10th ISCA Speech Synthesis Workshop
20-22 September 2019, Vienna, Austria

GlottDNN-based spectral tilt analysis of tense voice emotional styles for the expressive 3D numerical synthesis of vowel [a]

Marc Freixes, Marc Arnela, Francesc Alías and Joan Claudi Socoró

GTM – Grup de recerca en Tecnologies Mèdia, La Salle - Universitat Ramon Llull
Quatre Camins, 30, 08022 Barcelona, Spain
{marc.freixes,marc.arnela,francesc.alias,joanclaudi.socoro}@salle.url.edu

Abstract

Three-dimensional (3D) acoustic models allow for an accurate modelling of acoustic wave propagation in 3D realistic vocal tracts. However, voice generated by these approaches is still limited in terms of expressiveness, which could be improved through proper modifications of the glottal source excitation. This work aims at adding some expressiveness to a 3D numerical synthesis approach based on the Finite Element Method (FEM) that uses as input an LF (Liljencrants-Fant) model controlled by the glottal shape parameter R_d. To that effect, a parallel Spanish speech corpus containing neutral and tense voice emotional styles is analysed with the GlottDNN vocoder, obtaining $F0$ and spectral tilt parameters associated with the glottal excitation. The variations of these two parameters are computed for happy and aggressive styles with reference to neutral speech, differentiating between stressed and unstressed vowels [a]. From this analysis, $F0$ and R_d values are then derived and used in the LF-FEM based synthesis of vowels [a] to resemble the aforementioned expressive styles. Results show that it is necessary to increase $F0$ and decrease R_d with respect to neutral speech, with larger deviations for happy than aggressive style, especially for the stressed vowels.

Index Terms: expressive speech synthesis, glottal source modelling, LF model, voice production, finite element method, spectral tilt

1. Introduction

Three-dimensional (3D) acoustic models are currently being developed to generate synthetic voice. These models simulate the propagation of 3D acoustic waves through realistic vocal tracts, typically obtained from Magnetic Resonance Imaging (MRI) (see e.g., [1]). The classical plane wave assumption required by 1D models is thus avoided, increasing the accuracy of the generated voice especially above 5 kHz where higher order modes also propagate [2, 3]. Many 3D approaches can be found in literature. The most extended ones are those based on the Finite Element Method (FEM) [4], but also on finite differences [5], digital waveguides [6], and multimodal approaches [2]. However, to the authors' knowledge, 3D acoustic models still present limitations in the generation of expressive voice.

This expressiveness can be incorporated to a 3D acoustic model through the proper modification of the glottal excitation characteristics. The glottal flow is closely linked to some of the primary prosodic features, such as pitch and energy, which are important to reproduce a certain speaking style. However, expressiveness is also conveyed by secondary prosodic features associated to voice quality [7]. Given that the latter are difficult to obtain from speech signals [8], several works have studied the contribution of voice quality on the generation of expressive speaking styles by means of inverse filtering and copy-synthesis. In [9], the parameters of modal stimuli were modified with the KLSYN88 synthesiser to study the mapping of $F0$ contours and voice quality on affect for different languages. Similarly, a 1D articulatory synthesizer was used in [8] to analyse the impact of the phonation type on the perception of emotions in German vowels. Some approaches have introduced parametric glottal flow models in the copy-synthesis process. For instance, an LF (Liljencrants-Fant) model controlled by the R_d glottal shape parameter [10] was used in [11] to simulate the tense-lax continuum and explore its affective correlates. Likewise an Auto-Regressive eXogenous variant of the LF model was proposed in [12] to analyse the contribution of glottal source and vocal tract to the perception of emotions. The aforementioned approaches usually involve manual tuning in the inverse filtering process. However, recent advances in inverse filtering techniques [13] have allowed for competitive glottal vocoders [14]. These are able to automatically analyse a speech corpus, decompose glottal source and vocal tract response, and parameterise them independently. These parameters have been proved useful to capture expressive nuances [15]. In this context, it has been recently proposed a GlottDNN-based speaking style conversion from natural to Lombard speech [16].

In this work, we aim at incorporating some expressiveness to a 3D FEM-based acoustic model that uses an LF model as glottal excitation. In [17], the R_d parameter was considered to control the LF model in the generation of synthetic voice with lax, modal, and tense phonations. That preliminary work is here extended by investigating how the LF model could be configured to generate tense voice emotional styles. To that effect, we use the GlottDNN vocoder to analyse the glottal excitation characteristics of a parallel speech corpus composed of paired utterances in neutral, happy and aggressive speaking styles. Subsequently, the values derived from this analysis are translated to the LF-FEM based synthesis of vowel [a], and the results are compared in terms of the obtained long term average spectra.

The paper is organised as follows. Section 2 details the methodology followed to analyse the glottal source properties on an expressive speech corpus, and subsequently incorporate some of these characteristics in the LF-FEM based synthesis. Next, the obtained results are described and discussed in Section 3. Finally, Section 4 closes the paper with the conclusions.

2. Methodology

Figure 1 depicts a workflow diagram describing the methodology proposed to incorporate some expressiveness in the LF-FEM based synthesis approach. On the one hand, there is a natural speech parallel corpus of paired utterances, which contain N vowels for each of the K expressive styles (EXP_k) and for the neutral style (NEU). On the other hand, a synthetic

10.21437/SSW.2019-24

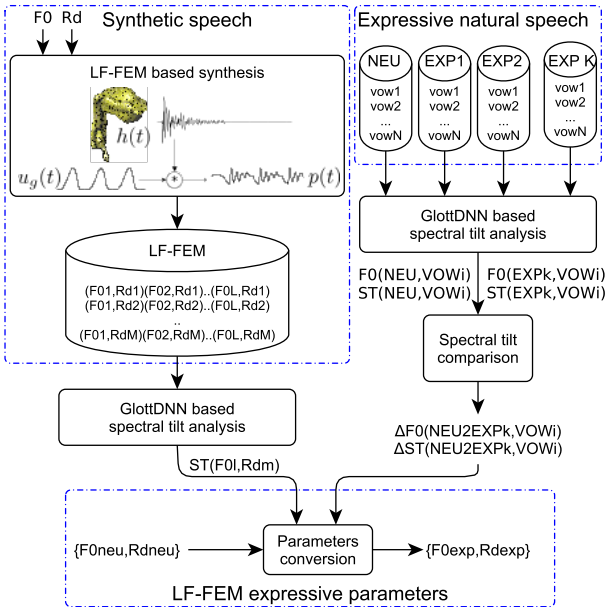

Figure 1: *Workflow diagram used for the analysis and comparison of expressive natural speech respect with synthetic speech generated with 3D FEM-based acoustic model that uses an LF model as glottal excitation.*

speech corpus (LF-FEM) is built using a 3D FEM-based acoustic model with the glottal source generated with an LF model, doing a sweep from $F0_1$ to $F0_L$ and from R_{d1} to R_{dM}. Both the natural and synthetic utterances are then inverse filtered by the GlottDNN [18], which parameterises the resulting glottal source signals. From the parameters of each analysed vowel a spectral tilt (ST) and an $F0$ value are obtained. In the synthetic speech corpus, each ST value is associated with the $F0$ and R_d used to generate that vowel. Regarding the natural speech corpus, for each vowel the increment of $F0$ and ST from neutral to each of the expressions is computed. Finally, when a pair of $F0$ and R_d neutral values is input, it is converted by applying the previously computed increments to obtain a pair of values with the target expressive style. The following subsections describe the processes appearing in Figure 1.

2.1. LF-FEM based synthesis

Synthetic speech is generated with a realistic vocal tract by combining a 3D FEM-based acoustic model with an LF model for the glottal source.

2.1.1. Vocal Tract Acoustic Model

The 3D acoustic model uses the FEM to simulate the propagation of 3D acoustic waves within a vocal tract [4]. In particular, it numerically solves the acoustic wave equation for the acoustic pressure $p(\boldsymbol{x}, t)$,

$$\partial_{tt}^2 p - c_0^2 \nabla^2 p = 0, \tag{1}$$

with $c_0 = 350$ m/s being the speed of sound and ∂_{tt}^2 denoting the second partial time derivative. This model also uses a Perfectly Matched Layer (PML) to absorb sound waves emanating from the mouth aperture, thus considering radiation losses. Wall losses are introduced through the boundary admittance coefficient $\mu = 0.005$, set on the vocal tract walls. Details about the implementation of this model can be found in [4].

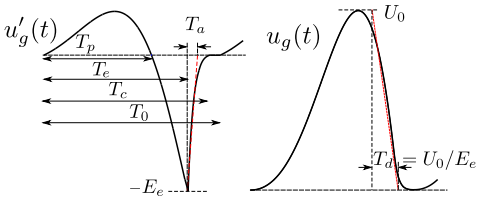

Figure 2: *Glottal flow $u_g(t)$ and its time derivative $u_g'(t)$ according to the LF model [19]. T_p is the rise time, T_e is the open phase duration, T_c corresponds to the complete closure, T_0 is the period, T_a is the effective duration of the return phase and T_d is the declination time. The maximum amplitudes of the glottal flow and its derivative are respectively U_0 and E_e.*

A vowel sound can be synthesised introducing a train of glottal pulses at the vocal tract entrance, i.e. at the glottal cross-section. However, that would require a new FEM simulation for every glottal source configuration. To circumvent it, the vocal tract impulse response $h(t)$ is computed instead, and convolved with the desired input signal $u_g(t)$ to generate the output sound $p_o(t)$,

$$p_o(t) = h(t) * u_g(t). \tag{2}$$

The impulse response $h(t)$ can be simulated by introducing at the glottal cross-section the Gaussian Pulse

$$g_p(t) = e^{-\left[(t-T_{gp})/0.29T_{gp}\right]^2} [\mathrm{m}^3/\mathrm{s}], \tag{3}$$

with $T_{gp} = 0.646/f_c$ and $f_c = 10$ kHz, while capturing the acoustic pressure $p_o(t)$ at the vocal tract exit. The vocal tract transfer function

$$H(f) = \frac{P_o(f)}{G_p(f)} \tag{4}$$

can next be computed, with $P_o(f)$ and $G_p(f)$ respectively denoting the Fourier Transform of $p_o(t)$ and $g_p(t)$. The impulse response $h(t)$ is finally obtained by applying the inverse Fourier transform to $H(f)$.

2.1.2. Glottal Source Model

An LF model [19] is used to generate the train of glottal pulses $u_g(t)$ needed in (2) to synthesise a vowel sound. In particular, the Kawahara's implementation [20] is adopted to obtain aliasing-free glottal flow derivative pulses $u_g'(t)$ according to the parameters T_p, T_e, T_a, T_c and T_0 (see Figure 2). The original code [1] has been adapted to introduce the glottal shape parameter R_d [10], defined as

$$R_d = \frac{U_0}{E_e}\frac{F0}{110}. \tag{5}$$

In Eq. (5), U_0 is the glottal flow peak, E_e is the negative amplitude of the differentiated glottal flow, and $F0$ the fundamental frequency. The R_d parameter greatly simplifies the control of the LF model [10]. For instance, high values of R_d generate a lax phonation, whereas low values of R_d produce a very abducted phonation, i.e., a tense voice [21]. The glottal flow $u_g(t)$ is obtained by performing the cumulative integration of $u_g'(t)$ using the composite trapezoidal rule. Finally, an SoX resampling [2] has been incorporated to adapt the signals originally generated at 44100 Hz to the sampling frequency of $h(t)$.

[1] `https://github.com/HidekiKawahara/SparkNG`
[2] `http://sox.sourceforge.net/SoX/Resampling`

2.2. Spectral tilt analysis

This section describes the spectral tilt analysis applied to both the natural speech corpus and the synthetic speech.

2.2.1. GlottDNN-based spectral tilt extraction

The GlottDNN vocoder [18] is used in this study. This glottal vocoder applies the quasi-closed phase (QCP) inverse filtering technique to decompose speech into glottal source and vocal tract filter, and parameterise their corresponding spectra with 10 and 30 Line Spectral Frequencies (LSF) per frame, respectively. The QCP method has a tendency to include some tilt in the vocal tract estimate. To compensate for this, the spectral tilt of the vocal tract filter is parameterised with a first order LP filter and transferred to the glottal source, as done in [16]. Finally, a glottal source LSF vector is computed for each vowel by averaging the vectors obtained at a frame level on its stable part, thus minimising coarticulation effects. Similarly, an $F0$ mean value is computed for each vowel.

2.2.2. Spectral tilt representation

Glottal source LSF can be used to derive a scalar meaningful representation of the glottal source spectral tilt. In this work, following [22, 23] a scalar-based measure of the spectral tilt, ST, has been computed as

$$ST = 10\log_{10}\left(\frac{\int_{f_3}^{f_4} S_{xx}(f)}{\int_{f_1}^{f_2} S_{xx}(f)}\right), \tag{6}$$

where Sxx is the power spectral density computed from the glottal excitation LSF, and the frequencies that delimit the bands where the energy is integrated are $f_1 = 50$ Hz, $f_2 = 1$ kHz, $f_3 = 1$ kHz and $f_4 = 5$ kHz.

2.3. Expressive LF-FEM based synthesis

2.3.1. Comparison of spectral tilt between expressive styles

The values obtained from the vowels in the parallel expressive corpus are compared with respect to the neutral style. Considering a vowel from the target expressive style and their corresponding in the neutral one, the increment of $F0$ (in semitones) is computed as

$$\Delta F0 = 12\log_2\left(\frac{F0_t}{F0_n}\right), \tag{7}$$

and the increment of spectral tilt (in dB) as

$$\Delta ST = ST_t - ST_n, \tag{8}$$

where $F0_t$ and ST_t are respectively the fundamental frequency (in Hz) and spectral tilt of the expressive vowel, while $F0_n$ and ST_n are those obtained from the neutral vowel.

2.3.2. Spectral tilt transplantation

The $\Delta F0$ and ΔST increments computed in the previous section are used to obtain LF parameters that can incorporate some expressiveness in the LF-FEM based synthesis (see Fig. 1, bottom). To this end, given an input pair $F0_{neu}$ and R_{dneu} corresponding to a neutral style, a vowel generated with these values is searched in the LF-FEM corpus to obtain its spectral tilt ST_{neu}. Then, the increments previously computed for the target expression are applied on $F0_{neu}$ and ST_{neu}, thus obtaining an $F0_{exp}$ and an ST_{exp}. Finally, looking for the LF-FEM vowel closest to these values, an R_{dexp} value can be derived.

3. Experiments and results

This section details the setup of the experiments and the results obtained from the conducted analyses.

3.1. Experiments setup

3.1.1. Expressive natural speech

This work has used an emotional Spanish speech corpus, which was explicitly designed to elicit expressive speech (see [24] for further details). To that effect, the corpus was built by recording a professional female speaker reading texts whose semantic content helped to express the desired style (stimulated speech). The audios were sampled at 16 kHz using a non-compressed pulse coded modulation and 16 bits per sample.

The study has focused on the analysis of tense voice emotional styles with respect to neutral speech. Accordingly, among the five expressive categories available in the corpus three have been selected: (i) neutral (NEU), which denotes certain maturity; (ii) happy (HAP), which transmits a feeling of extroversion; (iii) and aggressive (AGR), which express hardness.

A subset of 836 paired utterances from the NEU, HAP and AGR expressive styles has been selected for this work (i.e., totalling 2508 utterances), composed of one or two words with at least one vowel [a], either stressed or unstressed. In total, 679 [a] and 495 ['a] have been analysed for each style.

3.1.2. LF-FEM synthesis

Synthesis of vowel [a] has been done using the LF-FEM based model described in Section 2.1. For this purpose, we have used the 3D vocal tract geometry originally generated from MRI in [1] and latter adjusted in [3], in which the trachea and part of the face were removed, preserving the lips. This geometry was set on a rectangular baffle being part of a radiation space that allows sound waves emanate from the mouth aperture. Unstructured tetrahedral elements were used to mesh the computational domain, with an average size of 1 mm within the vocal tract and 3-4 mm in the radiation space.

FEM simulations were first performed to obtain the vocal tract impulse response $h(t)$, considering a time event of 20 ms and setting the sampling frequency to $f_s = 8000$ kHz. Such a high f_s was needed to ensure stability of the numerical schemes. The acoustic pressure $p_0(t)$ was captured at the vocal tract exit, 4 cm from the mouth aperture, which permits to first compute $H(f)$ using Eq. (4), and next $h(t)$ through its inverse Fourier transform. Finally, $h(t)$ was resampled to 16 kHz so as to match with the sampling frequency of the natural speech corpus.

Several vowels [a] have been then synthesised convolving $h(t)$ with the glottal pulses generated by the LF model. The latter has been configured to generate synthetic voice using different pairs of $F0$ and R_d. For the R_d 25 logarithmically spaced values covering the range from 0.3 to 2.7 [21] have been used. Regarding the $F0$, a pitch contour has been extracted from a real sustained vowel lasting for 2 seconds. This curve has been successively pitch-shifted from an F0 mean value of 71.4 Hz to 240 Hz in steps of 1 semitone Note, however, that Eq. (5) still requires to determine U_0 or E_e. U_0 has been deemed fixed through all synthesised vowels and adjusted to obtain realistic sound pressure levels in a modal phonation, as in [17].

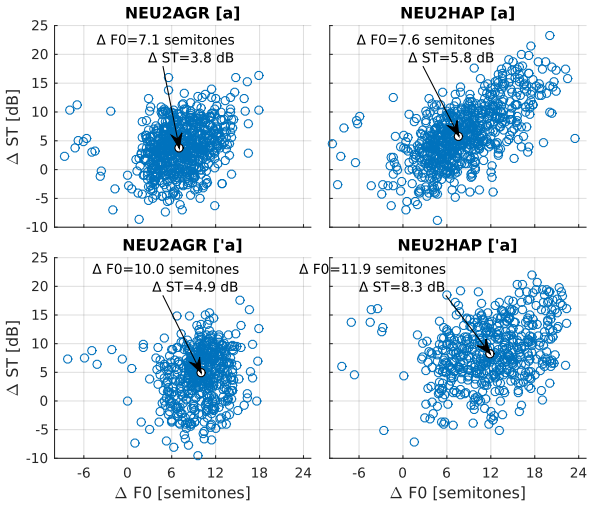

Figure 3: *Distributions of $\Delta F0$ and ΔST for stressed and unstressed [a] vowels from neutral to aggressive (NEU2AGR), and from neutral to happy (NEU2HAP). The centroid of each distribution is represented as a white dot indicated with an arrow.*

Table 1: *$F0$, spectral tilt (ST) and R_d values obtained for the LF-FEM synthesis of vowels [a] and ['a] in neutral, aggressive and happy styles.*

	Vow	NEU	AGR	HAP
$F0$ (Hz)	[a]	100.0	150.4	155.1
	['a]	106.3	189.7	211.2
ST (dB)	[a]	-25.6	-21.8	-19.8
	['a]	-24.5	-19.7	-16.3
R_d	[a]	1.00	0.82	0.74
	['a]	0.90	0.74	0.61

3.2. Results

The increments $\Delta F0$ and ΔST from neutral to both happy and aggressive styles have been computed for the stressed and unstressed vowels [a] of the parallel corpus. Figure 3 depicts the results, where each circle represents the $\Delta F0$ and ΔST from a neutral vowel to its corresponding expressive counterpart. As can be observed, the two expressive styles increase both the $F0$ and the ST with respect to the neutral speech. In the aggressive style, the $\Delta F0$ and ΔST with respect to the neutral speech are, in average, 7.1/10.0 semitones and 3.8/4.9 dB for the unstressed/stressed [a], respectively, whereas the happy ones are 7.6/11.9 semitones and 5.8/8.3 dB. Note then, on the one hand, that stressing a vowel produces a significant increase of the $F0$ and ST independently on the speaking style, although the variation is more prominent for the happy speech. On the other hand, comparing the two expressive styles, happy vowels entail higher values of ΔST and $\Delta F0$ than the aggressive ones.

Table 1 shows the values derived from the above analysis and that have been used for the synthesis of unstressed and stressed [a] in the neutral, aggressive and happy styles. First, a pair of LF parameters $F0 = 100$ Hz and $R_d = 1$ has been used as a reference for a neutral [a]. The vowel that was generated with these parameters has been retrieved from the LF-FEM corpus, and as Table 1 shows it has an ST=-25.6 dB. The $F0$ and ST values for a reference neutral ['a] have then been obtained according to the increments observed in the neutral speech corpus between the stressed and the unstressed [a]. This results in $F0 = 106.3$ Hz and ST=-24.5 dB, which correspond to an $R_d = 0.90$. From these neutral reference values (first column in Table 1), the $F0$s and STs for the expressive styles (second and third column) have been obtained applying the $\Delta F0$ and ΔST increments corresponding to the centroids in Fig. 3. Note, however, that ST values do not directly map with any of the input parameters of the LF glottal model. This link has been achieved by looking for those vowels in the LF-FEM corpus with the closest $F0$ and ST. As a result, we have derived the R_d parameters that, together with the $F0$ values, have been used to resemble the analysed expressive styles.

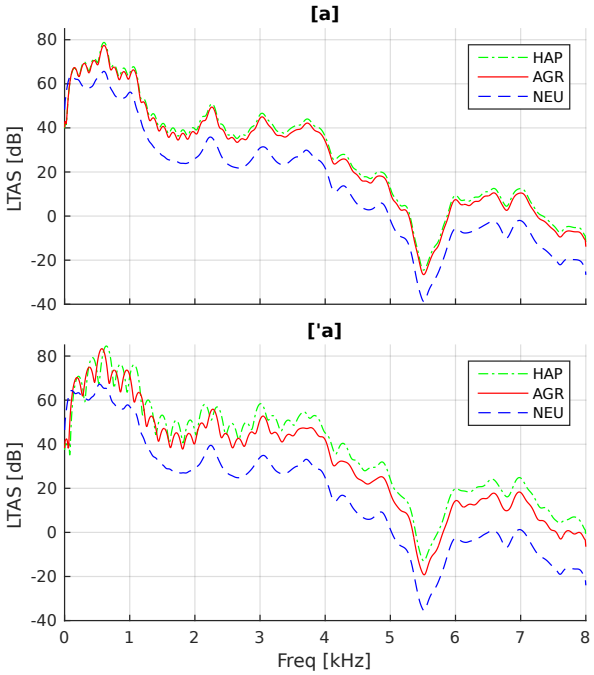

Figure 4: *Long Term Average Spectra of the unstressed (top) and stressed (bottom) [a] vowels synthesised with the LF-FEM model for the neutral, aggressive and happy styles.*

Figure 4 shows the Long Term Average Spectra (LTAS) computed from the synthesised vowels [a] with neutral, happy and aggressive styles, for both the unstressed (top) and stressed (bottom) versions. Observe that below 4-5 kHz the classical formants of vowel [a] are generated. However, beyond this frequency, some dips and asymmetrical modes are also produced. Most of them are the so called higher order modes, which as said in the introduction, can only be captured with a 3D acoustic model. Besides, it is to be mentioned that the strongest dip between 5 and 6 kHz is mainly generated by the piriform fossae –a pair of side branches located close to the larynx– although a higher order mode also contributes [3]. Focusing now on the comparison between expressive styles, as observed, the happy and aggressive not only increase the total sound pressure levels (SPL) with respect to the neutral one, but also reduce the relative differences between low and high frequencies. The latter are a direct consequence of increasing the ST, as one could expect. The curves are also very similar between the two tense

styles for the unstressed [a]. However, this is not the case when the stressed version is generated. As also observed in the distributions of Figure 3, the happy style entails a higher ST than the aggressive one, thus producing the observed increment in the high frequency range.

4. Conclusions

In this work, we have explored the glottal source variations of happy and aggressive emotional styles with respect to neutral speech. The analysis has focused on those features that could be translated to a 3D FEM-based acoustic model that uses as excitation an LF model controlled by the R_d parameter. In particular, we have considered the variations of $F0$ and spectral tilt associated with the glottal source, extracted from the corpus by means of the GlottDNN vocoder. These variations have then been translated into LF parameters for the expressive LF-FEM based synthesis of vowels [a] and ['a]. Results have shown that to generate aggressive and happy styles, it is necessary to increase the $F0$ and to decrease the R_d with respect to the neutral style, presenting larger deviations the happy emotion than the aggressive one. These differences of $F0$ and R_d values are even greater for the stressed version of the vowel. Future work will focus on extending the analysis to other vowels and emotional styles.

5. Acknowledgements

This research has been supported by the Agencia Estatal de Investigación (AEI) and FEDER, EU, through project GENIOVOX TEC2016-81107-P. The third author also acknowledges the support from the Obra Social "La Caixa" for grant ref. 2018-URL-IR2nQ-029.

6. References

[1] D. Aalto, O. Aaltonen, R.-P. Happonen, P. Jääsaari, A. Kivelä, J. Kuortti, J.-M. Luukinen, J. Malinen, T. Murtola, R. Parkkola, J. Saunavaara, T. Soukka, and M. Vainio, "Large scale data acquisition of simultaneous MRI and speech," *Applied Acoustics*, vol. 83, pp. 64–75, 2014.

[2] R. Blandin, M. Arnela, R. Laboissière, X. Pelorson, O. Guasch, A. V. Hirtum, and X. Laval, "Effects of higher order propagation modes in vocal tract like geometries," *The Journal of the Acoustical Society of America*, vol. 137, no. 2, pp. 832–8, 2015.

[3] M. Arnela, S. Dabbaghchian, R. Blandin, O. Guasch, O. Engwall, A. Van Hirtum, and X. Pelorson, "Influence of vocal tract geometry simplifications on the numerical simulation of vowel sounds," *The Journal of the Acoustical Society of America*, vol. 140, no. 3, pp. 1707–1718, 2016.

[4] M. Arnela and O. Guasch, "Finite element computation of elliptical vocal tract impedances using the two-microphone transfer function method," *The Journal of the Acoustical Society of America*, vol. 133, no. 6, pp. 4197–4209, 2013.

[5] H. Takemoto, P. Mokhtari, and T. Kitamura, "Acoustic analysis of the vocal tract during vowel production by finite-difference time-domain method," *The Journal of the Acoustical Society of America*, vol. 128, no. 6, pp. 3724–3738, 2010.

[6] M. Speed, D. Murphy, and D. Howard, "Modeling the vocal tract transfer function using a 3d digital waveguide mesh," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 22(2), pp. 453–464, 2014.

[7] P. Birkholz, L. Martin, Y. Xu, S. Scherbaum, and C. Neuschaefer-Rube, "Manipulation of the prosodic features of vocal tract length, nasality and articulatory precision using articulatory synthesis," *Computer Speech and Language*, vol. 41, pp. 116–127, 2017.

[8] P. Birkholz, L. Martin, K. Willmes, B. J. Kröger, and C. Neuschaefer-Rube, "The contribution of phonation type to the perception of vocal emotions in German: An articulatory synthesis study," *The Journal of the Acoustical Society of America*, vol. 137, no. 3, pp. 1503–1512, 2015.

[9] I. Yanushevskaya, C. Gobl, and C. A. Ní, "Cross-language differences in how voice quality and f0 contours map to affect." *The Journal of the Acoustical Society of America*, vol. 144, no. 5, p. 2730, 2018.

[10] G. Fant, "The LF-model revisited. Transformations and frequency domain analysis," *Speech Transmission Laboratory Quarterly Progress and Status Report (STL-QPSR)*, vol. 36, no. 2-3, pp. 119–156, 1995.

[11] A. Murphy, I. Yanushevskaya, A. N. Chasaide, and C. Gobl, "Rd as a Control Parameter to Explore Affective Correlates of the Tense-Lax Continuum," in *INTERSPEECH*, 2017, pp. 3916–3920.

[12] Y. Li, J. Li, and M. Akagi, "Contributions of the glottal source and vocal tract cues to emotional vowel perception in the valence-arousal space." *The Journal of the Acoustical Society of America*, vol. 144, no. 2, p. 908, 2018.

[13] Y. R. Chien, D. D. Mehta, J. Guenason, M. Zanartu, and T. F. Quatieri, "Evaluation of glottal inverse filtering algorithms using a physiologically based articulatory speech synthesizer," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 8, pp. 1718–1730, 2017.

[14] M. Airaksinen, L. Juvela, B. Bollepalli, J. Yamagishi, and P. Alku, "A comparison between straight, glottal, and sinusoidal vocoding in statistical parametric speech synthesis," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 9, pp. 1658–1670, Sep. 2018.

[15] J. Lorenzo-Trueba, R. Barra-Chicote, T. Raitio, N. Obin, P. Alku, J. Yamagishi, and J. M. Montero, "Towards Glottal Source Controllability in Expressive Speech Synthesis," in *INTERSPEECH*, 2012, pp. 2–5.

[16] S. Seshadri, L. Juvela, O. Räsänen, and P. Alku, "Vocal effort based speaking style conversion using vocoder features and parallel learning," *IEEE Access*, 2019.

[17] M. Freixes, M. Arnela, J. C. Socoró, F. Alías, and O. Guasch, "Influence of tense, modal and lax phonation on the three-dimensional finite element synthesis of vowel [A]," in *IberSPEECH2018*, 2018, pp. 132–136.

[18] M. Airaksinen, B. Bollepalli, L. Juvela, Z. Wu, S. King, and P. Alku, "Glottdnn - a full-band glottal vocoder for statistical parametric speech synthesis," in *INTERSPEECH*, 9 2016, pp. 2473–2477.

[19] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," *Speech Transmission Laboratory Quarterly Progress and Status Report (STL-QPSR)*, vol. 26, no. 4, pp. 1–13, 1985.

[20] H. Kawahara, K.-I. Sakakibara, H. Banno, M. Morise, T. Toda, and T. Irino, "A new cosine series antialiasing function and its application to aliasing-free glottal source models for speech and singing synthesis," in *INTERSPEECH*, 2017, pp. 1358–1362.

[21] C. Gobl, "Reshaping the Transformed LF Model : Generating the Glottal Source from the Waveshape Parameter Rd," in *INTERSPEECH*, 2017, pp. 3008–3012.

[22] P. Murphy, K. Mcguigan, M. Walsh, and M. Colreavy, "Investigation of a glottal related harmonics-to-noise ratio and spectral tilt as indicators of glottal noise in synthesized and human voice signals," *The Journal of the Acoustical Society of America*, vol. 123, pp. 1642–52, 04 2008.

[23] S. Kakouros, O. Räsänen, and P. Alku, "Evaluation of spectral tilt measures for sentence prominence under different noise conditions," in *INTERSPEECH*, 2017, pp. 3211–3215.

[24] I. Iriondo, S. Planet, J.-C. Socoró, E. Martínez, F. Alías, and C. Monzo, "Automatic refinement of an expressive speech corpus assembling subjective perception and automatic classification," *Speech Communication*, vol. 51, no. 9, pp. 744–758, 2009.