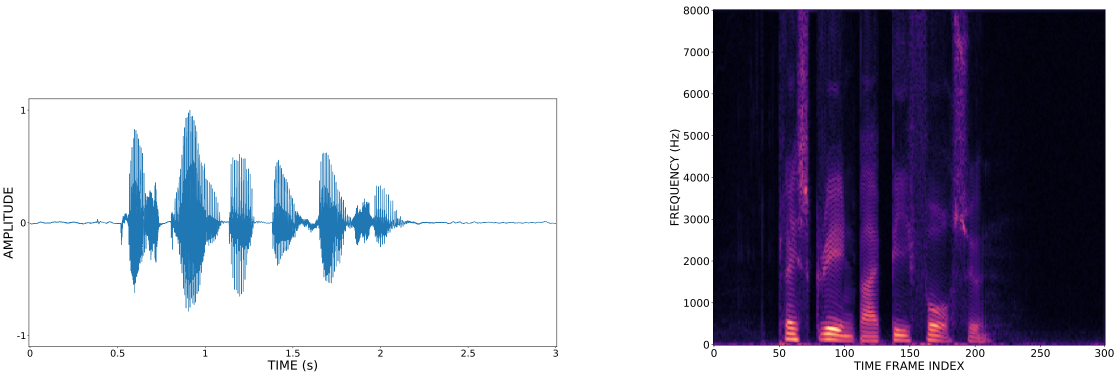
3.4.2 Optimization

In all experiments, the models are trained using the Adam optimizer [66]. The learning rate is set to 0.0001 for VL2M, and to 0.001 for the other models. Early stopping is applied when the error on the validation set does not decrease over 5 consecutive epochs.

VL2M, AV Concat and AV Concat-ref have 5, 3 and 3 stacked BLSTM layers, respectively. All BLSTMs have 250 units. Hyper-parameters selection is performed by using random search with a limited number of samples, therefore all the reported results may improve through a deeper hyper-parameters validation phase.

VL2M-ref and AV Concat-ref training is performed in 2 steps. We first discard the VL2M module and use the oracle TBM to train the systems. This step improves stability of gradients and reduce training times. Then we substitute the oracle masks with the pre-trained VL2M component. We freeze the parameters of the VL2M module and apply fine-tuning to obtain the final parameters.

Additionally, we employ an AO-SS model based on the utterance-level Permutation Invariant Training (uPIT) objective function [68] as baseline system. It is fed with the mixed-speech spectrogram and estimates two IAMs that are used to recover the two sources in the mixture. The architecture consists of 3 stacked BLSTMs with 250 hidden units.



(a) Waveform.

(b) Spectrogram.

Figure 3.10: Speech representations.

3.4.3 Audio Processing

As acoustic features we use a TF representation of input speech signals, that is a very common choice in both AO and AV speech enhancement and separation systems [85, 126].

The original waveform, $x[n]$, is resampled to 16 kHz. STFT, $Y(k, l)$, is computed using FFT size of 512, Hann window of length 25 ms (400 samples) and hop length of 10 ms (160 samples). The spectrogram, $Y_s(k, l)$, is obtained taking the STFT magnitude and performing power-law compression to simulate human loudness perception. Finally, we apply standard normalization using speaker-wise mean and standard deviation to obtain the normalized input features, $\overline{Y}_s(k, l)$. More formally:

$$\overline{Y}_s(k, l) = \frac{|Y(k, l)|^p - \mu_s^i(k)}{\sigma_s^i(k)}, \tag{3.8}$$

where $p = 0.3$, while $\mu_s^i(k)$ and $\sigma_s^i(k)$ are speaker-wise mean and standard deviation, respectively. They are obtained by taking every spectrogram of a speaker i, and by computing the mean and standard deviation of TF bins over the time axis:

$$\mu_s^i(k) = \frac{1}{\sum_{n=1}^{N_i} L_i^n} \sum_{n=1}^{N_i} \sum_{l=1}^{L_i^n} Y_s^n(k,l), \tag{3.9}$$

$$\sigma_s^i(k) = \sqrt{\frac{1}{\sum_{n=1}^{N_i} L_i^n} \sum_{n=1}^{N_i} \sum_{l=1}^{L_i^n} (Y_s^n(k,l) - \mu_s^i(k))^2}, \tag{3.10}$$

where N_i is the total number of samples of the speaker i, and L_i^n is the number of audio time frames in the n^{th} spectrogram of the speaker i. Fig. 3.10 shows an example of the waveform of a speech signal and its associated spectrogram.

In the post-processing stage, the enhanced waveform generated by the SE models is reconstructed by applying the inverse STFT to the estimated clean spectrogram and using the phase of the noisy input signal.

3.4.4 Visual Features Extraction

The face landmarks are extracted using a pre-trained algorithm available in Dlib framework [65]. The algorithm consists of two phases: face detection and face landmarks estimation. The face detector uses the classic Histogram of Gradients (HOGs) [25] features combined with a linear classifier, an image pyramid, and sliding window detection scheme. The face landmarks estimator implementation is based on an ensemble of regression trees [63] and is trained on the iBUG 300-W face landmark dataset [109], which contains 600 annotated faces. The pipeline outputs 68 2D points, for an overall of 136 values.

The video frame rate is upsampled from 25/29.97 fps (GRID/TCD-TIMIT) to 100 fps to match the audio spectrogram frame rate. Upsampling is carried out through linear interpolation over time. The motion vectors of face landmarks are computed by simply subtracting every frame with the previous one. The motion vector of the first frame is set to zero. Similarly to audio features, we perform speaker-wise normalization using mean and standard deviation to obtain the input visual feature vectors.

3.5 Evaluation Results

3.5.1 Evaluation Metrics

Our AV-SE models are designed to separate the speech of a target speaker from interfering speech sources. This task is very related to SS, which aims at estimating all the sources involved in the mixture. Therefore, we evaluate the performance of the proposed models using both SE and SS metrics. Specifically, we measure the capability of separating the target utterance from the concurrent utterances with the Source-to-Distortion Ratio (SDR) [104, 123]. The quality of estimated target speech is measured with the Perceptual Evaluation of Speech Quality (PESQ) [106] and the Virtual Speech Quality Objective Listener (ViSQOL) [55] metrics. For PESQ we use the narrow band mode, while for ViSQOL we use the wide band mode. For every metric, higher values correspond to better performance. The metrics are only computed on target speakers' speech to assure that each speaker provides the same contribution to the overall evaluation.

3.5.2 Results

As a very first experiment we compare face landmark position features vs. landmark motion features vectors. It turns out that landmark positions perform poorly, thus all results reported here refer to landmark motion feature vectors only.

We then carry out some speaker-dependent experiments to compare our models with previous studies as, to the best of our knowledge when we published this study [A], there were no reported results of speaker-independent systems trained

	SDR	PESQ	ViSQOL
Noisy	−1.06	1.81	2.11
VL2M	3.17	1.51	1.16
VL2M-ref	**6.50**	**2.58**	**2.99**
AV Concat	6.31	2.49	2.83
AV Concat-ref	6.17	**2.58**	2.96

Table 3.1: GRID results - speaker-dependent. The "Noisy" row refers to the evaluation values of the input mixed-speech signal.

	2 Speakers			3 Speakers		
	SDR	PESQ	ViSQOL	SDR	PESQ	ViSQOL
Noisy	0.21	1.94	2.58	−5.34	1.43	1.62
AO uPIT	7.39	2.59	3.00	Not Available		
VL2M	3.02	1.81	1.70	−2.03	1.43	1.25
VL2M-ref	6.52	2.53	3.02	2.83	2.19	2.53
AV Concat	7.37	2.65	3.03	3.02	2.24	2.49
AV Concat-ref	**8.05**	**2.70**	**3.07**	**4.02**	**2.33**	**2.64**

Table 3.2: GRID results - speaker-independent.

and tested on GRID and TCD-TIMIT to compare with. Table 3.1 reports the test set evaluation of speaker-dependent models on the GRID corpus with landmark motion vectors. Results are comparable with previous state-of-the-art studies in an almost identical setting [35, 36].

Table 3.2 and 3.3 show speaker-independent test set results on the GRID and TCD-TIMIT datasets, respectively. V2ML performs significantly worse than the other three models indicating that a successful mask generation has to depend on the acoustic context. The performance of the three models in the speaker-independent setting is comparable to that in the speaker-dependent setting.

AV Concat-ref outperforms V2ML-ref and AV Concat in both datasets. This supports the utility of a refinement strategy and suggests that the refinement is more effective when it directly refines the estimated clean spectrogram, rather than

	2 Speakers			3 Speakers		
	SDR	PESQ	ViSQOL	SDR	PESQ	ViSQOL
Noisy	0.21	2.22	2.74	−3.42	1.92	2.04
VL2M	2.88	2.25	2.62	−0.51	1.99	1.98
VL2M-ref	9.24	2.81	3.09	5.27	2.44	2.54
AV Concat	9.56	2.80	3.09	5.15	2.41	2.52
AV C-ref	**10.55**	**3.03**	**3.21**	**5.37**	**2.45**	**2.58**

Table 3.3: TCD-TIMIT results - speaker-independent.

refining the estimated mask.

Additionally, in order to assess the importance of masking, we create a modified version of the AV Concat model that reconstructs the target speaker spectrogram without going through any mask operation. During training, we observed a very unstable behaviour of the loss function and a SDR value just above 5 on the GRID test-set.

Regarding the AO baseline, we only report the results of the model trained on GRID for the two-speakers' case. The loss function tends to diverge rapidly when the training is carried out on TCD-TIMIT, as it has more acoustic complexity and does not contain enough audio data to solve the AO-SS task. As expected, the separation performance (SDR) is good, since PIT exploits all the clean sources during training. However, it is still lower than the best AV model. In general, compared to the AO baseline, AV approaches tend to favor speech quality (PESQ and ViSQOL), which is better for AV Concat and AV Concat-ref, and in par for VL2M-ref, although the last shows poorer separation capability.

Finally, we evaluate the systems in a more challenging testing condition where the target utterance is mixed with 2 utterances from 2 competing speakers. Even though the model is trained with mixtures of two speakers, the decrease of performance is not dramatic.

3.6 Event-Driven Camera Approach

So far we have demonstrated that face landmarks motion features are very effective for AV-SE task. Although our systems does not require huge AV datasets at training time, they need to acquire and process every pixel in each frame. However, a large number of pixels does not change throughout the scene and does not carry any information. This leads to a waste of computational resources. In order to overcome this limitation, we propose to use Event-Driven Cameras (EDCs) [96].

EDCs asynchronously measure the brightness change for each pixel, featuring a temporal resolution as high as 1 µs, extremely low latency, and data compression (as only active pixels communicate data). With such sensor, the visual features can be sampled with the same temporal discretization of the auditory pipeline (about 10 ms), removing the need of artificial upsampling which can generate distortions. All these features are desirable to pave the way towards online embedded AV speech processing.

Event-driven vision sensors were already widely used for object tracking [14, 41], detection [58] and segmentation [115], and for gesture recognition [80]. Recently, they were applied for speech processing tasks. In [110] lip movements detected by EDCs were used together with audio features to enable AV voice activity detection. Li et al. [73] proposed to use EDCs for visual speech recognition, i.e., lip-reading, using a DNN architecture.

We present an AV target speaker extraction system on multi-talker environments using event-driven sensors. We substitute the visual pipeline implemented with traditional frame-based cameras, face tracking and extraction of face landmarks motion with an equivalent pipeline based on EDCs, which compute motion at lower latency and computational cost. The work presented in this section has been accepted at the *2021 IEEE International Symposium on Circuits and Systems (ISCAS)* [C].

3.6.1 Methods and Experimental Setup

Since we aim at analyzing the contribution of the new EDC-based processing pipeline, we follow the architecture and the experimental setup of the AV Concat model described in Subsection 3.3.1.

Event-Driven Motion Features Extraction

EDCs output asynchronous events whenever a pixel detects changes in log intensity larger than a threshold. Each event has an associated timestamp, a 2D pixel position, and a polarity (log intensity increase or decrease). The events are emitted with high temporal resolution and low latency, only when there is relative motion between the camera and the scene. Fig. 3.11 shows a graphical reconstruction of a snapshot of a talking face. Only the motion of the person generate events, in particular his mouth and eyes, leading to a low amount of information to process. The different data structure and content from EDCs require specific algorithms for optical flow estimation, that can rely on the precise timing of each event and the continuous observation of the events produced by contrast edges moving from one pixel to its neighbors. State-of-the-art optical flow estimation from event-based data streams is based on deep learning [139]. However, lips and other facial landmarks move mostly horizontally and vertically, then we do not need an extremely complex optical flow estimator. For that reason, we employ another temporally and computationally efficient algorithm that works well in absence of motion in depth [10].

Dataset

We generate an event-driven version of the mixed-speech GRID dataset described in Subsection 3.4.1. The videos are upscaled to 60 fps to have more temporal information and avoid artifacts in the generation of events. The event-based data

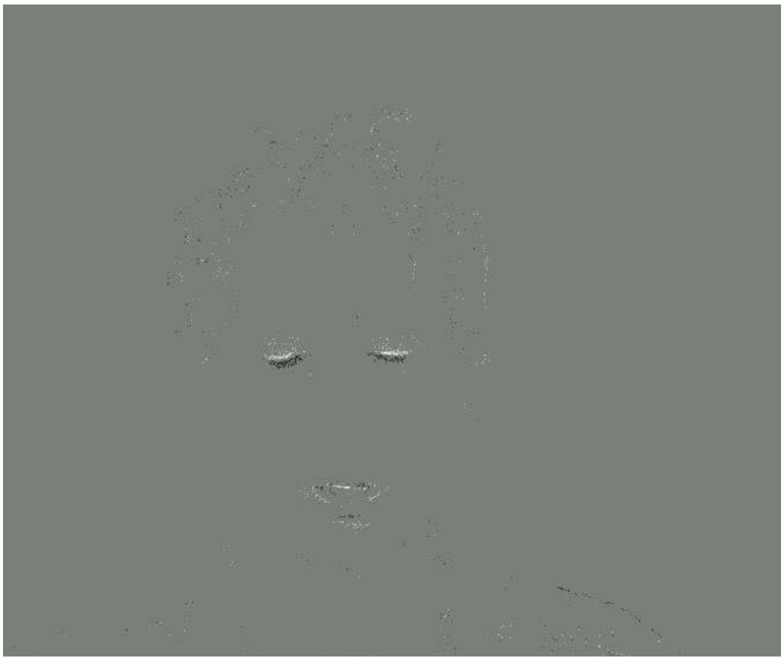


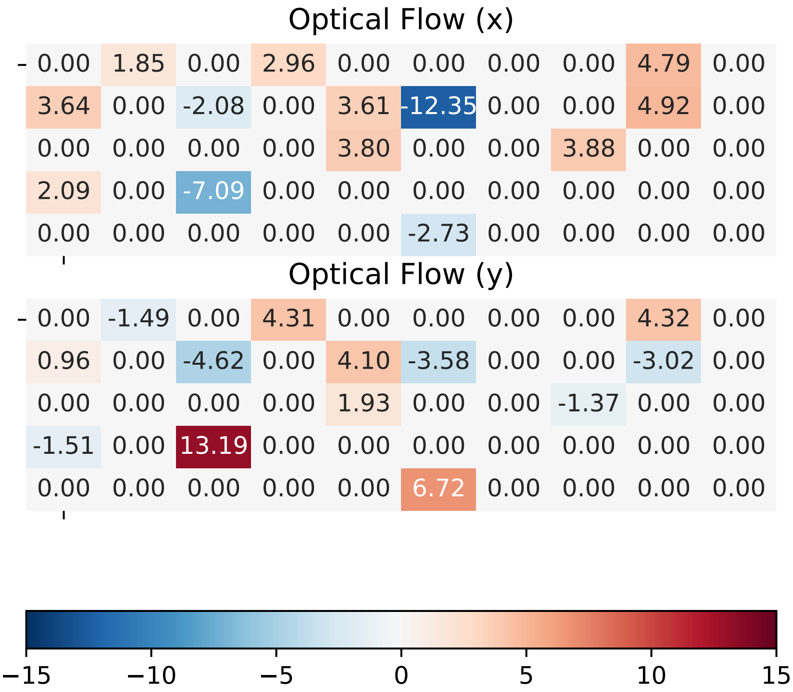
Figure 3.11: Reconstruction of a snapshot of a person talking in front of an EDC.

stream is generated by pointing the ATIS EDC [99] towards a high definition LED monitor while the upscaled videos are played. Due to the low quality of the original videos (360×288 pixels resolution) and in order to preserve the details of lip movements, we crop the mouth area over 100×50 pixels from the event stream.

Video Pre-Processing

To align the visual features to the audio frame rate, we accumulate events every 10 ms and compute the optical flow with the method explained in [10]. The optical flow estimator outputs 2D motion vectors for each event. Therefore, the number of pixels that generates optical flow in each frame can change, originating a different number of video features. Since the BLSTM accepts inputs of the same size for each time step, we apply a transformation that converts the event-based optical flow into fixed-sized regions across the 100×50 pixels.

For each region, we compute three values: the mean of the horizontal (x) and


Optical Flow (x)

0.00	1.85	0.00	2.96	0.00	0.00	0.00	0.00	4.79	0.00
3.64	0.00	-2.08	0.00	3.61	-12.35	0.00	0.00	4.92	0.00
0.00	0.00	0.00	0.00	3.80	0.00	0.00	3.88	0.00	0.00
2.09	0.00	-7.09	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	-2.73	0.00	0.00	0.00	0.00

Optical Flow (y)

0.00	-1.49	0.00	4.31	0.00	0.00	0.00	0.00	4.32	0.00
0.96	0.00	-4.62	0.00	4.10	-3.58	0.00	0.00	-3.02	0.00
0.00	0.00	0.00	0.00	1.93	0.00	0.00	-1.37	0.00	0.00
-1.51	0.00	13.19	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	6.72	0.00	0.00	0.00	0.00

Figure 3.12: Optical flow representation when using regions of 10 × 10 pixels.

of the vertical (y) components of optical flow, and the event-rate. The event-rate is the total number of events on each region at each frame. Fig. 3.12 shows an example of the x and y components of the optical flow for a specific frame.

Optimization

The AV target speaker enhancement model consists of 5 stacked BLSTM layers, with 250 hidden units in each layer. The inputs of the model are the concatenation of the power-law compressed noisy spectrogram and the extracted event-based visual features. The output of the network is the IAM (Eq. 3.5) and the loss function is J_{mr} (Eq. 3.7).

We train the model using the Adam optimizer [66] and 20% of dropout to avoid overfitting. The model is trained up to 500 epochs and early stopping is applied on the validation set when the loss does not decrease for 5 epochs.

3.6.2 Results

We use the SDR [123] and PESQ [106] metrics to measure the separation and the speech quality performance, respectively.

Table 3.4 shows the results from three different models. In the first model we extract visual features using regions of 10×10 pixels. In this case, we obtain 50 regions. Each region produces 3 values, resulting in a total of 150 video features. The model performs on par with the approach based on face landmarks and conventional cameras on PESQ metric, while the SDR is slightly worse, although the separation capability is still good. It is noteworthy that our dataset is not recorded live with subjects, but obtained by recording a movie played back on a high resolution monitor. That results in low quality video and can generate noisy artificial events. Nevertheless, the event-based approach reaches similar performance as the frame-based approach.

In the next experiment we decrease the size of each region to 5×5 in order to have more localized video features. However, the number of input features increases enormously. For that reason, we only use x and y components of the optical flow obtaining 400 video features. Although the results are decent, they are not close to those achieved with 150 input features. Additionally, the computational time increases due to higher input dimensionality.

The main drawback of BLSTMs used in previous experiments is that they need to pass all the features forward and backward before giving a prediction. They can only work in offline scenarios and have higher latency. We implement a causal system using unidirectional LSTM layers with the 150 video features used in the first model. In causal systems the output only depends on past and current inputs,

	SDR	PESQ
Noisy	0.21	1.94
Frame-based approach (face landmarks)	**7.37**	**2.65**
Event-based approach (150 features)	7.03	**2.65**
Event-based approach (400 features)	6.58	2.59
Event-based causal approach (150 features)	3.79	2.22

Table 3.4: GRID results - Event-based approach.

allowing real-time processing. We employ the 150 video features used in the first model. However, the performance of deep forward LSTM is far from that yielded by the BLSTM-based models. This finding suggests that LSTMs are not able to learn powerful AV representations only from past information. Forward LSTMs with look-ahead [121] can be a solution to reduce the gap with BLSTMs, although look-ahead causes a small processing latency. Additionally, other neural network architectures should be investigated to perform online processing.

As a final esperiment, we compare the computation time of the visual features extraction pipelines based on face landmarks and event-driven approaches [2]. The computation of the face landmarks movements for each video file (each video is 3 s long) using Dlib [65] takes 2.980 s on average with 0.825 s standard deviation. On the other hand, for the event-based approach with frame size of 10 ms and 10×10 regions the mean is 1.126 s with 0.212 s standard deviation, almost three times less than the frame-based approach. The computation time of the event-based approach is divided as follows: 0.679 s for optical flow computation and 0.447 s for optical flow-regions mapping. In Fig. 3.13 the computation time of optical flow for different frame sizes is shown. For all the cases the computation time is lower than the frame size by a large margin, leaving enough time to generate visual features, e.g., frame-based optical flow regions, before the next frame arrives without leaks.

[2] Measurements are made with an Intel® Core™ i7-7500U CPU @ 2.70GHz x 4 CPU.

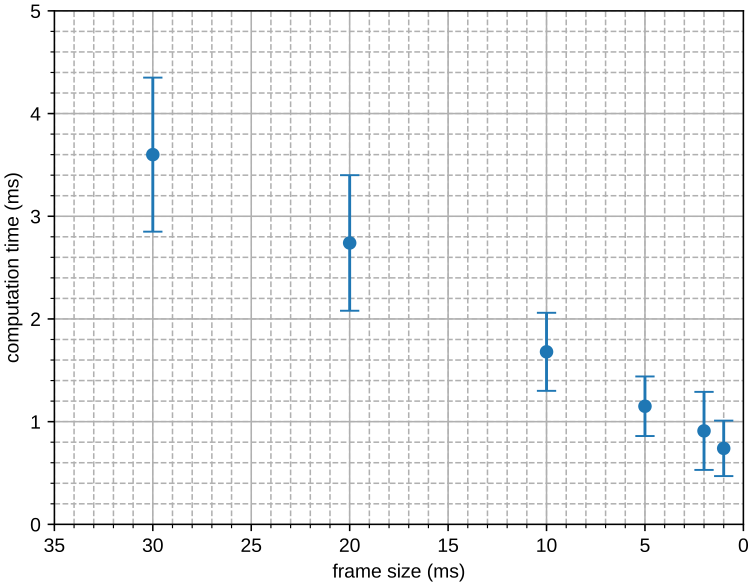


Figure 3.13: Computation time of optical flow for different frame sizes.

3.7 Concluding Remarks

Choosing suitable visual features is of great importance for AV speech processing tasks. In this chapter, we propose the use of face landmarks motion vectors for AV-SE in a single-channel multi-talker scenario. We explore different ways to map face landmarks motion vectors and noisy spectrograms to TF masks that are employed to extract the denoised spectrogram.

Since face landmarks are extracted with an efficient pre-trained model, we relieve our systems to learn useful visual features from huge AV datasets. Indeed, we carry out experiments on two limited size datasets, GRID and TCD-TIMIT. Experiments show that a masking method independent of the acoustic context (VL2M) performs significantly worse than the other approaches (VL2M-ref, AV Concat, AV Concat-ref). The acoustic-aware approaches are able to perform AV-SE in a speaker-independent scenario. In particular, we obtain the best performance with the AV Concat-ref model. It employs a two-stage enhancement where the clean spectrogram estimated from VL2M is further refined by exploiting the acoustic context.

Additionally, in Section 3.6 we propose to substitute the traditional frame-based visual feature extraction pipeline with one based on event-driven signals. The model trained on the event-driven GRID with the new features is almost on par with the one trained with native frame-based features. Besides, experiments demonstrate that event-driven features extraction requires one third of the time compared to face landmarks motion computation. To the best of our knowledge, this is the first study that use event-driven cameras for target speaker enhancement. Our results are very promising and can pave the way towards online AV-SE in embedded devices.

Chapter 4

Joint Audio-Visual Speech Enhancement and Recognition

In this chapter we analyze how the AV-SE models described in Chapter 3 can help to perform the ASR task in a cocktail party scenario. We have presented this study at the *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* [B].

4.1 Introduction

Robust ASR aims at recognizing speech in adverse conditions, where the performance of classical ASR systems can drop significantly. The first robust ASR systems attempt to tackle the problem by processing the input noisy signal through SE phase and passing it to the speech recognizer [89]. Other works follow a multi-task approach where SE/SS and ASR modules are concatenated and jointly trained [20, 90, 130].

The aforementioned works only deal with non-speech environmental noises. Recognizing the speech of a target speaker in multi-talker environments is an even more challenging task. Indeed, it is an ill-posed problem since many different hypotheses about what the target speaker says are consistent with the mixture. Similarly to SE in multi-talker setting, some additional information about the target speaker is needed. Delcroix et al. [27] exploit a reference utterance of the target speaker which is used to inform a DNN-based model about which speaker to extract. Chao et al. [18] and Yu et al. [135] use video as additional information in neural network architectures and obtain large improvements over AO counterparts with overlapped speech input.

We also address this problem by exploiting the visual information, i.e., face landmarks motion, associated to the speaker of interest in order to extract her speech from input mixed-speech signal. The extracted speech is then used as input to a speech recognizer to obtain the phonetic transcription. This task is slightly different from the classic ASR, which aims at generating a word-level transcription. In this chapter, we use the terms ASR and phone recognition interchangeably to denote the same task.

We consider two simple end-to-end LSTM-based models that perform single-channel AV-SE and phone recognition tasks, respectively. In contrast to [18] and [135], we study the interaction between the two tasks and analyze how it is advantageous to train them jointly within a MTL framework [15]. We propose several training strategies that reveal some interesting and unexpected behaviors.

Our systems are applied to a limited data scenarios, which are very common in real-world AV processing applications.

The rest of this chapter is organized as follows. In Section 4.2 we present the joint AV-SE and ASR model, and the different training strategies. The experimental setup is described in Section 4.3. Evaluation results are provided in Section 4.4. We conclude the chapter in Section 4.5.

4.2 System Overview

In this section we present the models used to analyze and study how AV-SE and ASR tasks can be combined to perform phone recognition in a cocktail party scenario. We aim at performing a fair analysis between different training strategies. Therefore, we use very simple and common model architectures based on deep BLSTM. These models are fed with the mixed-speech spectrogram, $Y_s(k, l)$, and/or the motion vectors computed from face landmarks of the speaker of interest, $V(p, l)$.

4.2.1 AV-SE Model

The AV-SE model is developed with the goal of extracting the speech of the target speaker, given the mixed-speech and the visual information of the target speaker. We use an architecture very similar to the AV Concat model described in Subsection 3.3.1.

We denote this model as a function, $\mathcal{F}_{av-se}(Y_s(k, l), V(p, l)) = \hat{X}_s(k, l)$, where $\hat{X}_s(k, l)$ is the estimated target spectrogram. The model consists of a stacked BLSTM and a final Fully Connected (FC) layer that projects the output onto $\mathbb{R}^K$. In order to obtain values in a scale comparable to the SE target, a sigmoid layer is applied and the output is multiplied by $k \cdot \mathbf{d}$, where k is a constant and $\mathbf{d} \in \mathbb{R}^K$ is a vector that contains the standard deviations of each output feature.

The model is trained using the MSE loss, $J_{SE}(X_s(k, l), \hat{X}_s(k, l))$, between the target and the estimated spectrogram.

4.2.2 ASR Model

Similarly to [47], our ASR models consist of end-to-end LSTM-based architectures. All the models are trained with the CTC method [46] which allows direct mapping from acoustic/visual to phonetic sequences. We develop three different versions of the ASR models that differ by the features used as input. We denote the input features as $I^{asr}(f,l)$, with f and l indicating a generic feature index and a time frame index, respectively.

The first version only uses audio input. As acoustic features we employ the mel-scale filter bank representation, $Y_{mel}(j,l)$, derived from the input spectrogram, $Y_s(k,l)$:

$$Y_{mel}(j,l) = \mathbf{W}^{mel} \cdot Y_s(k,l), \tag{4.1}$$

where $\mathbf{W}_{mel} \in \mathbb{R}^{J \times K}$ is the matrix that warps the spectrogram to the mel-filter banks representation.

The second version uses both audio and visual features. In that case, $I^{asr}(f,l)$ consists of a frame-by-frame concatenation of $Y_{mel}(j,l)$ and $V(p,l)$.

The last version of the ASR models is only fed with face landmark motion vectors, $V(p,l)$.

All the models employ the CTC training scheme [46] to map the input, $I^{asr}(f,l)$, to a phone label sequence, whose length is usually different from the number of input time frames. We indicate the generic ASR model as a function, $\mathcal{F}_{asr}(I^{asr}(f,l))$, and the process to obtain the phone label sequence is as follows. We feed the input, $I^{asr}(f,l)$, to a stacked BLSTM. The output of the last BLSTM layer is projected onto $\mathbb{R}^D$ using a FC layer, where D is the size of the phone dictionary. Finally, we apply a softmax layer to obtain a CTC probability mass function, $\hat{\mathbf{l}} = [\mathbf{p}_1(d), \ldots, \mathbf{p}_P(d)]$, where P is the number of phone labels in the utterance.

We use the CTC loss function, $J_{ASR}(\mathbf{l}, \hat{\mathbf{l}})$, to optimize the phone recognition task. The phone transcription is generated using a *beam search* decoder [45] with a beam width of 5.

4.2.3 Joint AV-SE and ASR Model

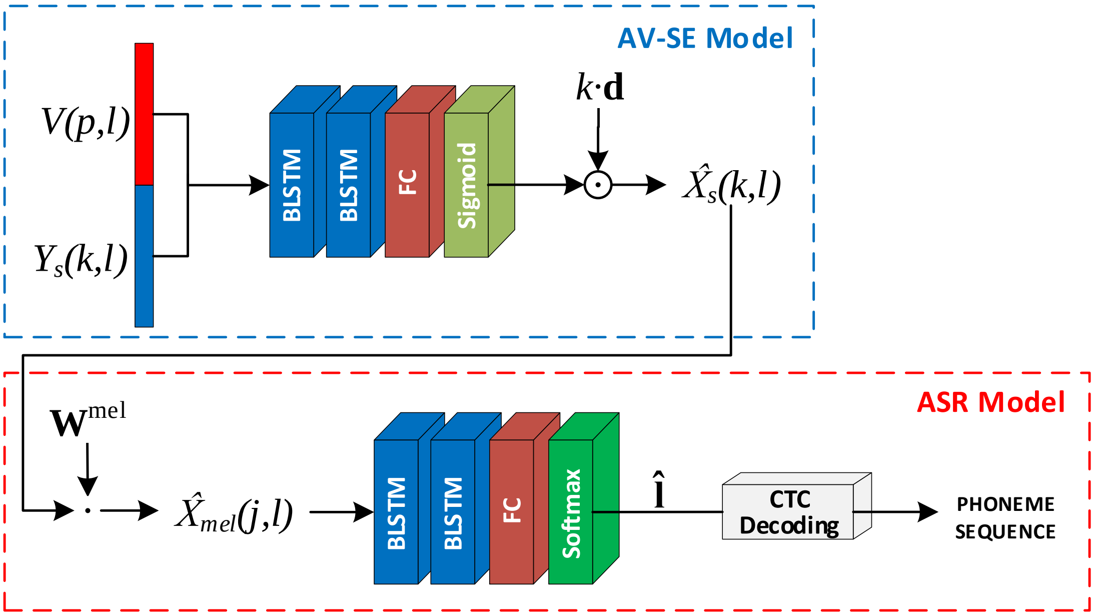


Figure 4.1: Diagram of the joint AV-SE and ASR architecture.

The joint AV-SE and ASR system, $\mathcal{F}_{joint}(Y_s(k,l), V(p,l))$, is depicted in Fig. 4.1. The two models are connected by using the output of the AV-SE model (Subsection 4.2.1) as the input of the ASR model (4.2.2). More formally:

$$\begin{aligned} \hat{\mathbf{l}} &\triangleq \mathcal{F}_{joint}(Y_s(k,l), V(p,l)) \\ &= \mathcal{F}_{asr}(\mathbf{W}^{mel} \cdot \mathcal{F}_{av-se}(Y_s(k,l), V(p,l))). \end{aligned} \tag{4.2}$$

In this architecture the visual information is only exploited by the AV-SE module, $\mathcal{F}_{av-se}(\cdot,\cdot)$, while the ASR module, $\mathcal{F}_{asr}(\cdot)$, is fed with the mel-scaled spectrogram, $\hat{X}_{mel}(j,l) = \mathbf{W}^{mel} \cdot \hat{X}_s(k,l)$, estimated by the AV-SE. The joint model is trained with different MTL schemes, which employ both the $J_{SE}(\cdot,\cdot)$ and the $J_{ASR}(\cdot,\cdot)$ losses. We present the training strategies in the following subsection.

4.2.4 Training Strategies

Our aim is to explore and study the behaviors of the two losses, $J_{SE}(\cdot,\cdot)$ and $J_{ASR}(\cdot,\cdot)$. Therefore, we explore different training techniques to analyze how the losses interact.

The first training technique, henceforth referred to as *joint loss*, attempts to minimize a weighted sum of $J_{SE}(\cdot,\cdot)$ and $J_{ASR}(\cdot,\cdot)$:

$$J_{joint}(X_s, \hat{X}_s, \mathbf{l}, \hat{\mathbf{l}}) = \lambda \cdot J_{SE}(X_s, \hat{X}_s) + J_{ASR}(\mathbf{l}, \hat{\mathbf{l}}), \tag{4.3}$$

where $\lambda \in \mathbb{R}$ is a hyperparameter that defines the weight of the SE task.

During training we observe that the ratio of the two losses significantly changes. To keep both the two losses at the same order of magnitude we do an additional experiment using an adaptive coefficient,

$$\lambda_{ada} = 10^{\lfloor \log_{10}(J_{ASR}) \rfloor - \lfloor \log_{10}(J_{SE}) \rfloor}, \tag{4.4}$$

where the exponent, $\lfloor \log_{10}(J_{ASR}) \rfloor - \lfloor \log_{10}(J_{SE}) \rfloor$, is the difference between the orders of magnitude of the two losses. Therefore, $\lambda_{ada} \neq 1$ only when the two losses do not have the same order of magnitude. This formulation is particularly beneficial in the case one of the two losses, e.g., the CTC loss, has large fluctuations in first training steps, before reaching values comparable to the other loss.

The second training method, *alternated training*, consists of alternation of the SE and ASR training phases. This training procedure performs a few steps of each phase several times. The SE phase uses $J_{SE}(\cdot,\cdot)$ as loss function. In that case, only the parameters of the function $\mathcal{F}_{av-se}(\cdot,\cdot)$ are updated. On the other hand, during the ASR phase the loss function is $J_{ASR}(\cdot,\cdot)$. A particular case of the *alternated training* is the *alternated two full phases training*, where the two phases are performed one time each for a large number of epochs.

In *alternated training* and *alternated two full phases training*, the ASR phase updates the parameters of both the $\mathcal{F}_{av-se}(\cdot, \cdot)$ and the $\mathcal{F}_{asr}(\cdot)$ functions. This configuration can generate many fluctuations in the parameters of the AV-SE model, penalizing the subsequent SE optimization phase. Therefore, for both techniques we develop a *weight freezing* version that minimizes $J_{ASR}(\cdot, \cdot)$ by only updating the parameters of the ASR model.

4.3 Experimental Setup

4.3.1 Datasets

All experiments are carried out using the GRID [24] and TCD-TIMIT [52] AV datasets. We use the mixed-speech speaker-independent versions of these two datasets introduced in Subsection 3.4.1. For each utterance, we add the phone transcription for the target speaker.

For both datasets, the provided text transcriptions are converted to phone sequences using the standard TIMIT [37] phone dictionary, which consists of 61 phones. However, only 33 phones are present in the GRID corpus because of its limited vocabulary. In the TCD-TIMIT corpus all the 61 TIMIT phones are present. Similarly to what is usually done with TIMIT, we apply a mapping operation to the original 61 phones after decoding, and before computing the Phone Error Rate (PER). This operation maps similar phones in the same category, generating a final set of 39 phones.

We follow the same pipelines for audio and video processing described in Subsection 3.4.3 and Subsection 3.4.4, respectively.

4.3.2 Optimization

We train 4 different baseline ASR models, each of them fed with a different input: *(i)* clean mel-scaled spectrogram, $X_{mel}(j,l)$; *(ii)* mixed-speech mel-scaled spectrogram, $Y_{mel}(j,l)$; *(iii)* frame-by-frame concatenation of mixed-speech mel-scaled spectrogram, $Y_{mel}(j,l)$, and visual features, $V(p,l)$; *(iv)* visual features, $V(p,l)$. All the ASR models consist of 2 stacked BLSTM layers with 250 units and are trained using Back-Propagation Through Time (BPTT).

For what concerns the joint model, we use 2 BLSTM for both AV-SE and ASR modules. Each layer consists of 250 hidden units with hyperbolic tangent (tanh) activation function. We employ the same set of hyperparameters used by the ASR models in order to carry out a fair evaluation of the benefits of the AV-SE stage. We perform a limited random search-based hyperparameter tuning, therefore all reported results may be slightly improved.

In addition, we create a robust AO baseline. We substitute the AV-SE model with an AO-SS model trained with an utterance-level PIT (uPIT) [68] MSE loss. Since the AO-SS model outputs all the sources in the mixture, we only pass to the ASR model the source with the lowest MSE compared to the target speech.

In all experiments, the models are trained using the Adam optimizer [66], setting the initial learning rate to 0.001. Early stopping is applied when the error on the validation set does not decrease over 5 consecutive epochs.

	GRID	TCD-TIMIT	
Training Method	PER	PER-61	PER-39
ASR Model			
- Noisy Audio	0.494	0.784	0.713
- Noisy Audio + Video	0.499	0.772	0.709
- Video	0.294	0.786	0.747
Joint AV-SE/ASR Model			
- Joint loss	0.154	0.531	0.477
- Alternated two full	0.160	0.456	0.412
- Alternated two full w. freezing	0.187	**0.443**	**0.400**
- Alternated	**0.139**	0.449	0.406
- Alternated w. freezing	0.181	0.613	0.555
- Alternated AO uPIT	0.433	0.671	0.624

Table 4.1: PER scores on GRID and TCD-TIMIT datasets. For the joint loss strategy, the reported results are obtained by using $\lambda = 1$ for GRID, and the λ_{ada} for TCD-TIMIT. The ASR baseline model trained and tested on clean audio of GRID reaches a PER of 0.058. For TCD-TIMIT, the PER scores are 0.467 (61 phones) and 0.406 (39 phones).

4.4 Evaluation Results

4.4.1 Results

Table 4.1 reports PER of all baseline models and of the joint models with different training strategies. For PER metric, lower values correspond to better performance.

The results reveal that performing the ASR task on TCD-TIMIT is much more challenging compared to GRID. This behavior can be explained by several factors. The GRID corpus has a smaller vocabulary size (52 words) and its sentences are more constrained. Additionally, TCD-TIMIT consists of variable-length utterances, while in GRID all utterances are 3 s long. The lower complexity of the GRID dataset is confirmed by the good performance of the ASR baseline model trained on clean speech, which reaches a PER of 0.058 on the clean test set. Unlike

GRID, the ASR model trained and tested on the clean samples of TCD-TIMIT performs significantly worse, obtaining a PER score of 0.467 and 0.406 on evaluations with 61 and 39 phones, respectively.

In both datasets, the joint models significantly outperform the ASR baselines models trained with noisy speech and/or visual inputs. In particular, the *alternated training* reaches the best results in GRID, while in TCD-TIMIT it is slightly outperformed by the *alternated two full phases training* with *weight freezing*.

Finally, we observe that the performance of the joint model with the AO uPIT separation module is far behind the best joint AV-SE and ASR approaches. This result confirms that vision provides an important contribution in recognizing the speech of a speaker of interest with mixed-speech input.

4.4.2 Training Analysis

In this subsection, we analyze the trends of the $J_{SE}(\cdot,\cdot)$ and $J_{ASR}(\cdot,\cdot)$ losses during training. The loss curves computed on the validation set of the GRID dataset are depicted in Fig. 4.3, 4.2, and 4.4. We observe similar trends in TCD-TIMIT.

Fig. 4.2 shows the trends of the two losses when the system is trained with the *joint loss* method. We experiment with different λ values and the adaptive λ_{ada} (see Eq. 4.4). In general, when the training starts the two losses both decrease, and subsequently the ASR loss tends to increase. For higher values of λ, which favor the AV-SE task, the ASR loss diverges rapidly. The use of the adaptive λ_{ada} relieves us from tuning manually the λ value. However, this approach does not always lead to the best performance. Indeed, the best performing models on the test set of GRID are obtained with $\lambda = 1$. The *joint loss* training shows the interesting property of obtaining good results for both the SE and ASR metrics. Nevertheless, its ASR capability, which is the main objective of our models, turns out to be lower compared to the *alternated training* methods.

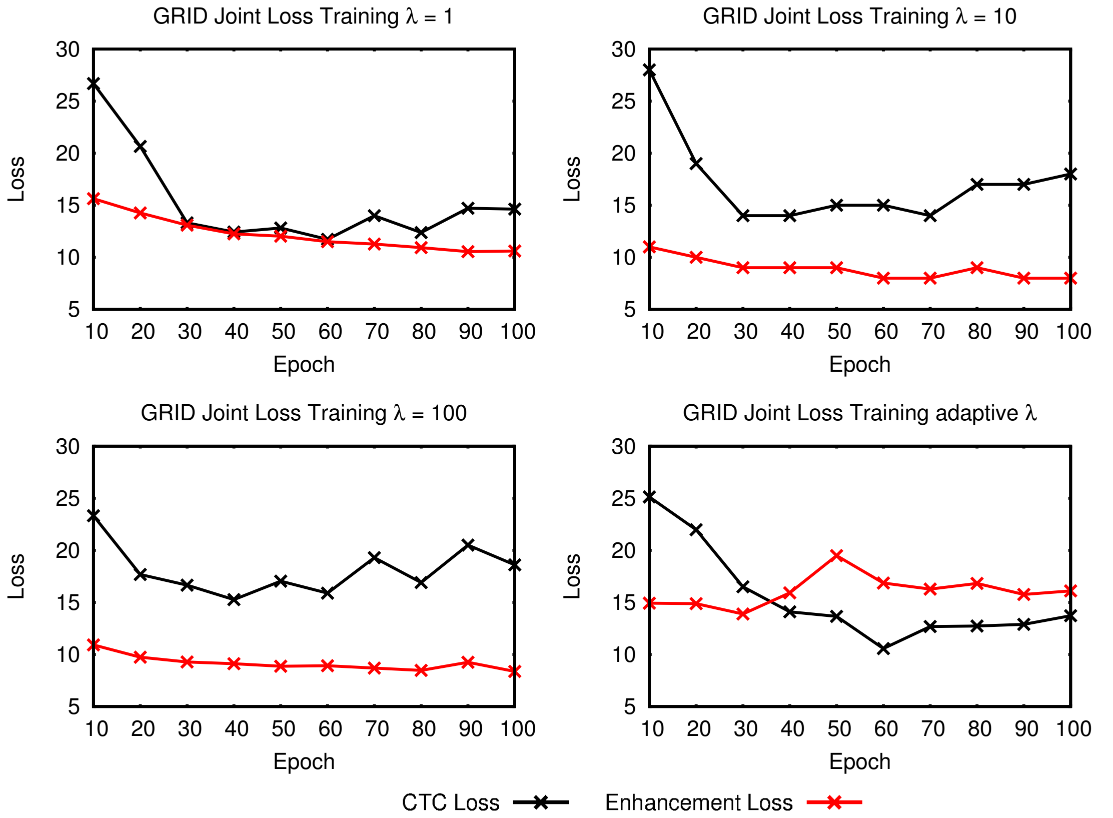


Figure 4.2: Trend of the two losses during training by using the *joint loss* with different λ values.

The first diagram in Fig. 4.3 shows the curves of the *alternated two full phases training*. Initially, this method optimizes the AV-SE model using the $J_{SE}(\cdot,\cdot)$ loss, until it reaches a plateau. The minimization of the ASR model starts from epoch 90, and involves the parameters of both AV-SE and ASR models. We observe that the SE loss remarkably diverges in few epochs. Therefore, the enhanced representation obtained by the AV-SE is not optimal to perform the phone recognition task. The *alternated two full phases training* with *weight freezing* confirms this result. In that case, the parameters of the AV-SE model are forced to not change during the ASR training phase. Although the $J_{SE}(\cdot,\cdot)$ loss does not diverge, the $J_{ASR}(\cdot,\cdot)$ loss is higher than the previous case. The second diagram in Fig. 4.3

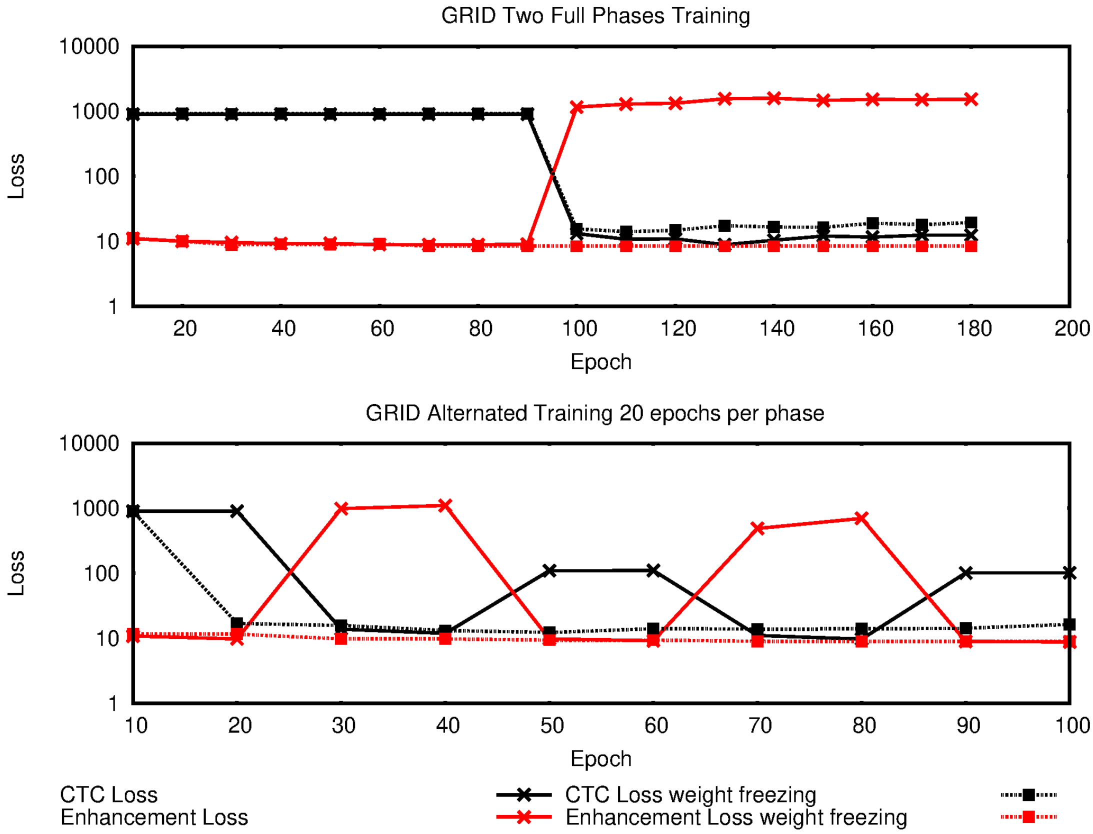


Figure 4.3: Trend of the two losses during training with and without *weight freezing* by using the *alternated two full phases training* and *alternate training.*

demonstrates that we can draw similar conclusions when we alternate the AV-SE and ASR phases every 20 epochs. This behavior is also reported in previous work [20, 90, 130]. However, we do not expect such divergence when we switch from AV-SE to ASR phases.

Experimental results obtained by the *alternated training* with different numbers of epochs per phase are showed in Fig. 4.4. In all cases, the decrease of the $J_{ASR}(\cdot,\cdot)$ loss coincides with a large increase of the value of $J_{SE}(\cdot,\cdot)$, and vice versa. Moreover, every repetition of the two phases gradually reduces the ASR loss. This finding suggests that the interaction between the two losses can be explored to get better performance. Finally, we analyze the *alternated training*

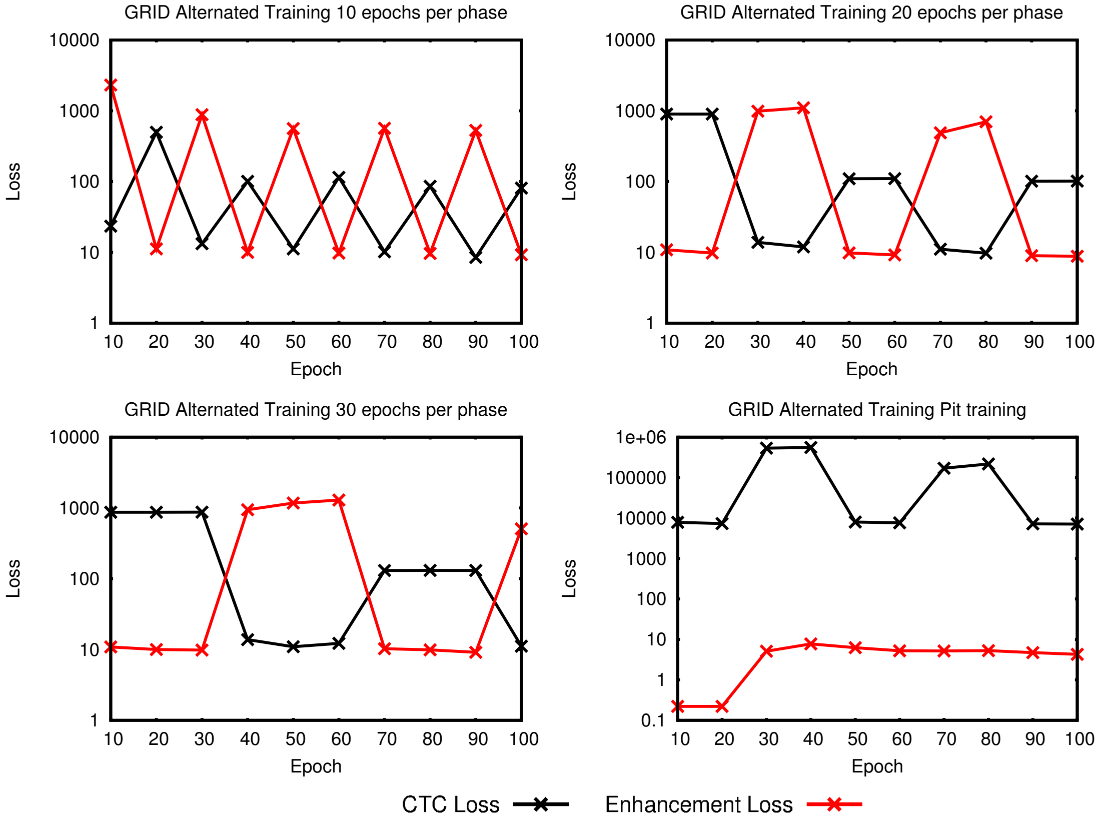


Figure 4.4: Trend of the two losses with *alternated training*, by using different number of epochs per phase.

with the AO uPIT-based enhancement. Contrary to AV methods, the uPIT loss significantly diverges in the first ASR phase and does not decrease anymore during the subsequent SE phases.

4.5 Concluding Remarks

In this chapter we study how single-channel AV-SE can help phone recognition when several people are talking simultaneously. The AV-SE and the phone recognition tasks are implemented using two end-to-end LSTM-based models. Additionally, we experiment with the limited size GRID and TCD-TIMIT datasets.

The analysis unveils that jointly minimizing the SE loss and the ASR loss may not the best strategy to improve ASR. Then we explore the trends of the loss functions when the training strategy consists of an alternation of the SE and ASR training phases. We observe that the loss function that is not considered for the training phase tends to diverge.

Finally, we find that the interaction between the two loss functions can be exploited in order to obtain better results. In particular, the *alternated training* method shows that PER can be gradually reduced by wisely alternating the two training phases.

Chapter 5

Audio-Visual Speech Inpainting

This chapter presents a deep learning-based framework for Audio-Visual Speech Inpainting (AV-SI), i.e., the task of restoring the missing time gaps of a speech signal from uncorrupted audio context and visual information. The study described in this chapter will be presented at the *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* [D] [3].

5.1 Introduction

In Chapter 3 we have dealt with the problem of extracting the speech of a target speaker from an input signal degraded by concurrent speakers' speech. One drawback of the proposed models it that they are not able to recover the target speech when the input information is completely lost. However, the use of vision might be useful in restoring the missing parts of an acoustic speech signal. Visual information has been successfully used in many speech-related tasks, such as ASR, SE, SS, etc. (cf. [85, 140] and references therein), but it has not been adopted for SI yet.

[3] Some testing examples of our SI models are available at `https://dr-pato.github.io/audio-visual-speech-inpainting/`.

Previous work try to solve the audio inpainting task both with traditional methods [2, 8, 97, 114, 132] and deep learning models [17, 28, 64, 78, 79]. In particular, the first deep learning-based SI systems are proposed by Chang et al. [17] and Kegler et al. [64]. They use convolutional encoder-decoder architectures to generate speech from masked speech signals. Zhou et al. [138] show that visual cues improve inpainting of music signals, but their models are not evaluated on speech data.

In this chapter we propose a first attempt to inpaint lost speech information by exploiting visual features, i.e., the face landmarks motion features presented in Chapter 3. We only focus on the case of restoring entire missing time segments because it is the most general case, and the one that has a direct use-case application, i.e., packet-loss. Indeed, other kinds of information loss, e.g., masking of irregular TF regions, do not occur frequently in real-world scenarios.

Our AV-SI systems are based on RNN architectures. They are able to generate new information and are designed to fill arbitrarily long missing gaps with coherent and plausible signals. In addition, we present a MTL [15] strategy where a phone recognition task is learned together with SI. The motivation of the MTL approach lies in previous work, which show that ASR can improve not only SE [33], but also speech reconstruction from silent videos [84].

The rest of the chapter is organized as follows. We first provide a formulation of the SI task in Section 5.2. The proposed algorithms are described in Section 5.3. Then the experimental setup is introduced in Section 5.4. In Section 5.5 results and comparisons are presented. We conclude the chapter in Section 5.6.

5.2 Problem Formulation

As done in previous work [2], we assume to know a priori the location of uncorrupted and lost data, and we use this information in the signal reconstruction stage. This scenario is referred to as *informed* SI.

Let $Y_s(k,l)$ denote the spectrogram of an observed acoustic speech signal, i.e., speech signal with missing parts. The information about the location of missing portions of the signal is encoded in a binary mask, $I(k,l)$, which indicates whether a TF bin of the spectrogram of the observed signal is lost, $I(k,l) = 1$, or uncorrupted, $I(k,l) = 0$. We assume that $Y_s(k,l) = 0$ if $I(k,l) = 1$.

We define the problem of Audio-Only Speech Inpainting (AO-SI) as the task of finding a function, $\mathcal{F}_{a-si}$, that estimates the spectrogram of the ground-truth speech signal, $X_s(k,l)$, from $I(k,l)$ and $Y_s(k,l)$.

When visual features, $V(p,l)$, are available, the task is denoted as AV-SI. In that case, the function $\mathcal{F}_{av-si}$ has to estimate $X_s(k,l)$ given $I(k,l)$, $Y_s(k,l)$, and $V(p,l)$.

5.3 System Overview

Fig. 5.1 shows the architecture of the proposed AV-SI system. The general system is described in the following subsection, while the MTL extension is presented in Subsection 5.3.2.

5.3.1 Model Architecture

We use a neural network architecture, indicated as a function, $\mathcal{F}_{av-si}(\cdot,\cdot,\cdot)$, to estimate the original spectrogram, $X_s(k,l)$. As audio and video input features we use audio context spectrogram, $Y_s(k,l)$, and face landmarks motion vectors,

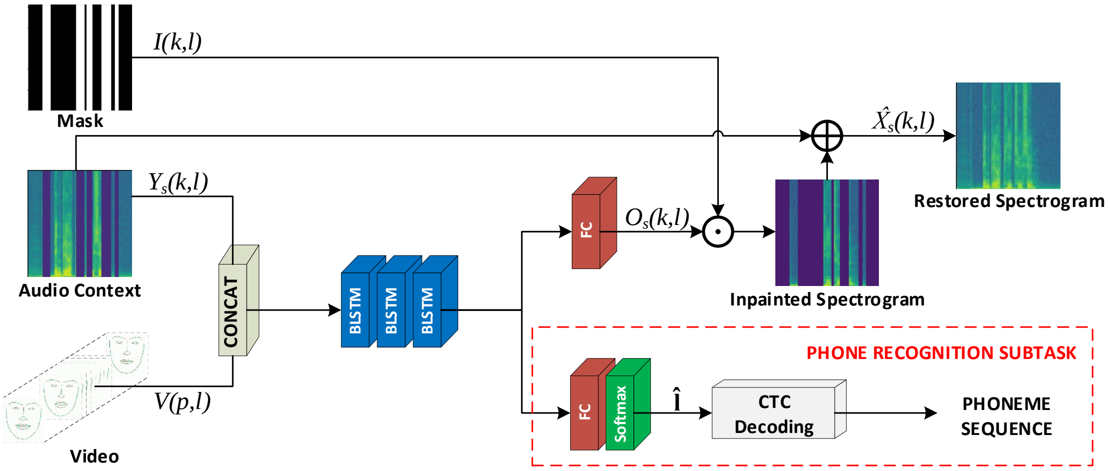


Figure 5.1: Overall architecture of the AV-SI system.

$V(p,l)$, respectively. The AV features are concatenated frame-by-frame and used as input of a stacked BLSTM that models the sequential nature of the data [47]. Then, a Fully Connected (FC) layer is fed with the output of the stacked BLSTM and outputs the inpainted spectrogram $O_s(k,l)$. To extract the inpainted spectrogram within the time gaps, $O_s(k,l)$ is element-wise multiplied with the input mask, $I(k,l)$. Finally, the fully restored spectrogram, $\hat{X}_s(k,l)$, is obtained by an element-wise sum between the observed input spectrogram, $Y_s(k,l)$, and the inpainted spectrogram. More formally:

$$\begin{aligned}\hat{X}_s(k,l) &\triangleq \mathcal{F}_{av}(I(k,l),Y_s(k,l),V(p,l))\\ &= O_s(k,l)\odot I(k,l)+Y_s(k,l),\end{aligned} \tag{5.1}$$

where $\odot$ is the element-wise product.

The model is trained to minimize the MSE loss, $J_{mse}(\cdot,\cdot)$, between the inpainted spectrogram, $\hat{X}_s(k,l)$, and the ground-truth spectrogram, $X_s(k,l)$:

$$J_{mse}(X_s, \hat{X}_s) = \frac{1}{N} \sum_{k=1}^{K} \sum_{l=1}^{L} (\hat{X}_s(k,l) - X_s(k,l))^2, \tag{5.2}$$

where N is the number of missing TF bins.

5.3.2 Multi-Task Learning with CTC

In addition to the plain AV-SI model, we devise a MTL approach, which attempts to perform SI and phone recognition simultaneously. Our MTL training makes use of a CTC loss [46] which is very similar to the one presented in [84] for the task of speech synthesis from silent videos. The block bounded by red dashed lines in Fig. 5.1 performs the *phone recognition subtask*. It is fed with the stacked BLSTM units' output and has a linear FC layer followed by a softmax layer which outputs a CTC probability mass function $\hat{\mathbf{l}} = [\mathbf{p}_1(d), \ldots, \mathbf{p}_P(d)]$, with $d \in [1, D]$, where D is the size of the phone dictionary and P is the number of phone labels in the utterance.

The MTL loss function is a weighted sum between the inpainting loss, $J_{MSE}(\cdot,\cdot)$, and the CTC loss, $J_{CTC}(\cdot,\cdot)$:

$$J_{MTL}(X_s, \hat{X}_s, \mathbf{l}, \hat{\mathbf{l}}) = J_{MSE}(X, \hat{X}_s) + \lambda \cdot J_{CTC}(\mathbf{l}, \hat{\mathbf{l}}), \tag{5.3}$$

with $\lambda \in \mathbb{R}$, where $\mathbf{l}$ is the sequence of ground truth phone labels. λ is a hyperparameter that controls the importance of the CTC loss in the overall cost function.

5.4 Experimental Setup

5.4.1 Dataset

We carry out our experiments on the GRID corpus [24], which we have introduced in Subsection 3.4.1 and in Subsection 4.3.1.

We generate a corrupted version of the dataset where random missing time gaps are introduced in the audio speech signals. Our models are designed to recover multiple variable-length missing gaps. Indeed, for each signal we draw the amount of total lost information from a normal distribution with a mean of 900 ms and a standard deviation of 300 ms. The total lost information is distributed between 1 to 8 time gaps and each time gap is randomly placed within the signal. To avoid unrealistic very short gaps, we set their minimum length to 36 ms. In addition, we assure the presence of audio context by limiting the total duration of the missing gaps to 2400 ms. Similarly to [64], the information loss is simulated by applying binary TF masking to the original spectrograms. The generation process is the same for training, validation, and test sets.

We evaluate our systems in a speaker-independent setting, with 25 speakers (s1-20, s22-25, s28) used for training, 4 speakers (s26-27, s29, s31) for validation, and 4 speakers (s30, s32-34) for testing. Furthermore, to evaluate the effect of the gap size, we generate additional versions of the test set, each of them containing a single gap of fixed size (100/200/400/800/1600 ms).

5.4.2 Optimization

The AV-SI models consist of 3 BLSTM layers, each of them with 250 units. The Adam optimizer [66] is used to train the systems, setting the learning rate to 0.001. We feed the models with mini-batches of size 8 and apply early stopping, when the validation loss does not decrease over 5 epochs. The λ weight of the MTL loss, $J_{mtl}(\cdot, \cdot, \cdot, \cdot)$, is set to 0.001. All the hyperparameters are tuned by using a random search and the best configuration in terms of the MSE loss, $J_{mse}(\cdot, \cdot)$, on the validation set is used for testing.

In addition, we implement an AO-SI baseline model by simply removing the video modality from the AV model, leaving the rest unchanged. We consider AO models both with and without the MTL approach described in Subsection 5.3.2. The baseline obtains slightly better results than a state-of-the-art AO-SI system [64]. In that work, the problem of AO-SI was tackled using an end-to-end deep learning model with the U-Net architecture [107], which was trained on the corrupted LibriSpeech corpus [95].

In order to assess the exact contribution of vision, we train a video-only SI system. Similarly to the AO-SI baseline, we discard the audio modality from the full AV model and analyze both the plain and the MTL approaches. Although this model can be regarded as a video-to-speech synthesizer, it differs from the systems presented at the end of Section 2.3, as it aims at generating only the missing time segments.

5.4.3 Pre- and Post-Processing

The original waveforms are downsampled to 16 kHz. A STFT is computed using a FFT size of 512 with Hann window of 384 samples (24 ms) and hop length of 192 samples (12 ms). Then, we compute the logarithm of the STFT magnitude and apply normalization with respect to global mean and standard deviation to get the acoustic input features.

The missing phase is recovered by applying the Local Weighted Sum (LWS) algorithm [71] to the restored spectrogram, setting the available context phase as the starting step. The LWS algorithm is a very efficient approximation of the Griffin-Lim algorithm [48], where the most consistent phase for a given magnitude spectrogram is obtained by repeatedly computing STFT and inverse STFT. Finally, we use the inpainted spectrogram and the estimated phase to compute the inverse STFT, which reconstructs the inpainted speech waveform.

We follow the pipeline described in Subsection 3.4.4 to extract the video features, i.e., 68 facial landmarks motion vectors. We upsample the video features from 25 to 83.33 fps to match the frame rate of the audio features.

During the testing of the MTL models, the estimated phone distribution is used to generate the best phone sequence. We find the phone transcription applying *beam search* decoding [45] with a beam width of 20.

5.5 Evaluation Results

5.5.1 Evaluation Metrics

We evaluate the systems using L1 loss, and two perceptual metrics, Short-Time Objective Intelligibility (STOI) [118] and PESQ [106], which provide an estimation of speech intelligibility and speech quality, respectively. Additionally, we evaluate the Phone Error Rate (PER) obtained with an AO phone recognizer trained on the original audio of the GRID dataset. The phone recognizer consists of 2 BLSTM layers (250 units) followed by a FC and a softmax layers. PER score of the ground-truth speech in the test set is 0.069.

While the L1 loss is computed only on the masked parts of the signals, the other three metrics are applied to the entire signals, as it is not possible to perform the evaluation on very short segments. Obviously, PER, STOI, and PESQ show lower sensitivity, when the masked part is small ($<$ 400 ms), since a large fraction of the original signal is unchanged in that case.

For L1 and PER, the lower their values the better, while for STOI and PESQ higher values correspond to better performance.

Audio	Video	MTL	L1 ▼	PER ▼	STOI ▲	PESQ ▲
Unprocessed			0.838	0.508	0.480	1.634
✗			0.482	0.228	0.794	2.458
	✗		0.549	0.162	0.787	2.379
✗	✗		0.452	0.151	0.811	2.506
✗		✗	0.476	0.214	0.799	2.466
	✗	✗	0.540	0.154	0.793	2.393
✗	✗	✗	**0.445**	**0.137**	**0.817**	**2.525**

Table 5.1: Results of the SI systems on the test set. The "Unprocessed" row refers to the evaluation values of the input corrupted speech.

5.5.2 Results and Discussion

The evaluation results of the proposed models on the test set are reported in the Table 5.1. On average, the masking process discards about half of the original speech information, as confirmed by the PER score of unprocessed data.

AV models outperform the AO counterparts on all metrics, demonstrating that visual features provide complementary information for SI. In particular, the PERs of AV models are lower by a considerable margin, meaning that generated signals are much more intelligible. The improvements in terms of STOI and PESQ are not as large as PER, mainly because the two perceptual metrics are less sensitive to silence than PER. Nonetheless, they are significantly better than the unprocessed data scores confirming the inpainting capability of our models.

Video-only models show an interesting behavior. They perform poorly on L1 and PESQ metrics compared to AO systems. This is motivated by the absence of audio context, which is crucial to allow good speech quality and signal reconstruction. On the other hand, they perform almost on par on STOI, and PER is way lower. These findings suggest that the major contribution of vision regards speech intelligibility rather than speech quality. The AV models can benefit from both audio and video modalities to improve speech quality and intelligibility, respectively.

The MTL strategy is also beneficial. Indeed, exploiting phonetic data during the training process is useful to improve the accuracy of SI. However, we observe just a small improvement of the AV-MTL model over the plain AV one. This might be explained by the fact that, unlike for the AV system, MTL strategy does not add any additional information at the inference stage. Note that the sentences in the GRID corpus follow a very constrained syntax, which can be easily learned by adding a phone recognition subtask. Therefore, the small benefits of the MTL might disappear with bigger and more complex datasets.

5.5.3 Gap Size Analysis

Table 5.1 reports the average results using multiple variable-length time gaps, not providing information about how the gap size affects the SI capability of our models. For this reason, we generate other test sets, each of them containing samples with a single time gap of fixed length (100/200/400/800/1600 ms). Fig. 5.2 shows the inpainting results for each metric on these test sets. As expected, while for short gaps ($\leq$ 200 ms) AO and AV models reach similar performance, their difference rapidly increases when missing time gaps get larger. The performance of AO models drops significantly with very long gaps ($\geq$ 800 ms). Therefore, the audio context does not contain enough information to correctly reconstruct missing audio signals without exploiting vision.

The diagram of L1 loss shows that video-only methods are very poor at inpainting very short segments compared to other models, although they are less affected by changes of missing gap lengths. In terms of PER, video-only models outperform AV models with 800 and 1600 ms gaps, confirming that vision provides a huge contribution in speech intelligibility in presence of extremely long gaps.

Regarding the models trained with the MTL approach, we can notice a good

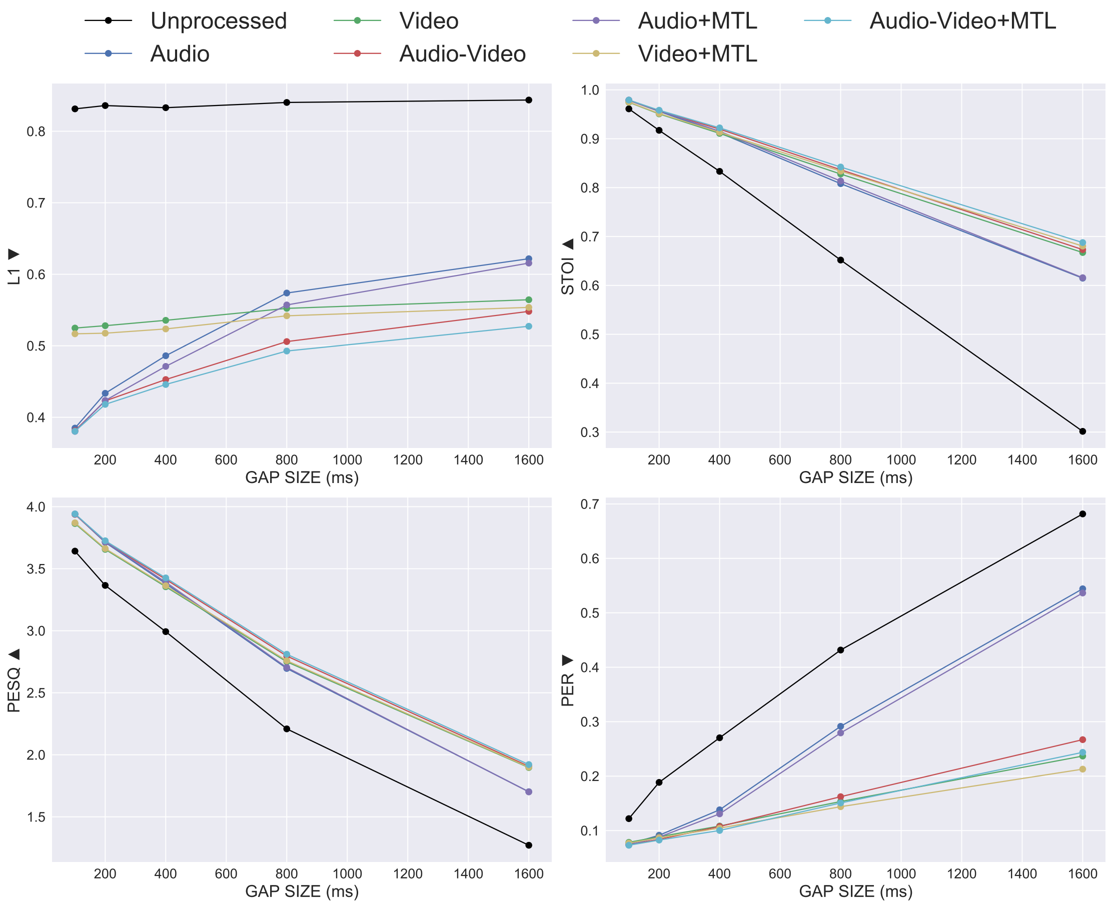
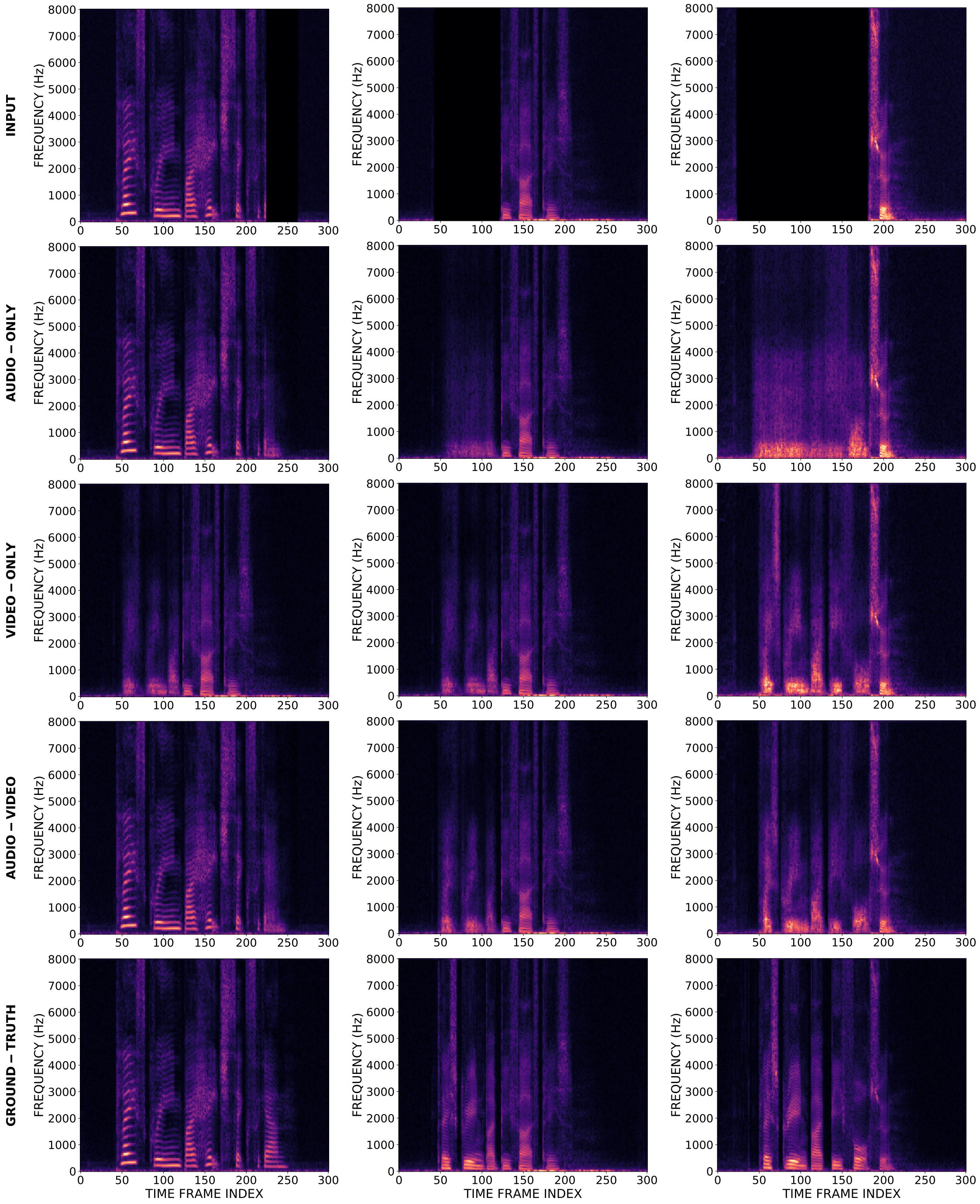


Figure 5.2: Effect of gap size on SI performance.

improvement in terms of L1 loss and PER, even if the contribution is not as high as the one provided by the visual modality.

Some examples of spectrograms inpainted by AO, video-only, and AV models trained with MTL are shown in Fig. 5.3. We do not present the examples with short gaps ($\leq$ 200 ms) since the performance of AO and AV models are similar. In general, AO models inpaint long gaps with stationary signals, whose energy is concentrated in the low frequencies. On the other hand, AV models are able to generate well-structured spectrograms, demonstrating the benefit that visual features bring to inpaint long gaps. However, they produce blurred TF regions,



(a) 400 ms gap.

(b) 800 ms gap.

(c) 1600 ms gap.

Figure 5.3: Examples of spectrogram inpainted by SI models trained with MTL.

resulting in robotic speech, although it is still intelligible. Finally, the spectrograms generated by vision-only models are very similar to AV ones, but sometimes they can fail in presence of shorter time gaps (see Fig. 5.3a).

5.6 Concluding Remarks

In addition to SE and SS, restoring missing information from audio signals, i.e., SI, is crucial to enable robust audio processing applications.

This chapter proposes the use of visual information, i.e., face landmark motion, for SI. We test our LSTM-based models on a speaker-independent setting using the GRID dataset and demonstrate that AV models strongly outperform their AO counterparts. In particular, the improvement due to the visual modality increases with duration of time gaps. Finally, we show that learning a phone recognition task together with the inpainting task leads to better results.

At the best of our knowledge, this is the first study that addresses the AV-SI problem. Our AV models can generate intelligible speech, although it sounds robotic. Future work will investigate other AV fusion techniques which make better use of the audio context in presence of long gaps. Furthermore, the audio context can be exploited to explicitly extract speaker characteristics which are useful to generate more natural speech.

Chapter 6

Conclusions and Future Work

6.1 Conclusions and Contributions

Background noisy speech and audio signal loss are two common distortions to speech signals in daily listening environments. These distortions can compromise an effective communication among people and between humans and machines. Speech perception is a multi-modal process, where visual information provides an important contribution to speech intelligibility. In this dissertation, we have presented several systems based on deep learning architectures which exploit both audio and visual modalities to improve the performances of speech-related tasks in adverse conditions.

In Chapter 3 we have proposed the use of face landmarks motion vectors for audio-visual speech enhancement in a single-channel multi-talker scenario. Different models are tested where face landmarks motion vectors are used to generate TF masks that extract the target speaker's spectrogram from the acoustic mixed-speech spectrogram. To the best of our knowledge, some of the proposed models are the first models trained and evaluated on the limited size GRID [24] and TCD-TIMIT [52] datasets that accomplish speaker-independent SE in the multi-talker setting, with a quality of enhancement comparable to that achieved in a speaker-dependent setting. In Section 3.6 we have described how to adapt our approach to work with visual features extracted with event-driven cameras in place of conventional frame-based features.

Motivated by the effectiveness of our audio-visual speech enhancement systems, in Chapter 4 we have studied how an audio-visual enhancement front-end can be combined with ASR to improve recognition accuracy in a multi-task learning framework. Surprisingly, the joint optimization of the two tasks does not lead to the best results. We explore the trends of the speech enhancement and ASR loss functions when the training strategy consists of an alternation of the enhancement and recognition training stages. We note that the loss function that is not considered for the training phase tends to diverge. However, the interaction between the two tasks can be exploited. In particular, the *alternated training* method shows that recognition error can be gradually reduced by wisely alternating the two training phases.

To deal with lost speech segments, in Chapter 5 we have proposed the use of visual information for speech inpainting. We test our models on a speaker-independent setting using the GRID [24] dataset and demonstrate that audio-visual models strongly outperform their audio-only counterparts. In particular, the improvement due to the visual modality increases with duration of time gaps. Finally, we show that learning a phone recognition task together with the inpainting is effective, although the largest contribution to performance comes from vision.

6.2 Future work

This dissertation explores different ways to use audio-visual information in speech processing tasks. In the last few years this research area has received growing attention from speech and computer vision communities. Even though we have witnessed a clear breakthrough in the field, a lot of problems remain unsolved. In the following, we suggest some possible future research directions:

- *Face landmarks features.* We propose to use face landmarks motion features for several speech-related tasks. However, we employ a landmarks extractor [63] which only works for frontal faces. Therefore, the performance of our systems drops in more challenging scenarios. This problem can be addressed by using different face detectors and landmarks extractors, which are able to deal with faces in profile or even partially occluded. Besides, face landmarks can be normalized with several geometric transformations, in order to move the landmarks points according to a fixed head pose (e.g., roll, yaw, pitch). Such transformations should decrease the variability of input visual features, reducing the need of complex models and of extremely large training datasets.

- *Unsupervised learning.* The majority of audio-visual speech enhancement and separation systems are based on supervised learning methods. Supervised models need both clean and noisy speech at training time. Since it is very difficult to collect clean and noisy signals pairs, they are usually synthetically generated by summing the clean speech with several kind of noises at various SNRs. This approach can reach excellent generalization performance, especially when the systems are trained with a large amount of data. However, it treats target speech and noises as independent signals, which does not reflect a real-world scenario (see *Lombard effect* [13]). In addition, the collection of clean videos during real conversations represents a big issue due to disturbing sources that are present in everyday environments. Audio-visual speech recognition systems presents similar issues. Moreover, the availability of audio-visual datasets with word transcriptions is very limited compared to audio-only counterparts, posing an important limitation to large-scale training of audio-visual speech recognition models. Unsupervised learning can be a solution to these drawbacks. Learning to enhance and recognize speech directly from real noisy recordings would enable a major advancement in the field.

- *Low latency and low energy processing.* All the systems developed in this dissertation operate in an offline fashion. Although it is acceptable for some applications, like keyword spotting or video editing, most applications in daily life require nearly real-time processing. Additionally, many devices have to meet power consumption requirements. For example, hearing aids need to guarantee low latency performance using low processing power and limited storage. However, deep neural networks generally requires millions of parameters and computations, resulting on high computational and memory consumption. This challenge is even bigger for audio-visual systems, which have to deal with visual data. Indeed, video has a higher dimensionality compared to audio. Some possible directions include better model design to reduce redundancy in computation and to decrease the number of parameters maintaining the same performance. The use of event-driven cameras is an early attempt to optimize the computational load of visual stream. Follow-up work will focus on the collection of native AV event-driven datasets. Additionally, different neural network architectures, e.g., spiking neural networks [39], can be exploited to process event-driven data directly. In this way, there is no need to convert events to frame-based data, resulting in time and computational savings.

- *Joint speech enhancement and recognition.* In Chapter 4 we have described two simple BLSTM to perform speech enhancement and phone recognition, respectively. The connection between the two models is made by inserting a hidden layer with fixed weights to compute a mel-scale spectrogram. The results show that the mel-scaled spectrogram is not the optimal input for the recognition task. We can replace the fixed hidden layer with learnable hidden layers, which can be seen as adaptive filters. Then, different optimization schemes can be tested in order to find the optimal combination of the three modules, i.e., speech enhancement, learnable filters, and speech recognition. Finally, future work will carry out experiments with other data to verify if the findings presented in Section 4.4 can be generalized to different dataset sizes and scenarios.

- *Audio-visual speech inpainting.* The system proposed in Chapter 5 represents a first attempt to solve the audio-visual speech inpainting task with simple deep learning models. Indeed, there is still a lot of room for improvement. Training on large-scale datasets and more complex deep learning architectures are the natural way to move forward. For example, different AV fusion techniques can be tested. In particular, more natural speech can be generated by exploiting speaker characteristics explicitly extracted from the uncorrupted audio context. Additionally, different kinds of information loss can be explored, e.g., irregular masked time-frequency regions, to improve model generalization.

- *Joint speech enhancement and inpainting.* Both speech enhancement and inpainting aim at restoring clean speech from signals affected by several levels of degradation. Although joint enhancement and inpainting audio-only algorithms already exist [50], there are no audio-visual approaches that address this problem. Audio-visual speech enhancement, audio-visual speech inpainting, and speech synthesis from silent videos can be tied together to pave the way towards universal robust speech processing systems.

Bibliography

[1] A. Adler, V. Emiya, M. G. Jafari, M. Elad, R. Gribonval, and M. D. Plumbley, “A constrained matching pursuit approach to audio declipping”, in *Proc. of ICASSP*, 2011.

[2] A. Adler, V. Emiya, M. G. Jafari, M. Elad, R. Gribonval, and M. D. Plumbley, “Audio Inpainting”, *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 922–932, 2012.

[3] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, “Deep audio-visual speech recognition”, *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[4] T. Afouras, J. S. Chung, and A. Zisserman, “The Conversation: Deep Audio-Visual Speech Enhancement”, *Proc. of Interspeech*, 2018.

[5] H. Akbari, H. Arora, L. Cao, and N. Mesgarani, “Lip2audspec: Speech reconstruction from silent lip movements video”, in *Proc. of ICASSP*, 2018.

[6] K. A. Al-Karawi, A. H. Al-Noori, F. F. Li, and T. Ritchings, “Automatic Speaker Recognition System in Adverse Conditions—Implication of Noise and Reverberation on System Performance”, *International Journal of Information and Electronics Engineering*, vol. 5, no. 6, pp. 423–427, 2015.

[7] M. C. Anzalone, L. Calandruccio, K. A. Doherty, and L. H. Carney, “Determination of the Potential Benefit of Time-Frequency Gain Manipulation”, *Ear Hear*, vol. 27, no. 5, pp. 480–492, 2006.

[8] Y. Bahat, Y. Y. Schechner, and M. Elad, “Self-content-based audio inpainting”, *Signal Processing*, vol. 111, pp. 61–72, 2015.

[9] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate”, *arXiv preprint arXiv:1409.0473*, 2014.

[10] R. Benosman, C. Clercq, X. Lagorce, S.-H. Ieng, and C. Bartolozzi, “Event-based visual flow”, *IEEE transactions on neural networks and learning systems*, vol. 25, no. 2, pp. 407–417, 2013.

[11] S. Boll, “Suppression of acoustic noise in speech using spectral subtraction”, *IEEE Transactions on acoustics, speech, and signal processing*, vol. 27, no. 2, pp. 113–120, 1979.

[12] A. W. Bronkhorst, “The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions”, *Acta Acustica united with Acustica*, vol. 86, no. 1, pp. 117–128, 2000.

[13] H. Brumm and S. A. Zollinger, “The evolution of the Lombard effect: 100 years of psychoacoustic research”, *Behaviour*, vol. 148, no. 11-13, pp. 1173–1198, 2011.

[14] L. A. Camuñas-Mesa, T. Serrano-Gotarredona, S.-H. Ieng, R. Benosman, and B. Linares-Barranco, “Event-driven stereo visual tracking algorithm to solve object occlusion”, *IEEE transactions on neural networks and learning systems*, vol. 29, no. 9, pp. 4223–4237, 2017.

[15] R. Caruana, “Multitask learning”, *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

[16] X. Chang, Y. Qian, K. Yu, and S. Watanabe, “End-to-end monaural multi-speaker ASR system without pretraining”, in *Proc. of ICASSP*, 2019.

[17] Y.-L. Chang, K.-Y. Lee, P.-Y. Wu, H.-Y. Lee, and W. Hsu, “Deep Long Audio Inpainting”, *arXiv preprint arXiv:1911.06476*, 2019.

[18] G.-L. Chao, W. Chan, and I. Lane, “Speaker-Targeted Audio-Visual Models for Speech Recognition in Cocktail-Party Environments”, in *Proc. of Interspeech*, 2016.

[19] Z. Chen, Y. Luo, and N. Mesgarani, “Deep attractor network for single-microphone speaker separation”, in *Proc. of ICASSP*, 2017.

[20] Z. Chen, S. Watanabe, H. Erdogan, and J. R. Hershey, “Speech enhancement and recognition using multi-task learning of long short-term memory recurrent neural networks”, in *Proc. of Interspeech*, 2015.

[21] E. C. Cherry, “Some Experiments on the Recognition of Speech, with One and with Two Ears”, *The Journal of the Acoustical Society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[22] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, “Lip reading sentences in the wild”, in *Proc. of CVPR*, 2017.

[23] S.-W. Chung, S. Choe, J. S. Chung, and H.-G. Kang, “FaceFilter: Audio-Visual Speech Separation Using Still Images”, in *Proc. of Interspeech*, 2020.

[24] M. Cooke, J. Barker, S. Cunningham, and X. Shao, “An audio-visual corpus for speech perception and automatic speech recognition”, *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[25] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection”, in *Proc. of CVPR*, 2005.

[26] T. Darrell, J. W. Fisher, and P. Viola, “Audio-visual segmentation and “the cocktail party effect””, in *International Conference on Multimodal Interfaces*, Springer, 2000, pp. 32–40.

[27] M. Delcroix, K. Zmolikova, K. Kinoshita, A. Ogawa, and T. Nakatani, “Single Channel Target Speaker Extraction and Recognition with Speaker Beam”, in *Proc. of ICASSP*, 2018.

[28] P. P. Ebner and A. Eltelt, “Audio inpainting with generative adversarial network”, *arXiv preprint arXiv:2003.07704*, 2020.

[29] Y. Ephraim and D. Malah, “Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator”, *IEEE Transactions on acoustics, speech, and signal processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[30] A. Ephrat, T. Halperin, and S. Peleg, “Improved speech reconstruction from silent video”, in *Proc. of ICCVW*, 2017.

[31] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, “Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation”, *ACM Transactions on Graphics*, vol. 37, no. 4, pp. 1–11, 2018.

[32] A. Ephrat and S. Peleg, “Vid2speech: speech reconstruction from silent video”, in *Proc. of ICASSP*, 2017.

[33] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, “Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks”, in *Proc. of ICASSP, 2015.*

[34] W. Etter, “Restoration of a discrete-time signal segment by interpolation based on the left-sided and right-sided autoregressive parameters”, *IEEE transactions on signal processing*, vol. 44, no. 5, pp. 1124–1135, 1996.

[35] A. Gabbay, A. Ephrat, T. Halperin, and S. Peleg, “Seeing through noise: Visually driven speaker separation and enhancement”, in *Proc. of ICASSP*, 2018.

[36] A. Gabbay, A. Shamir, and S. Peleg, “Visual Speech Enhancement”, *Proc. of Interspeech*, 2018.

[37] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, “DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1”, *NASA STI/Recon technical report n*, vol. 93, p. 27 403, 1993.

[38] T. Gerkmann, M. Krawczyk-Becker, and J. Le Roux, “Phase processing for single-channel speech enhancement: History and recent advances”, *IEEE signal processing Magazine*, vol. 32, no. 2, pp. 55–66, 2015.

[39] S. Ghosh-Dastidar and H. Adeli, “Spiking neural networks”, *International journal of neural systems*, vol. 19, no. 04, pp. 295–308, 2009.

[40] L. Girin, J.-L. Schwartz, and G. Feng, “Audio-visual enhancement of speech in noise”, *The Journal of the Acoustical Society of America*, vol. 109, no. 6, pp. 3007–3020, 2001.

[41] A. Glover and C. Bartolozzi, “Robust visual tracking with a freely-moving event camera”, in *Proc. of IROS*, 2017.

[42] M. Gogate, K. Dashtipour, A. Adeel, and A. Hussain, “CochleaNet: A Robust Language-independent Audio-Visual Model for Speech Enhancement”, *Information Fusion*, 2020.

[43] E. Z. Golumbic, G. B. Cogan, C. E. Schroeder, and D. Poeppel, “Visual input enhances selective speech envelope tracking in auditory cortex at a “cocktail party””, *Journal of Neuroscience*, vol. 33, no. 4, pp. 1417–1426, 2013.

[44] D. Goodman, G Lockhart, O Wasem, and W.-C. Wong, “Waveform substitution techniques for recovering missing speech segments in packet voice communications”, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 6, pp. 1440–1448, 1986.

[45] A. Graves, “Sequence transduction with recurrent neural networks”, *arXiv preprint arXiv:1211.3711*, 2012.

[46] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks”, in *Proc. of ICML*, 2006.

[47] A. Graves, A.-R. Mohamed, and G. Hinton, “Speech recognition with deep recurrent neural networks”, in *Proc. of ICASSP*, 2013.

[48] D. Griffin and J. Lim, “Signal estimation from modified short-time Fourier transform”, *IEEE Transactions on acoustics, speech, and signal processing*, vol. 32, no. 2, pp. 236–243, 1984.

[49] R. Gu, S.-X. Zhang, Y. Xu, L. Chen, Y. Zou, and D. Yu, “Multi-modal multi-channel target speech separation”, *IEEE Journal of Selected Topics in Signal Processing*, 2020.

[50] X. Hao, X. Su, S. Wen, Z. Wang, Y. Pan, F. Bao, and W. Chen, “Masking and Inpainting: A Two-Stage Speech Enhancement Approach for Low SNR and Non-Stationary Noise”, in *Proc. of ICASSP*, 2020.

[51] R. W. Harris and D. W. Swenson, “Effects of reverberation and noise on speech recognition by adults with various amounts of sensorineural hearing impairment”, *Audiology*, vol. 29, no. 6, pp. 314–321, 1990.

[52] N. Harte and E. Gillen, “TCD-TIMIT: An Audio-Visual Corpus of Continuous Speech”, *IEEE Transactions on Multimedia*, vol. 17, no. 5, pp. 603–615, 2015.

[53] K. S. Helfer and L. A. Wilber, “Hearing loss, aging, and speech perception in reverberation and noise”, *Journal of Speech, Language, and Hearing Research*, vol. 33, no. 1, pp. 149–155, 1990.

[54] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, “Deep clustering: Discriminative embeddings for segmentation and separation”, in *Proc. of ICASSP*, 2016.

[55] A. Hines, J. Skoglund, A. Kokaram, and N. Harte, “ViSQOL: The virtual speech quality objective listener”, in *Proc. of IWAENC*, 2012.

[56] J.-C. Hou, S.-S. Wang, Y.-H. Lai, J.-C. Lin, Y. Tsao, H.-W. Chang, and H.-M. Wang, “Audio-visual speech enhancement using deep neural networks”, in *Proc. of APSIPA*, 2016.

[57] J.-C. Hou, S.-S. Wang, Y.-H. Lai, Y. Tsao, H.-W. Chang, and H.-M. Wang, “Audio-Visual Speech Enhancement Using Multimodal Deep Convolutional Neural Networks”, *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 117–128, 2018.

[58] M. Iacono, S. Weber, A. Glover, and C. Bartolozzi, “Towards event-driven object detection with off-the-shelf deep learning”, in *Proc. of IROS*, 2018.

[59] E. Ideli, B. Sharpe, I. V. Bajić, and R. G. Vaughan, “Visually assisted time-domain speech enhancement”, in *Proc. of GlobalSIP*, 2019.

[60] Y. Isik, J. L. Roux, Z. Chen, S. Watanabe, and J. R. Hershey, “Single-Channel Multi-Speaker Separation Using Deep Clustering”, in *Proc. of Interspeech*, 2016.

[61] I. Kauppinen, J. Kauppinen, and P. Saarinen, “A method for long extrapolation of audio signals”, *Journal of the Audio Engineering Society*, vol. 49, no. 12, pp. 1167–1180, 2001.

[62] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne, “Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds”, *Speech communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

[63] V. Kazemi and J. Sullivan, “One Millisecond Face Alignment with an Ensemble of Regression Trees”, in *Proc. of CVPR*, 2014.

[64] M. Kegler, P. Beckmann, and M. Cernak, “Deep Speech Inpainting of Time-Frequency Masks”, in *Proc. of Interspeech*, 2020.

[65] D. E. King, “Dlib-ml: A Machine Learning Toolkit”, *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[66] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization”, *arXiv preprint arXiv:1412.6980*, 2014.

[67] U. Kjems, J. B. Boldt, M. S. Pedersen, T. Lunner, and D. Wang, “Role of mask pattern in intelligibility of ideal binary-masked noisy speech”, *The Journal of the Acoustical Society of America*, vol. 126, no. 3, pp. 1415–1426, 2009.

[68] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, “Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks”, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.

[69] T. Le Cornu and B. Milner, “Reconstructing intelligible audio speech from visual speech features”, in *Proc. of Interspeech*, 2015.

[70] J. Le Roux, H. Kameoka, N. Ono, A. De Cheveigne, and S. Sagayama, “Computational auditory induction as a missing-data model-fitting problem with Bregman divergence”, *Speech Communication*, vol. 53, no. 5, pp. 658–676, 2011.

[71] J. Le Roux, H. Kameoka, N. Ono, and S. Sagayama, “Fast signal reconstruction from magnitude STFT spectrogram based on spectrogram consistency”, in *Proc. of DAFx, 2010.*

[72] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, “An overview of noise-robust automatic speech recognition”, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 745–777, 2014.

[73] X. Li, D. Neil, T. Delbruck, and S.-C. Liu, “Lip reading deep network exploiting multi-modal spiking visual and auditory sensors”, in *Proc. of ISCAS*, 2019.

[74] F. Lieb and H.-G. Stark, “Audio inpainting: Evaluation of time-frequency representations and structured sparsity approaches”, *Signal Processing*, vol. 153, pp. 291–299, 2018.

[75] J. S. Lim and A. V. Oppenheim, “Enhancement and bandwidth compression of noisy speech”, *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

[76] P. C. Loizou, *Speech enhancement: theory and practice.* CRC press, 2013.

[77] Y. Luo and N. Mesgarani, "TasNet: time-domain audio separation network for real-time, single-channel speech separation", in *Proc. of ICASSP*, 2018.

[78] A. Marafioti, P. Majdak, N. Holighaus, and N. Perraudin, "GACELA-A generative adversarial context encoder for long audio inpainting of music", *IEEE Journal of Selected Topics in Signal Processing*, 2020.

[79] A. Marafioti, N. Perraudin, N. Holighaus, and P. Majdak, "A Context Encoder For Audio Inpainting", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 2362–2372, 2019.

[80] J.-M. Maro, G. Lenz, C. Reeves, and R. Benosman, "Event-based Visual Gesture Recognition with Background Suppression running on a smartphone", in *Proc. of FG*, 2019.

[81] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 2, pp. 137–145, 1980.

[82] J. H. McDermott, "The cocktail party problem", *Current Biology*, vol. 19, no. 22, R1024–R1027, 2009.

[83] H. McGurk and J. MacDonald, "Hearing lips and seeing voices", *Nature*, vol. 264, no. 5588, pp. 746–748, 1976.

[84] D. Michelsanti, O. Slizovskaia, G. Haro, E. Gómez, Z.-H. Tan, and J. Jensen, "Vocoder-Based Speech Synthesis from Silent Videos", en, in *Proc. of Interspeech, 2020.*

[85] D. Michelsanti, Z.-H. Tan, S.-X. Zhang, Y. Xu, M. Yu, D. Yu, and J. Jensen, "An overview of deep-learning-based audio-visual speech enhancement and separation", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.

[86] O. Mokrỳ and P. Rajmic, "Audio inpainting: Revisited and reweighted", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2906–2918, 2020.

[87] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications", *IEICE Transactions on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[88] Y. Mroueh, E. Marcheret, and V. Goel, "Deep multimodal learning for audio-visual speech recognition", in *Proc. of ICASSP*, 2015.

[89] A. Narayanan and D. Wang, "Investigation of speech separation as a front-end for noise robust speech recognition", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 826–835, 2014.

[90] ——, “Improving robustness of deep neural network acoustic models via speech separation and joint adaptive training”, *IEEE/ACM transactions on audio, speech, and language processing*, vol. 23, no. 1, pp. 92–101, 2015.

[91] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, “Multimodal deep learning”, in *Proc. of ICML*, 2011.

[92] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, “Audio-visual speech recognition using deep learning”, *Applied Intelligence*, vol. 42, no. 4, pp. 722–737, 2015.

[93] T. Ochiai, M. Delcroix, K. Kinoshita, A. Ogawa, and T. Nakatani, “Multimodal SpeakerBeam: Single Channel Target Speech Extraction with Audio-Visual Speaker Clues”, *Proc. of Interspeech, 2019*,

[94] A. Owens and A. A. Efros, “Audio-visual scene analysis with self-supervised multisensory features”, in *Proc. of ECCV*, 2018.

[95] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An ASR corpus based on public domain audio books”, in *Proc. of ICASSP*, 2015.

[96] L. Patrick, C. Posch, and T. Delbruck, “A 128x 128 120 dB 15μ s Latency Asynchronous Temporal Contrast Vision Sensor”, *IEEE journal of solid-state circuits*, vol. 43, pp. 566–576, 2008.

[97] N. Perraudin, N. Holighaus, P. Majdak, and P. Balazs, “Inpainting of Long Audio Segments With Similarity Graphs”, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 6, pp. 1083–1094, 2018.

[98] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic, “End-to-end audiovisual speech recognition”, in *Proc. of ICASSP*, 2018.

[99] C. Posch, D. Matolin, and R. Wohlgenannt, “A QVGA 143 dB dynamic range frame-free PWM image sensor with lossless pixel-level video compression and time-domain CDS”, *IEEE Journal of Solid-State Circuits*, vol. 46, no. 1, pp. 259–275, 2010.

[100] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, “Purely sequence-trained neural networks for ASR based on lattice-free MMI.”, in *Proc. of Interspeech*, 2016.

[101] P. Prablanc, A. Ozerov, N. Q. K. Duong, and P. Perez, “Text-informed speech inpainting via voice conversion”, in *Proc. of EUSIPCO*, 2016.

[102] K. R. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar, “Learning Individual Speaking Styles for Accurate Lip to Speech Synthesis”, in *Proc. of CVPR*, 2020.

[103] Y. Qian, X. Chang, and D. Yu, “Single-channel multi-talker speech recognition with permutation invariant training”, *Speech Communication*, vol. 104, pp. 1–11, 2018.

[104] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, D. P. Ellis, and C. C. Raffel, “mir_eval: A transparent implementation of common MIR metrics”, in *Proc. of ISMIR*, 2014.

[105] B. Rivet, W. Wang, S. M. Naqvi, and J. A. Chambers, “Audiovisual speech source separation: An overview of key methodologies”, *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 125–134, 2014.

[106] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, “Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs”, in *Proc. of ICASSP*, 2001.

[107] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation”, in *Proc. of MICCAI*, 2015.

[108] M. Sadeghi, S. Leglaive, X. Alameda-Pineda, L. Girin, and R. Horaud, “Audio-Visual Speech Enhancement Using Conditional Variational Auto-Encoders”, *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 28, pp. 1788–1800, 2020.

[109] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, “300 faces in-the-wild challenge: Database and results”, *Image and vision computing*, vol. 47, pp. 3–18, 2016.

[110] A. Savran, R. Tavarone, B. Higy, L. Badino, and C. Bartolozzi, “Energy and computation efficient audio-visual voice activity detection driven by event-cameras”, in *Proc. of FG*, 2018.

[111] H. Seki, T. Hori, S. Watanabe, J. Le Roux, and J. R. Hershey, “A Purely End-to-End System for Multi-speaker Speech Recognition”, in *Proc. of ACL*, 2018.

[112] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, *et al.*, “Natural tts synthesis by conditioning wavenet on mel spectrogram predictions”, in *Proc. of ICASSP*, 2018.

[113] B. G. Shinn-Cunningham and V. Best, “Selective attention in normal and impaired hearing”, *Trends in amplification*, vol. 12, no. 4, pp. 283–299, 2008.

[114] P. Smaragdis, B. Raj, and M. Shashanka, “Missing data imputation for time-frequency representations of audio signals”, *Journal of signal processing systems*, vol. 65, no. 3, pp. 361–370, 2011.

[115] T. Stoffregen, G. Gallego, T. Drummond, L. Kleeman, and D. Scaramuzza, “Event-based motion segmentation by motion compensation”, in *Proc. of ICCV*, 2019.

[116] W. H. Sumby and I. Pollack, “Visual contribution to speech intelligibility in noise”, *The journal of the acoustical society of america*, vol. 26, no. 2, pp. 212–215, 1954.

[117] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to Sequence Learning with Neural Networks”, *Advances in Neural Information Processing Systems*, vol. 27, pp. 3104–3112, 2014.

[118] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech”, *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[119] Z. Tang, L. Li, and D. Wang, “Multi-task recurrent model for speech and speaker recognition”, in *Proc. of APSIPA*, 2016.

[120] I. Toumi and V. Emiya, “Sparse non-local similarity modeling for audio inpainting”, in *Proc. of ICASSP*, 2018.

[121] J. Turek, S. Jain, V. Vo, M. Capotă, A. Huth, and T. Willke, “Approximating Stacked and Bidirectional Recurrent Architectures with the Delayed Recurrent Neural Network”, in *Proc. of ICML*, 2020.

[122] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is All you Need”, *Advances in Neural Information Processing Systems*, vol. 30, pp. 5998–6008, 2017.

[123] E. Vincent, R. Gribonval, and C. Fevotte, “Performance measurement in blind audio source separation”, *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[124] K. Vougioukas, P. Ma, S. Petridis, and M. Pantic, “Video-Driven Speech Reconstruction Using Generative Adversarial Networks”, in *Proc. of Interspeech*, 2019.

[125] D. Wang and G. J. Brown, *Computational auditory scene analysis: Principles, algorithms, and applications.* Wiley-IEEE press, 2006.

[126] D. Wang and J. Chen, “Supervised Speech Separation Based on Deep Learning: An Overview”, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[127] D. Wang and J. Lim, “The unimportance of phase in speech enhancement”, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 30, no. 4, pp. 679–681, 1982.

[128] Q. Wang, H. Muckenhirn, K. Wilson, P. Sridhar, Z. Wu, J. R. Hershey, R. A. Saurous, R. J. Weiss, Y. Jia, and I. L. Moreno, "VoiceFilter: Targeted Voice Separation by Speaker-Conditioned Spectrogram Masking", in *Proc. of Interspeech*, 2019.

[129] Y. Wang, X. Fan, I.-F. Chen, Y. Liu, T. Chen, and B. Hoffmeister, "End-to-end Anchored Speech Recognition", in *Proc. of ICASSP*, 2019.

[130] Z.-Q. Wang and D. Wang, "Joint training of speech separation, filterbank and acoustic model for robust automatic speech recognition", in *Proc. of Interspeech*, 2015.

[131] F. Weninger, J. R. Hershey, J. Le Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation", in *Proc. of GlobalSIP*, 2014.

[132] P. J. Wolfe and S. J. Godsill, "Interpolation of missing data values for audio signal restoration using a Gabor regression model", in *Proc. of ICASSP, 2005.*

[133] Z. Wu, S. Sivadas, Y. K. Tan, M. Bin, and R. S. M. Goh, "Multi-modal hybrid deep neural network for speech enhancement", *arXiv preprint arXiv:1606.04750*, 2016.

[134] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation", in *Proc. of ICASSP*, 2017.

[135] J. Yu, S.-X. Zhang, J. Wu, S. Ghorbani, B. Wu, S. Kang, S. Liu, X. Liu, H. Meng, and D. Yu, "Audio-visual recognition of overlapped speech for the LRS2 dataset", in *Proc. of ICASSP*, 2020.

[136] Yuxuan Wang, A. Narayanan, and DeLiang Wang, "On Training Targets for Supervised Speech Separation", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[137] Y. Zhang and Q. Yang, "A survey on multi-task learning", *arXiv preprint arXiv:1707.08114*, 2017.

[138] H. Zhou, Z. Liu, X. Xu, P. Luo, and X. Wang, "Vision-Infused Deep Audio Inpainting", in *Proc. of ICCV, 2019.*

[139] A. Z. Zhu, L. Yuan, K. Chaney, and K. Daniilidis, "Unsupervised event-based learning of optical flow, depth, and egomotion", in *Proc. of CVPR*, 2019.

[140] H. Zhu, M. Luo, R. Wang, A. Zheng, and R. He, "Deep Audio-Visual Learning: A Survey", *arXiv preprint arXiv:2001.04758*, 2020.

[141] K. Zmolikova, M. Delcroix, K. Kinoshita, T. Higuchi, A. Ogawa, and T. Nakatani, “Speaker-Aware Neural Network Based Beamformer for Speaker Extraction in Speech Mixtures.”, in *Proc. of Interspeech*, 2017.