PARAMETRIC MODEL OF SPECTRAL ENVELOPE TO SYNTHESIZE REALISTIC INTENSITY VARIATIONS IN SINGING VOICE

Emilio Molina, Isabel Barbancho, Ana M. Barbancho, Lorenzo J. Tardón

Universidad de Málaga, Andalucía Tech, ATIC Research Group,
ETSI Telecomunicación, Campus de Teatinos s/n, 29071 Málaga, SPAIN
emm@ic.uma.es, ibp@ic.uma.es, abp@ic.uma.es, lorenzo@ic.uma.es

ABSTRACT

In this paper, we propose a method to synthesize the natural variations of spectral envelope as intensity varies in singing voice. To this end, we propose a parametric model of spectral envelope based on novel 4-pole resonators as formant filters. This model has been used to analyse 60 vowels sung at different intensities in order to define a set of functions describing the global variations of parameters along intensity. These functions have been used to modify the intensity of 16 recorded vowels and 8 synthetic vowels generated with Vocaloid. The realism of the transformations performed with our approach has been evaluated by four amateur musicians in comparison to Melodyne for real sounds and to Vocaloid for synthetic sounds. The proposed approach has been proved to achieve more realistic sounds than Melodyne and Vocaloid, especially for loud-to-weak transformations.

Index Terms— Human voice, Acoustic signal processing, Speech Synthesis

1. INTRODUCTION

In recent years, the development of software to process and/or synthesize singing voice with creative purposes has become trendy [1, 2, 3]. In this context, a commonly addressed problem is related to the definition of singing processing algorithms (e.g. pitch shifting) able to preserve the spectral envelope of the original sound in order to avoid uncontrolled changes of timbre [4]. However, the spectral envelope has been proved to vary in different contexts of intensity and pitch [5, 6]. This fact motivated the development of the work presented in this paper. Indeed, timbre variations along F0 or intensity are commonly annotated by phoniatricians with different colors in the phonetogram [7] (a graph that displays the dynamic range of a given singer in terms of fundamental frequency and intensity).

This work has been funded by the Ministerio de Economía y Competitividad of the Spanish Government under Project No. TIN2010-21089-C03-02 and Project No. IPT-2011-0885-430000, by the Junta de Andalucía under Project No. P11-TIC-7154 and by the Ministerio de Educación, Cultura y Deporte through the 'Programa Nacional de Movilidad de Recursos Humanos del Plan Nacional de I-D+i 2008-2011, prorrogado por Acuerdo de Consejo de Ministros de 7 de octubre de 2011'. The work has been done in the context of Campus de Excelencia Internacional Andalucía Tech, Universidad de Málaga.

Prior research on singing processing/synthesis generally focus on artifact reduction [8], formant preservation [4][9], realistic f_0 evolution [10][11], etc. since these aspects are important to achieve a natural result. However, to the best of our knowledge, these references do not address the natural changes of the spectral envelope along intensity or pitch, and we consider that this gap limits the naturalness of the processed sounds.

In this paper, we address the modelling and synthesizing schemes of the natural variations of the spectral envelope as the intensity of the singing voice varies. We propose a parametric model of the spectral envelope, which has been implemented in a freely available software tool able to analyse and process recorded vowels (Section 2). Making use of this tool, we have analysed a set of 60 sung vowels at different intensities and we have defined a set of functions that describe the variations of parameters along intensity (Section 3). We have used these functions to modify the intensity of 24 sung notes and we have evaluated the naturalness of the processed sounds though a subjective analysis (Section 4). The results of this evaluation are presented in Section 5. Some conclusions about our study are given in Section 6.

2. PROPOSED PARAMETRIC MODEL OF SPECTRAL ENVELOPE

In this section, we describe a novel parametric model of the spectral envelope (Section 2.1), which is used to parametrize voiced sounds. Additionally, in Section 2.2 we describe the used procedure to estimate the parameters from real sounds.

2.1. Proposed parametric model of spectral envelope

Regarding the existing approaches to parameterize the spectral envelope of speech, Linear Prediction Coding (LPC) is one the most used techniques [12]. LPC efficiently fits the input signal using a N-order all-pole filter, which is appropriate to model the vocal tract acoustic response [13]. However, the optimal order of the filter is hard to obtain, and LPC technique contains systematic errors that are specially manifested in high-pitched signals [14]. Moreover, LPC coefficients are hard to manipulate to perform timbre transformations. Later related approaches, like Line Spectral Frequencies (LSF) [15]

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

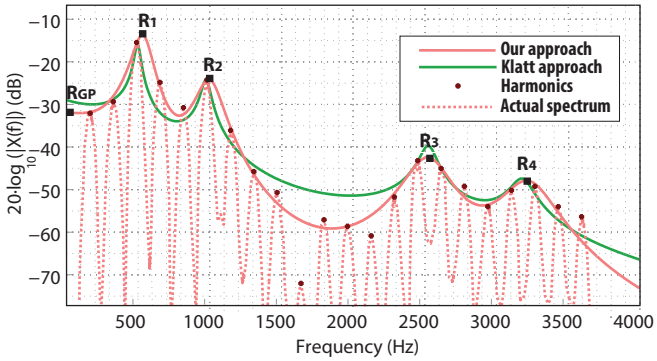

Fig. 1. Comparison between our approach (4-poles resonators) and Klatt's approach (2-poles resonators) to fit the harmonic spectrum of a */o/* vowel sung by a male singer.

or cepstrum-based techniques [14], partially improve some of these drawbacks. However, the parameters of these models are hard to manipulate in order to obtain natural timbre variations, since they are not directly related to physical aspects of the human vocal production. Therefore, we have discarded LPC, LSF or cepstrum-based techniques and we have focused on formant-based models, whose parameters are harder to obtain, but easier to manipulate to synthesize natural timbre variations.

The finally chosen approach is a formant-based parametric model of the spectral envelope, inspired by previous systems for speech/singing synthesis like [16] or [17]. This type of model consider that the spectral envelope of speech signals can be synthesized with several resonator filters in parallel (equivalent to the acoustic formants) with a certain overall slope (determined by the glottal source).

In our approach, we model the spectrum of sung vowels in the frequency band [0, 5000 Hz] with a source filter (as mentioned in [17], determining the overall slope of the spectrum) in cascade with a set of five parallel resonators (the glottal resonator R_{GP} + the first four formants of the vocal tract: R_1, R_2, R_3 and R_4). Therefore, the final envelope $|E(f)|_{dB}$ is defined by the following expression:

$$|E(f)|_{\text{dB}} = \text{Source}(f)_{\text{dB}} + \text{Resonances}(f)_{\text{dB}} \quad (1)$$

where:

$$\text{Source}(f)_{\text{dB}} = \text{Gain}_{\text{dB}} + \text{SlopeDepth}_{\text{dB}} \left(e^{\alpha \cdot f} - 1\right) \quad (2)$$

$$\text{Resonances}(f)_{\text{dB}} = 20 \log_{10} \left(|R_{GP}(f)| + \sum_{i=1}^{4} |R_i(f)| \right) \quad (3)$$

and where the constant $\alpha = -4 \cdot 10^{-4}$, and $R_i(f)$ is the frequency response of resonator i. In the literature, vocal formants are typically modelled with 2-poles resonators, which is the approach of well-known Klatt's speech synthesizer[16]. However we propose the use of 4-poles resonators, which proved to obtain better results at modelling the voiced sounds of our dataset. In Figure 1, we show an example in which our approach fits a natural spectral envelope more accurately than Klatt's approach. Specifically, the proposed 4-poles resonator consists of two identical 2nd-order filters in cascade, with poles $p_1 = p_2 = \rho \cdot e^{\theta}$ and $p_3 = p_4 = \rho \cdot e^{-\theta}$, defined by the following equation:

$$R_x(z) = \frac{K \cdot (1 - 2 \cdot \rho \cdot cos(\theta) + \rho^2)}{(1 - 2 \cdot \rho \cdot cos(\theta) z^{-1} + \rho^2 z^{-2})^2 \, (1 - \rho)^2} \quad (4)$$

$$\text{being: } K = 10^{-4} \quad \rho = e^{-\pi B_x / f_s} \quad \theta = \frac{2 \cdot \pi \cdot f_x}{f_s} \quad (5)$$

where f_x is the central frequency in Hz of resonator x, B_x is the 6 dB-bandwidth in Hz (it should not be confused with the typical 3 dB-bandwidth) and f_s is the sampling rate in Hz. In Table 1 we provide all the specific values of our model (based on [16]) to make our experiments reproducible.

Parameter	Range	Parameter	Range
Gain_{dB}	$[-200, 0]$ dB	f_2	$[500, 3000]$ Hz
$\text{SlopeDepth}_{\text{dB}}$	$[-50, 100]$ dB	B_2	$[40, 1000]$ Hz
f_{GP}	$[0, 600]$ Hz	f_3	$[500, 3000]$ Hz
B_{GP}	$[100, 2000]$ Hz	B_3	$[40, 1000]$ Hz
f_1	$[150, 1100]$ Hz	f_4	$[3000, 5000]$ Hz
B_1	$[40, 1000]$ Hz	B_4	$[100, 1000]$ Hz

Table 1. Range of values used to fit the spectrum of real sounds with the proposed model.

2.2. Estimation of parameters

The automatic estimation of formant frequencies and bandwidths from recorded audio is not straightforward, and such is one of the drawbacks of the proposed model. In the literature we can find many algorithms for formants estimation [18, 19], some of which are included in software tools like Praat [20]. However, we realized that they are not free of errors and sometimes may require human intervention for a reliable annotation of sounds [21]. Due to this, we have manually corrected the parameters obtained with Praat in order to accurately fit the target envelope. The harmonic and the residual component have been separated using the algorithms described in [22, 23] in order to analyse them independently .

3. ANALYSIS OF THE VARIATION OF PARAMETERS ALONG INTENSITY

We have analysed a dataset of sung notes (described in Section 3.1) in order to model the variation of parameters along intensity for both the harmonic (Section 3.2) and the residual component (Section 3.3). Then, we use this information to create a model that emulates the natural changes of spectral envelope for different degrees of intensity (Section 3.4).

3.1. Analysis dataset

We have annotated a total of 60 sustained notes sung by two male and two female singers with ages between 20 and 40

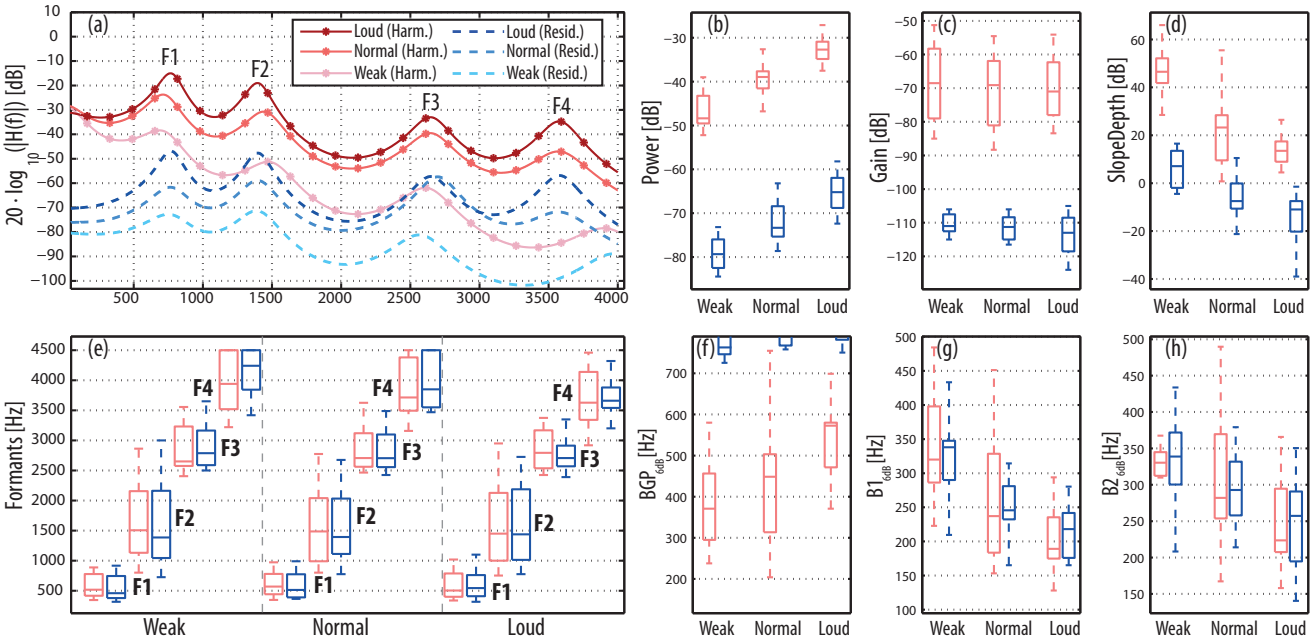

Fig. 2. Information about the harmonic component (light red color) and the residual component (dark blue color) at different degrees of intensity **(2.a)** Spectral envelope of an */a/* vowel sung by a male singer **(2.b)** Average power **(2.c)** Average Gain **(2.d)** Average SlopeDepth **(2.e)** Average frequency values of the first four formants **(2.f)** Average bandwidths of the glottal resonator R_{GP} **(2.g)** Average bandwidths of the first formant R_1 **(2.h)** Average bandwidths of the second formant R_2

years old. All of them have some experience as singers in amateur pop bands, but they do not have academic vocal training. The 60 sung notes correspond to 5 different sustained vowels (/a/, /e/, /i/, /o/ and /u/), in 3 different intended intensities (weak, normal, loud) for 4 different singers. All the notes were sung in a comfortable pitch register for all singers. The recordings were made in a semi-anechoic chamber with a microphone Neumann TLM103, a pop shield and a Onyx-Blackbird firewire interface with a sample rate of 11025 Hz (since we are modelling the frequency band [0,5000] Hz). Each singer was told to keep the same distance to the microphone for all the recordings.

3.2. Analysis for the harmonic envelope

In the analysis, the parameters were averaged over all the notes of the dataset in order to study their global tendency along intensity (see Fig. 2). According to the results, the parameter most noticeably affected by the changes of intensity is the slope depth (Figure 2.d). Indeed, vocal effort has been proved to affect the slope of the source spectrum in prior studies [24]. Moreover, we observed that the bandwidths of the various formants (Figures 2.f, 2.g and 2.h) are reduced as the intensity increases. This phenomenon, which is clearly manifested in our experiments, is coherent with the nonlinear energy damping model of vocal resonators proposed in [25], in which the Q of the filters depends on the input signal. Surprisingly, the gain is not strongly affected by the degree of intensity (2.c), since the measured variations of power (2.b) are mainly due to more prominent formants (decrease of bandwidths) and an increase of high frequencies (decrease of SlopeDepth). Additionally, the shifting of formants frequencies along intensity is slight, and there is not such a clear pattern.

3.3. Analysis for the residual component

The residual and the harmonic component behaves in a similar way as the intensity increases: the slope depth and the bandwidth of the formants decrease, the gain remains rather stable and the formants frequencies are not strongly modified. Additionally, the ratio between harmonic and residual power remains surprisingly stable for different intensity levels (when the whole bandwidth is considered). However, we have observed that such ratio is lower at high frequencies, since the slope depth is generally higher for the harmonic component (this effect is especially noticeable in weak notes). Additionally, we observed that the residual component of loud notes sounds rather "creaky", whereas weak notes has a more breathy texture.

3.4. Proposed model to modify the intensity of the notes

We have defined the *variation of intensity* ΔI as the desired change of intensity to produce. A positive variation produces an increase of intensity, and a negative variation a decrease. The parameter has been normalized so that a step $\Delta I = \pm 10$ produces a complete change from weak to loud, or viceversa.

Typically, $\Delta I = \pm 1$ is a reasonable unit to gradually increase or decrease the intensity of the signal.

Each parameter is modified according to the following expression:

$$p'_x = p_x + \Delta I \cdot w_x \quad (6)$$

where p'_x is the new value of parameter x, p_x is the old value of parameter x, and w_x is the specific weight of paramter x. Additionally, p'_x must always be limited to the range presented in Table 1. In Table 2, we show the specific weights w_x associated to all the parameters for both the residual and the harmonic component. These values have been obtained through linear regression on the analysis dataset described in Section 3.1.

$\mathbf{p_x}$	$\mathbf{w_x}$ (harm.)	$\mathbf{w_x}$ (res.)	$\mathbf{p_x}$	$\mathbf{w_x}$(harm.)	$\mathbf{w_x}$(res.)
Gain_{dB}	0.00 dB	−0.30 dB	F_{R2}	0 dB	0.95 Hz
$\text{SlopeDepth}_{\text{dB}}$	−3.00 dB	−2.04 Hz	B_{R2}	−5.20 dB	−8.26 Hz
F_{R0}	−8.15 Hz	2.33 Hz	F_{R3}	−1.70 Hz	−14.16 Hz
B_{R0}	15.50 Hz	−9.59 Hz	B_{R3}	−2.25 Hz	−8.53 Hz
F_{R1}	0 Hz	5.83 Hz	F_{R4}	−21.76 Hz	−42.16 Hz
B_{R1}	−8.00 Hz	−10.91 Hz	B_{R4}	−2.31 Hz	−9.28 Hz

Table 2. Proposed weights of parameters to modify the perceived intensity of sung notes

4. EVALUATION

In this section, we describe the dataset of sung vowels used for the evaluation (Section 4.1), as well as the used evaluation methodology (Section 4.2).

4.1. Evaluation dataset

We have collected 12 pairs of weak-loud sung vowels in mono audio with a sample rate of 11025 Hz: 4 weak-loud pairs sung by two singers (male M1 and female F1) taken from the analysis dataset (Section 3.1), 4 sung by two singers (male M2 and female F2) not analysed before, and 4 pairs synthesized with "Bruno" (VM) and "Clara" (VF) singers in Vocaloid 3.0. Each singer (either real or synthetic) has sung a weak-loud pair using both an open vowel (*/a/*) and a closed vowel (*/i/*)) in a comfortable register.

4.2. Evaluation methodology

In the case of natural vowels, we have compared our approach ($\Delta I = \pm 10$) with Melodyne Editor (state-of-the-art commercial software). In the case of synthetic vowels, we have compared our approach ($\Delta I = \pm 10$) with Vocaloid 3.0 by setting the parameter *Dynamics* to 127 (loud vowels) and 32 (weak vowels). It makes a total of 48 pairs of weal-loud or loud-weak changes[1]. The evaluation has been performed by four amateur musicians, who listened (with high quality headphones) the different systems in random order, and they were asked to evaluate how close to a real change of intensity was the applied processing.

5. RESULTS

In Figure 3 we show the perceived closeness to a real change of intensity for each of the 48 pairs described in Section 4.2.

[1] Available at: http://www.atic.uma.es/icassp2014singing

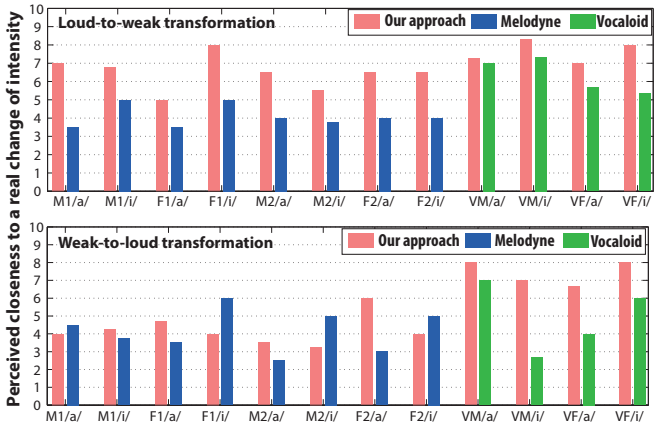

Fig. 3. Mean perceived closeness to a real change of intensity. Each specific combination of singer/vowel (see Section 4.2) has been evaluated with various approaches, represented with different colours.

In general, our approach achieves better results for loud-to-weak transformations, whereas in the case of weak-to-loud transformations, the results are less realistic. Indeed, we have observed that formants are less defined in weak sounds (see example in Figure 2.a), and therefore they are harder to analyse and manipulate. Regarding the results with synthetic vowels, our approach achieves more realism than Vocaloid at modifying the intensity for all cases.

6. CONCLUSIONS

In this paper, we have presented a novel parametric model of spectral envelope to produce realistic variations of intensity in recorded or synthetic vowels. Our model is inspired by previous approaches like [16] or [26], but we introduce some improvements to fit more accurately the spectral envelope of real sounds. Specifically, we propose the use of 4-poles resonators to synthesize the vocal formants, instead of 2-poles resonators. Using our parametric model, a set of 60 sung vowels (natural and synthesized with Vocaloid 3.0) at different intensities have been analysed with Praat (combined with manual annotation) in order to define a set of functions describing the variation of parameters along intensity in singing voice. These functions have been applied to 16 recorded and 8 synthetic vowels (generated with Vocaloid 3.0) to modify their intensity. We have also modified the intensity of the 16 natural vowels with Melodyne Editor, and of the 8 synthetic ones with Vocaloid. The realism of the transformations has been evaluated by four amateur musicians through a survey. The results showed that our approach is especially good when dealing with synthetic vowels, but it also performs well in loud-to-weak transformations with real sounds. In the future, we plan to apply our approach to model the natural changes of spectral envelope along pitch, since it could contribute to extend the idea of *envelope preservation* [4], which is recurrent in many state-of-the-art pitch shifting algorithms.

7. REFERENCES

[1] "Celemony Software: Melodyne Editor," Official website: //http://www.celemony.com.

[2] H. Kenmochi and H. Ohshita, "VOCALOID-commercial singing synthesizer based on sample concatenation.," in *INTERSPEECH*, 2007, pp. 4009–4010.

[3] C. Roig, I. Barbancho, E. Molina, L. J. Tardón, and A. M. Barbancho, "Rumbator: A flamenco rumba cover version generator based on audio processing at note-level," in *DAFx-13*, 2013.

[4] A. Röbel and X. Rodet, "Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation," in *Proc. DAFx*, 2005.

[5] J. E. Huber, E. T. Stathopoulos, G. M. Curione, T. A. Ash, and K. Johnson, "Formants of children, women, and men: The effects of vocal intensity variation," *The Journal of the Acoustical Society of America*, vol. 106, pp. 1532, 1999.

[6] K. Chládková, P. Boersma, and V. J. Podlipskỳ, "On-line formant shifting as a function of F0.," in *INTERSPEECH*, 2009, pp. 464–467.

[7] P. Pabon, "Manual of Voice Profiler Version 4.0," 2010, http://kc.koncon.nl/staff/pabon.

[8] J. Laroche, "Frequency-domain techniques for high quality voice modification," *Proc. of DAFx-03*, pp. 328–322, 2003.

[9] D. Arfib, F. Keiler, U. Zölzer and V. Verfaille "DAFX: Digital Audio Effects, 2nd Edition", p.279-320, Willey, Chichester, UK, 2011.

[10] M. Saitou, T. Unoki and M. Akagi, "Development of an F0 control model based on F0 dynamic characteristics for singing-voice synthesis," *Speech communication*, vol. 46, no. 3, pp. 405–417, 2005.

[11] Y. Ohishi, H. Kameoka, D. Mochihashi, K. Kashino, "A Stochastic Model of Singing Voice F0 Contours for Characterizing Expressive Dynamic Components," *INTERSPEECH*, 2012.

[12] D. O'Shaughnessy, "Linear predictive coding," *Potentials, IEEE*, vol. 7, no. 1, pp. 29–32, 1988.

[13] B. Atal and J. Remde, "A new model of LPC excitation for producing natural-sounding speech at low bit rates," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'82*. IEEE, 1982, vol. 7, pp. 614–617.

[14] Axel Röbel, Fernando Villavicencio, and Xavier Rodet, "On cepstral and all-pole based spectral envelope modeling with unknown model order," *Pattern Recognition Letters*, vol. 28, no. 11, pp. 1343–1350, 2007.

[15] P. Kabal and R. P. Ramachandran, "The computation of line spectral frequencies using Chebyshev polynomials," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 34, no. 6, pp. 1419–1426, 1986.

[16] D. H. Klatt, "Software for a cascade/parallel formant synthesizer," *the Journal of the Acoustical Society of America*, vol. 67, pp. 971, 1980.

[17] J. Bonada, O. Celma, A. Loscos, J. Ortolà, X. Serra, Y. Yoshioka, H. Kayama, Y. Hisaminato, and H. Kenmochi, "Singing voice synthesis combining excitation plus resonance and sinusoidal plus residual models," in *Proceedings of International Computer Music Conference*. Citeseer, 2001.

[18] R. C. Snell and F. Milinazzo, "Formant location from LPC analysis data," *Speech and Audio Processing, IEEE Transactions on*, vol. 1, no. 2, pp. 129–134, 1993.

[19] C. Glaser, M. Heckmann, F. Joublin, and C. Goerick, "Combining auditory preprocessing and Bayesian estimation for robust formant tracking," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 2, pp. 224–236, 2010.

[20] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 5.1. 05)[computer program]," *online: http://www. praat. org*, 2009.

[21] S. A. Fulop, "Accuracy of formant measurement for synthesized vowels using the reassigned spectrogram and comparison with linear prediction," *The Journal of the Acoustical Society of America*, vol. 127, pp. 2114, 2010.

[22] J. Bonada, X. Serra, X. Amatriain, and A. Loscos, "Spectral processing," *DAFX Digital Audio Effects*, pp. 393–444, 2011.

[23] E. Molina, A. Barbancho, L. Tardon, and I. Barbancho, "Dissonance reduction in polyphonic audio using harmonic reorganization," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, 2013.

[24] J. Sundberg and M. Nordenberg, "Effects of vocal loudness variation on spectrum balance as reflected by the alpha measure of long-term-average spectra of speech," *The Journal of the Acoustical Society of America*, vol. 120, pp. 453, 2006.

[25] H. Ohmura and K. Tanaka, "Speech synthesis using a nonlinear energy damping model for the vocal folds vibration effect," in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*. IEEE, 1996, vol. 2, pp. 1241–1244.

[26] J. Bonada, "High quality voice transformations based on modeling radiated voice pulses in frequency domain," in *Proc. Digital Audio Effects (DAFx)*, 2004.