A GRAPHICAL USER INTERFACE FOR GENERATING SYNTHETIC VOWELS WITH PREDEFINED ACOUSTIC PARAMETERS

D. Gasperini[1], S. Orlandi[2,3], A. Bandini[1,4,5]
[1]Health Science Interdisciplinary Research Center, Scuola Superiore Sant'Anna, Pisa, Italy
[2]Department of Electrical, Electronic, and Information Engineering "Guglielmo Marconi" – DEI, University of Bologna, Bologna, Italy; Health Sciences and Technologies, Interdepartmental Center for Industrial Research (CIRI-SDV), University of Bologna, Italy
[3]IRCCS Istituto delle Scienze Neurologiche di Bologna, Bologna, Italy
[4]The BioRobotics Institute and Department of Excellence in Robotics and AI, Scuola Superiore Sant'Anna, Pisa, Italy
[5]KITE - Toronto Rehabilitation Institute, University Health Network, Toronto, ON, Canada
daniela.gasperini@santannapisa.it; silvia.orlandi9@unibo.it; andrea.bandini@santannapisa.it

***Abstract*: Dysarthric speech is an important biomarker for clinical assessment and diagnostic support in neurological diseases. However, speech recordings are often collected in uncontrolled and noisy environments, such as clinics and home settings. Speech enhancement and denoising tools can help with this challenge, but their effects on important acoustic features, like fundamental frequency (F_0) and the first two formants (F_1, F_2), are not well understood. This uncertainty raises concerns about possible distortions. To address this, a Graphical User Interface (GUI) was developed to create synthetic American English vowels with predefined fundamental frequency and first two formants, providing a controlled ground-truth reference. The fundamental frequency is derived from Gaussian distributions. The first and second formants are sampled using a kernel density estimation technique to ensure physiologically plausible vowels. The glottal source is modulated with jitter and shimmer to mimic variations in speech. The GUI allows flexible control over vowel type, noise condition, signal duration, and modulation parameters. This enables reproducible benchmarking of speech processing algorithms under controlled noise scenarios. Planned extensions include adding support for Italian vowels, higher-order formants, and validating quality indices for signal denoising and speech enhancement algorithms.**

Keywords: **synthetic speech, vowels, phonetics, GUI**

I. INTRODUCTION

Neurological diseases such as Parkinson's disease (PD) and amyotrophic lateral sclerosis (ALS) commonly lead to dysarthria, a motor speech disorder characterized by impaired speech execution [1], [2]. Acoustic parameters, particularly fundamental frequency (F_0) and formants (F_1, F_2), serve as valuable biomarkers for monitoring disease progression [3].

F_0 reflects the periodic vibration of the vocal folds and differs between males and females. In contrast, F_1 and F_2 correspond to vocal tract resonances during phonation, characterizing different vowels. These formants are strongly correlated with jaw and tongue muscular activity, making them sensitive indicators of neuromotor degeneration [4]. By mapping vowels in the F_1–F_2 plane, it is possible to derive the vowel space area (VSA), which is a critical acoustic biomarker for monitoring neurological diseases due to its sensitivity to articulatory impairments. Patients with PD and ALS typically present a reduced VSA relative to healthy controls [5], [6].

Estimating these parameters in patients with neurological diseases presents significant challenges. The inherent variability in neurological voice signals, especially when affected by dysarthria, necessitates robust analysis methods that can accommodate irregular vocal patterns while maintaining measurement accuracy. Furthermore, audio recordings are typically acquired in noisy environments (e.g., hospitals or at home), where background noise from medical equipment or power lines can corrupt the voice signal [7].

Recent advances in machine learning and deep learning have improved speech signal analysis [8]. For instance, Jolad & Khanai proposed a Competitive Crow Search Algorithm-based Speech Enhancement Generative Adversarial Network (FCCSA-SEGAN) to enhance dysarthric speech, boosting quality and intelligibility across noise conditions [9], while Wang et al. used a Convolutional Neural Network (CNN) to improve the intelligibility of dysarthric speech, achieving over 10% improvement in automatic speech recognition (ASR) and subjective human intelligibility tests [10]. Nonetheless, evaluating the performance of these algorithms remains challenging, as it requires simultaneous consideration of multiple factors. Traditional metrics such as Signal-to-Noise Ratio (SNR) are insufficient because they are too general and can apply to non-speech signals. A comprehensive evaluation should consider various metrics including

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Lorenzo Frassineti, Antonio Lanatà, Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications. 14th International Workshop, December, 16-17, 2025*, © 2025 Author(s), CC BY 4.0, published by Firenze University Press, ISBN 979-12-215-0821-5, DOI 10.36253/979-12-215-0821-5

Perceptual Evaluation of Speech Quality (PESQ), and Short-Time Objective Intelligibility (STOI), as different algorithms excel in different aspects [11].
The analysis of acoustic features is further complicated by the absence of ground-truth clean signals for direct comparison. Although tools such as Praat and BioVoice can estimate F_0 and formants [12], [13], these remain approximations and do not address the absence of reference data. Importantly, vowel segments are especially critical for speech assessment, but among the most challenging signals to denoise. Vowel spectral envelopes are affected by noise, particularly in the mid-frequency band (1–2.7 kHz), which includes the third formant band. This sensitivity makes vowels more susceptible to distortion and harder to recover accurately in noisy conditions [14]. To overcome this limitation, we present a vowel synthesizer that produces controlled vowel signals with known acoustic parameters, thereby enabling the creation of a reliable ground-truth database for benchmarking speech processing algorithms. Building on the work of Orlandi et al. [15], which employed synthetic signals of infant cries to evaluate and compare various analysis methods, the present study employed a public dataset of American English vowels collected by Hillenbrand [16]. This dataset includes acoustic measurements of F_0, F_1, and F_2 for male, female, and child speakers, and was used as a reference for generating synthetic signals. Specifically, we developed a MATLAB-based Graphical User Interface (GUI) that allows users to select vowels within the F_1-F_2 plane, control signal duration, and introduce realistic sources such as power line (50 Hz) and fan noise, and a theoretical one (i.e., white noise). The vowel space is sampled via kernel density estimation (KDE) to constrain synthesized vowels to physiologically plausible regions, resulting in a synthetic database with known ground-truth parameters. This tool provides a flexible framework to compare denoising algorithms and conduct controlled studies on vowel acoustics.

II. METHODS

According to the source–filter theory [17], the glottal acoustic signal is filtered by the resonant properties of the vocal tract, producing the characteristic formants of speech sounds. To parameterize the synthesizer, we relied on the widely used dataset of American English vowels [18], which reports acoustic measures of F_0, F_1, and F_2 for men, women, and children. In this preliminary study, we focused exclusively on adult speakers. To simulate realistic recording conditions, additive noise was included in the model according to:

$$y[k]=s[k]+n[k],\ k = 1,\ldots., N \quad (1)$$

where $y[k]$ is the noisy speech signal, $s[k]$ is the clean synthesized vowel signal, $n[k]$ is one of the three types of additive noise, and N is the number of samples.
Each formant was modeled as a second-order band-pass filter. Variations in speech were simulated by introducing jitter and shimmer, which reflect relevant characteristics for pathological speech [19]. Candidate values of F_1 and F_2 were initially sampled uniformly within the observed ranges of the dataset. A KDE model, fitted on the original data, was then used to evaluate the likelihood of each (F_1, F_2) pair in the vowel space. Only candidates exceeding the 95-th percentile of the KDE distribution were retained, ensuring that synthesized vowels remained within high-probability regions of the acoustic space. From the accepted points, a single pair was randomly selected for synthesis. Corresponding values for F_0 were sampled from Gaussian distributions parameterized by the mean and standard deviation of the respective vowel in the original dataset, with clamping applied to maintain physiological plausibility.

III. RESULTS

Fig. 1 shows the GUI. Users select the vowel to synthesize, the sex of the speaker, set the duration of the signal, and optionally introduce shimmer and jitter.

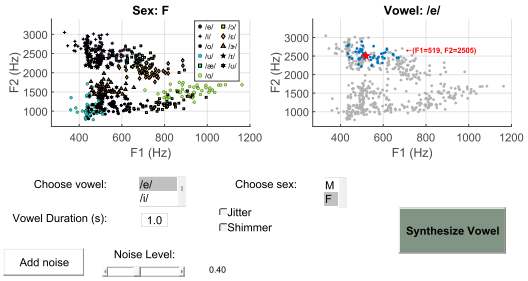

Fig. 1: Graphical User Interface for vowel synthesis.

A noise level slider lets users add background noise from a predefined library. When the "Synthesize Vowel" button is pressed, the GUI selects formant values from the dataset, generates an excitation signal, applies the chosen parameters, and synthesizes the vowel through the resonator model described in the previous section. It then saves both clean and noisy audio and updates interactive plots displaying the vowel space and the selected vowel detail.
Fig. 2 shows F_1-F_2 plane for male speakers in the Hillenbrand dataset. By sampling F_1 and F_2 values using the KDE approach, the synthesized vowels are constrained to high-probability regions, closely reflecting the distributions observed in real speakers. The legend is reported in IPA (International Phonetic

Alphabet), which for American English includes 11 phonemes for vowels *(/i/, /ɪ/, /e/, /ɛ/, /ɝ/, /æ/, /ɑ/, /ɔ/, /o/, /ʊ/, /u/)*.

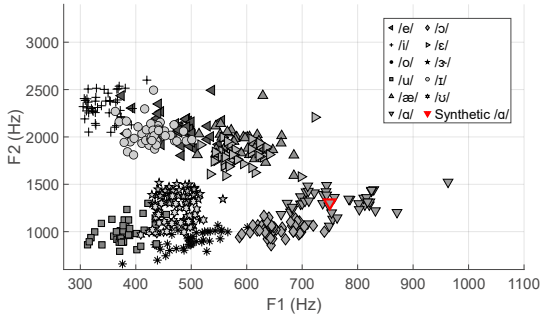

Fig. 2: F_1-F_2 plane for male speakers in the dataset. Each marker represents a different vowel. The red marker indicates the synthetically generated vowel.

Table 1 reports the mean and standard deviation of F_0, F_1, and F_2 values for 48 women and 45 men from the dataset, which were used as reference parameters for generating the synthesized vowels.

Table 1. Mean and standard deviation (in Hz) of fundamental frequency (F_0) and formants (F_1–F_2) for female and male speakers.

Parameter (Hz)	Male	Female
F_0	131 ± 22	220 ± 23
F_1	515 ± 123	602 ± 160
F_2	1538 ± 499	1795 ± 623

IV. Discussion

Neurological diseases often lead to speech impairments such as dysarthria, which compromise communication. Acoustic markers, including formant-related measures, have proven valuable for tracking disease severity and progression. However, their extraction from natural recordings remains challenging, as speech signals are often corrupted by environmental noise, such as background sounds from medical equipment or household environments.

In this paper, we presented an easy tool for generating synthetic vowels with known acoustic parameters to create a synthetic F_1-F_2 plane. In this direction the proposed tool enables controlled testing of denoising algorithms by isolating variables such as SNR and noise type, providing a reliable framework for algorithm evaluation under controlled conditions. Previous works have also explored vowel synthesis for the analysis of pathological voices. In [20], the authors proposed a model that reproduces alterations in acoustic parameters such as jitter and shimmer, using sustained vowels from a database of speakers without speech impairments and individuals affected by various voice disorders. Their method focuses on capturing irregularities in phonation to simulate pathological conditions and evaluate acoustic correlations of vocal disorders. By contrast, our approach does not aim to reproduce pathology-related perturbations but rather to generate clean, parameter-controlled vowels that can be subsequently corrupted with realistic noise sources. More recently, VSpace, a browser-based tool for vowel synthesis, was presented in [21]. It allows exploration of the universal vowel space by selecting formant frequencies within a trapezoid scaled to different speaker ranges. While this tool is valuable for educational and perceptual studies, its design primarily focuses on accessibility and interactive exploration rather than systematic dataset generation. Conversely, in this work we sample the vowel space through a KDE strategy, ensuring that synthesized vowels remain within physiologically plausible regions. This enables the construction of a synthetic database with known ground-truth parameters, specifically tailored for benchmarking denoising and speech enhancement algorithms under controlled conditions. The proposed tool is publicly available on GitHub (https://github.com/Gasp-sh/VowelSynthGUI.git).

V. Conclusion

This work presents a simple vowel synthesizer designed to evaluate algorithms developed for neurological patients with speech impairments. The database is based on real male and female American English speakers and employs a KDE-based approach to generate realistic vowel signals. By allowing precise control over synthesis parameters, including SNR and noise type, the tool provides a reproducible framework for benchmarking of denoising and speech enhancement algorithms under controlled conditions. It is important to note that this framework is currently limited to vowels and the first two formants. While it allows controlled testing of algorithms on synthetic signals, whether the same algorithms perform similarly on natural speech from patients remains to be evaluated. Nevertheless, the tool could represent a promising starting point for controlled evaluations, and future work will extend the synthesis to higher-order formants and additional languages, such as Italian.

References

[1] S. Sapir, «Multiple Factors Are Involved in the Dysarthria Associated With Parkinson's Disease: A Review With Implications for Clinical Practice and Research», *J. Speech Lang. Hear. Res.*, vol.

57, fasc. 4, pp. 1330–1343, ago. 2014, doi: 10.1044/2014_JSLHR-S-13-0039.

[2] B. Tomik e R. J. Guiloff, «Dysarthria in amyotrophic lateral sclerosis: A review», *Amyotroph. Lateral Scler.*, vol. 11, fasc. 1–2, pp. 4–15, gen. 2010, doi: 10.3109/17482960802379004.

[3] P. Gómez-Vilda *et al.*, «Neurological Disease Detection and Monitoring from Voice Production», in *Advances in Nonlinear Speech Processing*, vol. 7015, C. M. Travieso-González e J. B. Alonso-Hernández, A c. di, in Lecture Notes in Computer Science, vol. 7015. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 1–8. doi: 10.1007/978-3-642-25020-0_1.

[4] P. Gómez-Vilda *et al.*, «Neuromechanical Modelling of Articulatory Movements from Surface Electromyography and Speech Formants», *Int. J. Neural Syst.*, vol. 29, fasc. 02, p. 1850039, mar. 2019, doi: 10.1142/S0129065718500399.

[5] S. Skodda, W. Grönheit, e U. Schlegel, «Impairment of Vowel Articulation as a Possible Marker of Disease Progression in Parkinson's Disease», *PLOS ONE*, vol. 7, fasc. 2, p. e32132, feb. 2012, doi: 10.1371/journal.pone.0032132.

[6] G. S. Turner, K. Tjaden, e G. Weismer, «The Influence of Speaking Rate on Vowel Space and Speech Intelligibility for Individuals With Amyotrophic Lateral Sclerosis», *J. Speech Lang. Hear. Res.*, vol. 38, fasc. 5, pp. 1001–1013, ott. 1995, doi: 10.1044/jshr.3805.1001.

[7] E. E. Ryherd, K. P. Waye, e L. Ljungkvist, «Characterizing noise and perceived work environment in a neurological intensive care unit», *J. Acoust. Soc. Am.*, vol. 123, fasc. 2, pp. 747–756, feb. 2008, doi: 10.1121/1.2822661.

[8] J. Wang *et al.*, «Automatic prediction of intelligible speaking rate for individuals with ALS from speech acoustic and articulatory samples», *Int. J. Speech Lang. Pathol.*, vol. 20, fasc. 6, pp. 669–679, ott. 2018, doi: 10.1080/17549507.2018.1508499.

[9] B. Jolad e R. Khanai, «An approach for speech enhancement with dysarthric speech recognition using optimization based machine learning frameworks», *Int. J. Speech Technol.*, vol. 26, fasc. 2, pp. 287–305, lug. 2023, doi: 10.1007/s10772-023-10019-y.

[10] S. Wang *et al.*, «Dysarthric Speech Enhancement Based on Convolution Neural Network», in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, lug. 2022, pp. 60–64. doi: 10.1109/EMBC48229.2022.9871531.

[11] Y. Hu e P. C. Loizou, «Subjective comparison and evaluation of speech enhancement algorithms», *Speech Commun.*, vol. 49, fasc. 7–8, pp. 588–601, lug. 2007, doi: 10.1016/j.specom.2006.12.006.

[12] «Praat: Doing Phonetics by Computer», *Ear Hear.*, vol. 32, fasc. 2, p. 266, apr. 2011, doi: 10.1097/AUD.0b013e31821473f7.

[13] M. S. Morelli, S. Orlandi, e C. Manfredi, «BioVoice: A multipurpose tool for voice analysis», *Biomed. Signal Process. Control*, vol. 64, p. 102302, feb. 2021, doi: 10.1016/j.bspc.2020.102302.

[14] G. Parikh e P. C. Loizou, «The influence of noise on vowel and consonant cues», *J. Acoust. Soc. Am.*, vol. 118, fasc. 6, pp. 3874–3888, dic. 2005, doi: 10.1121/1.2118407.

[15] S. Orlandi, A. Bandini, F. F. Fiaschi, e C. Manfredi, «Testing software tools for newborn cry analysis using synthetic signals», *Biomed. Signal Process. Control*, vol. 37, pp. 16–22, ago. 2017, doi: 10.1016/j.bspc.2016.12.012.

[16] J. Hillenbrand, L. A. Getty, M. J. Clark, e K. Wheeler, «Acoustic characteristics of American English vowels», *J. Acoust. Soc. Am.*, vol. 97, fasc. 5, pp. 3099–3111, mag. 1995, doi: 10.1121/1.411872.

[17] I. Tokuda, «The Source–Filter Theory of Speech», in *Oxford Research Encyclopedia of Linguistics*, 2021. doi: 10.1093/acrefore/9780199384655.013.894.

[18] «OSF | A practical guide to calculating vocal tract length and scale-invariant formant patterns». Consultato: 16 settembre 2025. [Online]. Disponibile su: https://osf.io/4c2r9/

[19] H. F. Wertzner, S. Schreiber, e L. Amaro, «Analysis of fundamental frequency, jitter, shimmer and vocal intensity in children with phonological disorders», *Braz. J. Otorhinolaryngol.*, vol. 71, fasc. 5, pp. 582–588, ott. 2015, doi: 10.1016/S1808-8694(15)31261-1.

[20] G. A. Alzamendi, G. Schlotthauer, H. L. Rufiner, e M. E. Torres, «Evaluation of a new model for vowels synthesis with perturbations in acoustic parameters», *Lat. Am. Appl. Res.*, vol. 43, fasc. 3, pp. 225–230, lug. 2013.

[21] M. I. Proctor, «VSpace: A browser-based vowel synthesiser», *J. Acoust. Soc. Am.*, vol. 154, fasc. 4_supplement, p. A203, ott. 2023, doi: 10.1121/10.0023276.