IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

Contribution of vocal tract and glottal source spectral cues in the generation of happy and aggressive [a] vowels

Marc Freixes, 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,francesc.alias,joanclaudi.socoro}@salle.url.edu

Abstract

At present, three-dimensional (3D) acoustic models allow for the numerical simulation of vowels, diphthongs and some vowel-consonant-vowel sequences using realistic vocal tract geometries. While research is being done to generate more phonemes and short utterances, some attempts have been made to incorporate expressiveness into the 3D numerical simulation of isolated vowels. However they are very preliminary and still far from the generation of expressive utterances. To move towards this goal, this work analyses the contribution of vocal tract (VT) and glottal source spectral (GSS) cues to the production of happy and aggressive vowels with respect to neutral vowels. After parameterising with the GlottDNN vocoder the paired neutral-expressive utterances from a Spanish database, neutral utterances are transplanted with the target expressive prosody as baseline, and subsequently resynthesised considering also the GSS and/or VT from their expressive pairs. Objective and subjective evaluations show that, both GSS and VT have a statistically significant contribution to convey the tense voice target emotions. VT prevails over GSS specially for aggressive. Best results are achieved when considering both GSS and VT, which compared to the baseline permits an increase in the perceived emotional intensity of 55.3% for happy and 62.8% for aggressive utterances.

Index Terms: expressive speech synthesis, emotional corpora, speech analysis, inverse filtering, glottal source, vocal tract, numerical voice production

1. Introduction

Speech synthesis methods that rely on the acoustic theory of voice production have traditionally considered simplified source-filter models [1], such as the ones based on one-dimensional (1D) representations of the vocal tract [2]. Recently, the increase in computing power has allowed the development of three-dimensional (3D) acoustic models, which have been applied to generate vowels [3], diphtongs [4] and some vowel-consonant-vowel sequences [5], overcoming the limitations of 1D-based models [6]. While researchers keep investigating on the production of other phonemes and short utterances [7], preliminary attempts to synthesise expressive vowels have been made, but limited to basic modifications of glottal source signals [8].

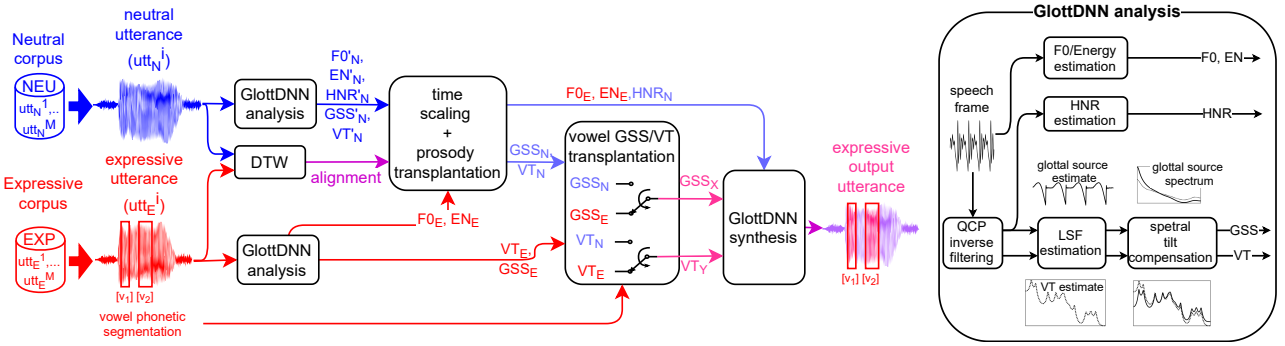
The relevance of glottal source and/or vocal tract cues in the production of expressive speech has been explored in several studies using traditional source-filter based approaches on an analysis-by-synthesis scheme. For example, the contribution of phonation types to the perception of emotions was analysed in [9]. To this end, a set of utterances was resynthesised with different phonation types through an articulatory-based synthesiser that incorporates a self-oscillating model of the vocal folds. A similar approach was followed in [10] considering a formant-based synthesiser with a modified Liljencrants-Fant (LF) glottal flow model [11]. This synthesis approach was also considered in [12] to study the mapping of F0 contours and voice quality on affect for different languages by modifying the parameters of modal stimuli. In [13], the LF model was controlled by modifying the R_d glottal shape parameter [14] to simulate the tense-lax continuum and explore its influence on emotion perception. Similarly, an auto-regressive exogenous LF model was proposed in [15] to analyse the contribution of glottal source and vocal tract to the perception of emotions in a valence-arousal space. Nevertheless, the study only evaluated isolated vowels, and suffered from the considered prosody *neutralisation* process.

The aforementioned approaches have focused on the analysis and resynthesis of a small set of vowels or utterances, some of them involving costly manual tuning processes. Nonetheless, recent advances in inverse filtering and glottal source processing techniques have facilitated the automatic decomposition of the speech signal into glottal source and vocal tract features [16]. For instance, GFM-Voc (Glottal Flow Model-based Vocoder) allows real-time voice manipulations, such as vowel formants shifting and voice quality modifications related to the glottal source [17]. Also included in this strand are glottal vocoders like GlottHMM, whose features proved effective in the analysis of expressive nuances in [18]. More recently, its successor, GlottDNN [19], was used to perform speaking style conversion to mimic the Lombard effect from natural speech [20]. A GlottDNN-based analysis of the glottal source spectral tilt was proposed [21] to introduce expressiveness in the 3D numerical generation of isolated vowels through modifications of the R_d parameter of the LF model. However, this proof-of-concept is still far from generating natural expressive utterances.

As a next step towards this goal, in this work we analyse the contribution of vocal tract (VT) and glottal source spectral (GSS) cues in the generation of emotional styles with tense phonation. For this purpose, paired neutral, happy and aggressive utterances from a Spanish speech database are inverse filtered and parameterised using the GlottDNN vocoder. Then, neutral utterances transplanted with prosody, and GSS and/or VT from the expressive pairs are resynthesised and evaluated through both objective and subjective tests, focused on vowels [a]; the most common vowel in the database.

The paper is organised as follows. Section 2 presents the methodology proposed for the GlottDNN-based analysis and synthesis of expressive utterances to study the contribution of GSS and VT on tense voice emotional styles. Next, the conducted experiments are described and the obtained results discussed in Sections 3 and 4, respectively. Finally, conclusions and future work are presented in Section 5.

10.21437/IberSPEECH.2021-51



(a) *Overall diagram.* (b) *Detail of the GlottDNN analysis.*

Figure 1: *Framework proposed to study the contribution of vocal tract and glottal source spectral cues (VT and GSS) in the generation of tense voice expressive (EXP) vowels, depicting the overall diagram of the process in (a) and the main elements of the GlottDNN analysis in (b). M pairs of neutral and expressive utterances from parallel speech corpora are analysed using the GlottDNN vocoder, and aligned through dynamic time warping (DTW). According to this alignment, the features of each neutral utterance utt_N^i (marked with $'$) are time-scaled and transplanted with the prosody of the corresponding expressive pair utt_E^i ($F0_E$, EN_E). Finally, different expressive output utterances are obtained by applying GSS_X and VT_Y to the vowels –being X, Y either neutral (N) or expressive (E).*

2. Methodology

Figure 1 depicts the framework proposed to analyse the contribution of the spectral cues from glottal source and vocal tract to the synthesis of tense voice expressive styles with respect to parallel neutral speech data. The study takes the neutral utterances with the prosody transplanted from their expressive pairs as the baseline. The contribution of GSS and VT is analysed through different synthesis configurations, which are denoted as $GSS_X VT_Y$, where X and Y indicate the origin of the GSS and VT applied to the vowels: N for the neutral utterance and E for the expressive utterance (hereafter, this subindex notation is applied for all the variables).

Firstly, each neutral-expressive utterance pair is parameterised using the GlottDNN vocoder. As depicted in Figure 1b, for each input speech frame (either neutral or expressive), the GlottDNN estimates its fundamental frequency (F0 in Hz) and energy (EN in dB), and applies quasi-closed phase (QCP) inverse filtering to obtain the corresponding glottal source and VT estimates, which are parameterised using Line Spectral Frequencies (LSF). Moreover, it computes the Harmonic-to-Noise Ratio (HNR) of the glottal source estimate. Further details of this process can be found in [19]. Next, a spectral tilt compensation module is included to compensate the tendency of QCP to include residual spectral cues from the glottal source on the vocal tract estimation [20]. In this module, the spectral tilt of the vocal tract estimate is modelled with a first order linear prediction filter, and subsequently transferred to the GSS to adjust it following [20].

Secondly, the prosody of the neutral utterance is transplanted by that from the expressive pair. On one hand, the GlottDNN features of the neutral utterance (marked with $'$ on Figure 1a) are linearly interpolated to time-scale them to the expressive target according to the alignment obtained through dynamic time warping. On the other hand, $F0_N$ and EN_N are replaced by those from the expressive utterance, that is, $F0_E$ and EN_E, respectively. Next, GSS_E and/or VT_E are transplanted into the vowels depending on the selected synthesis configuration.

The GlottDNN synthesis process is briefly described below (for further details the reader is referred to [19]). The GlottDNN vocoder uses white noise sequences as input to generate the unvoiced frames. Conversely, the vocoder features of the voiced frames are input into a simple feed-forward deep neural network to generate zero-padded two pitch-period glottal flow derivative pulses. These signals are scaled according to EN_E, and concatenated through the classic Pitch-Synchronous Overlap and Add (PSOLA) algorithm [22]. This initial excitation is then processed adding noise in the spectral domain as specified by the HNR_N besides modifying its spectra to match GSS_X. Finally, the excitation signal is filtered according to the considered VT_Y to obtain the synthetic speech output. It should be noted that since this work focuses on the spectral characteristics of the glottal source (i.e., GSS), the synthesis has been performed considering HNR_N and using the GlottDNN pulse generation model trained with neutral speech. The relevance of these two aspects of the glottal source in the generation of expressive speech will be analysed in future studies.

3. Experiments

This section describes the conducted experiments, detailing the main characteristics of the emotional speech database and the configuration of the GlottDNN vocoder, together with the design of the objective and subjective evaluations.

3.1. Emotional speech database

The experiments have been conducted on an emotional Spanish speech database explicitly designed to elicit expressive speech and recorded by a female professional speaker at a sampling frequency of 16 kHz [23].

Three out of the five expressive styles of that database have been chosen as the ones characterised by a modal or a tense phonation, namely: (i) neutral; (ii) happy; (iii) and aggressive. These corpora count with a set of 1250 paired short utterances (with an average of 1.2 words per utterance) that ensure phonetic coverage for Spanish text-to-speech synthesis purposes. Out of them, 841 utterances have been used in this work, specifically those containing at least a vowel [a]; the most common vowel in the database. As a result, a total of 1171 paired vowels have been considered in the experiments.

Table 1: *Mean values of the spectral distances (either Itakura-Saito, or $\overline{d_{\mathrm{IS}}}$, and Kullback-Leibler, or $\overline{d_{\mathrm{KL}}}$) computed from the analysed configurations to the expressive configurations for happy and aggressive [a] vowels.*

	Happy	Aggressive
$\overline{d_{\mathrm{IS}}}(\mathrm{GSS_N}, \mathrm{GSS_E})$	0.14	0.08
$\overline{d_{\mathrm{IS}}}(\mathrm{VT_N}, \mathrm{VT_E})$	3.11	3.79
$\overline{d_{\mathrm{KL}}}(\mathrm{GSS_N VT_N}, \mathrm{GSS_E VT_E})$	2.27	2.42
$\overline{d_{\mathrm{KL}}}(\mathrm{GSS_E VT_N}, \mathrm{GSS_E VT_E})$	1.17	1.90
$\overline{d_{\mathrm{KL}}}(\mathrm{GSS_N VT_E}, \mathrm{GSS_E VT_E})$	0.45	0.17

3.2. GlottDNN-based analysis and synthesis

The analysis and synthesis of utterances have been done using the default GlottDNN settings[1], parameterising GSS and VT with 10 and 30 LSF coefficients per frame, respectively, and considering voiced frames of 25 ms, and unvoiced frames of 15 ms. The whole neutral corpus (of 2.4 h length) has been used to train the GlottDNN pulse generation model [19].

3.3. Objective and subjective evaluation

The objective contribution of GSS and VT on the generation of happy and aggressive emotions has been evaluated through the computation of spectral distances between the [a] vowel pairs taking the expressive vowel as the target reference.

Given a neutral-expressive vowel pair, 2 GSSs and 2 VTs are obtained: from the neutral vowel ($\mathrm{GSS_N}$, $\mathrm{VT_N}$) and from the expressive vowel ($\mathrm{GSS_E}$, $\mathrm{VT_E}$). The similarity between the $\mathrm{GSS_N}$ and the $\mathrm{GSS_E}$ is computed as the Itakura-Saito LPC-based spectral distance, i.e., $d_{\mathrm{IS}}(\mathrm{GSS_N}, \mathrm{GSS_E})$. The same is done for the VT, i.e., $d_{\mathrm{IS}}(\mathrm{VT_N}, \mathrm{VT_E})$. GSS and VT are parameterised by the GlottDNN as LSF vectors at a frame level, from which LSF vectors at vowel level are obtained using the median to reduce coarticulation effects. Finally, LSF are translated into LPC to compute the Itakura-Saito distances [24].

Regarding the synthesis, a total of 4 configurations per emotion have been considered to evaluate the contribution of GSS and VT to the production of the tense voice target emotion, considering in all of them the target expressive prosody: (i) $\mathrm{GSS_N VT_N}$ as the baseline configuration; (ii) $\mathrm{GSS_E VT_N}$; (iii) $\mathrm{GSS_N VT_E}$; and (iv) $\mathrm{GSS_E VT_E}$ as the expressive target configuration. In order to evaluate how close is each vowel version to the expressive target, their long term average spectrum (LTAS) have been computed as the Welch's power spectral estimate, with a 15 ms hamming window, 50% of overlap and a 2048-point FFT [8]. Then, the similarity of each vowel obtained from configurations (i) to (iii) with the expressive target has been measured as the symmetrical Kullback-Leibler spectral distance [25] between its LTAS and the corresponding one in configuration (iv), i.e., $d_{\mathrm{KL}}(\mathrm{GSS}_X \mathrm{VT}_Y, \mathrm{GSS_E VT_E})$.

Regarding the subjective evaluation, the perceived emotional intensity for the different synthesis configurations has been assessed through a MUSHRA (MUltiple Stimuli with Hidden Reference and Anchor) perceptual test [26]. Six words from the speech database subset containing only [a] vowels were used to evaluate the contribution of GSS and VT to the perception of the two target emotions through the four aforementioned configurations. The words included in the test are *pala*, *taza*, *capa*, *gastar*, *agravar* and *MACBA* (in English, they correspond to shovel, cup, layer, spend, aggravate and MACBA –the name of a museum in Barcelona). Two versions of the test were prepared, each one consisting of 7 evaluation sets (three words per emotion plus one control point to validate the evaluator consistency according to the Pearson correlation coefficient r). In each set, the participants were asked to rate the perceived emotion intensity for each one of the four versions of the word on a 0 to 100 scale. A GlottDNN-based resynthesised utterance different from those evaluated was included in the test as an example of the target happy/aggressive emotion. In order to determine if the differences between the evaluated configurations are statistically significant, the Wilcoxon signed-rank test [27] has been applied to both objective and subjective results.

Forty-four Spanish native speakers with an average age of 28.9 took one of the two versions of the online test using headphones and the Web Audio Evaluation Tool [28]. Among them, 61.9% of the participants are engineering students in their final year, 42.9% have experience in playing and/or producing music, 21.4% in audio software/hardware design and the 28.6% in audio or speech research. Once the perceptual test was concluded, the responses of eight participants were discarded since they presented significant criteria inconsistencies (i.e., with $r < 0.5$).

[1] https://github.com/ljuvela/GlottDNN

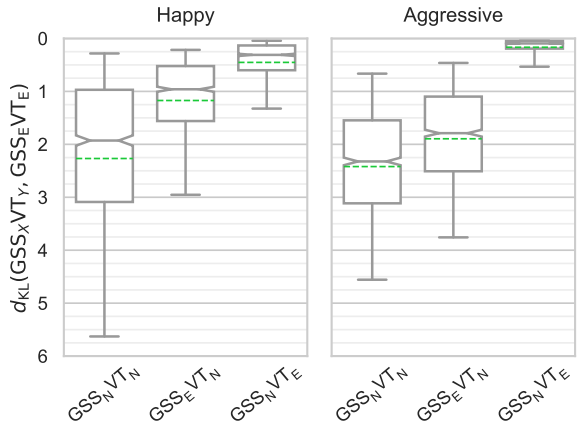


Figure 2: *Boxplots of the Kullback-Leibler distances from the analysed configurations to the target configuration ($\mathrm{GSS_E VT_E}$) for happy and aggressive. Dotted lines represent the mean of the distributions, and whiskers are set to 5th and 95th percentiles.*

4. Results and discussion

In this section, the results of both the objective and subjective experiments are presented and discussed in detail.

4.1. Objective results

Table 1 lists the mean values of the spectral distances computed from the analysed GSS and VT configurations for happy and aggressive [a] vowels. It is to note that the differences between all the configurations are statistically significant according to the Wilcoxon signed-rank test (with $p < 0.01$).

Regarding the comparison of $\mathrm{GSS_N}$ and $\mathrm{VT_N}$ with respect to $\mathrm{GSS_E}$ and $\mathrm{VT_E}$, it can be observed that GSS differences are higher for happy than for aggressive, while the opposite is happening in the case of the VT. Looking at the spectral Kullback-Leibler distances between the synthesised vowels (see the bottom of the Table 1), it can be observed that the GSS contribu-

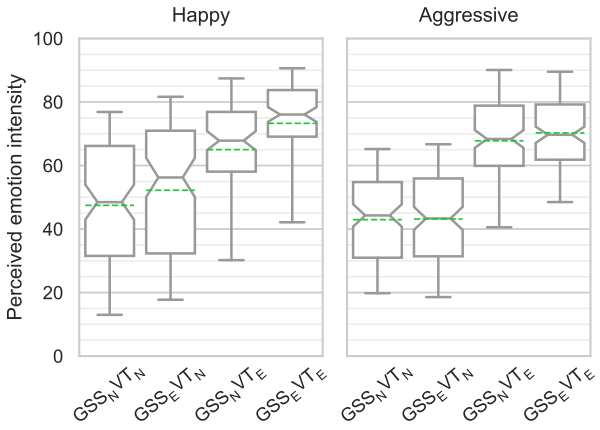


Figure 3: *Results of the MUSHRA perceptual test. Boxplots depict the perceived emotion intensity scores reported by the participants. The dotted lines represent the mean of the distributions, and whiskers are set to 5th and 95th percentiles.*

tion is more relevant in happy than in aggressive vowels. Thus, compared to the baseline, the incorporation of GSS_E is able to reduce the spectral distances to the target by 48% and 21%, respectively. The VT has a greater contribution than GSS, reducing the spectral distances by 80% for happy and by 93% for aggressive vowels.

The distributions of the computed Kullback-Leibler distances are depicted in Figure 2. It can be observed that both GSS and VT have a relevant contribution to the generation of happy and aggressive vowels, which is statistically significant according to the Wilcoxon test results. However, it is worth mentioning that VT dominates over GSS for both emotions, specially for aggressive vowels.

4.2. Perceptual evaluation results

Figure 3 depicts the results obtained from the MUSHRA test. According to the computed Wilcoxon signed-rank test, the differences between the four configurations are statistically significant with $p < 0.01$, except between $GSS_N VT_N$ and $GSS_E VT_N$ in aggressive utterances. The baseline configuration (with expressive prosody, GSS_N and VT_N) obtains the lowest perceived emotional intensity (mean score of 47 for happy and 43 for aggressive). For happy utterances, GSS_E and VT_E do significantly contribute to increase the perceived emotional intensity by 10.6% and 38.3%, respectively (from 47 to 52 and 65 points in the MUSHRA scale). When both are considered the increase is of 55.3%, thus reaching a mean score of 73 points. Regarding aggressive utterances, while GSS_E does not increase the perceived emotional intensity, VT_E leads to an increase of 58.1% (the mean score increases from 43 to 68). When both are incorporated the increase is of 62.8%, achieving a MUSHRA mean score of 70.

4.3. Discussion

Several issues can be discussed from the results obtained through the conducted objective and subjective evaluations. On the one hand, when GSS_E is applied, the distance to the target is significantly reduced for happy vowels, increasing the perceived emotion intensity for happy utterances. GSS contribution for aggressive is also objectively significant, but it is not perceived as such in the MUSHRA test unless VT_E is also transplanted. This result could be explained by analysing the differences between GSS_N and GSS_E in Table 1, which are more subtle in aggressive than in happy vowels. On the other hand, VT contribution is significantly relevant for both happy and aggressive, as observed in both the objective and perceptual analyses.

Although our work has been somehow inspired by that presented in [15], there are some important differences. Since our aim was to study the relevance of GSS and VT in resembling the target emotion, the effect of prosody has been *neutralised* as done in the second experiment of [15]. Nevertheless, in contrast to that work, GSS and VT contributions have been evaluated with a prosodic pattern coherent with the target emotion instead of using the neutral one, thereby avoiding the undesired neutralisation of the conveyed emotion as observed in the MUSHRA results. Moreover, using short utterances has not only allowed us to study vowels in their phonetic context, but also to ask evaluators about the perceived emotional intensity, instead of only evaluating isolated vowels in the arousal-valence space.

Finally, although the contribution of GSS and VT have been analysed through both objective and perceptual relative comparisons, these preliminary analyses should be completed in order to generalise the obtained results. In future works, we plan to consider more vowels and expressive styles, such as those with lax phonation, as well as other speakers covering different genders and ages.

5. Conclusions

In this work, the contribution of the GSS and VT to the generation of happy and aggressive emotional vowels has been studied on vowels [a] from a Spanish database composed of paired utterances by means of GlottDNN-based analysis and resynthesis. The objective and subjective evaluations with respect to the baseline reference (with expressive prosody, GSS_N and VT_N) show that both GSS and VT have a statistically significant contribution to convey the tense voice target emotions. Specifically, VT prevails over GSS specially for aggressive, where GSS perceptual contribution is statistically significant only when VT_E is also transplanted. Finally, it is to note that the best results are achieved when both GSS_E and VT_E are applied. When they are compared to the baseline the perceived emotional intensity is increased by 55.3% for happy and 62.8% for aggressive utterances, respectively. Properly modelling of both GSS and VT seems therefore instrumental for the upcoming 3D numerical generation of happy and aggressive vowels.

To that effect, future work will be focused on developing further analyses to extend the results obtained on vowel [a], by considering more phonemes and other expressive speaking styles and phonation types. Moreover, we envision the integration of the results within a 3D-based numerical synthesis workflow by introducing the observed relevant subtle changes in the glottal flow waveform together with the proper variations of the 3D vocal tract geometry in order to generate the desired expressive speaking style.

6. Acknowledgements

The research that has led to the results reported in this work has been funded by the SUR/DEC from the Government of Catalonia and the Ramon Llull University (ref. 2020-URL-Proj-056). The authors also would like to thank the participants on the perceptual test for their collaboration in this work.

7. References

[1] P. Taylor, *Text-to-Speech Synthesis*. Cambridge, UK: Cambridge University Press, 2009.

[2] B. H. Story, I. R. Titze, and E. A. Hoffman, "Vocal tract area functions from magnetic resonance imaging," *The Journal of the Acoustical Society of America*, vol. 100, no. 1, pp. 537–554, 1996.

[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, S. Dabbaghchian, O. Guasch, and O. Engwall, "Mri-based vocal tract representations for the three-dimensional finite element synthesis of diphthongs," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, pp. 2173–2182, 2019.

[5] M. Arnela and O. Guasch, "Finite element simulation of /asa/ in a three-dimensional vocal tract using a simplified aeroacoustic source model," in *International Congress on Acoustics (ICA)*, Aachen, Germany, sep 2019, pp. 1802–1809.

[6] 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.

[7] A. Pont, O. Guasch, and M. Arnela, "Finite element generation of sibilants /s/ and /z/ using random distributions of kirchhoff vortices," *International Journal for Numerical Methods in Biomedical Engineering*, vol. 36, no. 2, p. e3302, 2020.

[8] M. Freixes, M. Arnela, J. C. Socoró, F. Alías, and O. Guasch, "Glottal Source Contribution to Higher Order Modes in the Finite Element Synthesis of Vowels," *Applied Sciences*, vol. 9, no. 21, p. 4535, 2019.

[9] 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.

[10] F. Burkhardt, "Rule-Based Voice Quality Variation with Formant Synthesis," in *Proc. INTERSPEECH-2009*, Brighton, UK, 2009, pp. 2659–2662.

[11] 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.

[12] I. Yanushevskaya, C. Gobl, and C. A. Ní, "Cross-language differences in how voice quality and f_0 contours map to affect," *The Journal of the Acoustical Society of America*, vol. 144, no. 5, p. 2730, 2018.

[13] 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 *Proc. InterSpeech*, Stockholm, Sweden, Aug. 2017, pp. 3916–3920.

[14] 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.

[15] 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.

[16] T. Drugman, P. Alku, A. Alwan, and B. Yegnanarayana, "Glottal source processing: From analysis to applications," *Computer Speech and Language*, vol. 28, no. 5, pp. 1117–1138, 2014.

[17] O. Perrotin and I. McLoughlin, "GFM-Voc: A Real-Time Voice Quality Modification System," in *Proc. InterSpeech*, Graz, Austria, Sep. 2019, pp. 3685–3686.

[18] 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 *Proc. InterSpeech*, Portland, OR, USA, Sep. 2012, pp. 1620–1623.

[19] 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 Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1658–1670, Sep. 2018.

[20] S. Seshadri, L. Juvela, O. Räsänen, and P. Alku, "Vocal effort based speaking style conversion using vocoder features and parallel learning," *IEEE Access*, vol. 7, pp. 17 230–17 246, 2019.

[21] M. Freixes, M. Arnela, F. Alías, and J. C. Socoró, "GlottDNN-based spectral tilt analysis of tense voice emotional styles for the expressive 3D numerical synthesis of vowel [a]," in *Proc. 10th ISCA Speech Synthesis Workshop (SSW)*, Vienna, Austria, Sep. 2019, pp. 132–136.

[22] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech communication*, vol. 9, no. 5-6, pp. 453–467, 1990.

[23] 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.

[24] L. Rabiner, *Fundamentals of speech recognition*. PTR Prentice Hall, 1993.

[25] E. Klabbers and R. Veldhuis, "Reducing audible spectral discontinuities," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 1, pp. 39–51, 2001.

[26] R. ITU, "ITU-R BS.1534-1: Method for the subjective assessment of intermediate quality level of coding systems," *International Telecommunication Union*, 2003.

[27] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

[28] N. Jillings, B. De Man, D. Moffat, and J. D. Reiss, "Web audio evaluation tool: A browser-based listening test environment," in *12th International Conference in Sound and Music Computing (SMC 2015)*. Maynooth, Ireland: SMC network, Jul. 2015, pp. 147–152.