applied sciences

MDPI

Article

Contribution of Vocal Tract and Glottal Source Spectral Cues in the Generation of Acted Happy and Aggressive Spanish Vowels †

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

GTM—Grup de Recerca en Tecnologies Mèdia, La Salle—Universitat Ramon Llull, Sant Joan de la Salle, 42, 08022 Barcelona, Spain

* Correspondence: marc.freixes@salle.url.edu (M.F.); joanclaudi.socoro@salle.url.edu (J.C.S.); francesc.alias@salle.url.edu (F.A.); Tel.: +34-932-902-440 (M.F. & J.C.S. & F.A.)

† This paper is an extended version of our paper published in the conference IberSPEECH2020.

Abstract: The source-filter model is one of the main techniques applied to speech analysis and synthesis. Recent advances in voice production by means of three-dimensional (3D) source-filter models have overcome several limitations of classic one-dimensional techniques. Despite the development of preliminary attempts to improve the expressiveness of 3D-generated voices, they are still far from achieving realistic results. Towards this goal, this work analyses the contribution of both the the vocal tract (VT) and the glottal source spectral (GSS) cues in the generation of happy and aggressive speech through a GlottDNN-based analysis-by-synthesis methodology. Paired neutral expressive utterances are parameterised to generate different combinations of expressive vowels, applying the target expressive GSS and/or VT cues on the neutral vowels after transplanting the expressive prosody on these utterances. The conducted objective tests focused on Spanish [a], [i] and [u] vowels show that both GSS and VT cues significantly reduce the spectral distance to the expressive target. The results from the perceptual test show that VT cues make a statistically significant contribution in the expression of happy and aggressive emotions for [a] vowels, while the GSS contribution is significant in [i] and [u] vowels.

Keywords: expressive speech synthesis; emotional database; speech analysis; inverse filtering; glottal source; vocal tract; numerical voice production; GlottDNN

Citation: Freixes, M.; Socoró, J.C.; Alías, F. Contribution of Vocal Tract and Glottal Source Spectral Cues in the Generation of Acted Happy and Aggressive Spanish Vowels. *Appl. Sci.* **2022**, *12*, 2055. https://doi.org/10.3390/app12042055

Academic Editor: Douglas O'Shaughnessy

Received: 31 December 2021
Accepted: 13 February 2022
Published: 16 February 2022

1. Introduction

Speech is an incredibly powerful means of communication, as it not only codifies linguistic information, i.e., a message, but also provides paralinguistic cues about the emotional state of the speaker [1]. Emotion is, therefore, a key element for seamless human–computer interaction (HCI), both in the input channel, by means of emotional speech recognition [2] and in its output, through expressive speech synthesis [3], among others. In this context, emotions have been traditionally represented: (i) using a dimensional space, such as the circumplex model, which is defined by arousal, valence, and dominance [4]; or, (ii) as discrete categories, such as those defined in [5] and denoted as the *big six* basic emotions, namely, anger, disgust, fear, happiness, sadness, and surprise. According to these categories, expressive speech databases are built containing spontaneous speech, elicited speech, or acted speech [2].

In order to model the specific characteristics of spoken emotions, acoustic features can be extracted from these databases and analysed to describe them with respect to neutral speech [6], which is typically considered as reference of inexpressiveness. In this regard, several studies focusing on basic emotions have found specific prosodic patterns of F0, energy and speech rate. Angry and happy speech, for instance, present higher F0, energy and speech rate values compared with neutral speech, while the opposite occurs with sadness [7].

Alternatively, the emotional content of speech can be studied by manipulating the extracted features following an analysis-by-synthesis scheme [8], as illustrated in Figure 1. To this aim, neutral speech utterances are analysed to extract features that are subsequently modified according to the analysed expressive utterances to obtain resynthesised expressive stimuli. The obtained results can be objectively and/or subjectively evaluated with respect to the non-expressive reference and to what extent they resemble the expressive target. In this regard, speech synthesis approaches that are based on the voice production theory are of particular interest since they can provide valuable insights into the mechanisms involved in the generation of emotional speech. Most of these approaches follow the so-called source-filter model [9], which considers the speech signal to be composed of an excitation (glottal pulses) amplified at certain frequencies through the vocal tract (formants).

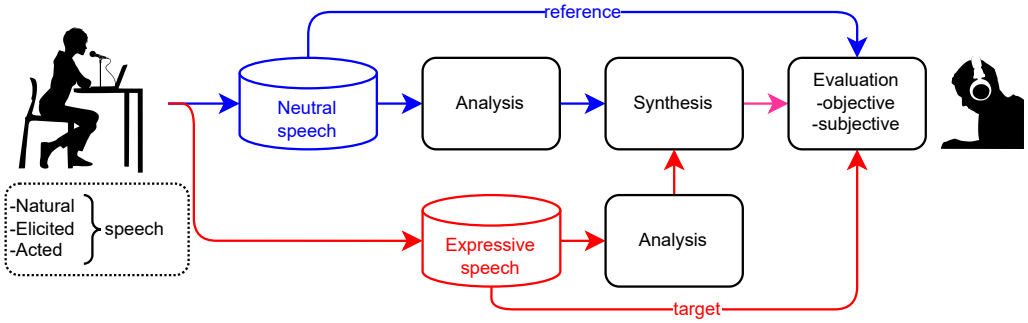

Figure 1. Workflow diagram of the analysis-by-synthesis approach applied to expressive speech. Neutral speech utterances are analysed to extract features that are manipulated according to the analysed expressive utterances to obtain resynthesised expressive stimuli. The results are objectively and/or subjectively evaluated in comparison with the non-expressive reference and the expressive target.

The relevance of both the glottal source and the vocal tract in the production of expressive speech has been explored in different investigations, considering a source-filter model and following an analysis-by-synthesis scheme. In [10], the authors analysed to what extent different phonation types contribute to the perception of emotion by resynthesising a set of utterances using the VocalTractLab synthesizer 2.0 [11]. This articulatory-based synthesiser is based on a 3D vocal tract geometric model and integrates a self-oscillating model of the vocal folds. It allows simulating the pressed-to-breathy phonation continuum by varying the degree of glottal abduction. In that work, the authors manually controlled the synthesiser using different gestural scores (to account for both supraglottal and laryngeal articulation) to mimic pre-recorded sentences as well as to produce certain variations of their phonation. The portrayed emotional utterances were resynthesised with their original phonation type, as well as with purely breathy, modal, and pressed phonations, which were evaluated perceptually. The obtained results showed that the recognition of fear, anger, and neutral speech demands specific phonation types, while it mainly relies on prosodic parameters for sadness, happiness, boredom, and disgust.

In [12], a similar study was conducted using a Klatt formant-based synthesiser with a modified Liljencrants–Fant (LF) glottal flow model [13] as input. A set of target sentences were copy-synthesised and subsequently resynthesised with different phonation types (breathy, creaky, falsetto, modal, and tense) through a rule-based model that controlled the glottal source model parameters. A forced choice perceptual test was conducted, in which the participants were asked to assign each stimulus to either neutral, anger, joy, sadness, frightened, or bored speech. The phonation types were distinguished by the listeners and lead to a satisfactory emotional impression. In [14], a similar synthesis scheme was considered to explore the interaction of voice quality and F0 contours with the listener's language background (Irish, English, Russian, Spanish, and Japanese) in the perception of affect. Three types of synthetic stimuli were generated from a short Swedish utterance spoken with modal voice: (i) stimuli varied in voice quality, (ii) modal stimuli incorporating

affect-related F0 contours, and (iii) stimuli combining certain non-modal voice qualities with the affect-related F0 contours from (ii). A perceptual test was conducted asking participants to rate the stimuli on six bipolar scales defined with contrastive adjectives at each end: indignant–apologetic, interested–bored, formal–intimate, stressed–relaxed, happy–sad, and fearless–scared. The results suggested that stimuli incorporating non-modal voice qualities, with or without F0 variation, were generally more effective in conveying emotions than stimuli only varying F0.

In [15], the R_d glottal shape parameter [16] was used to control the LF model so as to produce expressive voice along the tense–lax continuum, exploring its influence on the perception of emotions. Moreover, the relative importance of glottal source and vocal tract on emotional vowel perception was examined in [17] using an auto-regressive exogenous LF model. Glottal source and vocal tract parameters were estimated from isolated vowels [a] recorded by two professional male actors using four emotions (neutral, joy, anger, and sadness). Besides the eight resynthesised original vowels, sixteen synthetic vowels per speaker were obtained by combining the averaged glottal source parameters from one emotion and the averaged vocal tract parameters from another emotion. The synthesised vowels were perceptually evaluated using a valence–arousal model. The glottal source was found to have a dominant role, mainly due to the predominant effect of the prosodic features, i.e., F0, intensity, and duration. These cues were subsequently *neutralised* in a second experiment using the values from the neutral vowels. As a consequence, the relevance of the vocal tract was higher, but the resulting vowels were perceived in a similar way to their neutral counterparts, negatively affecting the expressiveness analysis.

The aforementioned approaches have focused on the analysis and resynthesis of a small set of isolated vowels or short utterances at most, involving typically costly manual tuning processes when developing the analysis-and-resynthesis approach. However, recent advances in inverse filtering have shown promising results in the automatic decomposition of the speech signal into its glottal source and vocal tract components [18]. Among these techniques, the glottal flow model-based vocoder allows real-time voice manipulations, such as shifting vowel formants and modifying voice quality through the manipulation of the glottal source model [19]. The GlottHMM approach, based on hidden Markov models, showed good results for the analysis of expressive nuances, beyond prosodic parameters such as F0 and rhythm [20]. More recently, GlottDNN [21] a successor of GlottHMM based on deep neural networks, has been used to to embed the Lombard effect on regular speech [22], obtaining good results.

The source-filter models has traditionally considered a representation of the vocal tract in one dimension (1D) [23]. Although the 1D-based approach has been widely used in the literature (see e.g., [11,24,25]), the simplification it entails makes it only capable of modelling the propagation of plane waves along the vocal tract midline, thus, limiting the accuracy of 1D-based approaches up to about 4–5 kHz [26]. During the last decade, and thanks to the significant increase in computing power, several attempts have been developed to overcome this limitation by means of three-dimensional (3D) vocal tract models [27–29], which allow the propagation of higher order modes [26]. This characteristic is especially relevant for the production of vowels with a tense phonation [30]. So far, such models have been already used for the synthesis of vowels [29], diphthongs [31] and vowel–consonant–vowel sequences [32], including sibilants [33] that entail considering aeroacoustic sources [34], as well as works focused on tuning the vocal tract resonances [35] to be able to simulate effects such as the so-called singing formant [36].

In order to introduce some expressiveness in the 3D numerical synthesis of vowels with a tense phonation, a GlottDNN-based analysis of the spectral tilt of the glottal source was proposed in [37]. To that effect, the R_d parameter of an LF model was modified accordingly, and the results were evaluated on isolated aggressive and happy [a] vowels. Although that work showed interesting preliminary results, it was only focused on modelling one parameter of the excitation. Therefore, further investigations are still necessary to be able to

generate natural expressive utterances within a 3D-based finite element methods synthesis scheme [38].

In order to advance in the modelling and integration of tense-expressive styles within a speech synthesizer based on a source-filter model, this work analyses the contribution of vocal tract (VT) and glottal source spectral (GSS) cues in the generation of happy and aggressive speech styles. To this aim, an GlottDNN-based analysis-by-synthesis approach is proposed and evaluated using a parallel Spanish speech corpus composed of neutral and expressive utterances. The contributions of GSS and/or VT cues on each target emotion are analysed through objective and perceptual tests by considering neutral utterances transplanted with the prosody (baseline), and GSS and/or VT cues from their expressive pairs vowels [a], [i] and [u], differentiating if they are stressed or not.

The paper is organised as follows. Section 2 presents the proposed methodology based on the GlottDNN analysis and synthesis of expressive utterances to study the contribution of GSS and VT cues to the generation of happy and aggressive tense emotional vowels from neutral speech. Next, Section 3 describes the conducted experiments, whose results are detailed and discussed in Section 4. Finally, the derived conclusions from this work, as well as future research lines, are presented in Section 5.

2. Analysis-by-Synthesis Methodology

Figure 2 depicts the proposed GlottDNN-based analysis-by-synthesis methodology designed to evaluate the contribution of both the vocal tract and the glottal source spectral cues to the synthesis of expressive speech vowels. Specifically, the analysis of expressive speech is based on the inverse filtering of a parallel speech corpus composed of utterances in the expressive style of interest, as well as their neutral counterparts (as reference). In order to focus exclusively on the voice quality differences between expressive and neutral speech, the prosody from the expressive utterance (F0, duration, and energy) is transplanted into the corresponding neutral pair. The resulting utterance is considered as the baseline in the comparisons between the different synthesis configurations, labelled as $GSS_X VT_Y$, where X and Y denote the origin of the GSS and VT cues applied to the utterance vowels, respectively, which can be either borrowed from the time-scaled neutral utterance (N), or from the target expressive component (E).

The main steps of the process are described as follows. As a first step, the GlottDNN vocoder is applied to parameterise each speech frame of the neutral–expressive utterance pair (see Figure 2b), obtaining their fundamental frequency (F0 in Hz) and energy (EN in dB). Moreover, the corresponding glottal source and VT cues are estimated through a quasi-closed phase (QCP) inverse filtering approach, whose result is subsequently parameterised using line spectral frequencies (LSF). Moreover, the harmonic-to-noise ratio (HNR) of the glottal source estimate is also computed at the frame level (the reader is referred to [21] for further details of this process). It is worth noting that the obtained QCP-based spectral tilt is compensated for to minimise the presence of residual spectral cues from the glottal source on the vocal tract estimation, following [22].

Secondly, the prosody of the neutral utterance is replaced by the one borrowed from its expressive pair. To this aim, the neutral GlottDNN-based features (marked with $'$ on Figure 2a) are time scaled to fit the expressive target according to the temporal alignment obtained through dynamic time warping. Next, $F0_N$ and EN_N are respectively replaced by the corresponding values from the expressive utterance, i.e., $F0_E$ and EN_E. Finally, depending on the tested combination, GSS_E and/or VT_E are also transplanted into the neutral vowels.

The final step corresponds to the GlottDNN-based expressive synthesis. Following [21], the parameters of voiced frames are inserted into a simple feed-forward deep neural network to generate a zero-padded, two-pitch-period, glottal flow-derivative pulse, while a white noise excitation is used for unvoiced frames. The amplitude of these excitation signals is adjusted to fit EN_E and concatenated through the classic pitch-synchronous overlap and add algorithm [39]. Subsequently, the HNR_N parameter is used to adjust the

noise component of the excitation in the spectral domain, besides modifying its spectrum to match the target GSS_X. Finally, the excitation signal is filtered according to the considered VT_Y to obtain the synthetic speech output. The GlottDNN-based pulse generation model trained with the neutral speech corpus and HNR_N are considered in the synthesis of the expressive utterances. This is due to the fact that the designed analysis methodology only focuses on evaluating the spectral characteristics of the glottal source, leaving the analysis of these two features of the glottal source for future works.

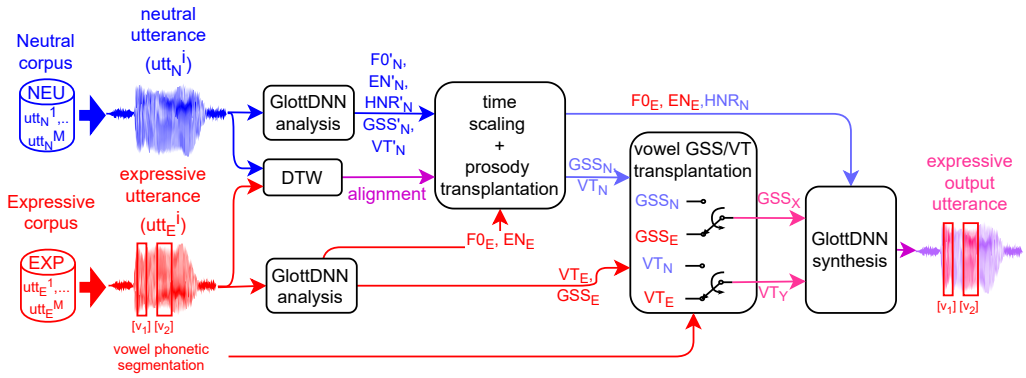

Figure 2. Block diagram of the analysis-by-synthesis methodology designed to evaluate the contributions of vocal tract and glottal source spectral cues (VT and GSS) in the generation of voice-expressive (EXP) vowels. M pairs of neutral and expressive utterances from a parallel speech corpus are parameterised with the GlottDNN vocoder, and aligned through dynamic time warping (DTW). Next, the features of each neutral utterance utt_N^i (marked with $'$) are time-scaled according to the DTW alignment, and transplanted with the $F0_E$ and EN_E of the expressive pair utt_E^i counterpart. The final step is GlottDNN-based synthesis, which generates different expressive output utterances by applying GSS_X and VT_Y to their vowels—being X, Y either neutral (N) or expressive (E). (**a**) Overall diagram of the analysis-by-synthesis methodology. (**b**) Details of the GlottDNN analysis.

3. Experiments

This section presents the main elements of the conducted experiments, that is, the considered emotional speech database, the GlottDNN-based inverse filtering features, and the objective and subjective evaluations.

The expressive data used in the experiments were extracted from a parallel emotional Spanish speech database recorded by a professional female speaker (acted speech), downsampled to 16 kHz [40], and subsequently labelled for expressive text-to-speech speech synthesis purposes [41]. Among the five speaking styles composing the expressive database, three of them were chosen for this work. Specifically, 1250 neutral, happy, and aggressive paired short utterances were retrieved from the database, thus, covering modal and tense phonation styles, respectively. These utterances—with an average of 1.2 words per utterance—ensure a phonetic coverage of the Spanish language. Out of them, 1035 ut-

terances were used in this work, specifically those containing at least one of the vowels [a], [i] or [u], as they represent the three extreme vowels of the Spanish vowel triangle. As a result, a total of 1852 paired vowels were considered in the experiments. Neutral vowels had an F0 mean of 152 Hz, while the happy and aggressive vowels presented higher values, of 268 Hz and 257 Hz, respectively.

Regarding the GlottDNN-based analysis-by-synthesis approach, the default settings of the GlottDNN (https://github.com/ljuvela/GlottDNN, accessed on 25 October 2021) were considered. That is, the GSS and the VT cues were parameterised using 10 and 30 LSF coefficients per frame, respectively, considering frames of 25 ms and 15 ms in length for voiced and unvoiced frames, accordingly. The GlottDNN pulse generation model was trained using the whole neutral corpus (of 2.4 h in length) following [21]. For the GlottDNN-based synthesis stage, after transplanting the target expressive prosody in its neutral pairs, the contributions of GSS and VT cues to the production of each target emotion were evaluated by considering four different configurations, denoted as GSS_NVT_N (the baseline configuration), GSS_EVT_N, GSS_NVT_E, and GSS_EVT_E (the expressive target configuration).

The contributions of GSS and VT cues to the generation of happy and aggressive expressive styles were evaluated through both objective measures and subjective tests. The former was determined by computing the spectral distances between the neutral–expressive vowel pairs, considering the expressive vowel as the target reference. To that effect, GSS and VT cues were obtained from the neutral vowel (GSS_N, VT_N), and from the expressive vowel (GSS_E, VT_E). Specifically, the similarity between the GSS_N and the GSS_E was computed using the Itakura–Saito LPC-based spectral distance denoted as $d_{IS}(GSS_N, GSS_E)$; a metric that was also used to compare the VT cues, that is, $d_{IS}(VT_N, VT_E)$. To that effect, the GlottDNN parameterises GSS and VT cues as LSF vectors, at a frame level, from which LSF vectors at the vowel level were obtained using the median to reduce coarticulation effects. Finally, LSFs were translated into LPC to compute the Itakura–Saito distances [42]. Moreover, the similarities of each version of the [a], [i] and [u] vowels obtained from one of the three possible configurations with respect to the expressive target configuration were measured as the symmetrical Kullback–Leibler spectral distance [43] between their long term average spectra (LTAS) and the corresponding one of their expressive target pairs, i.e., $d_{KL}(GSS_XVT_Y, GSS_EVT_E)$. To do so, LTAS were computed as the Welch's power spectral estimate, considering a 15-ms Hamming window with 50% overlap and a 2048-point FFT [30].

Finally, a multiple stimuli with hidden reference and anchor (MUSHRA) test [44] was conducted to evaluate the contribution of GSS and VT cues to the perception of the two target emotions through the four aforementioned synthesis configurations on the six vowels under study ($\Psi = \{$[a], ['a], [i], ['i], [u], ['u]$\}$). To this end, a total of 24 words were selected from the speech database subset containing only one vowel from the modified set of vowels Ψ plus possibly other vowels (i.e., [e] and/or ['e] and/or [o] and/or ['o]). In this way, for each type of vowel from Ψ four words were obtained, in which only the respective vowel was modified in terms of GSS and/or VT cues using the aforementioned analysis-by-synthesis methodology, while the other vowels were kept with the neutral version of both GSS and VT. Four versions of the test were prepared, each one consisting of 13 evaluation sets (6 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 emotional 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 to support the listener evaluation.

Twenty-five Spanish native speakers with an average age of 29.1 took one of the four versions of the online test using headphones and the Web Audio Evaluation Tool [45]. Among them, 56.0% of the participants had experience in playing and/or producing music, 32.0% in audio software/hardware design, and the 40.0% in audio or speech research.

According to the results of the control sets, the responses of nine participants were discarded since they presented significant criteria inconsistencies (i.e., with $r < 0.5$).

In order to determine if the differences between the evaluated configurations were statistically significant, the Wilcoxon signed-rank test [46] was applied to both the objective and subjective results. The obtained p-values were corrected according to the Holm–Bonferroni method.

4. Results and Discussion

This section presents the results obtained from both the objective comparisons and the conducted subjective evaluation.

4.1. Objective Results

Figure 3 depicts the distributions of the Itakura–Saito spectral distances of GSS_N and VT_N with respect to GSS_E and VT_E, respectively. When analysing unstressed and stressed vowels (first two boxplots), it can be observed that GSS differences are higher for happy than for aggressive, while the opposite occurs with the VT cues. Moreover, stressed vowels present higher differences than the unstressed ones.

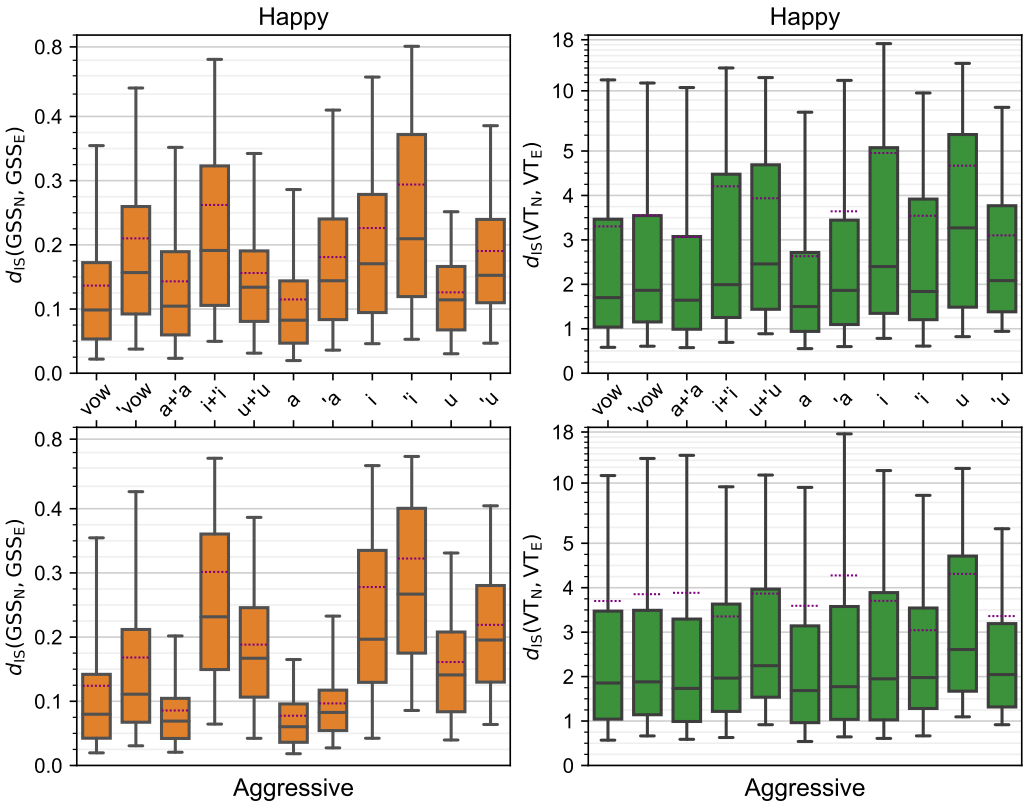

Figure 3. Boxplots of the Itakura–Saito distances from GSS_N to GSS_E (first column) and from VT_N to VT_E (second column), for happy (first row) and aggressive (second row) vowels. Dotted lines represent the mean of the distributions, and whiskers are set to 5th and 95th percentiles.

If we consider the vowels, higher GSS differences are observed on [i] + ['i], while [a] + ['a] present the lower ones. In between we find [u] + ['u], whose differences are halfway through the other vowels for aggressive speech, and closer to [a] + ['a] for happy. VT differences are higher for [u] + ['u], then for [i] + ['i], and finally for [a] + ['a].

Regarding the effect of the stress, GSS distances are higher for stressed vowels than for the unstressed ones in both expressive configurations. Similarly, VT distances for ['a] are higher than [a]. Conversely, VT distances for ['i] and ['u] are lower than [i] and [u],

respectively. The stress increases the VT differences for ['a], while ['i] and ['u] present lower differences with respect to their unstressed counterparts.

Figure 4 depicts the distributions of the spectral Kullback–Leibler distances from the vowels synthesised with the different configurations to the target GSS_EVT_E. According to the Wilcoxon test results, all the differences between configurations are statistically significant except for the GSS_NVT_N–GSS_EVT_N pair on happy [i] vowels and GSS_EVT_N–GSS_NVT_E on aggressive [u] vowels. Therefore, it can be observed that both GSS and VT cues make a relevant contribution to the generation of happy and aggressive vowels except in the aforementioned cases.

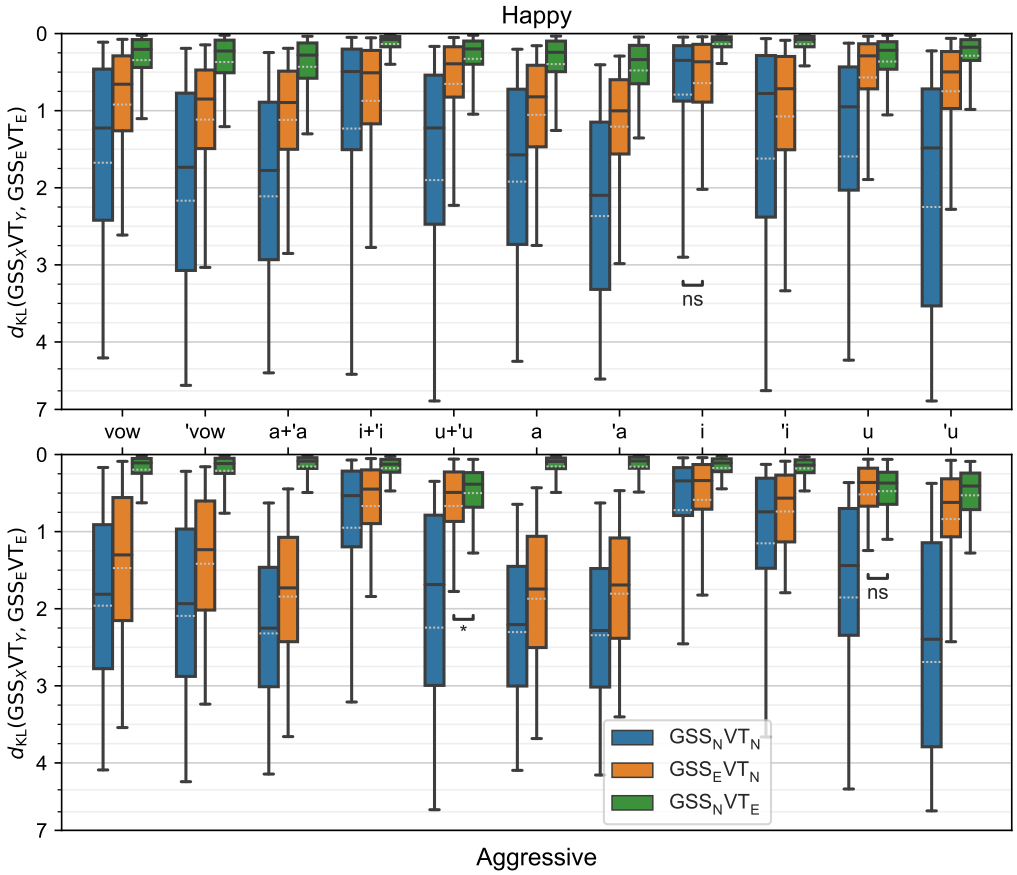

Figure 4. Boxplots of the Kullback–Leibler distances from the analysed configurations to the target configuration (GSS_EVT_E) for happy and aggressive vowels. Dotted lines represent the mean of the distributions, and whiskers are set to the 5th and 95th percentiles. The differences between configurations are statistically significant ($p < 0.01$) except those marked with * ($p < 0.05$) and ns (no significant).

4.2. Perceptual Evaluation Results

Figure 5 depicts the results obtained from the MUSHRA test. First, results of grouped unstressed and stressed vowels (vow and 'vow) are analysed (i.e., the first two groups of boxplots). The baseline configuration (with expressive prosody, GSS_N and VT_N) obtains the lowest perceived emotional intensity (median score of 47 and 52 for happy unstressed and stressed vowels, respectively, and 54 and 52 for the aggressive ones).

On the one hand, the incorporation of GSS_E significantly increases the perceived emotional intensity of happy unstressed and stressed vowels by 6.4% and 11.5%, respectively (moving from 47 to 50 and from 52 to 58 points in the MUSHRA scale). Regarding the aggressive emotion, the contribution of GSS_E is only significant for the stressed vowels, with an 19.2% increase (from 52 to 62). On the other hand, the sole contribution of the VT_E

with respect to the baseline is only significant for the aggressive stressed vowels, increasing the perceived emotion by 23.1% (from 52 to 64).

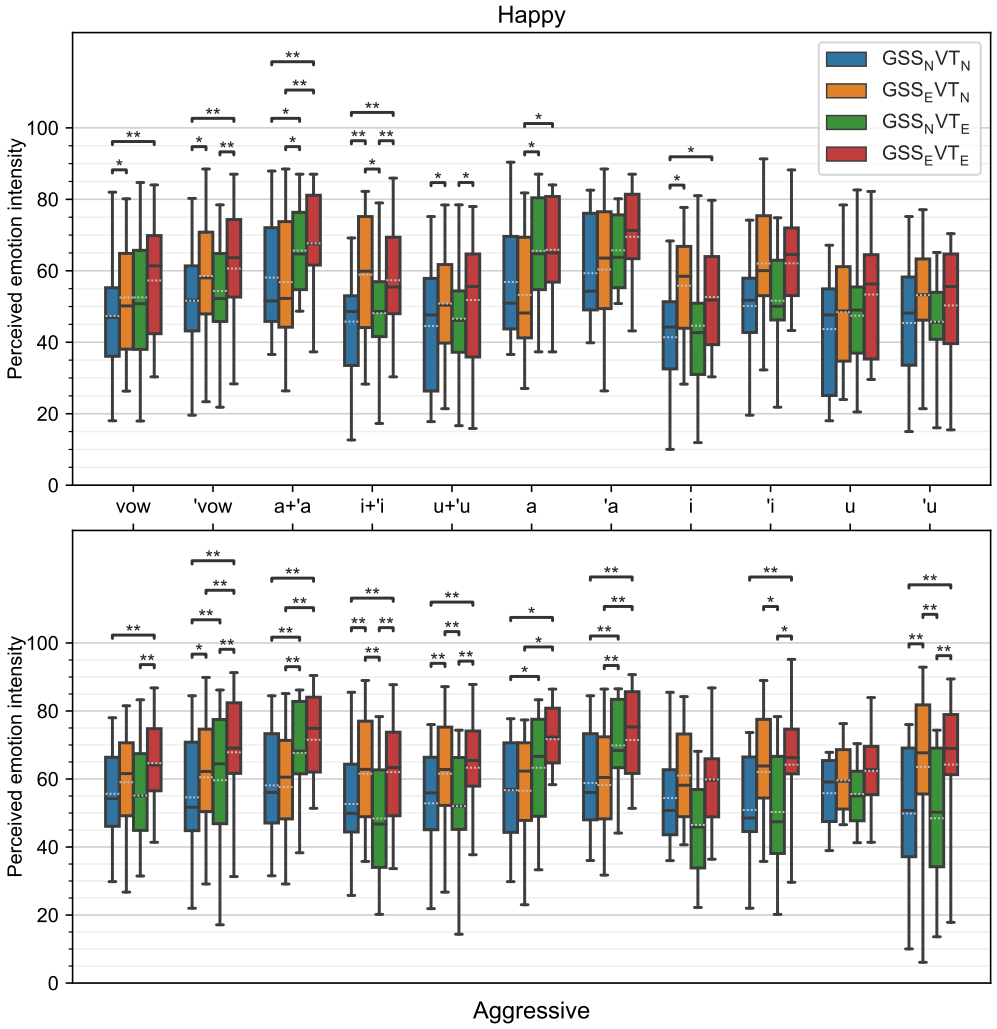

Figure 5. Results of the MUSHRA perceptual test. Boxplots depict the perceived emotion intensity scores reported by the participants. The dotted lines represent the mean of the distributions, and whiskers are set to 5th and 95th percentiles. The statistically significant differences are marked with ** ($p < 0.01$) and * ($p < 0.05$).

The best results are achieved when both GSS_E and VT_E are considered in the synthesis stage. The increase with respect to the baseline for unstressed and stressed happy vowels is 29.8% (from 47 to 61) and 23.1% (from 52 to 64), respectively, and 18.5% (from 54 to 64) and 32.78% (from 52 to 69) for the aggressive ones.

With regard to the vowels, the contribution of GSS_E is significant for [i] + ['i] and [u] + ['u]. In the latter, the increase is 4.2% (from 48 to 50) and 12.5% (from 56 to 63) for happy and aggressive speech, respectively. The higher differences are found in [i] + ['i], with increases of 22.4% (from 49 to 60) and 26.0% (from 50 to 63). Conversely, the contribution of VT_E is significant for happy and aggressive [a] + ['a], with increases of 25% (from 52 to 65) and 21.4% (from 56 to 68), respectively.

4.3. Discussion

Several aspects of the results obtained from the objective and subjective evaluations may be the subject of discussion. On the one hand, when only GSS_E is considered in the generation of the expressive vowel, the spectral distance to the target is significantly reduced for all the vowels and emotions except happy [i] (see Figure 4). However, according

to the results of the perceptual test (see Figure 5), the GSS contribution is only significant on vowels [i] + ['i] and [u] + ['u]. This could be explained by looking at the differences between GSS_N and GSS_E (see left part of Figure 3), which are more pronounced for these vowels, especially for the aggressive ones.

On the other hand, considering only VT_E significantly reduces the spectral distances to the target in all the cases, as can be seen in Figure 4. Nevertheless, from the perceptual point of view, such differences are only significant for vowels [a] + ['a]. As Figure 3 shows, these are exactly the vowels which present the largest differences between VT_N and VT_E.

The aforementioned differences between vowels can be also analysed by looking at the averaged spectra of the happy and aggressive vowels (see Figures 6 and 7, respectively). It can be observed how GSS differences are more pronounced for [i] + ['i] and [u] + ['u] than for [a] + ['a] vowels. Conversely, VT spectral differences are particularly relevant in vowels [a] + ['a], where a clear shift of the first two formants can be observed.

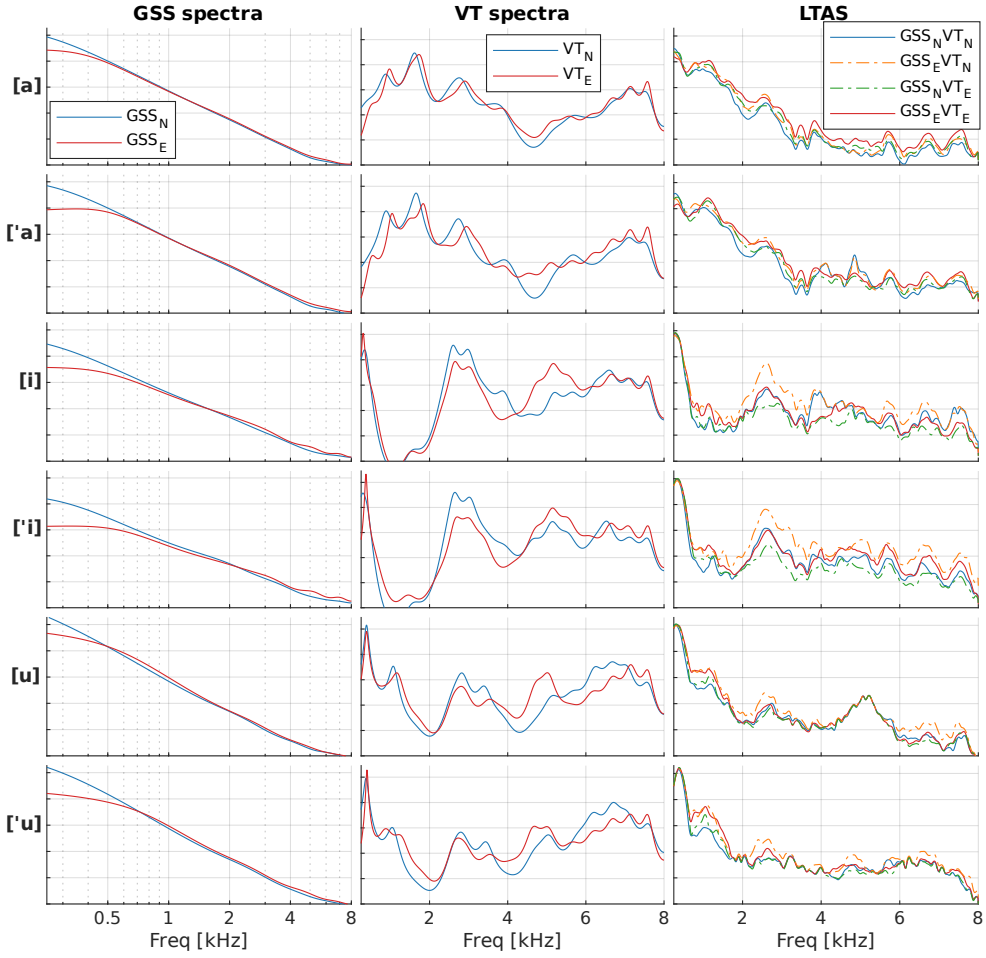

Figure 6. Averaged spectra of the analysed configurations for the happy emotion. The first and the second column show the spectra obtained from the mean vectors of glottal source spectra and vocal tract, respectively, for the neutral (GSS_N, VT_N) and expressive (GSS_E, VT_E) configurations. The third column depicts the mean LTAS of the samples synthesised with the considered combinations of GSS and VT.

Although this work bears some resemblance to [17], there are several differences that should be pointed out. In order to study the relevance of GSS and VT cues in resembling the target emotion, the contribution of prosody was minimised. To this end, the synthesised utterances were transplanted with the prosody of their expressive pairs instead of using the neutral prosody, as in [17], biasing the obtained results significantly. On the other hand, short utterances were used instead of isolated vowels. This has allowed us to study vowels

in their phonetic context, and to ask for the perceived emotional intensity instead of using the arousal–valence space, as done in [17].

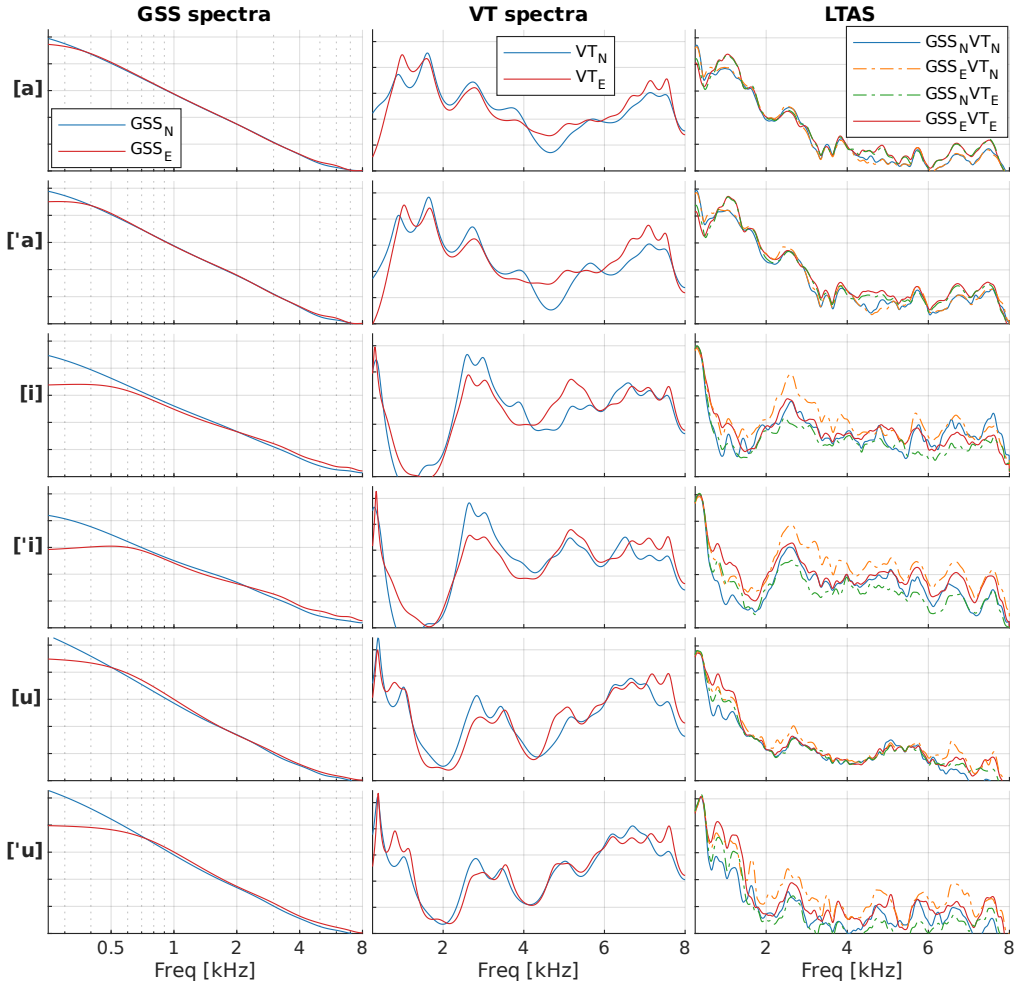

Figure 7. Averaged spectra of the analysed configurations for the aggressive emotion. The first and the second column show the spectra obtained from the mean vectors of glottal source spectra and vocal tract, respectively, for the neutral (GSS_N, VT_N) and expressive (GSS_E, VT_E) configurations. The third column depicts the mean LTAS of the samples synthesised with the considered combinations of GSS and VT.

The present work has focused on two expressive speaking styles characterised by a tense phonation, using happy and aggressive acted speech. Further work could be done to extend the study to other acted speaking styles (e.g., sad acted speech, with more relaxed phonation types) as well as expressive speech obtained in other different conditions (e.g., real or elicited speech). Another aspect that should be addressed in future studies is the inverse filtering process. High-pitched female speech and closed vowels (with low F1) represent a challenge to this process [47]. In this regard, future advances in inverse filtering methods could be considered in order to improve the precision of the GSS and VT cue estimations.

Finally, as previously mentioned, the inclusion of expressive nuances in numerical voice production is still an open issue. In [37], the authors attempted to add some expressiveness to a 3D numerical synthesis of the vowel [a] based on the finite element method (FEM) by modifying the R_d parameter of a LF glottal flow model. In that research important simplifications were assumed, that is, the GSS was parameterised using a single scalar value to represent its spectral tilt, and the comparisons between styles were based on the mean differences of this spectral tilt value. Moreover, it is worth mentioning that no modifications

of the VT were considered. After analysing the contributions of both GSS and VT cues, we envision the integration of these results within the 3D FEM-based numerical synthesis workflow by introducing the observed changes in the glottal flow waveform together with the proper variations of the realistic 3D vocal tract geometry [48].

5. Conclusions

This work has analysed the contribution of the GSS and VT cues to the generation of happy and aggressive emotional vowels by means of GlottDNN-based analysis-by-synthesis methodology, considering a parallel neutral-expressive speech database. The experiments have considered the Spanish [a], [i] and [u] vowels from aggressive, happy, and neutral paired utterances. Results from the objective evaluation show that both GSS and VT cues have a statistically significant contribution to convey these tense voice emotions when compared with the baseline reference (with expressive prosody, GSS_N and VT_N). The subjective evaluations show that VT cues significantly affect the perceived emotion of vowels [a] + ['a], while GSS cues have a significant effect on the vowels [i] + ['i] and [u] + ['u], especially on the former.

Future work will focus on developing further analyses to validate the obtained results on other phonemes and will also consider other expressive speaking styles and/or phonation types, such as the ones present in sadness. Furthermore, we also plan to input the obtained results within our 3D-based numerical synthesizer by introducing the observed relevant variations of both the glottal source and the vocal tract to improve the current naturalness of the happy and aggressive speaking styles generated.

Author Contributions: Conceptualisation, methodology, writing—original draft preparation, writing—reviewing, and editing, M.F., J.C.S. and F.A.; software, formal analysis and investigation, M.F. and J.C.S.; validation, data curation, and visualisation, M.F.; supervision, J.C.S. and F.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research has been partially funded by the Agencia Estatal de Investigación (AEI) through the FEMVoQ project (PID2020-120441GB-I00/AEI/10.13039/501100011033).

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable.

Data Availability Statement: The data generated in this study are available on request from the corresponding author.

Acknowledgments: The authors would like to thank the participants on the perceptual test for their collaboration in this work.

Conflicts of Interest: The authors declare no conflict of interest.

Sample Availability: The data generated and analysed during the current study is available from the corresponding author on reasonable request.

Abbreviations

The following abbreviations are used in this manuscript:

DNN	Deep Neural Networks
DTW	Dynamic Time Warping
FEM	Finite element Method
GSS	Glottal Source Spectra
HMM	Hidden Markov Models
HNR	Harmonic-to-Noise Ratio
LF	Liljencrants–Fant model
LPC	Linear Predictive Coding

LSF	Line Spectral Frequencies
LTAS	Long-Term Average Spectrum
MUSHRA	MUltiple Stimuli with Hidden Reference and Anchor
QCP	Quasi-Closed Phase
VT	Vocal Tract

References

1. Schuller, D.M.; Schuller, B.W. A Review on Five Recent and Near-Future Developments in Computational Processing of Emotion in the Human Voice. *Emot. Rev.* **2021**, *13*, 44–50. [CrossRef]
2. Wani, T.M.; Gunawan, T.S.; Qadri, S.A.A.; Kartiwi, M.; Ambikairajah, E. A Comprehensive Review of Speech Emotion Recognition Systems. *IEEE Access* **2021**, *9*, 47795–47814. [CrossRef]
3. Zhou, K.; Sisman, B.; Liu, R.; Li, H. Emotional voice conversion: Theory, databases and ESD. *Speech Commun.* **2022**, *137*, 1–18. [CrossRef]
4. Russell, J.A. A circumplex model of affect. *J. Personal. Soc. Psychol.* **1980**, *39*, 1161. [CrossRef]
5. Ekman, P. An argument for basic emotions. *Cogn. Emot.* **1992**, *6*, 169–200. [CrossRef]
6. Eyben, F.; Scherer, K.R.; Schuller, B.W.; Sundberg, J.; Andre, E.; Busso, C.; Devillers, L.Y.; Epps, J.; Laukka, P.; Narayanan, S.S.; Truong, K.P. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Trans. Affect. Comput.* **2016**, *7*, 190–202. [CrossRef]
7. Scherer, K.R. Vocal communication of emotion: A review of research paradigms. *Speech Commun.* **2003**, *40*, 227–256. [CrossRef]
8. Arias, P.; Rachman, L.; Liuni, M.; Aucouturier, J.J. Beyond Correlation: Acoustic Transformation Methods for the Experimental Study of Emotional Voice and Speech. *Emot. Rev.* **2021**, *13*, 12–24. [CrossRef]
9. Taylor, P. *Text-to-Speech Synthesis*; Cambridge University Press: Cambridge, UK, 2009; pp. 1–626. [CrossRef]
10. Birkholz, P.; Martin, L.; Willmes, K.; Kröger, B.J.; Neuschaefer-Rube, C. The contribution of phonation type to the perception of vocal emotions in German: An articulatory synthesis study. *J. Acoust. Soc. Am.* **2015**, *137*, 1503–1512. [CrossRef]
11. Birkholz, P. Modeling Consonant-Vowel Coarticulation for Articulatory Speech Synthesis. *PLoS ONE* **2013**, *8*, e60603. [CrossRef]
12. Burkhardt, F. Rule-Based Voice Quality Variation with Formant Synthesis. In Proceedings of the InterSpeech 2009, Brighton, UK, 6–10 September 2009; pp. 2659–2662.
13. Fant, G.; Liljencrants, J.; Lin, Q. A four-parameter model of glottal flow. *Speech Transm. Lab. Q. Prog. Status Rep. (STL-QPSR)* **1985**, *26*, 1–13. [CrossRef]
14. Yanushevskaya, I.; Gobl, C.; Ní Chasaide, A. Cross-language differences in how voice quality and f_0 contours map to affect. *J. Acoust. Soc. Am.* **2018**, *144*, 2730–2750. [CrossRef] [PubMed]
15. Murphy, A.; Yanushevskaya, I.; Ní Chasaide, A.; Gobl, C. Rd as a Control Parameter to Explore Affective Correlates of the Tense-Lax Continuum. In Proceedings of the InterSpeech 2017, Stockholm, Sweden, 20–24 August 2017; pp. 3916–3920. [CrossRef]
16. Fant, G. The LF-model revisited. Transformations and frequency domain analysis. *Speech Transm. Lab. Q. Prog. Status Rep. (STL-QPSR)* **1995**, *36*, 119–156.
17. Li, Y.; Li, J.; Akagi, M. Contributions of the glottal source and vocal tract cues to emotional vowel perception in the valence-arousal space. *J. Acoust. Soc. Am.* **2018**, *144*, 908–916. [CrossRef] [PubMed]
18. Drugman, T.; Alku, P.; Alwan, A.; Yegnanarayana, B. Glottal source processing: From analysis to applications. *Comput. Speech Lang.* **2014**, *28*, 1117–1138. [CrossRef]
19. Perrotin, O.; McLoughlin, I. GFM-Voc: A Real-Time Voice Quality Modification System. In Proceedings of the InterSpeech 2019, Graz, Austria, 15–19 September 2019; pp. 3685–3686.
20. Lorenzo-Trueba, J.; Barra-Chicote, R.; Raitio, T.; Obin, N.; Alku, P.; Yamagishi, J.; Montero, J.M. Towards Glottal Source Controllability in Expressive Speech Synthesis. In Proceedings of the InterSpeech 2012, Portland, OR, USA, 9–13 September 2012; pp. 1620–1623.
21. Airaksinen, M.; Juvela, L.; Bollepalli, B.; Yamagishi, J.; Alku, P. A Comparison between STRAIGHT, Glottal, and Sinusoidal Vocoding in Statistical Parametric Speech Synthesis. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2018**, *26*, 1658–1670. [CrossRef]
22. Seshadri, S.; Juvela, L.; Räsänen, O.; Alku, P. Vocal Effort based Speaking Style Conversion using Vocoder Features and Parallel Learning. *IEEE Access* **2019**, *7*, 17230–17246. [CrossRef]
23. Story, B.H.; Titze, I.R.; Hoffman, E.A. Vocal tract area functions from magnetic resonance imaging. *J. Acoust. Soc. Am.* **1996**, *100*, 537–554. [CrossRef]
24. Story, B.H. Phrase-level speech simulation with an airway modulation model of speech production. *Comput. Speech Lang.* **2013**, *27*, 989–1010. [CrossRef]
25. Stone, S.; Marxen, M.; Birkholz, P. Construction and evaluation of a parametric one-dimensional vocal tract model. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2018**, *26*, 1381–1392. [CrossRef]
26. Blandin, R.; Arnela, M.; Laboissière, R.; Pelorson, X.; Guasch, O.; Hirtum, A.V.; Laval, X. Effects of higher order propagation modes in vocal tract like geometries. *J. Acoust. Soc. Am.* **2015**, *137*, 832–843. [CrossRef] [PubMed]
27. Vampola, T.; Horáček, J.; Švec, J.G. FE Modeling of Human Vocal Tract Acoustics. Part I: Production of Czech vowels. *Acta Acust. United Acust.* **2008**, *94*, 433–447. [CrossRef]

28. Takemoto, H.; Adachi, S.; Mokhtari, P.; Kitamura, T. Acoustic interaction between the right and left piriform fossae in generating spectral dips. *J. Acoust. Soc. Am.* **2013**, *134*, 2955–2964. [CrossRef] [PubMed]
29. Arnela, M.; Dabbaghchian, S.; Blandin, R.; Guasch, O.; Engwall, O.; Van Hirtum, A.; Pelorson, X. Influence of vocal tract geometry simplifications on the numerical simulation of vowel sounds. *J. Acoust. Soc. Am.* **2016**, *140*, 1707–1718. [CrossRef] [PubMed]
30. Freixes, M.; Arnela, M.; Socoró, J.C.; Alías, F.; Guasch, O. Glottal Source Contribution to Higher Order Modes in the Finite Element Synthesis of Vowels. *Appl. Sci.* **2019**, *9*, 4535. [CrossRef]
31. Arnela, M.; Dabbaghchian, S.; Guasch, O.; Engwall, O. MRI-based vocal tract representations for the three-dimensional finite element synthesis of diphthongs. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2019**, *27*, 2173–2182. [CrossRef]
32. Arnela, M.; Guasch, O. Finite element simulation of /asa/ in a three-dimensional vocal tract using a simplified aeroacoustic source model. In Proceedings of the 23rd International Congress on Acoustics (ICA), Aachen, Germany, 9–13 September 2019; pp. 1802–1809.
33. Pont, A.; Guasch, O.; Arnela, M. Finite element generation of sibilants /s/ and /z/ using random distributions of Kirchhoff vortices. *Int. J. Numer. Methods Biomed. Eng.* **2020**, *36*, e3302. [CrossRef]
34. Schoder, S.; Maurerlehner, P.; Wurzinger, A.; Hauser, A.; Falk, S.; Kniesburges, S.; Döllinger, M.; Kaltenbacher, M. Aeroacoustic sound source characterization of the human voice production-perturbed convective wave equation. *Appl. Sci.* **2021**, *11*, 2614. [CrossRef]
35. Guasch, O.; Arnela, M.; Pont, A. Resonance tuning in vocal tract acoustics from modal perturbation analysis instead of nonlinear radiation pressure. *J. Sound Vib.* **2021**, *493*, 115826. [CrossRef]
36. Arnela, M.; Guasch, O.; Freixes, M. Finite element generation of sung vowels tuning 3D MRI-based vocal tracts. In Proceedings of the 27th International Congress on Sound and Vibration (ICSV27), Graz, Austria, 11–16 July 2021; pp. 1–8.
37. Freixes, M.; Arnela, M.; Alías, F.; Socoró, J.C. GlottDNN-based spectral tilt analysis of tense voice emotional styles for the expressive 3D numerical synthesis of vowel [a]. In Proceedings of the 10th ISCA Speech Synthesis Workshop (SSW), Vienna, Austria, 20–22 September 2019; pp. 132–136. [CrossRef]
38. Guasch, O.; Alías, F.; Arnela, M.; Socoró, J.C.; Freixes, M.; Pont, A. GENIOVOX Project: Computational generation of expressive voice. In Proceedings of the IberSPEECH2021, Valladolid, Spain, 24–25 March 2021.
39. Moulines, E.; Charpentier, F. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Commun.* **1990**, *9*, 453–467. [CrossRef]
40. Iriondo, I.; Planet, S.; Socoró, J.C.; Martínez, E.; Alías, F.; Monzo, C. Automatic refinement of an expressive speech corpus assembling subjective perception and automatic classification. *Speech Commun.* **2009**, *51*, 744–758. [CrossRef]
41. Alías, F.; Sevillano, X.; Socoró, J.C.; Gonzalvo, X. Towards high-quality next-generation text-to-speech synthesis: A multidomain approach by automatic domain classification. *IEEE Trans. Audio Speech Lang. Process.* **2008**, *16*, 1340–1354. [CrossRef]
42. Rabiner, L.; Biing-Hwang, J. *Fundamentals of Speech Recognition*; Prentice Hall: Englewood Cliffs, NJ, USA, 1993.
43. Klabbers, E.; Veldhuis, R. Reducing audible spectral discontinuities. *IEEE Trans. Speech Audio Process.* **2001**, *9*, 39–51. [CrossRef]
44. ITU-R. *ITU-R BS.1534-1: Method for the Subjective Assessment of Intermediate Quality Level of Coding Systems*; International Telecommunication Union: Geneva, Switzerland, 2003.
45. Jillings, N.; De Man, B.; Moffat, D.; Reiss, J.D. Web audio evaluation tool: A browser-based listening test environment. In Proceedings of the 12th International Conference in Sound and Music Computing (SMC 2015), Maynooth, Ireland, 26 July–1 August 2015; pp. 147–152.
46. Wilcoxon, F. Individual Comparisons by Ranking Methods. *Biom. Bull.* **1945**, *1*, 80–83. [CrossRef]
47. Perrotin, O.; McLoughlin, I. A Spectral Glottal Flow Model for Source-filter Separation of Speech. In Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 7160–7164. [CrossRef]
48. Arnela, M.; Guasch, O. Tuning MRI-based vocal tracts to modify formants in the three-dimensional finite element production of vowels. In Proceedings of the 12th International Conference on Voice Physiology and Biomechanics, Grenoble, France, 18–20 March 2020.