Flexible parametric implantation of voicing in whispered speech under scarce training data

1st João Silva
University of Porto-FEUP-DEEC
Porto, Portugal

2nd Marco Oliveira
University of Porto-FEUP-DEEC
Porto, Portugal

3rd Aníbal Ferreira
University of Porto-FEUP-DEEC
Porto, Portugal

***Abstract*—Whispered-voice to normal-voice conversion is typically achieved using codec-based analysis and re-synthesis, using statistical conversion of important spectral and prosodic features, or using data-driven end-to-end signal conversion. These approaches are however highly constrained by the architecture of the codec, the statistical projection, or the size and quality of the training data. In this paper, we presume direct implantation of voiced phonemes in whispered speech and we focus on fully flexible parametric models that i) can be independently controlled, ii) synthesize natural and linguistically correct voiced phonemes, iii) preserve idiosyncratic characteristics of a given speaker, and iv) are amenable to co-articulation effects through simple model interpolation. We use natural spoken and sung vowels to illustrate these capabilities in a signal modeling and re-synthesis process where spectral magnitude, phase structure, F0 contour and sound morphing can be independently controlled in arbitrary ways.**

***Index Terms*—whispers, voice conversion, parametric models**

I. Introduction

Whispered speech is a form of voice communication that does not include vibration of the vocal folds as in normal speech [1]. This communication modality may be intentionally adopted in environments where silence is recommended or privacy is desired, or may result from a health condition affecting the vocal folds, including laryngectomy. In these cases, the voice lacks projection, clarity and individuality. Oral communication becomes very difficult because whispered speech is easily corrupted by competing signals or noises.

A classical technology trying to improve the intelligibility and projection of whispered speech is the electrolarynx. It is a device that generates a vibration pattern that adds periodicity (or voicing) to the whispered speech when it is pressed against the throat. However, the resulting speech sounds robotic and unpleasant because the vibration pattern is monotonous and affects all phonemes. More advanced technological approaches to the problem have included adaptations of voice codecs in such a way as to explicitly control the source excitation signals and the vocal tract filter models [2], [3]. However, the success of these approaches has been strongly limited by the structural constraints of the codecs, namely in terms of model parameters control and synthesis of natural prosody.

Other approaches are based on statistical voice conversion such as non-audible murmur (NAM) that uses several GMM to convert whispered speech characteristics to voiced speech characteristics [4], namely in terms of spectral envelope conversion and prosody generation. However, these methods depend critically on high-quality training data and good match between 'source' and 'target' data.

More recent approaches follow a data-driven paradigm and are typically based on machine learning structures such as Deep Neural Networks [5]–[7]. Typically, these approaches require large databases of pairs of whispered speech and normal speech, require either prior data time-alignment or dynamic time warping and, most importantly, do not allow explicit and independent control of individual voice characteristics such as F0 contour or glottal signature.

In this paper, we adopt a signal processing approach that is tailored for real-time operation and that uses the whispered voice signal as a baseline in order to just implant the missing voicing in carefully selected phonemic regions [8]. This way, significant parts of the whispered speech are left unchanged, namely in plosive and unvoiced fricative regions, which helps to preserve a significant degree of naturalness and idiosyncratic information. In fact, the articulatory gestures, in both whispered and normal speech realizations of those regions, are essentially the same for the same speaker. This approach relies on scarce data since only parametric models of representative voiced phonemes, notably vowels, are required to be implanted in the whispered voice signal. This is a likely scenario when someone looses the ability to phonate, for example, after a sudden accident or disease, and only just a few old recordings of his/her voice are available.

We show in this paper that a careful selection of five parametric models makes it possible not only to correctly represent linguistic information, but also to convey idiosyncratic elements of a desired voice signature. In addition to being independently controllable, that selection of five parametric models permits co-articulation effects through simple model interpolation. To the best knowledge of the authors, this is the first time such a fully flexible parametric representation is developed that allows the synthesis of high-quality voiced sounds with independent control of such important signal attributes as phase structure, spectral magnitude and F0 trajectories.

This work was financed by FEDER - Fundo Europeu de Desenvolvimento Regional funds through the COMPETE 2020-Operacional Programme for Competitiveness and Internationalization (POCI), and by Portuguese funds through FCT-Fundação para a Ciência e a Tecnologia in the framework of the project POCI-01-0145-FEDER-029308. Support was also received from the FCT-POCH Program under the Sabbatical Leave Fellowship SFRH/BSAB/150372/2019.

In Sec. II, we describe our database, in Sec. III we detail the five types of parameters we use for full flexible and high-quality parametric synthesis of voiced sounds and, in Sec. IV, we present and discuss our test results. Section V concludes this paper.

II. Vowel database

Our database consists of the five tonic Portuguese vowels (/a/, /e/, /i/, /o/ and /u/) that were spoken, and sung, by a professional Portuguese singer. Thus, in total, we have 10 vowel realizations that were sustained for at least 1 second while avoiding *vibrato*. Their average fundamental frequency and associated standard deviation are indicated in Table I. We opted for a female singer in order to deal with high fundamental frequencies which are known to represent a significant challenge to LPC modeling, namely due to the well-known 'pitch-locking' phenomenon [1].

III. Parametric analysis/synthesis framework

Figure 1 represents the general block diagram of the analysis front-end of our framework and consists in an enhanced version of a recent development [9]. It relies on a 50% overlap-add analysis scheme using the Odd-frequency DFT [10] (ODFT) and the half-sine window as the discrete-time analysis window ($h[n]$) [11]. Subsequently, a window-aware spectral analysis implements accurate fundamental frequency (F0) estimation and harmonic analysis. Accurate F0 estimation is borrowed from our audio coding research [12] and delivers the estimated ω_0. Accurate harmonic analysis is also window-aware and delivers the estimated spectral magnitudes (A_ℓ) and phases (ϕ_ℓ) of all harmonics in a harmonic structure [11], [13]. Subsequently, a detailed parametric-oriented analysis and modeling takes place that, in addition to ω_0, represents any voiced signal by three magnitude-oriented features (HLPC, NLPC and D_ℓ) and a shift-invariant phase-related feature that describes the holistic phase structure of all harmonics [14]. These parametric models are addressed and illustrated next.

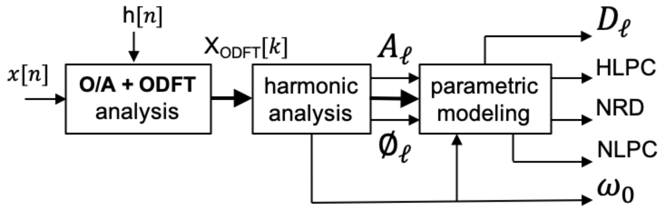

Fig. 1. Block diagram of the analysis front-end of full parametric modeling of a voiced signal. After ODFT transformation, all harmonics are identified. Then, the magnitude structure is represented by the HLPC, NLPC and D_ℓ models, and the shift-invariant phase structure is represented by the NRD model.

The synthesis part of our framework consists of the reverse block-based processing steps that are illustrated in Fig. 1, and are all frequency-domain, although an equivalent time-domain reconstruction alternative also exists. Most details are provided in [9] and it should be emphasized that a major novelty in our work is that, in LPC modeling, only the magnitude part is used given that the phase part is independently synthesized according to the NRD model. This approach ensures high quality signal reconstruction as the group delay of LPC models tends to be deleterious to the delicate harmonic phase structure of natural voiced sounds [15].

A. NRD shift-invariant spectral phase-related model

As we describe in [9], [14], after ODFT transformation of a windowed voice frame (N samples long), ϕ_ℓ denotes the phase of the ℓ-th harmonic, $\ell = 0, 1, \ldots, L-1$ relative to a reference point in that frame and that corresponds to the group delay of the frequency response of $h[n]$. A holistic phase structure of all harmonics in that frame can then be obtained as [14]

$$\mathrm{NRD}_\ell = \frac{n_\ell - n_0}{2\pi/\left((\ell+1)\omega_0\right)} = \frac{\phi_\ell - (\ell+1)\phi_0}{2\pi} \;, \qquad (1)$$

where $n_\ell = \phi_\ell/\left((\ell+1)\omega_0\right)$ represents the delay of the onset of the ℓ-th harmonic relative to the frame reference point. The phase structure is first computed as the relative delay (in samples) between the ℓ-th harmonic onset and the onset of the fundamental frequency. Then, this relative delay is further normalized by the period (in samples) of the ℓ-th harmonic. The resulting phase-related feature is named Normalized Relative Delay (NRD) and consists in a number in the range $[0.0,\ 1.0[$ for each harmonic. A convenient and powerful property of the NRD is that, by definition, $\mathrm{NRD}_0 = 0$, which means that NRD is time-shift invariant. In fact, the NRD is very informative because it is idiosyncratic [14] and, when combined with spectral magnitude information, it expresses the shape-invariance of the periodic waveform it represents [14], [16], independently of the fundamental frequency. Moreover, since NRD preserves the property of normalized phase, it can be wrapped and unwrapped, which facilitates modeling and interpretation. Other harmonic phase features that ressemble NRD were proposed by Stylianou [17], Di Federico [18], and Saratxaga [19].

In order to illustrate the shift-invariance and robustness of unwrapped NRD vectors, we analyze a natural voice signal in our database corresponding to the spoken /e/ vowel (as in 'leg'). The spectrogram of this signal is represented in the left-most panel in Fig. 7. The sampling frequency is 22050 Hz and the frame length is 1024 samples. Figure 2 represents an overlay of all ODFT magnitude spectra up to 7.5 kHz. These spectra are shift-invariant by nature and the smearing that is observed, especially for harmonics higher than the 9th harmonic, just reflects the fact that ω_0 is affected by micro-variations that occur in natural voice sounds as Fig. 7 clearly denotes. Figure 3 represents an overlay of all NRD vectors for the same signal, up to the 35th harmonic. This figure also represents the average NRD vector up to harmonic 20. It can be seen that the NRD vectors are much more stable than magnitude spectra are, which is a consequence of the fact that NRD vectors are independent of the fundamental frequency and, therefore, are immune to ω_0 micro-variations as long the shape of the waveform remains consistent. This is what happens in a sustained vowel since the glottal source excitation

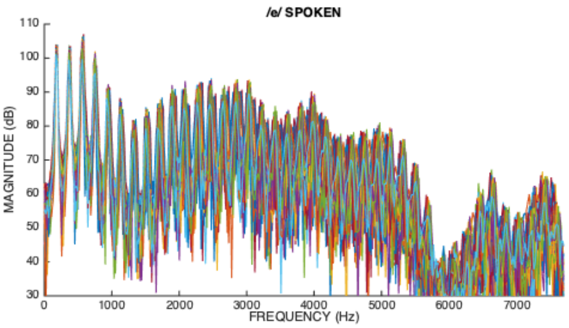

Fig. 2. Overlay of all magnitude spectra (up to 7.5 kHz) found in a sustained /e/ vowel spoken by a professional female singer.

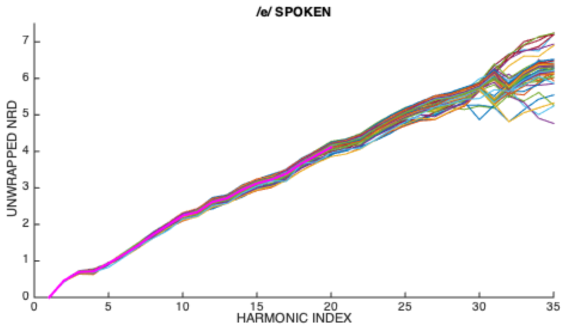

Fig. 3. Overlay of all unwrapped NRD vectors (first 35 harmonics) in a sustained /e/ vowel spoken by a professional female singer. The thick magenta line represents the average model up to harmonic 20.

and the supra-laryngeal filter remain essentially steady. Figure 3 also shows that after the 30th harmonic the NRD vectors start to disperse. This is the result of harmonics reaching the noise-floor which makes phase estimation less reliable. This, however, is not problematic since the NRD vector for the first 20 harmonics is quite stable encompassing the first four formant frequencies whose linguistic and idiosyncratic information is strong. Thus, the NRD of upper harmonics can seamlessly be set as the first-order extrapolation from that NRD model [9].

Figure 4 represents the average NRD models of all 10 vowel realizations in our database. It can be seen that spoken vowels give rise to more consistent NRD models than sung vowels do. This outcome is believed to be related to the fact that the sub-glottal pressure is stronger in singing which is likely to generate more irregular glottal pulse excitations, and to accentuate non-linear effects in the acoustic coupling between the glottis and supra-laryngeal structures.

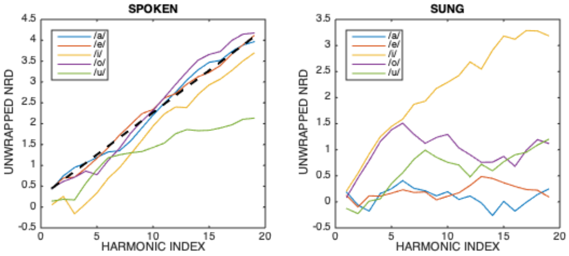

Fig. 4. Average NRD models pertaining to the 5 spoken vowels (left panel) and to the 5 sung vowels (right panel) in our database. The dashed line represents the average between /a/, /e/ and /o/ models.

B. *Spectral magnitude models*

Figure 5 represents an instance (blue line) of the magnitude spectra represented in Fig. 2. All harmonics in this magnitude spectrum are signaled by means of red triangles. As detailed in [9], a first spectral envelope model is obtained by using a harmonic interpolation approach that is similar to that proposed by Hermansky et al. [20] and to the non-iterative part of Discrete All Pole modeling [21]. This harmonic spectral envelope model is represented in Fig. 1 as HLPC.

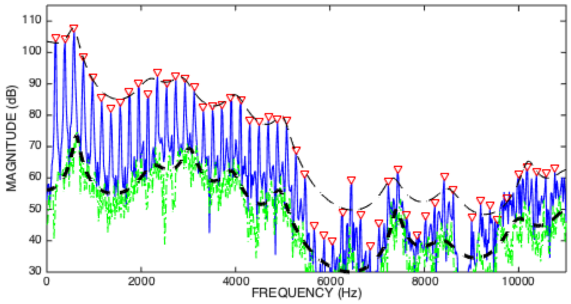

Fig. 5. Magnitude spectrum of a sustained /e/ vowel frame (solid blue line). Harmonics are signaled by vertical red triangles. The dashed black line represents a 22nd-order harmonic envelope model and the thick dashed black line represents the spectral envelope model of the noise-floor residual (dash-dotted green line) after accurate harmonic spectral subtraction.

Using the estimated spectral magnitudes (A_ℓ) and phases (ϕ_ℓ) and using window-aware accurate harmonic reconstruction (9 ODFT bins are reconstructed for each harmonic, as detailed in [9]), a noise-floor residual is obtained by subtracting the reconstructed harmonic structure from the original ODFT spectrum, in the complex ODFT domain, and in a non-iterative way. This represents a novelty in our approach and that is not found in other vocoders such as HMPD, STRAIGHT, or WORLD [22]. The magnitude of the spectral residual is represented in Fig. 5 by the green dash-dotted line. As in [9], autocorrelation coefficients are then computed from this residual taking advantage of the Wiener-Khintchine theorem. Finally, the noise-floor LPC model is obtained using the Levinson-Durbin recursion [23], [24]. This noise-floor spectral envelope model is represented in Fig. 1 as NLPC.

Both HLPC and NLPC spectral envelope models convey important linguistic and idiosyncratic information for natural signal reconstruction. However, due to the smooth properties of HLPC, very localized idiosyncratic vocal tract (anti-) resonances, and that provoke a (negative or) positive deviation of a given harmonic magnitude relative to HLPC, need to be captured such that the signal synthesis conveys the natural voice signature of a given speaker. Figure 6 illustrates the average magnitude deviation (and corresponding 95% Conf.

Int.) of individual harmonics in the case of the sustained /e/ vowel. We highlight that although a higher-order LPC model could give rise to smaller deviations in individual harmonic magnitudes, that is not desirable because a smooth harmonic envelope model captures linguistic information in a more representative and suitable way, whereas the individual magnitude deviations capture idiosyncratic features of the voice signature of a given speaker.

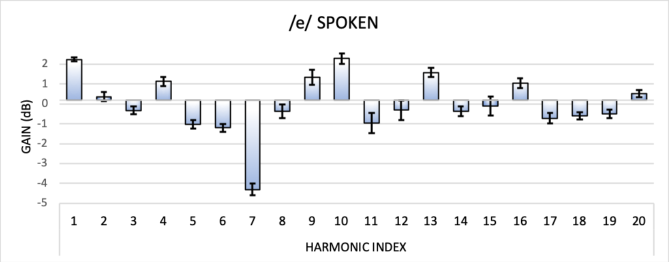

Fig. 6. Average and 95% C.I. of the dB deviation of the first 20 harmonics relative to the HLPC spectral envelope model (/e/ vowel).

The average individual harmonic magnitude deviations are represented in Fig. 1 as D_ℓ. This also represents represents a novelty in our approach that we believe is not used in other vocoders. All three spectral magnitude models (HLPC, NLPC and D_ℓ) have been obtained for all 5 spoken vowels and all 5 sung vowels in our research.

C. F0 models

We extracted the F0 contour of all five spoken vowels and all 5 sung vowels in our database for parametric modeling and re-synthesis purposes. Each contour is an F0 micro-variation model. Figure 7 illustrates the F0 contour in the case of the spoken /e/ vowel. Informal listening tests revealed that F0 micro-variations are not specific of a given vowel, or speaker. In fact, we have concluded that implanting the F0 micro-variation pattern from one vowel into another one, does not change the subjective perception of that vowel. Table I shows the average fundamental frequency and standard deviation for all sustained vowel realizations in our database. In all cases, the standard deviation is relatively low because all vowels have been produced in a sustained way, included in singing, where *vibrato* has been avoided.

TABLE I
AVERAGE F0 (AVGF0) AND F0 STANDARD DEVIATION (STDF0), IN HERTZ, OF ALL 5 SPOKEN AND ALL 5 SUNG VOWELS IN THE DATABASE.

	SPOKEN		SUNG	
	avgF0	stdF0	avgF0	stdF0
/a/	190.24	1.79	420.19	4.44
/e/	190.53	1.69	416.17	4.64
/i/	194.00	2.08	417.69	3.51
/o/	188.66	2.55	417.85	2.88
/u/	191.60	1.49	418.07	2.63

IV. RESULTS AND DISCUSSION

After extracting all five parametric models, as described above, for all 10 vowel realizations, we created synthetic versions using for each vowel its average HLPC, NLPC, D_ℓ and NRD models, and its estimated F0 contour. In each synthetic realization, only F0 and the noise floor change in time. The noise floor is obtained by shaping white Gaussian noise according to the NLPC spectral model. In addition, the overall signal magnitude was adjusted to be similar to that of the original vowel. This synthetic version was named "A". An anchor version, "B", was also created for each vowel by using the exact same synthesis conditions, except for the F0 contour which was forced to be constant and equal the the average F0 of that vowel throughout its duration. Then, we conducted listening tests[1] in which participants were asked to listen to both versions and to rate their quality when compared to the original version. Figure 7 illustrates the spectrograms of the original and A and B versions in the case of the spoken /e/ vowel. Rating was performed by using Rec. ITU-R BS.1116 for subjective assessment of small impairments [25], and using the 100-point Continuous Quality Scale (CQS), as specified in Rec. ITU-R BS.1284 [26]. This scale adopts the following intuitive ranges in the grading of the sound differences heard: imperceptible (80%-100%), perceptible but not annoying (60%-80%), slightly annoying (40%-60%), annoying (20%-40%), and very annoying (0%-20%). Headphone listening was also advised since some of the impairments are subtle. Twelve subjects participated in the tests, 7 male and 5 female. The average age is 32 (min. 16 and max. 50). The results are represented in Fig. 8. Results clearly indicate that accurate frequency-domain re-synthesis using all five parametric models delivers a quality that is essentially indistinguishable from the original signal, in both linguistic and speaker identity perspectives. These results essentially confirm our best results in [9], despite the fact that in each realization only F0 and the synthetic noise truly vary with time. In contrast, when the F0 is forced to be flat, which is a rather subtle modification, the synthetic versions sound

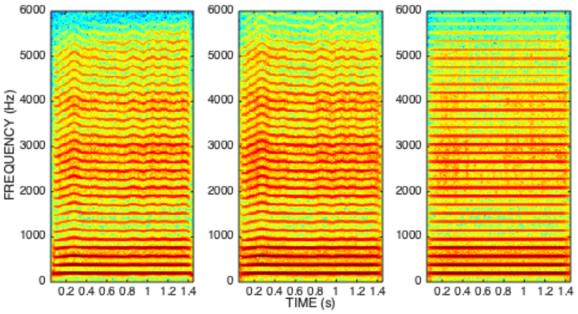

Fig. 7. Spectrogram of the original spoken /e/ vowel (left panel), and of the synthetic versions when the original F0 contour is preserved (centre panel), and when the average F0 is enforced (right panel).

[1] Instructions and all audio signals are available: `http://fe.up.pt/~ajf/EUSIPCO2020_subjTEST.pptx.zip`

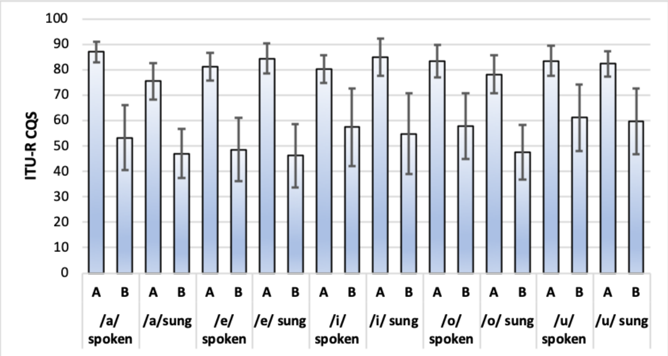

Fig. 8. Means and 95% C.I. of the listening test results. A versions preserve the original F0 contour and B versions have a flat F0.

clearly different from the original. Differences between the A and B results are all statistically significant ($p < 0.01$). This confirms that F0 micro-variations are very important in conveying a strong sense of naturalness. After these results were obtained, informal listening tests were conducted comparing our synthetic signals and the re-synthesized versions of the same input audio material using three vocoders: HMPD, STRAIGHT, and WORLD [22]. In all cases, we concluded that our approach clearly produced higher-quality synthetic signals. New extended listening tests are currently being prepared.

Finally, we created synthetic versions with the same duration and F0 contour of the spoken and sung /i/ vowels, but where all five vowels were synthesized, in sequence, for 1/5 of the signal duration, and where transitions were implemented as a simple linear interpolation between the parametric models of adjacent vowels. Listeners reported that the synthetic versions did not present objectionable artifacts and that the sung version sounded more natural than the spoken version, probably because the high F0 in this case helps to mitigate the absence of the /i/ vowel 'intrinsic pitch' in the first case. This feedback confirmed that smooth vowel co-articulation is obtained through simple interpolation between the phase-related (NRD) and spectral magnitude-related (HLPC, NLPC, D_ℓ) parametric models, which represents a very important flexibility in parametric-oriented artificial voicing of whispered speech.

V. Conclusion

In this paper we described an accurate procedure that relies on five independently controllable parametric features and that is able to synthesize high-quality voiced regions, ensuring linguistic correctness and preserving idiosyncratic characteristics. Subjective listening test results confirmed that the quality of synthetic spoken and sung vowels is essentially indistinguishable from the original signals. The procedure requires only minimal voice recordings such that parametric models can be built for a representative diversity of voiced phonemes, and is tailored for real-time operation by directly implanting synthetic voicing on selected regions of whispered speech.

Acknowledgment

The authors would like to thank the acclaimed Portuguese Artist Teresa Salgueiro for her kind collaboration, and all the subjects for their participation in the listening tests.

References

[1] L. Rabiner and B-H Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, Inc., 1993.

[2] R. W. Morris and M. A. Clements, "Reconstruction of speech from whispers," *Med. Eng. & Physics*, vol. 24, no. 7, pp. 515–520, 2002.

[3] I. V. Mcloughlin, H. R. Sharifzadeh, et al., "Reconstruction of phonated speech from whispers using formant-derived plausible pitch modulation," *ACM Trans. Access. Comput.*, vol. 6, no. 4, 2015.

[4] T. Toda, M. Nakagiri, and K. Shikano, "Statistical voice conversion techniques for body-conducted unvoiced speech enhancement," *IEEE TASSP*, vol. 20, no. 9, pp. 2505–2517, Nov. 2012.

[5] H. Lian, Y. Hu, et al., "Whisper to normal speech based on deep neural networks with MCC and F0 features," in *2018 IEEE 23rd International Conference on Digital Signal Processing (DSP)*, 2018, pp. 1–5.

[6] P. K. Ghosh G. N. Meenakshi, "Whispered speech to neutral speech conversion using bidirectional lstms," in *Interspeech*, 2018.

[7] H. Konno, M. Kudo, et al., "Whisper to normal speech conversion using pitch estimated from spectrum," *Speech Com.*, vol. 83, pp. 10–20, 2016.

[8] A. Ferreira, "Implantation of voicing on whispered speech using frequency-domain parametric modelling of source and filter information," in *ISIVC*, 2016, pp. 159–166, Tunis, Tunisia.

[9] A. Ferreira, J. Silva, F. Brito, and D. Sinha, "Impact of a shift-invariant harmonic phase model in fully parametric harmonic voice representation and time/frequency synthesis," in *IEEE ICASSP*, 2020.

[10] M. Bellanger, *Digital Processing of Signals*, John Willey & Sons, 1989.

[11] Aníbal J. S. Ferreira and Ricardo Sousa, "DFT-based frequency estimation under harmonic interference," in *4th International Symposium on Communications, Control and Signal Processing*, March 2010.

[12] Aníbal Ferreira, Filipe Abreu, and Deepen Sinha, "Stereo ACC real-time audio communication," *125th Convention of the Audio Engineering Society*, October 2008, Paper 7502.

[13] A. Ferreira and D. Sinha, "Accurate and robust frequency estimation in the ODFT domain," in *2005 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, October 2005, pp. 203–206.

[14] Aníbal J. Ferreira and José M. Tribolet, "A holistic glotal phase related feature," in *21st International Conference on Digital Audio Effects (DAFx-18)*, 2018, Aveiro, Portugal.

[15] Aníbal Ferreira, "On the physiological validity of the group delay response of all-pole vocal tract modeling," *145th Convention of the Audio Engineering Society*, October 2018, Paper 10038.

[16] Thomas. F. Quatieri and Robert J. McAulay, "Shape invariant time-scale and pitch modification of speech," *IEEE Transactions on Signal Processing*, vol. 40, no. 3, pp. 497–510, March 1992.

[17] Ioannis Stylianou, *Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification*, Ph.D. thesis, École Nationale Supérieure des Télécom., France, 1996.

[18] Riccardo Di Federico, "Waveform preserving time stretching and pitch shifting for sinusoidal models of sound," in *COST-G6 Digital Audio Effects Workshop*, 1998, pp. 44–48.

[19] I. Saratxaga, I Hernaez, D. Erro, et al., "Simple representation of signal phase for harmonic speech models," *El. Letters*, vol. 45, no. 381, 2009.

[20] H. Hermansky, H. Fujisaki, and Y. Sato, "Spectral envelope sampling for interpolation in linear predictive analysis of speech," in *IEEE ICASSP*, 1984, pp. 2.2.1–2.2.4.

[21] Amro El-Jaroudi and John Makhoul, "Discrete all-pole modeling," *IEEE Trans. on Signal Processing*, vol. 39, no. 2, pp. 411–423, February 1991.

[22] Masanori Morise and Yusuke Watanabe, "Sound quality comparison among high-quality vocoders by using resynthesized speech," *Acoust. Sci. and Tech.*, vol. 39, no. 3, pp. 263–265, 2018.

[23] Alan V. Oppenheim and Ronald W. Schafer, *Discrete-Time Signal Processing*, Pearson Higher Education, Inc., 2010.

[24] Monson H. Hayes, *Statistical Digital Signal Processing and Modeling*, John Wiley & Sons Inc., 1996.

[25] ITU-R Recommendation BS.1116-3, "Methods for the subjective assessment of small impairments in audio systems," February 2015.

[26] ITU-R Recommendation BS.1282-2, "General methods for the subjective assessment of sound quality," January 2019.