Adding expressiveness to unit selection speech synthesis and to numerical voice production

Marc Freixes Guerreiro

UNIVERSITAT RAMON LLULL

DOCTORAL THESIS

Title	Adding expressiveness to unit selection speech synthesis and to numerical voice production
Presented by	Marc Freixes Guerreiro
Centre	La Salle International School of Commerce and Digital Economy
Department	Engineering
Directed by	Dr. Francesc Alías Dr. Joan Claudi Socoró

Abstract

Speech is one of the most natural and direct forms of communication between human beings, as it codifies both a message and paralinguistic cues about the emotional state of the speaker, its mood, or its intention, thus becoming instrumental in pursuing a more natural Human Computer Interaction (HCI). In this context, the generation of expressive speech for the HCI output channel is a key element in the development of assistive technologies or personal assistants among other applications.

Synthetic speech can be generated from recorded speech using corpus-based methods such as Unit-Selection (US), which can achieve high quality results but whose expressiveness is restricted to that available in the speech corpus. In order to improve the quality of the synthesis output, the current trend is to build ever larger speech databases, especially following the so-called End-to-End synthesis approach based on deep learning techniques. However, recording ad-hoc corpora for each and every desired expressive style can be extremely costly, or even unfeasible if the speaker is unable to properly perform the styles required for a given application (e.g., singing in the storytelling domain). Alternatively, new methods based on the physics of voice production have been developed in the last decade thanks to the increase in computing power. For instance, vowels or diphthongs can be obtained using the Finite Element Method (FEM) to simulate the propagation of acoustic waves through a 3D realistic vocal tract geometry obtained from Magnetic Resonance Imaging (MRI). However, since the main efforts in these numerical voice production methods have been focused on improving the modelling of the voice generation process, little attention has been paid to its expressiveness up to now. Furthermore, the collection of data for such simulations is very costly, besides requiring manual time-consuming postprocessing like that needed to extract 3D vocal tract geometries from MRI.

The aim of the thesis is to add expressiveness into a system that generates neutral voice, without having to acquire expressive data from the original speaker. One the one hand, expressive capabilities are added to a Unit-Selection Text-to-Speech (US-TTS) system fed with a neutral speech corpus, to address specific and timely needs in the storytelling domain, such as for singing or in suspenseful situations. To this end, speech is parameterised using a harmonic-based model and subsequently transformed to the target expressive style according to an expert system. A first approach dealing with the synthesis of storytelling increasing suspense shows the viability of the proposal in terms of naturalness and storytelling quality. Singing capabilities are also added to the US-TTS system through the integration of Speech-to-Singing (STS) transformation modules into the TTS pipeline, and by incorporating an expressive prosody generation module that allows the US to select units closer to the target singing prosody obtained from the input score. This results in a Unit Selection based Text-to-Speech-and-Singing (US-TTS&S) synthesis framework that can generate both speech

and singing from the same neutral speech small corpus (∼2.6 h). According to the objective results, the score-driven US strategy can reduce the pitch scaling factors required to produce singing from the selected spoken units, but its effectiveness is limited regarding the time-scale requirements due to the short duration of the spoken vowels. Results from the perceptual tests show that although the obtained naturalness is obviously far from that given by a professional singing synthesiser, the framework can address eventual singing needs for synthetic storytelling with a reasonable quality.

The incorporation of expressiveness is also investigated in the 3D FEM-based numerical simulation of vowels through modifications of the glottal flow signals following a source-filter approach of voice production. These signals are generated using a Liljencrants-Fant (LF) model controlled with the glottal shape parameter R_d, which allows exploring the tense-lax continuum of phonation besides the spoken vocal range of fundamental frequency values, $F0$. The contribution of the glottal source to higher order modes in the FEM synthesis of cardinal vowels [a], [i] and [u] is analysed through the comparison of the High Frequency Energy (HFE) values obtained with realistic and simplified 3D geometries of the vocal tract. The simulations indicate that higher order modes are expected to be perceptually relevant according to reference values stated in the literature, particularly for tense phonations and/or high $F0$s. Conversely, vowels with a lax phonation and/or low $F0$s can result in inaudible HFE levels, especially if aspiration noise is not present in the glottal source. After this preliminary study, the excitation characteristics of happy and aggressive vowels from a Spanish parallel speech corpus are analysed with the aim of incorporating this tense voice expressive styles into the numerical production of voice. To that effect, the GlottDNN vocoder is used to analyse $F0$ and spectral tilt variations associated with the glottal excitation on vowels [a]. These variations are mapped through the comparison with synthetic vowels into $F0$ and R_d values to simulate vowels resembling happy and aggressive styles. Results show that it is necessary to increase $F0$ and decrease R_d with respect to neutral speech, with larger variations for happy than aggressive style, especially for the stressed [a] vowels.

The results achieved in the conducted investigations validate the possibility of adding expressiveness to both corpus-based US-TTS synthesis and FEM-based numerical simulation of voice. Nevertheless, there is still room for improvement. For instance, the strategy applied to the numerical voice production could be improved by studying and developing inverse filtering approaches as well as incorporating modifications of the vocal tract, whereas the developed US-TTS&S framework could benefit from advances in voice transformation techniques including voice quality modifications, taking advantage of the experience gained in the numerical simulation of expressive vowels.

Acknowledgements

Voldria començar amb unes paraules d'agraïment pels meus directors de tesi. Moltes gràcies Francesc per confiar en mi abans i durant el doctorat, molts cops més que jo mateix. No ha estat un procés fàcil i sempre has estat allà *keep pushing*. Moltes gràcies Joan Claudi per la teva ajuda, pel teu optimisme i entusiasme i per arremangar-te i baixar a les catacumbes del Matlab quan ha fet falta. Moltes gràcies també a la Núria i el Xavi que des del programa de doctorat m'han anat seguint i acompanyant durant tot el procés.

Un agraïment a tota la gent que ha participat de forma totalment altruista en els tests perceptius. Gràcies als companys amb els que he treballat en els articles que conformen aquesta tesi. Marc i Oriol, desitjo que seguim estretant els llaços entre acústica i processament (reconeixeu-ho, vosaltres també feu processament). Raúl, fue un placer trabajar contigo, lástima que te pasaras al lado oscuro. Gracies Rosa per les xerrades/reunions al bar i per comptar amb mi per projectes, papers i altres aventures. Als altres companys que contribuïu a crear un magnífic entorn de treball: Alejandro, Ferran, Sevillano, David, Roger, Paula... I als excompanys: Ramon, Davide, Arnau, i de forma especial a l'Àngels, gràcies per ser-hi sempre.

Thanks to the researchers of the University of Crete. You are a great example of Cretan hospitality. Thank you Yannis for opening the doors of your lab. And thank you George for your help during the internship.

Voldria agrair el suport de la Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement de la Generalitat de Catalunya i el Fons Social Europeu pels ajuts per a la contractació de personal investigador novell (FI) [2014FI_B 00868, 2015FI_B1 00069 i 2016FI_B2 000904]. La recerca d'aquesta tesi també ha estat parcialment finançada pel SUR/DEC mitjançant l'ajut 2014-SGR-0590 al Grup de recerca en Tecnologies Mèdia; i per la Agencia Estatal de Investigación (AEI) i FEDER, EU, a través del projecte GENIOVOX TEC2016-81107-P.

Finalment, gràcies als meus amics. Xavi, sembla que ara sí, ara va de debò. Moltes gràcies a la meva família, i especialment als meus pares pel seu recolzament, per haver-me ajudat a arribar fins aquí. Voldria acabar amb un agraïment molt especial a la Laia, que ha estat qui ha patit de més aprop les conseqüències de la tesi. Gràcies pel teu recolzament i la teva paciència. Sé que a vegades no ha estat fàcil, però sembla que ho hem aconseguit. Amb moltes ganes de començar nous projectes al teu costat.

List of Papers

Indexed articles

Paper I

Marc Freixes, Joan Claudi Socoró, Francesc Alías. 'Adding Singing Capabilities to Unit Selection TTS through HNM-Based Conversion'. In: *Advances in Speech and Language Technologies for Iberian Languages. IberSPEECH 2016. Lecture Notes in Computer Science.* volume 10077, pp. 33–43. DOI: 10.1007/978-3-319-49169-1_4.

Paper II

Marc Freixes, Joan Claudi Socoró, Francesc Alías. 'A Unit Selection Text-to-Speech-and-Singing Synthesis Framework from Neutral Speech: Proof of concept'. In: *EURASIP Journal on Audio, Speech and Music Processing.* December 2019, volume 2019, article number 22. DOI: 10.1186/s13636-019-0163-y.

Paper III

Marc Freixes, Marc Arnela, Joan Claudi Socoró, Francesc Alías, Oriol Guasch. 'Glottal Source Contribution to Higher Order Modes in the Finite Element Synthesis of Vowels'. In: *Applied Sciences - Special Issue "IberSPEECH 2018: Speech and Language Technologies for Iberian Languages".* October 2019, volume 9(21), pp. 4535. DOI: 10.3390/app9214535.

Other articles

Paper IV

Raul Montaño, Marc Freixes, Francesc Alías, Joan Claudi Socoró. 'Generating Storytelling Suspense from Neutral Speech using a Hybrid TTS Synthesis framework driven by a Rule-based Prosodic Model'. In: *Proceedings of IberSPEECH 2016.* November 2016, pp. 129–138.

Paper V

Marc Freixes, Marc Arnela, Joan Claudi Socoró, Francesc Alías, Oriol Guasch. 'Influence of tense, modal and lax phonation on the three-dimensional finite element synthesis of

vowel [ɑ]'. In: *Proceedings of IberSPEECH 2018*. November 2018, pp. 132–136. DOI: 10.21437/IberSPEECH.2018-28.

Paper VI

Marc Freixes, Marc Arnela, Francesc Alías, Joan Claudi Socoró. 'GlottDNN-based spectral tilt analysis of tense voice emotional styles for the expressive 3D numerical synthesis of vowel [a]'. In: *Proceedings of 10th ISCA Speech Synthesis Workshop (SSW10)*. September 2019, pp. 132–136. DOI: 10.21437/SSW.2019-24.

Contents

List of Figures

List of Tables

List of Abbreviations

1D one-dimensional. 3, 5, 10, 12, 83

3D three-dimensional. ii, 3–5, 11, 12, 14, 82, 83, 86

aHM adaptive Harmonic Model. 16, 82, 85

ARX Auto-Regressive eXogenous. 12

DCGAN Deep Convolutions Generative Adversarial Networks. 8

E2E End-to-End. i, 4, 8, 9, 84, 87

FEM Finite Element Method. i, ii, 4, 5, 11, 14, 15, 82–85, 87

FOF Forme d'Onde Formatique. 6

GAN Generative Adversarial Network. 8

GTM Grup de recerca en Tecnologies Mèdia. 1, 3

HCI Human Computer Interaction. i, 2, 3

HFE High Frequency Energy. ii, 10, 11, 82, 83

HMM Hidden Markov Model. 7, 9

HNM Harmonic plus Noise Model. 9, 15, 81, 85

LF Liljencrants-Fant. ii, 12, 14–16, 82, 83, 86

LTAS Long Term Average Spectra. 82

MBROLA Multi-Band Resynthesis OverLap Add. 6

MRI Magnetic Resonance Imaging. i, 3, 4, 11, 82, 86

MUSSE Music and Singing Synthesis Equipment. 6

SPSS Statistical Parametric Speech Synthesis. 7

STS Speech-to-Singing. i, 10, 14, 15, 81, 87

TTS Text-to-Speech. i, xi, 2–6, 8–10, 13–16, 81, 84, 85, 87

US Unit-Selection. i, ii, xi, 4, 5, 7, 9, 13–16, 81, 82, 84, 85, 87

US-TTS&S Unit Selection based Text-to-Speech-and-Singing. i, ii, 14, 15, 81, 85, 87

VoQ Voice Quality. 82, 84, 85, 87

Chapter 1

Introduction

This thesis has been developed under the doctoral program in "Information Technologies and its Application in Management, Architecture and Geophysics" of La Salle–Universitat Ramon Llull (LS-URL). It has been carried out within the Grup de recerca en Tecnologies Mèdia (GTM) of LS-URL under the supervision of Dr. Francesc Alías and Dr. Joan Claudi Socoró.

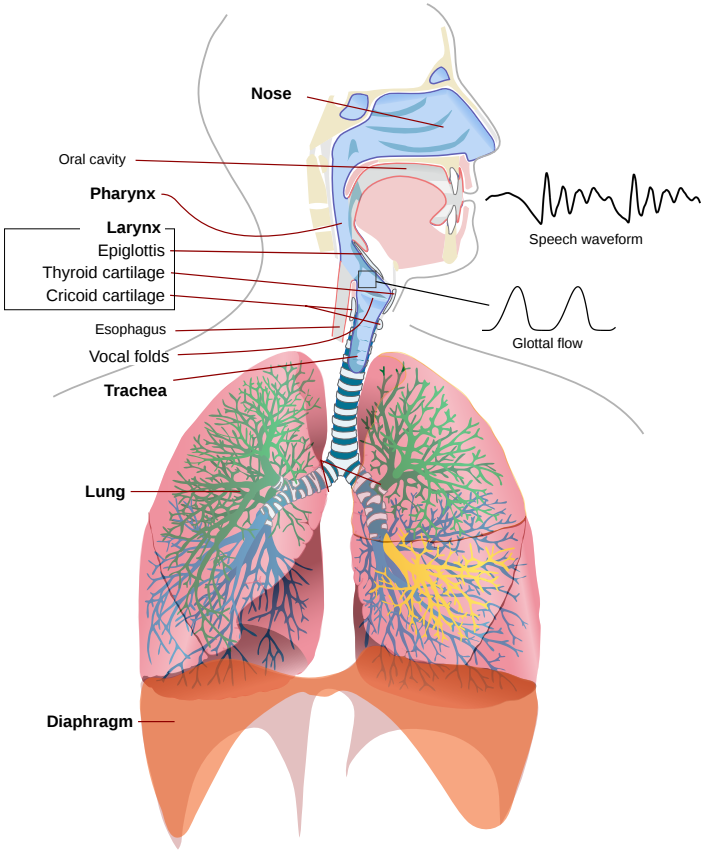


Figure 1.1: Phonatory system.

1.1 Motivation and objectives

This section briefly introduces the ultimate motivation behind this thesis: the synthesis of natural expressive speech (subsection 1.1.1). Then, it gives a short description of the work done by the GTM research group in this field (subsection 1.1.2). Finally, the objectives addressed in this thesis are outlined (subsection 1.1.3).

1.1.1 Expressive speech synthesis

Speech is an incredibly powerful mean of communication, which not only codifies linguistic information, i.e., the message, but also provides paralinguistic cues about the emotional state of the speaker, its mood or its intention (Schuller et al. 2013). This makes speech one of the most natural and direct forms of communication between human beings. Hence, it has become a key element in the development of Human Computer Interaction (HCI), by incorporating speech recognition and speech synthesis for the input and output channels of interaction, respectively. The first steps towards the integration of speech synthesis into HCI systems were mainly done in the field of assistive technologies to make computer systems accessible to people with special needs (e.g., Brodwin et al. 2004; Yamagishi et al. 2012). Nevertheless, nowadays, synthetic speech has become ubiquitous and it is present in our daily lives in smartphones and other devices through applications such as personal assistants (Barcelos Silva et al. 2020).

To better understand how speech can be generated, let us see how human speech is produced by the phonatory system (see Figure 1.1). When the airflow from the lungs cross the vocal folds, these vibrate producing a train of glottal pulses. This signal, known as glottal flow, is modified along its propagation path through the vocal tract and emanates from the mouth as a speech wave. The vocal tract shape, which is controlled by the articulators (e.g., tongue, palade, etc.), determines its acoustic behaviour and therefore which phoneme is produced. Speech can be synthesised by imitating this process. Nevertheless, since closely mimicking the human system is very complex and computationally demanding (Taylor 2009), simplifications are generally applied (e.g., Story et al. 1996; Birkholz 2013). In this respect, the classical acoustic theory of voice production proposes the source-filter model, where speech is obtained as the combination of a sound source, such as the glottal flow produced by the vocal folds, and the vocal tract, which can be modelled as a linear acoustic filter (Fant 1970).

Alternatively, several synthesis techniques have focused on reproducing the spoken output instead of the voice production process itself. To this end, a speaker is recorded to build a speech corpus. Most of nowadays speech synthesis systems follow this strategy and therefore are known as corpus-based approaches. Within this category, the most popular systems are those based on unit-selection, on statistical parametric models, and more recently on deep learning (see section 1.2).

Corpus-based approaches can produce intelligible and quite natural speech. However, as mentioned above, speech does not only convey linguistic information but also paralinguistic. Expressiveness is therefore a crucial component of speech communication and, as such, must be considered in speech synthesis systems. In this sense, the expressiveness achieved by corpus-based systems is mainly restricted to the style of the recorded corpus (Taylor 2009). Therefore, the expressive capabilities of a general-purpose Text-to-Speech (TTS) system could be expanded by recording new corpora covering the desired speaking styles (Alías et al. 2008). This approach, however, would be very costly and difficult to scale up, given the difficulty of gathering data to cover the wide range of expressive registers present in human communication.

Although corpus-based approaches have dominated the field of speech synthesis during the last decades (especially in the development of TTS systems), the increase in computing power has allowed the development of new methods of voice production. This is the case, for example, of numerical acoustic simulations considering three-dimensional (3D) vocal tract geometries to overcome restrictions of classic 1D simplifications (e.g., Blandin et al. 2015; Arnela et al. 2016b). These simulations, which try to characterise the production of human voice (see Figure 1.1) from an articulatory and acoustic point of view, have been mainly focused on improving the modelling of the voice generation process. As a consequence, little attention has been paid on providing this approaches with expressiveness. The acquisition of glottal source and vocal tract data using technologies such as Magnetic Resonance Imaging (MRI), electromagnetic articulograph or high-speed videoendoscopy is very costly (Y. Li et al. 2018), and it entails subsequent time-consuming postprocessing stages (Arnela et al. 2016b). Therefore, the collection of expressive data for numerical simulations may be even more difficult than building speech corpora. In this context, it might be interesting to consider the incorporation of glottal source processing techniques into the numerical simulations as an alternative means of adding expressiveness to the generated voice.

1.1.2 Research group

The Grup de recerca en Tecnologies Mèdia (GTM) has long experience in the analysis and synthesis of speech. This is fundamental in the development of natural HCI, which can help people with hearing and/or visual impairments (e.g., see the INREDIS project, CEN-2007-2011), and to interact with people with special needs (e.g., see the IntegraTV-4all project, FIT-350301-2004-2). A special focus has been placed on incorporating expressiveness into corpus-based speech synthesis through classic speech signal processing techniques in different projects (e.g., CreaVeu 2010-VALOR-00164, SALERO, FP6-027122 and SAVE, TEC2006-08043/TCM). On the other hand, the GTM has also been working on the numerical simulation of the physics involved in voice production (e.g., see the EUNISON project, EU-FET 308874). With the increase of computational power, it has become possible to simulate the propagation of acoustic waves through 3D geometries of the vocal tract obtained from MRI. Up to now, this method has been used to generate vowels, diphtongs and some vowel-consonant-vowel utterances (e.g., Arnela et al. 2016b; Arnela et al. 2019; Arnela and Guasch 2019). However, the voice produced with this approach is still limited in terms of expressiveness. This was precisely the motivation behind the GENIOVOX project (TEC2016-81107-P), within which part of the research presented in this thesis was developed. The GENIOVOX project aimed at the computational generation of expressive voice from the parameterisation of recorded expressive speech. To this end, a hybrid approach was proposed by incorporating speech processing techniques into numerical voice production. The key idea of this approach is to identify those parameters responsible for expressive effects through the analysis of speech signals to subsequently map the variations of such parameters into the glottal pulse models and vocal tract geometries used in the simulations.

1.1.3 Thesis objectives

The aim of the thesis is to add expressiveness into a system that generates neutral voice, without having to acquire expressive data from the original speaker. To this end, the system is extended with additional modules and/or controls which allow for speech signal transformations to approach the desired expressive styles, as depicted in Figure 1.2.

On the one hand, the thesis departs from a speech synthesiser developed in our research group. This is a Unit-Selection TTS (US-TTS) system fed with a small-sized neutral speech corpus (see Formiga et al. 2010 for further details). We want to extend the expressive capabilities of this system to address specific and timely needs in the storytelling domain. The storytelling style poses particular challenges for the generation of natural synthetic speech, such as the subtle expressive nuances in indirect speech, which differ between different storytelling categories (Montaño and Alías 2016; Montaño and Alías 2017). Another need that may occasionally arise is to generate singing when one of the characters of the story sings a song.

As mentioned above, adding expressive capabilities to a corpus-based TTS system by recording additional corpora would be very costly if we want to cover each and every expressive nuance we want to elicit (e.g., happiness, sadness, suspense, surprise,...). Therefore, applying this strategy to address sporadic expressive needs, like in the storytelling domain, would be difficult to justify. Furthermore, the speaker who was recorded to build the neutral speech corpus may not be a good storyteller or singer. In this respect, this thesis addresses the following research question: *is it possible to synthesise expressive speech or singing in the storytelling domain from a small-size neutral speech corpus through speech processing achieving a reasonable quality*? It may be observed that some End-to-End (E2E) approaches outperforming US-based systems have appeared during the course of this research. However, such approaches require large amounts of data and might not be the most appropriate to address the research question. Moreover, the classical US pipeline allows us not only to incorporate additional modules into the TTS but also to study their contribution.

In a similar vein, we aim at introducing expressiveness into the numerical simulation of voice. In this case, the starting point of our research is a 3D acoustic model based on the Finite Element Method (FEM), which simulates the propagation of acoustic waves through MRI-based vocal tract geometries developed within the EUNISON and GENIOVOX projects (Arnela et al. 2016b). As explained above, the acquisition and postprocessing of data for numerical simulations of expressive voice may be even more difficult than recording

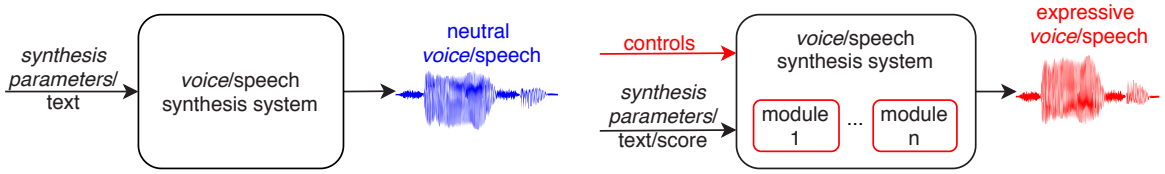


Figure 1.2: Expressive capabilities are added to a neutral voice/speech synthesis system by incorporating additional controls and modules.

speech corpora. That is why we ask ourselves: *is it possible to incorporate expressiveness in the numerical simulation of voice without explicitly acquiring expressive data?*

In accordance with the aforementioned research questions, the investigations conducted in this thesis pursue the following two main objectives:

- O1. Adding expressive capabilities to a US-TTS system fed with a small size neutral corpus to address the synthesis of singing and storytelling suspense, without recording additional voice samples.
- O2. Adding expressiveness to the 3D FEM-based numerical simulation of voice without having to acquire expressive data.

1.2 State of the art

This section introduces the state of the art related to the objectives addressed in this thesis. First, a brief review of the evolution of singing and speech synthesis is presented in subsection 1.2.1. Then, subsection 1.2.2 outlines some works focused on adding expressiveness to corpus-based TTS systems. Finally, subsection 1.2.3 presents investigations related to the goal of adding expressiveness to numerical voice production.

1.2.1 Singing and speech synthesis evolution

1.2.1.1 First generation synthesis systems

Synthesis of both speech and singing has attracted the attention of the speech research community. For instance, in the presentation of Dudley's VODER (Dudley 1939) apart from spoken utterances some short singing examples were performed by the operator to show the capabilities of the vocoder. Indeed, most of the early works facing singing synthesis were closely linked to speech synthesis (see Cook 1996 and references therein). This was particularly true in first generation synthesis systems (see left part of Figure 1.3). These synthesisers were typically built on a source-filter model according to the classical acoustic theory of voice production (Fant 1970). This model was driven by a rule-based control according to a detailed low-level synthesis specification, including a phonetic representation along with the duration of each phoneme as well as a $F0$ contour (Taylor 2009). These systems can be classified into three categories: articulatory synthesis, formant synthesis and classical linear prediction systems.

In articulatory synthesis, voice is generated by modelling the human articulator behaviour. Kelly and Lochbaum developed the first digital physical model of the voice where the vocal tract was simulated as a series of one-dimensional tubes (Kelly and Lochbaum 1962). Their collaboration with Max Mathews resulted in one of the first synthetic singing examples [1]. Afterwards this model was extended by means of digital waveguide synthesis in the SPASM/Singer system (Cook 1993), which could be used for both text-to-speech and singing synthesis purposes through control files (Cook et al. 1993).

[1] https://ccrma.stanford.edu/~jos/wav/daisy-klm.wav

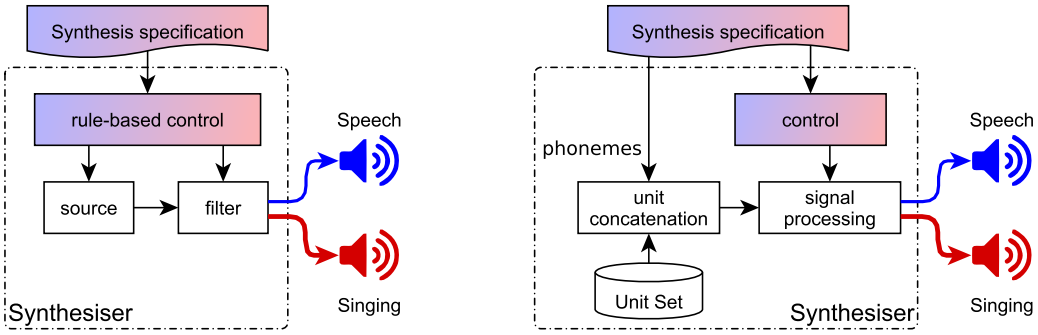


Figure 1.3: First generation (left) and second generation (right) synthesis systems.

Regarding formant synthesis approaches, an impulse train for voiced sounds and noise for unvoiced are passed through a bank of filters reproducing the resonances (formants) of the vocal tract transfer function. One of the most popular examples is the Klatt synthesiser (Klatt 1980), used by Charles Dodge in 1973 to produce singing synthesis. Inspired by Gunnar Fant's speech synthesiser (Fant 1970), the KTH developed the Music and Singing Synthesis Equipment (MUSSE) and the subsequent MUSSE DIG (Sundberg 2006). In a similar vein, the CHANT project (Rodet et al. 1984) used formant wave functions (in French, *Forme d'Onde Formatique* or FOF), so each formant is represented by its impulse response, and is excited by a pseudo-periodic controlling source. Very good singing results could be achieved but a costly process of refinement of the control parameters was required. The first generation of rule-based synthesis systems evolved to data-driven approaches due to the difficulty of determining the control parameters to get natural and high quality results. Nonetheless, it is worth mentioning that formant synthesis is still used in the context of performative singing synthesis (Feugère et al. 2017), where flexibility and real-time are the main requirements.

1.2.1.2 Second generation synthesis systems

Second generation synthesis systems (see right part of Figure 1.3) use a data driven approach to generate the verbal content of the signal and explicit algorithms to determine the prosodic content. To this end, a set of units (typically diphones) is recorded guaranteeing that there is one unit for each unique type. Pitch and timing of diphones are modified applying signal processing techniques to match the synthesis specification which includes the verbal specification, one or more $F0$ values and duration (Moulines and Charpentier 1990). Some works have exploited signal processing capabilities to generate singing from a spoken database. Flinger (Flinger 2001) for instance uses residual LPC synthesis and provides several modules allowing the Festival TTS (Festival 2016) to sing. Multi-Band Resynthesis OverLap Add (MBROLA, Dutoit and Leich 1993) has been used to generate both speech and singing from speech units (Uneson 2002) even in real-time as in MAXMBROLA (D'Alessandro et al. 2005). More recently, Ramcess synthesiser (D'Alessandro et al. 2008) generates singing by convolving vocal tract impulse responses from a database with an interactive model of the glottal source. The manipulation of both excitation and filter parameters allows to resemble singing even the database content is somewhere in between speech and singing.

The deployment of the data-driven paradigm led to the creation of databases explicitly designed for singing. In this context, sinusoidal models, that had been extensively used in speech synthesis (McAulay and Quatieri 1986) and speech modification (Quatieri and McAulay 1992), were extended to the synthesis of singing voice in the LYRICOS system (Macon et al. 1997) using a singing voice dataset. A bigger database is used in (Bonada and Serra 2007) trying to better cover the sonic space of a singer by means of performance sampling and spectral models. This approach is able to produce good quality synthetic singing and has successfully resulted in the Vocaloid (Kenmochi 2012).

1.2.1.3 Corpus-based synthesis systems

Speech and singing synthesis systems evolved towards corpus-based approaches, also known as third generation techniques (Taylor 2009). These techniques build on larger speech (or singing) corpora and they can be based on unit selection (US), statistical parametric models, and more recently deep learning.

In unit selection synthesis, units (usually diphones) are selected from a speech database according to their matching to a specification and how well they can be concatenated (Hunt and Black 1996; Clark et al. 2007; Alías et al. 2011). For a long time, US has been widely used in commercial speech synthesis systems because of the high degree of naturalness that can be achieved (King 2014). The US based approach was early adopted for the synthesis of singing in (Meron 1999). More recently, a singing synthesiser combining US and a Wideband Harmonic model was proposed in (Bonada et al. 2016). In this system, units from a database of expressive singing vowels were concatenated to obtain an expressive vowel performance. This performance together with the lyrics was introduced as input control of the synthesis, which used a timbre database of monotonic singing of a set of sentences.

Statistical Parametric Speech Synthesis (SPSS) pursued flexibility by training a model of speech (Taylor 2009). Speech is represented by vocal tract parameters and excitation parameters using a simplified speech production model known as vocoder. The parameters corresponding to phoneme sequences and linguistic specification context are modelled using a time-series stochastic generative model, being Hidden Markov Model (HMM) the most popular approach to build such models (Zen et al. 2007). HMM-based synthesis have also been applied to the generation of singing (Nose et al. 2015), resulting in systems like Sinsy (Oura and Mase 2010). The main advantage of statistical parametric speech synthesis is the flexibility in changing voice characteristics, speaking styles or emotions (Tokuda et al. 2013). Nevertheless, the naturalness achieved by this approach is limited mainly by vocoding artifacts and oversmoothing of the generated acoustic parameters (King 2014).

Deep learning was firstly introduced into speech synthesis with the aim of replacing HMM-based acoustic models (Zen et al. 2013). Nevertheless, the release of WaveNet (Oord et al. 2016) demonstrated the ability of deep learning autoregressive models to directly handle the generation of raw waveforms, outperforming the naturalness achieved by SPPS systems or even US-based synthesis. Moreover, Wavenet can be conditioned by acoustic parameters to be used as a vocoder. However, their high computational requirements have motivated the development

of other neural vocoders such as SampleRNN (Mehri et al. 2016), FFTNet (Jin et al. 2018), LPCNet (Valin and Skoglund 2019) or WaveGlow (Prenger et al. 2019). Autoregressive models were also applied to singing synthesis by (Blaauw and Bonada 2017), who proposed a model for singing synthesis based on a modified version of WaveNet architecture. The proposal included a parametric vocoder to separate the contribution of pitch and timbre, thus allowing for the pitch modification and training with smaller datasets.

Deep learning has been used to replace not only the acoustic model or the vocoder, but every component of a classical TTS pipeline, as is the case in DeepVoice (Arik et al. 2017). Furthermore, some approaches have abandoned the typical TTS pipeline by proposing E2E architectures like Tacotron (Wang et al. 2017), which use sequence-to-sequence models and attention mechanisms, thus avoiding the need for pre-aligned data. In this respect, (Blaauw and Bonada 2020) presented a sequence-to-sequence singing synthesiser, where a simple duration model yields an initial alignment that is subsequently refined by a decoder based on a feed-forward variant of the Transformer model to obtain the target acoustic features. Waveform generation in E2E synthesis systems can be tackled by neural vocoders. Nevertheless, Generative Adversarial Networks (GANs) have been successfully applied to this end for the synthesis of speech (Kumar et al. 2019). GANs have been used also in singing synthesis as in (Chandna et al. 2019), which proposed a multi-singer singing synthesiser inspired by the Deep Convolutions Generative Adversarial Networks (DCGAN) architecture and optimised by means of the Wasserstein-GAN algorithm (Martin Arjovsky and Bottou 2017). Vocoder features are used to separate the influence of pitch and timbre. Linguistic and $F0$ features, together with global singer identity are trained in a block-wise approach. Synthesis is performed using overlap-and-add to concatenate inferred sequential blocks. In (Lee et al. 2019), a Korean E2E singing voice synthesis system was proposed, where a GAN converts the input information into a Mel-spectrogram that is subsequently upsampled into a linear-spectrogram by a super-resolution network.

1.2.2 Adding expressiveness to a corpus-based TTS

Corpus-based TTS systems can achieve quite natural synthetic speech if they are asked for contents that are well represented in the corpus. However, there is still room for improvement regarding expressiveness (Alías et al. 2008). Until the beginning of the 21st century, works on analysis and synthesis of expressive speech were primarily focused on emotions either in a dimensional space of arousal and valence or considering a small number of discrete emotions such as "the big six" (see Scherer 2003; Schröder 2001, and references therein). From that time on, interest in other expressive speaking styles has been growing, especially through corpus-based approaches (see, for example, Schröder 2009). Nevertheless, building ad-hoc corpora for each desired expressive style (e.g., Iriondo et al. 2007; Alías et al. 2008) can be extremely costly or even unfeasible if the speaker is unable to properly perform all the required styles. To address this issue, several works have presented alternative methods to generate expressive synthetic speech.

Some studies proposed to transform neutral speech into expressive synthetic speech according to a set of fixed acoustic rules (Theune et al. 2006; Zovato et al. 2004; Montaño et al. 2013). For the same purpose, adaptation techniques have been used in HMM-based synthesizers, thus allowing the interpolation between statistical models trained on different expressive corpora (Yamagishi and Kobayashi 2007). Hybrid approaches can also be found in the literature. In (Erro et al. 2010) a Harmonic plus Noise Model (HNM) was incorporated into a US-based conversion system for the generation of emotions from neutral speech. (Lorenzo-Trueba et al. 2015) proposed an emotion transplantation approach, where HMM-based models were modified through adaptation functions used as pseudo-rules. These two approaches require 6-30 min and 10 min of speech data per style, respectively. Although this amount of data is quite small it is non-negligible. Moreover, these works present other limitations, such as the need of parallel recordings in (Erro et al. 2010) or the over-smooth quality typical of speech generated with statistical based approaches (Barra-Chicote et al. 2010). In recent E2E approaches like Tacotron (Wang et al. 2017), prosody is implicitly learned from the training speech data. Nevertheless, prosody can be transferred to Tacotron by conditioning the system with prosody embeddings learned from expressive reference signals (Skerry-Ryan et al. 2018). In that work, a single-speaker database of 147h and a multi-speaker dataset of 296h were used.

Among the different expressive speaking styles, storytelling is especially challenging to model and synthesise because of its high variability and degree of expressiveness. It is not strange then that audiobooks have been used in several editions of the Blizzard Challenge[2]. Some works have directly used audiobooks containing stories to generate expressive synthetic speech through corpus-based strategies (see e.g., Jauk et al. 2015; Charfuelan and Steiner 2013; Prahallad and Black 2011). Although these approaches can deliver expressive speech with good quality on average, storytelling style has several subtle expressive nuances that require deeper analysis to meet the needs of storytelling applications (see e.g., Leite et al. 2015; Alofs et al. 2015). Specific prosodic characteristics have been studied in detail for the storytelling style (Theune et al. 2006; Doukhan et al. 2011; Montaño and Alías 2016). Some works have modelled the characteristics of particular types of storytelling to synthesise them from neutral speech. (Theune et al. 2006) defined a set of *fixed* prosodic rules for global storytelling style and suspense that were applied in a diphone-based TTS synthesiser, thereby obtaining a significant improvement of storytelling quality and higher suspense scores. It should be noted that since suspense is rarely found in stories the prosodic rules for this style were derived from very few sentences (e.g., only *one* sentence for increasing suspense and *two* sentences for sudden suspense). (Montaño et al. 2013) analysed speaking rate, mean pitch, pitch standard deviation and mean intensity of several sentences for different storytelling categories. Moreover, a hybrid US-HNM framework was considered to transform the prosody of neutral speech to the different categories according to mean values of each category. Although converted utterances were preferred over the neutral ones, the use of constant conversion factors was found insufficient to accurately capture subtle expressive nuances.

[2] https://www.synsig.org/index.php/Blizzard_Challenge

Besides entailing the generation of subtle expressive nuances, storytelling may also require the synthesis of singing if one of the characters begins to sing (Montaño and Alías 2016; Fridin 2014). Therefore, a TTS with singing capabilities could be very useful in storytelling, but also for other applications. For instance, in voice output communication aid devices for individuals with vocal disabilities (Yamagishi et al. 2012) to allow them not only to talk, but also to sing. It could be also incorporated in assistive technologies, where the use of songs can improve the engagement of autistic children (Wood et al. 2017), reduce the procedural distress in children with cancer (Jibb et al. 2018), or to augment the positive memories of people with dementia (Khosla et al. 2017), to name a few. In order to enable a corpus-based TTS system to sing, a supplementary singing database would be required, among other things. However, building an additional corpus would lead to high costs that would not be justified by eventual singing needs and it may become even unfeasible if the original speaker is unavailable or unable to sing properly (Blanco et al. 2016). Alternatively, singing could be generated from speech following the so-called Speech-to-Singing (STS) conversion approach (see e.g., Röbel and Fineberg 2007; Saitou et al. 2007; Dong et al. 2014). STS conversion can be applied to the output of a TTS system, thus transforming the synthetic speech into singing while maintaining the identity of the speaker (J. Li et al. 2011). However, this straightforward approach is suboptimal in terms of flexibility and computational costs (J. Li et al. 2011). It is worth mentioning that STS is the subject of active research, as evidenced by some recent works such as (Parekh et al. 2020), where a encoder-decoder framework is proposed to perform the STS.

This section has been built around some relevant works that have proposed alternative methods to generate expressive synthetic speech without having to record ad-hoc corpora for each desired expressive style. Specifically, since our goal is the synthesis of speech and singing for the storytelling domain, the focus has been placed on studies dealing with this style and on STS conversion.

1.2.3 Adding expressiveness to numerical voice production

Human voice production is a very complex mechanism. For this reason, the first generation of synthesis systems opted to consider simplified source-filter models inspired by the classical acoustic theory of voice production (Fant 1970). Thus, for many years, articulatory speech synthesis approaches have considered a one-dimensional (1D) representation of the vocal tract, the so-called vocal tract area function (see e.g., Story et al. 1996). This function, which describes the cross-sectional area variations along the vocal tract center midline, has been widely used to generate synthetic speech (see e.g., Story 2013; Birkholz 2013; Stone et al. 2018). However, 1D approaches can only reproduce the propagation of plane waves along the vocal tract midline, and therefore, their accuracy is limited up to about 4–5 kHz. Beyond this frequency, higher order modes also propagate, so it is important to model them if the aim is to adequately characterise the voice production process. These modes produce resonances and anti-resonances that cannot be obtained with 1D models and which strongly affects the High Frequency Energy (HFE) content of the spectrum (Blandin et al. 2015; Arnela et al. 2016b).

For years, requirements of storage and speed have prevented high sampling rates being used when dealing with speech signals. Moreover, it has been shown that listeners can discriminate sounds by only the first three formants (Taylor 2009). This has motivated that little attention has been traditionally paid to the high frequency range of speech. Nevertheless, HFE has been found important for speech localisation and speaker recognition, and in the intelligibility and quality of the voice (see Monson et al. 2014 and references therein).

The increase of computational resources has allowed for the development of three-dimensional (3D) acoustic models, which can handle with higher order modes propagation. To date, 3D-based approaches have been applied to generate vowels (Arnela et al. 2016b), diphtongs (Arnela et al. 2019) and some vowel-consonant-vowel sequences (Arnela and Guasch 2019). For instance, in (Vampola et al. 2008) FEM was used to simulate the production of Czech vowels using 3D vocal tracts, obtained from MRI data. Similarly, in (Takemoto et al. 2010), the MRI-based vocal tracts of Japanese vowels were analysed using a finite-difference time-domain method. Moreover, the results of the simulations were contrasted with experiments conducted with physical models built from the same MRI data. In a similar vein, in (Arnela et al. 2016a), measurements on 3D-printed mechanical replicas yielded very similar results to those obtained from 3D FEM acoustic simulations on MRI-based vocal tract geometries. Nevertheless, the use of a 3D acoustic model does not necessarily imply that higher order modes propagate. For instance, these modes will rarely appear in a straight and axisymmetric vocal tract geometry excited at the glottis, as observed in (Arnela et al. 2016b). Indeed, several geometric simplifications were analysed in (Arnela et al. 2016b). Specifically, their cross-sectional shapes and midline curvature were varied while preserving their cross-sectional areas. Results obtained for the analysed configurations were similar in frequencies below 4–5 kHz but very large deviations appeared beyond that value. This highlights the limit of the plane wave assumption and also show that variations of the vocal tract shape modify the HFE content. Nonetheless, the HFE levels strongly depend on other factors such as phonation type. For instance, in (Monson et al. 2011) significant differences in HFE content were observed between loud and soft phonations of sustained vowels. Moreover, results also showed that modifications of HFE levels are more easily detected by listeners in a loud phonation case.

Despite the progress made in this area, to the authors' knowledge, expressiveness has not been central in the research conducted in this field. The generation of expressive voice would require to properly model the two key elements in the voice production approximation: the source and the filter. Glottal source and vocal tract characteristics of expressive voice production can be derived from experimental measurements using technologies such as MRI, electromagnetic articulograph or high-speed videoendoscopy (Y. Li et al. 2018). However, data acquisition with this technologies is very costly (Y. Li et al. 2018). Furthermore, they typically require of subsequent data processing, like that needed to obtain a fine 3D geometry of the vocal tract from MRI (Arnela et al. 2016b), which implies time-consuming manual tuning.

Alternatively, emotional speech characterisation can be done through the extraction of several acoustic features from the speech signal (Eyben et al. 2016). In this respect, some works propose an analysis inspired by the acoustic model of voice production. To that effect, they inverse filter the speech waveform to estimate the glottal flow and subsequently parameterise it using attributes such as the normalised amplitude quotient or the maximum flow declination rate. This strategy was applied in (Sundberg et al. 2011) to analyse the interdependencies between voice source parameters on sustained vowels uttered with different emotions. A similar approach was followed in (Waaramaa et al. 2010) to study the role of voice source and formant frequencies on the perception of emotional valence. The recent progress in inverse filtering and glottal source processing techniques (see Drugman et al. 2014 and references therein) have enabled the development of glottal vocoders. These vocoders decompose the speech signal into glottal source and vocal tract, which are independently parameterised (Drugman et al. 2014). Moreover, the features obtained from these vocoders have been proven effective in the analysis of expressive nuances (Lorenzo-Trueba et al. 2012). Although the aforementioned works have shed light on the production of emotional voice, they have not been explicitly focused on the generation of emotional speech.

Several works have proposed to bridge the gap between the analysis and synthesis of emotional speech through source-filter based approaches. For instance, (Birkholz et al. 2015) analysed the contribution of phonation types to the perception of emotions. To that effect, a set of utterances was resynthesised with different phonation types using an articulatory-based synthesiser, which incorporates a self-oscillation model of the vocal folds. With the same aim, (Burkhardt 2009) considered a formant-based synthesiser with a modified Liljencrants-Fant (LF) glottal flow model (Fant et al. 1985). This synthesiser was also used in (Yanushevskaya et al. 2018) to study how $F0$ contours and voice quality maps on affect for different languages, by varying the parameters of modal stimuli (e.g., amplitude of voicing, open quotient, spectral tilt, etc.). The LF model was controlled in (Murphy et al. 2017) by modifying the R_d glottal shape parameter (Fant 1995) to simulate the tense-lax continuum and explore its influence to emotion perception. Similarly, (Y. Li et al. 2018) proposed an Auto-Regressive eXogenous LF (ARX-LF) model to analyse the contribution of glottal source and vocal tract to the perception of emotions in a valence-arousal space. Nevertheless, the study only considered isolated vowels, and suffered from the considered prosody *neutralisation* process.

This section covers two main topics related to the goal of adding expressiveness to numerical voice production. On the one hand, the development of 3D acoustic models that, unlike 1D models, can handle with higher order modes propagation. On the other hand, the analysis of expressive speech, especially those studies inspired on the source-filter model and that therefore are closest to voice production methods.

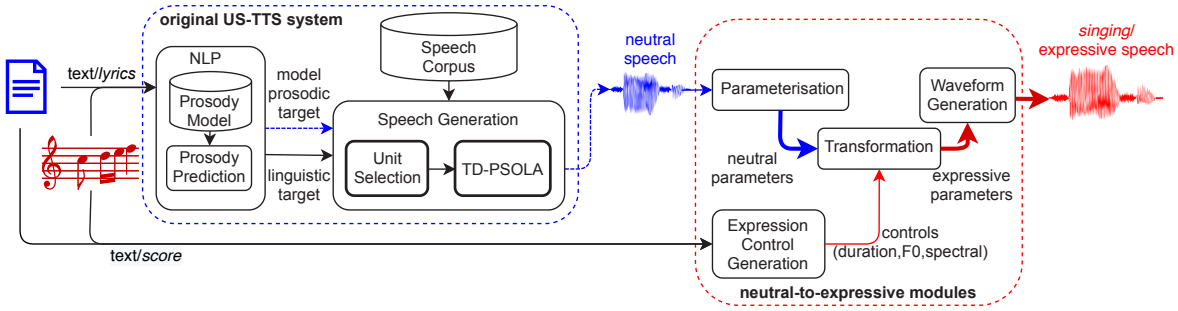
1.3 Contributions

This thesis, presented in the form of a collection of articles, addresses the objectives described in subsection 1.1.3. Besides the three publications which form the compendium (Paper I,

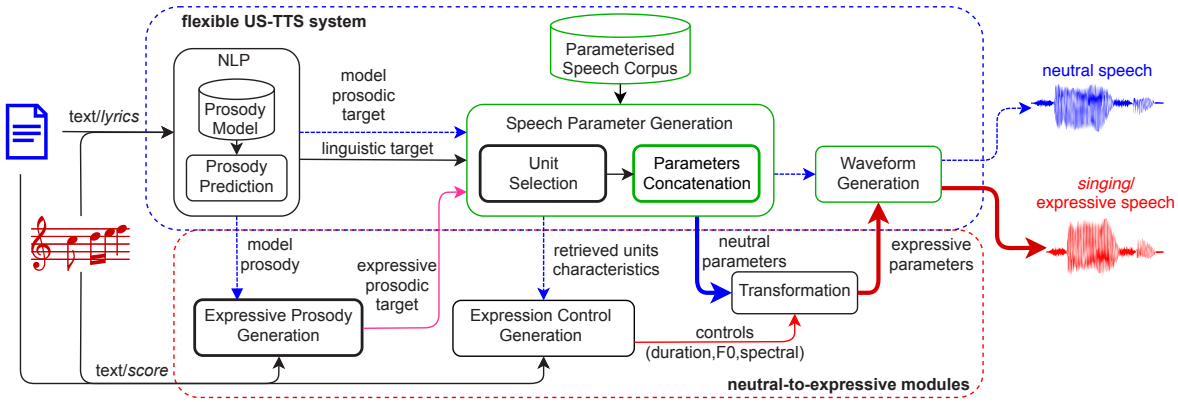
Paper II and Paper III), three complementary articles are also included as an appendix (Paper IV, Paper V and Paper VI). This section overviews the main contributions of these works.

1.3.1 Adding expressiveness to a unit-selection TTS system

The starting point for the objective O1 (see subsection 1.1.3) is a classical US-TTS system fed with a small corpus of neutral speech (see left part of Figure 1.4a). A preliminary approach to add expressive capabilities to that system consists of appending neutral-to-expressive modules at the output of the TTS system as depicted in Figure 1.4b. The synthetic neutral speech yield by the TTS system is parameterised and subsequently modified to obtain the synthetic expressive speech. This transformation is driven by expression controls derived from the input text/score according to an expert system that captures the desired expressive style. A second approach that pursues a higher degree of integration is depicted in Figure 1.4b. Corpus utterances are not directly stored as waveforms but as parameters. In this way, units are selected according to a raw input text specification as in a standard US-TTS system, but instead of retrieving waveforms a sequence of parameters is obtained. These parameters are concatenated and transformed to get the synthetic singing/expressive speech.



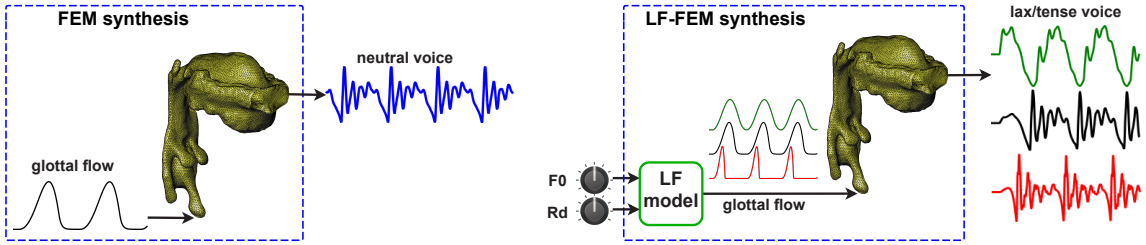
(a) The neutral-to-expressive transformation modules are appended to the original US-TTS system output.



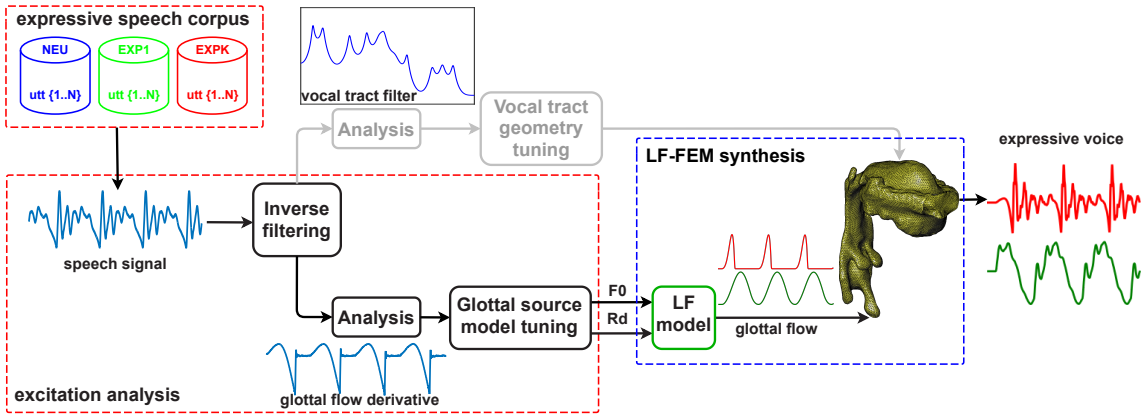
(b) The neutral-to-expressive transformation modules are integrated into the TTS pipeline. The US-TTS system has been adapted (modified/new modules in green) to work with parameterised speech units that can be transformed before the waveform generation. Moreover, an expressive prosody generation module has been also added to select units according to the expressive prosodic target.

Figure 1.4: Extending the expressive capabilities of a US-TTS system.

Paper I aims at adding singing capabilities to a neutral US-TTS system through the implementation of both the approach depicted in Figure 1.4a and a preliminary integration of STS transformation modules into the TTS pipeline. The lyrics of the score are used as input of the TTS. Duration, $F0$ and spectral controls are generated from the input score following a STS approach. Paper II presents the Unit Selection based Text-to-Speech-and-Singing (US-TTS&S) synthesis framework designed to generate both speech and singing using the same neutral speech corpus, following the approach depicted in Figure 1.4b. In this framework, STS integration goes a step further by incorporating an expressive prosody generation module so the prosody from the score can be considered in the unit selection process as shown in Figure 1.4b. Finally, Paper IV presents a complementary research in the storytelling domain, in particular it addresses the synthesis of storytelling increasing suspense by means of an expert system derived from a small but representative set of sentences of this style.



(a) An LF model is incorporated in order to generate the glottal flow signals required as the excitation of the FEM synthesis. This allows for exploration of the lax-tense continuum by modifying the R_d parameter.



(b) The excitation characteristics of expressive speech are analysed by means of inverse filtering to derive the LF parameters to incorporate these expressive styles into the LF-FEM synthesis.

Figure 1.5: Adding expressiveness to the numerical simulation of voice.

1.3.2 Adding expressiveness to numerical voice production

The research conducted under objective O2 is based on the numerical voice production system represented in the left part of Figure 1.5a, which generates neutral voice by means of a FEM-based simulation of a glottal flow signal propagating through a 3D vocal tract geometry. In order to add expressiveness to these simulations, we propose to modify the glottal flow signals by means of an LF model as shown in the right part of Figure 1.5a. This model allows

controlling the $F0$ but also the shape of the glottal pulses by means of the R_d parameter, which permits the reproduction of the lax-tense continuum of phonation (see glottal flow illustrative examples in green, black and red in Figure 1.5a). Paper V presents a preliminary analysis of the influence of tense, modal and lax phonation on the synthesis of vowel [ɑ], focusing on the effect on the higher order modes. To this end, an LF model is implemented and incorporated in the FEM-based simulation. In Paper III the analysis is extended to vowels [i] and [u] and covering the spoken vocal range of $F0$ and the complete R_d range from lax to tense. Moreover, aspiration noise is incorporated into the LF model to analyse its contribution to higher order modes.

Figure 1.5b depicts the approach proposed to add specific expressive styles to the simulations. Expressive vowels from a parallel corpus are inverse filtered and analysed following a source-filter based strategy. According to this analysis on real speech data the $F0$ and R_d parameters of the LF model are tuned to resemble the desired expressive style. This approach is applied to the generation of happy and aggressive vowels [a] in Paper VI. To that effect, the GlottDNN vocoder is used to analyse $F0$ and spectral tilt variations associated with the glottal source, which are mapped through the comparison with synthetic vowels to $F0$ and R_d parameters to perform the LF-FEM synthesis of vowels resembling the tense voice expressive styles. The analysis of the vocal tract response and the corresponding tuning of the vocal tract geometry is left for future works.

1.3.3 Author's contribution

Indexed articles

Paper I: Marc Freixes, Joan Claudi Socoró, Francesc Alías. 'Adding Singing Capabilities to Unit Selection TTS through HNM-Based Conversion'. In: *Advances in Speech and Language Technologies for Iberian Languages. IberSPEECH 2016. Lecture Notes in Computer Science.* volume 10077, pp. 33–43. DOI: 10.1007/978-3-319-49169-1_4.
Author's contributions: Marc Freixes significantly contributed in the design and implementation of the STS conversion and its integration with the US-TTS system. Moreover, besides leading the manuscript writing, his contribution to the generation and analysis of perceptual tests and results was particularly relevant.

Paper II: Marc Freixes, Joan Claudi Socoró, Francesc Alías. 'A Unit Selection Text-to-Speech-and-Singing Synthesis Framework from Neutral Speech: Proof of concept'. In: *EURASIP Journal on Audio, Speech and Music Processing.* December 2019, volume 2019, article number 22. DOI: 10.1186/s13636-019-0163-y.
Author's contributions: Marc Freixes had a significant participation in the design of the US-TTS&S framework besides leading the generation and analysis of results, the perceptual tests conduction and the manuscript writing. His work on the US-TTS&S framework, included the implementation of a new HNM model and the expressive prosody generation module, and the adaptation of the original US-TTS system to integrate all the components within it.

Paper III: Marc Freixes, Marc Arnela, Joan Claudi Socoró, Francesc Alías, Oriol Guasch. 'Glottal Source Contribution to Higher Order Modes in the Finite Element Synthesis of Vowels'. In: *Applied Sciences - Special Issue "IberSPEECH 2018: Speech and Language Technologies for Iberian Languages".* October 2019, volume 9(21), pp. 4535. DOI: 10.3390/app9214535.

Author's contributions: Marc Freixes significantly contributed in the design of the approach and leaded the manuscript writing and the generation, analysis and visualisation of results. He implemented and incorporated the aspiration noise model and the R_d control on the LF model to generate the glottal flow signals and synthesised the vowels according with the vocal tract responses obtained from the simulations performed by Marc Arnela and Oriol Guasch.

Other articles

Paper IV: Raul Montaño, Marc Freixes, Francesc Alías, Joan Claudi Socoró. 'Generating Storytelling Suspense from Neutral Speech using a Hybrid TTS Synthesis framework driven by a Rule-based Prosodic Model'. In: *Proceedings of IberSPEECH 2016.* November 2016, pp. 129–138.

Author's contributions: Marc Freixes significantly contributed in the design and implementation of the increasing suspense approach and in the conduction of the perceptual tests, besides participating in the manuscript writing. His contribution to the implementation of the expert system and its integration together with the aHM model within the US-TTS system was particularly relevant.

Paper V: Marc Freixes, Marc Arnela, Joan Claudi Socoró, Francesc Alías, Oriol Guasch. 'Influence of tense, modal and lax phonation on the three-dimensional finite element synthesis of vowel [ɑ]'. In: *Proceedings of IberSPEECH 2018.* November 2018, pp. 132–136.

Author's contributions: Marc Freixes had a significant participation in the design of the approach and leaded the manuscript writing and the generation and analysis of results. He worked on the LF model to generate the glottal flow signals and synthesised the vowels using the vocal tract responses obtained from the simulations performed by Marc Arnela and Oriol Guasch.

Paper VI: Marc Freixes, Marc Arnela, Francesc Alías, Joan Claudi Socoró. 'GlottDNN-based spectral tilt analysis of tense voice emotional styles for the expressive 3D numerical synthesis of vowel [a]'. In: *Proceedings of 10th ISCA Speech Synthesis Workshop (SSW10).* September 2019, pp. 132–136. DOI: 10.21437/SSW.2019-24.

Author's contributions: Marc Freixes significantly contributed in the design of the approach and leaded the manuscript writing and the generation and analysis of results. He generated the glottal flow signals and obtained the synthetic vowels using the vocal tract responses from the simulations performed by Marc Arnela and Oriol Guasch.

References

Alías, Francesc, Formiga, Lluís, and Llorá, Xavier (May 2011). "Efficient and reliable perceptual weight tuning for unit-selection text-to-speech synthesis based on active interactive genetic algorithms: A proof-of-concept". In: *Speech Communication* vol. 53, no. 5, pp. 786–800.

Alías, Francesc, Sevillano, Xavier, Socoró, Joan Claudi, and Gonzalvo, Xavier (2008). "Towards high quality next-generation Text-to-Speech synthesis: a Multidomain approach by automatic domain classification". In: *IEEE Transactions on Audio, Speech and Language Processing* vol. 16, no. 7, pp. 1340–1354.

Alofs, Thijs, Theune, Mariët, and Swartjes, Ivo (2015). "A tabletop interactive storytelling system: Designing for social interaction". In: *International Journal of Arts and Technology* vol. 8, no. 3, pp. 188–211.

Arik, Sercan O. et al. (Feb. 2017). "Deep Voice: Real-time Neural Text-to-Speech". In: arXiv: 1702.07825.

Arnela, Marc, Blandin, Rémi, Dabbaghchian, Saeed, Guasch, Oriol, Alías, Francesc, Pelorson, Xavier, Van Hirtum, Annemie, and Engwall, Olov (2016a). "Influence of lips on the production of vowels based on finite element simulations and experiments". In: *The Journal of the Acoustical Society of America* vol. 139, no. 5, pp. 2852–2859.

Arnela, Marc, Dabbaghchian, Saeed, Blandin, Rémi, Guasch, Oriol, Engwall, Olov, Van Hirtum, Annemie, and Pelorson, Xavier (2016b). "Influence of vocal tract geometry simplifications on the numerical simulation of vowel sounds". In: *The Journal of the Acoustical Society of America* vol. 140, no. 3, pp. 1707–1718.

Arnela, Marc, Dabbaghchian, Saeed, Guasch, Oriol, and Engwall, Olov (2019). "MRI-based vocal tract representations for the three-dimensional finite element synthesis of diphthongs". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* vol. 27, pp. 2173–2182.

Arnela, Marc and Guasch, Oriol (Sept. 2019). "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, pp. 1802–1809.

Barcelos Silva, Allan de, Gomes, Marcio Miguel, Costa, Cristiano André da, Rosa Righi, Rodrigo da, Barbosa, Jorge Luis Victoria, Pessin, Gustavo, De Doncker, Geert, and Federizzi, Gustavo (2020). "Intelligent personal assistants: A systematic literature review". In: *Expert Systems with Applications* vol. 147, p. 113193.

Barra-Chicote, Roberto, Yamagishi, Junichi, King, Simon, Montero, Juan M., and Macias-Guarasa, Javier (2010). "Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech". In: *Speech Communication* vol. 52, no. 5, pp. 394–404.

Birkholz, Peter (2013). "Modeling Consonant-Vowel Coarticulation for Articulatory Speech Synthesis". In: *PLoS ONE* vol. 8, no. 4, e60603.

Birkholz, Peter, Martin, Lucia, Willmes, Klaus, Kröger, Bernd J., and Neuschaefer-Rube, Christiane (2015). "The contribution of phonation type to the perception of vocal emotions

in German: An articulatory synthesis study". In: *The Journal of the Acoustical Society of America* vol. 137, no. 3, pp. 1503–1512.

Blaauw, Merlijn and Bonada, Jordi (2017). "A Neural Parametric Singing Synthesizer Modeling Timbre and Expression from Natural Songs". In: *Applied Sciences* vol. 7, no. 12, p. 1313.

Blaauw, Merlijn and Bonada, Jordi (May 2020). "Sequence-to-Sequence Singing Synthesis Using the Feed-Forward Transformer". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 7229–7233.

Blanco, Eder del, Hernaez, Inma, Navas, Eva, Sarasola, Xabier, and Erro, Daniel (Sept. 2016). "Bertsokantari: a TTS Based Singing Synthesis System". In: *Interspeech*. San Francisco, CA, USA: ISCA, pp. 1240–1244.

Blandin, Rémi, Arnela, Marc, Laboissière, Rafael, Pelorson, Xavier, Guasch, Oriol, Van Hirtum, Annemie, and Laval, Xavier (2015). "Effects of higher order propagation modes in vocal tract like geometries". In: *The Journal of the Acoustical Society of America* vol. 137, no. 2, pp. 832–8.

Bonada, Jordi and Serra, Xavier (Mar. 2007). "Synthesis of the Singing Voice by Performance Sampling and Spectral Models". In: *IEEE Signal Processing Magazine* vol. 24, no. 2, pp. 67–79.

Bonada, Jordi, Umbert, Martí, and Blaauw, Merlijn (2016). "Expressive Singing Synthesis Based on Unit Selection for the Singing Synthesis Challenge 2016". In: *Interspeech*, pp. 1230–1234.

Brodwin, Martin G., Star, Tristen, and Cardoso, Elizabeth (2004). "Computer assistive technology for people who have disabilities: Computer adaptations and modifications". In: *Journal of rehabilitation* vol. 70, no. 3, p. 28.

Burkhardt, Felix (Sept. 2009). "Rule-based voice quality variation with formant synthesis". In: *Interspeech*. Brighton, UK.

Chandna, Pritish, Blaauw, Merlijn, Bonada, Jordi, and Gomez, Emilia (Sept. 2019). "WGANSing: A Multi-Voice Singing Voice Synthesizer Based on the Wasserstein-GAN". In: *2019 27th European Signal Processing Conference (EUSIPCO)*. IEEE, pp. 1–5.

Charfuelan, Marcela and Steiner, Ingmar (2013). "Expressive speech synthesis in MARY TTS using audiobook data and EmotionML". In: *Interspeech*, pp. 1564–1568.

Clark, Robert A.J., Richmond, Korin, and King, Simon (Apr. 2007). "Multisyn: Open-domain unit selection for the Festival speech synthesis system". In: *Speech Communication* vol. 49, no. 4, pp. 317–330.

Cook, Perry R. (1993). "SPASM, a Real-Time Vocal Tract Physical Model Controller; and Singer, the Companion Software Synthesis System". In: *Computer Music Journal* vol. 17, no. 1, pp. 30–44.

Cook, Perry R. (1996). "Singing voice synthesis: History, current work, and future directions". In: *Computer Music Journal* vol. 20, no. 3, pp. 38–46.

Cook, Perry R., Kamarotos, Dimitris, Diamantopoulos, Taxiarchis, and Philippis, Giorgos (1993). "IGDIS (Instrument for Greek Diction and Singing): A Modern Greek Text to

Speech/Singing Program for the SPASM/Singer Instrument". In: *International Computer Music Conference (ICMC)*. Tokyo, Japan: ICMA, pp. 387–389.

D'Alessandro, Nicolas, Babacan, Onur, Bozkurt, Baris, Dubuisson, Thomas, Holzapfel, Andre, Kessous, Loic, Moinet, Alexis, and Vlieghe, Maxime (2008). "RAMCESS 2.X framework—expressive voice analysis for realtime and accurate synthesis of singing". In: *Journal on Multimodal User Interfaces* vol. 2, pp. 133–144.

D'Alessandro, Nicolas, Sebbe, Raphael, Bozkurt, Baris, and Dutoit, Thierry (Sept. 2005). "MaxMBROLA: A Max/MSP MBROLA-Based Tool for Real-Time Voice Synthesis". In: *13th European Signal Processing Conference (EUSIPCO)*. Antalya, Turkey: EURASIP.

Dong, Minghui, Lee, S. W., Li, Haizhou, Chan, Paul, Peng, Xuejian, Ehnes, Jochen Walter, and Huang, Dongyan (Sept. 2014). "I2R Speech2Singing Perfects Everyone's Singing". In: *Interspeech*. Singapore: ISCA, pp. 2148–2149.

Doukhan, David, Rilliard, Albert, Rosset, Sophie, Adda-Decker, Martine, and d'Alessandro, Christophe (2011). "Prosodic analysis of a corpus of tales". In: *Interspeech*, pp. 3129–3132.

Drugman, Thomas, Alku, Paavo, Alwan, Abeer, and Yegnanarayana, Bayya (2014). "Glottal source processing: From analysis to applications". In: *Computer Speech and Language* vol. 28, no. 5, pp. 1117–1138.

Dudley, Homer (Oct. 1939). "Remaking Speech". In: *The Journal of the Acoustical Society of America* vol. 11, no. 2, pp. 169–177.

Dutoit, Thierry and Leich, H. (1993). "MBR-PSOLA: Text-To-Speech synthesis based on an MBE re-synthesis of the segments database". In: *Speech Communication* vol. 13, no. 3-4, pp. 435–440.

Erro, Daniel, Navas, Eva, Hernáez, Inma, and Saratxaga, Ibon (2010). "Emotion conversion based on prosodic unit selection". In: *IEEE Transactions on Audio, Speech, and Language Processing* vol. 18, pp. 974–983.

Eyben, Florian et al. (2016). "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing". In: *IEEE Transactions on Affective Computing* vol. 7, no. 2, pp. 190–202.

Fant, Gunnar (1970). *Acoustic theory of speech production*. Ed. by Gruyter, Walter de. Mouton.

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

Fant, Gunnar, Liljencrants, Johan, and Lin, Qi-guang (1985). "A four-parameter model of glottal flow". In: *Speech Transmission Laboratory Quarterly Progress and Status Report (STL-QPSR)* vol. 26, no. 4, pp. 1–13.

Festival (2016). *The Festival speech synthesis system*.

Feugère, Lionel, D'Alessandro, Christophe, Doval, Boris, and Perrotin, Olivier (Dec. 2017). "Cantor Digitalis: chironomic parametric synthesis of singing". In: *EURASIP Journal on Audio, Speech, and Music Processing* vol. 2017, p. 2.

Flinger (2001). *Flinger: Festival Singer*.

Formiga, Lluís, Trilla, Alexandre, Alías, Francesc, Iriondo, Ignasi, and Socoró, Joan Claudi (Nov. 2010). "Adaptation of the URL-TTS system to the 2010 Albayzin evaluation campaign". In: *FALA 2010, Jornadas en Tecnología del Habla and Iberian SLTech Workshop*. Vigo, Spain: ISCA IL-SIG, pp. 363–370.

Fridin, Marina (Jan. 2014). "Storytelling by a kindergarten social assistive robot: A tool for constructive learning in preschool education". In: *Computers & Education* vol. 70, pp. 53–64.

Hunt, A.J. and Black, Alan W. (May 1996). "Unit selection in a concatenative speech synthesis system using a large speech database". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Vol. 1. Atlanta, GA, USA: IEEE, pp. 373–376.

Iriondo, Ignasi, Socoró, Joan Claudi, and Alías, Francesc (Apr. 2007). "Prosody Modelling of Spanish for Expressive Speech Synthesis". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Vol. 4, pp. 821–824.

Jauk, Igor, Bonafonte, Antonio, Lopez-Otero, Paula, and Docio-Fernandez, Laura (2015). "Creating Expressive Synthetic Voices by Unsupervised Clustering of Audiobooks". In: *Interspeech*, pp. 3380–3384.

Jibb, Lindsay A., Birnie, Kathryn A., Nathan, Paul C., Beran, Tanya N., Hum, Vanessa, Victor, J. Charles, and Stinson, Jennifer N. (Sept. 2018). "Using the MEDiPORT humanoid robot to reduce procedural pain and distress in children with cancer: A pilot randomized controlled trial". In: *Pediatric Blood & Cancer* vol. 65, no. 9, e27242.

Jin, Zeyu, Finkelstein, Adam, Mysore, Gautham J., and Lu, Jingwan (Apr. 2018). "FFTnet: A Real-Time Speaker-Dependent Neural Vocoder". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 2251–2255.

Kelly, John L. and Lochbaum, Carol C. (Aug. 1962). "Speech synthesis". In: *4th International Congress on Acoustics (ICA)*. Copenhagen, Denmark: ICA, G42.

Kenmochi, Hideki (Mar. 2012). "Singing synthesis as a new musical instrument". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Kyoto, Japan: IEEE, pp. 5385–5388.

Khosla, Rajiv, Nguyen, Khanh, and Chu, Mei-Tai (June 2017). "Human Robot Engagement and Acceptability in Residential Aged Care". In: *International Journal of Human–Computer Interaction* vol. 33, no. 6, pp. 510–522.

King, Simon (June 2014). "Measuring a decade of progress in Text-to-Speech". In: *Loquens* vol. 1, no. 1, e006.

Klatt, Dennis H (1980). "Software for a cascade/paralell formant synthesizer". In: *The Journal of the Acoustical Society of America* vol. 67, pp. 971–995.

Kumar, Kundan, Kumar, Rithesh, Boissiere, Thibault de, Gestin, Lucas, Teoh, Wei Zhen, Sotelo, Jose, Brebisson, Alexandre de, Bengio, Yoshua, and Courville, Aaron (2019). "MelGAN: Generative adversarial networks for conditional waveform synthesis". In: *Advances in Neural Information Processing Systems*.

Lee, Juheon, Choi, Hyeong-Seok, Jeon, Chang-Bin, Koo, Junghyun, and Lee, Kyogu (Sept. 2019). "Adversarially Trained End-to-End Korean Singing Voice Synthesis System". In: *Interspeech*. ISCA: ISCA, pp. 2588–2592.

Leite, Iolanda, McCoy, Marissa, Lohani, Monika, Ullman, Daniel, Salomons, Nicole, Stokes, Charlene K., Rivers, Susan, and Scassellati, Brian (2015). "Emotional Storytelling in the Classroom: Individual versus Group Interaction between Children and Robots". In: *ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 75–82.

Li, Jinlong, Yang, Hongwu, Zhang, Weizhao, and Cai, Lianhong (2011). "A lyrics to singing voice synthesis system with variable timbre". In: *Applied Informatics and Communication. ICAIC 2011. Communications in Computer and Information Science*. Vol. 225, pp. 186–193.

Li, Yongwei, Li, Junfeng, and Akagi, Masato (2018). "Contributions of the glottal source and vocal tract cues to emotional vowel perception in the valence-arousal space." In: *The Journal of the Acoustical Society of America* vol. 144, no. 2, p. 908.

Lorenzo-Trueba, Jaime, Barra-Chicote, Roberto, Raitio, Tuomo, Obin, Nicolas, Alku, Paavo, Yamagishi, Junichi, and Montero, Juan M. (Sept. 2012). "Towards Glottal Source Controllability in Expressive Speech Synthesis". In: *Interspeech*. Portland, USA, pp. 2–5.

Lorenzo-Trueba, Jaime, Barra-Chicote, Roberto, San-Segundo, Rubén, Ferreiros, Javier, Yamagishi, Junichi, and Montero, Juan M. (2015). "Emotion transplantation through adaptation in HMM-based speech synthesis". In: *Computer Speech & Language* vol. 34, no. 1, pp. 292–307.

Macon, Michael W., Jensen-Link, Leslie, Oliverio, James, Clements, Mark A., and George, E. Bryan (Apr. 1997). "A singing voice synthesis system based on sinusoidal modeling". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Vol. 1. Munich, Germany: IEEE, pp. 435–438.

Martin Arjovsky, SC and Bottou, Leon (2017). "Wasserstein generative adversarial networks". In: *34th International Conference on Machine Learning*, pp. 214–223.

McAulay, Robert J. and Quatieri, Thomas F. (Aug. 1986). "Speech analysis/Synthesis based on a sinusoidal representation". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* vol. 34, no. 4, pp. 744–754.

Mehri, Soroush, Kumar, Kundan, Gulrajani, Ishaan, Kumar, Rithesh, Jain, Shubham, Sotelo, Jose, Courville, Aaron, and Bengio, Yoshua (Dec. 2016). "SampleRNN: An Unconditional End-to-End Neural Audio Generation Model". In: arXiv: 1612.07837.

Meron, Yoram (1999). "High quality singing synthesis using the selection-based synthesis scheme". PhD thesis. University of Tokyo.

Monson, Brian B., Hunter, Eric J., Lotto, Andrew J., and Story, Brad H. (2014). "The perceptual significance of high-frequency energy in the human voice". In: *Frontiers in Psychology*.

Monson, Brian B., Lotto, Andrew J., and Ternström, Sten (2011). "Detection of high-frequency energy changes in sustained vowels produced by singers". In: *The Journal of the Acoustical Society of America* vol. 129, no. 4, pp. 2263–2268.

Montaño, Raúl and Alías, Francesc (Dec. 2016). “The role of prosody and voice quality in indirect storytelling speech: Annotation methodology and expressive categories”. In: *Speech Communication* vol. 85, pp. 8–18.

Montaño, Raúl and Alías, Francesc (2017). “The role of prosody and voice quality in indirect storytelling speech: A cross-narrator perspective in four European languages”. In: *Speech Communication* vol. 88, pp. 1–16.

Montaño, Raúl, Alías, Francesc, and Ferrer, Josep (2013). “Prosodic analysis of storytelling discourse modes and narrative situations oriented to Text-to-Speech synthesis”. In: *8th ISCA Workshop on Speech Synthesis (SSW)*, pp. 171–176.

Moulines, Eric and Charpentier, Francis (1990). “Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones”. In: *Speech Communication* vol. 9, no. 5-6, pp. 453–467.

Murphy, Andy, Yanushevskaya, Irena, Chasaide, Ailbhe Ní, and Gobl, Christer (Aug. 2017). “Rd as a Control Parameter to Explore Affective Correlates of the Tense-Lax Continuum”. In: *Interspeech*. Stockholm, Sweden, pp. 3916–3920.

Nose, Takashi, Kanemoto, Misa, Koriyama, Tomoki, and Kobayashi, Takao (2015). “HMM-based expressive singing voice synthesis with singing style control and robust pitch modeling”. In: *Computer Speech & Language* vol. 34, no. 1, pp. 308–322.

Oord, Aaron van den, Dieleman, Sander, Zen, Heiga, Simonyan, Karen, Vinyals, Oriol, Graves, Alex, Kalchbrenner, Nal, Senior, Andrew, and Kavukcuoglu, Koray (Sept. 2016). “WaveNet: A Generative Model for Raw Audio”. In: pp. 1–15.

Oura, Keiichiro and Mase, Ayami (2010). “Recent development of the HMM-based singing voice synthesis system-Sinsy”. In: *7th ISCA Workshop on Speech Synthesis (SSW)*, pp. 211–216.

Parekh, Jayneel, Rao, Preeti, and Yang, Yi-Hsuan (2020). “Speech-to-singing conversion in an encoder-decoder framework”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 261–265.

Prahallad, Kishore and Black, Alan W. (2011). “Segmentation of Monologues in Audio Books for Building Synthetic Voices”. In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* vol. 19, no. 5, pp. 1444–1449.

Prenger, Ryan, Valle, Rafael, and Catanzaro, Bryan (May 2019). “Waveglow: A Flow-based Generative Network for Speech Synthesis”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 3617–3621.

Quatieri, Thomas F. and McAulay, Robert J. (Mar. 1992). “Shape invariant time-scale and pitch modification of speech”. In: *IEEE Transactions on Signal Processing* vol. 40, no. 3, pp. 497–510.

Röbel, Axel and Fineberg, Joshua (Aug. 2007). “Speech to chant transformation with the phase vocoder”. In: *Interspeech*. Antwerp, Belgium: ISCA, pp. 4007–4008.

Rodet, Xavier, Potard, Yves, and Barriere, Jean-baptiste (1984). “The CHANT Project: From the Synthesis of the Singing Voice to Synthesis in General”. In: *Computer Music Journal* vol. 8, no. 3, pp. 15–31.

Saitou, Takeshi, Goto, Masataka, Unoki, Masashi, and Akagi, Masato (Oct. 2007). "Speech-to-Singing Synthesis: Converting Speaking Voices to Singing Voices by Controlling Acoustic Features Unique to Singing Voices". In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. New Paltz, NY, USA: IEEE, pp. 215–218.

Scherer, Klaus R. (2003). "Vocal communication of emotion: A review of research paradigms". In: *Speech Communication* vol. 40, no. 1-2, pp. 227–256.

Schröder, Marc (2001). "Emotional Speech Synthesis: A review". In: *Interspeech*. Aalborg, Denmark, pp. 561–564.

Schröder, Marc (2009). "Expressive speech synthesis: Past, present, and possible futures". In: *Affective information processing*, pp. 111–126.

Schuller, Björn, Steidl, Stefan, Batliner, Anton, Burkhardt, Felix, Devillers, Laurence, Müller, Christian, and Narayanan, Shrikanth (Jan. 2013). "Paralinguistics in speech and language—State-of-the-art and the challenge". In: *Computer Speech & Language* vol. 27, no. 1, pp. 4–39.

Skerry-Ryan, RJ, Battenberg, Eric, Xiao, Ying, Wang, Yuxuan, Stanton, Daisy, Shor, Joel, Weiss, Ron J., Clark, Rob, and Saurous, Rif A. (2018). "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron". In: *arXiv preprint arXiv:1803.09047*.

Stone, Simon, Marxen, Michael, and Birkholz, Peter (2018). "Construction and evaluation of a parametric one-dimensional vocal tract model". In: *IEEE/ACM Transactions on Audio Speech and Language Processing* vol. 26, no. 8, pp. 1381–1392.

Story, Brad H. (2013). "Phrase-level speech simulation with an airway modulation model of speech production". In: *Computer Speech & Language* vol. 27, no. 4, pp. 989–1010.

Story, Brad H., Titze, Ingo R., and Hoffman, E. A. (1996). "Vocal tract area functions from magnetic resonance imaging". In: *The Journal of the Acoustical Society of America* vol. 100, no. 1, pp. 537–554.

Sundberg, Johan (2006). "The KTH synthesis of singing". In: *Advances in Cognitive Psychology* vol. 2, no. 2-3, pp. 131–143.

Sundberg, Johan, Patel, Sona, Bjorkner, Eva, and Scherer, Klaus R. (2011). "Interdependencies among Voice Source Parameters in Emotional Speech". In: *IEEE Transactions on Affective Computing* vol. 2, no. 3, pp. 162–174.

Takemoto, Hironori, Mokhtari, Parham, and Kitamura, Tatsuya (2010). "Acoustic analysis of the vocal tract during vowel production by finite-difference time-domain method". In: *The Journal of the Acoustical Society of America* vol. 128, no. 6, pp. 3724–3738.

Taylor, Paul (2009). *Text-to-Speech Synthesis*. Cambridge, UK: Cambridge University Press, p. 626.

Theune, Mariët, Meijs, Koen, Heylen, Dirk, and Ordelman, Roeland (2006). "Generating expressive speech for storytelling applications". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* vol. 14, pp. 1137–1144.

Tokuda, Keiichi, Nankaku, Yoshihiko, Toda, Tomoki, Zen, Heiga, Yamagishi, Junichi, and Oura, Keiichiro (May 2013). "Speech Synthesis Based on Hidden Markov Models". In: *Proceedings of the IEEE* vol. 101, no. 5, pp. 1234–1252.

Uneson, Marcus (2002). “Outlines of Burcas - A simple MIDI-to-singing voice synthesis system”. In: *Fonetik*. Vol. 44. 1, pp. 133–136.

Valin, Jean-Marc and Skoglund, Jan (May 2019). “LPCNET: Improving Neural Speech Synthesis through Linear Prediction”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1. Brighton, UK: IEEE, pp. 5891–5895.

Vampola, Tomáš, Horáček, Jaromír, and Švec, Jan G. (2008). “FE Modeling of Human Vocal Tract Acoustics. Part I: Production of Czech vowels”. In: *Acta acustica united with Acustica* vol. 94, no. 5, pp. 433–447.

Waaramaa, Teija, Laukkanen, Anne Maria, Airas, Matti, and Alku, Paavo (2010). “Perception of Emotional Valences and Activity Levels from Vowel Segments of Continuous Speech”. In: *Journal of Voice* vol. 24, no. 1, pp. 30–38.

Wang, Yuxuan et al. (Aug. 2017). “Tacotron: Towards End-to-End Speech Synthesis”. In: *Interspeech*. Stockholm, Sweden: ISCA, pp. 4006–4010.

Wood, Luke, Dautenhahn, Kerstin, Robins, Ben, and Zaraki, Abolfazl (Aug. 2017). “Developing child-robot interaction scenarios with a humanoid robot to assist children with autism in developing visual perspective taking skills”. In: *26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. Lisbon, Portugal: IEEE, pp. 1055–1060.

Yamagishi, Junichi and Kobayashi, Takao (Feb. 2007). “Average-Voice-based Speech Synthesis using HSMM-based Speaker Adaptation and Adaptive Training”. In: *IEICE Transactions on Information and Systems* vol. E90–D, no. 2, pp. 533–543.

Yamagishi, Junichi, Veaux, Christophe, King, Simon, and Renals, Steve (2012). “Speech synthesis technologies for individuals with vocal disabilities: Voice banking and reconstruction”. In: *Acoustical Science and Technology* vol. 33, no. 1, pp. 1–5.

Yanushevskaya, Irena, Gobl, Christer, and Chasaide, Ailbhe Ní (2018). “Cross-language differences in how voice quality and f_0 contours map to affect”. In: *The Journal of the Acoustical Society of America* vol. 144, no. 5, p. 2730.

Zen, Heiga, Nose, Takashi, Yamagishi, Junichi, Sako, Shinji, Masuko, Takashi, Black, Alan W, and Tokuda, Keiichi (2007). “The HMM-based Speech Synthesis System (HTS) Version 2.0”. In: *6th ISCA Workshop on Speech Synthesis (SSW)*. Bonn, Germany, pp. 294–299.

Zen, Heiga, Senior, Andrew, and Schuster, Mike (May 2013). “Statistical parametric speech synthesis using deep neural networks”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, pp. 7962–7966.

Zovato, Enrico, Pacchiotti, Alberto, Quazza, Silvia, and Sandri, Stefano (2004). “Towards Emotional Speech Synthesis: A rule based approach”. In: *5th ISCA Workshop Speech Synthesis (SSW)*, pp. 219–220.

Indexed articles

Paper I

Adding Singing Capabilities to Unit Selection TTS through HNM-Based Conversion

Marc Freixes, Joan Claudi Socoró, Francesc Alías

Published in *Advances in Speech and Language Technologies for Iberian Languages. IberSPEECH 2016. Lecture Notes in Computer Science.*, November 2016, volume 10077, pp. 33–43. DOI: 10.1007/978-3-319-49169-1_4.

Abstract

Adding singing capabilities to a corpus-based concatenative text-to-speech (TTS) system can be addressed by explicitly collecting singing samples from the previously recorded speaker. However, this approach is only feasible if the considered speaker is also a singing talent. As an alternative, we consider appending a Harmonic plus Noise Model (HNM) speech-to-singing conversion module to a Unit Selection TTS (US-TTS) system. Two possible text-to-speech-to-singing synthesis approaches are studied: applying the speech-to-singing conversion to the US-TTS synthetic output, or implementing a hybrid US+HNM synthesis framework. The perceptual tests show that the speech-to-singing conversion yields similar singing resemblance than the natural version, but with lower naturalness. Moreover, no statistically significant differences are found between both strategies in terms of naturalness nor singing resemblance. Finally, the hybrid approach allows reducing more than twice the overall computational cost.

L'interval de pàgines 28-37 s'ha extret per conflictes de confidencialitat

L'interval de pàgines 28-37 s'ha extret per conflictes de confidencialitat

L'interval de pàgines 28-37 s'ha extret per conflictes de confidencialitat

L'interval de pàgines 28-37 s'ha extret per conflictes de confidencialitat

L'interval de pàgines 28-37 s'ha extret per conflictes de confidencialitat

L'interval de pàgines 28-37 s'ha extret per conflictes de confidencialitat

L'interval de pàgines 28-37 s'ha extret per conflictes de confidencialitat

L'interval de pàgines 28-37 s'ha extret per conflictes de confidencialitat

L'interval de pàgines 28-37 s'ha extret per conflictes de confidencialitat

L'interval de pàgines 28-37 s'ha extret per conflictes de confidencialitat

L'interval de pàgines 28-37 s'ha extret per conflictes de confidencialitat

Paper II

A Unit Selection Text-to-Speech-and-Singing Synthesis Framework from Neutral Speech: Proof of concept

II

Marc Freixes, Joan Claudi Socoró, Francesc Alías

Published in *EURASIP Journal on Audio, Speech and Music Processing*, December 2019, volume 2019, article number 22. DOI: 10.1186/s13636-019-0163-y.

Abstract

Text-to-speech (TTS) synthesis systems have been widely used in general-purpose applications based on the generation of speech. Nonetheless, there are some domains, such as storytelling or voice output aid devices, which may also require singing. To enable a corpus-based TTS system to sing, a supplementary singing database should be recorded. This solution, however, might be too costly for eventual singing needs, or even unfeasible if the original speaker is unavailable or unable to sing properly. This work introduces a Unit Selection based Text-to-Speech-and-Singing (US-TTS&S) synthesis framework, which integrates Speech-to-Singing (STS) conversion to enable the generation of both speech and singing from an input text and a score, respectively, using the same neutral speech corpus. The viability of the proposal is evaluated considering three vocal ranges and two tempos on a proof-of-concept implementation using a 2.6h Spanish neutral speech corpus. The experiments show that challenging STS transformation factors are required to sing beyond the corpus vocal range and/or with notes longer than 150 ms. While score-driven US configurations allow the reduction of pitch-scale factors, time-scale factors are not reduced due to the short length of the spoken vowels. Moreover, in the MUSHRA test, text-driven and score-driven US configurations obtain similar naturalness rates of around 40 for all the analysed scenarios. Although these naturalness scores are far from those of Vocaloid, the singing scores of around 60 which were obtained validate that the framework could reasonably address eventual singing needs.

II.1 Introduction

Text-to-speech (TTS) synthesis systems have been widely used to generate speech in several general-purpose applications, such as call-centre automation, reading emails or news, or

providing travel directions, among others (Taylor 2009). However, there are other domains that may require the eventual generation of singing in addition to speech. For instance, in storytelling (Montaño and Alías 2016; Fridin 2014), when one of the characters sings at one point in the story, or in voice output communication aid devices for individuals with vocal disabilities (Yamagishi et al. 2012) to allow them not only to talk, but also to sing. Moreover, a TTS with singing capabilities could also be useful in assistive technologies, where the incorporation of songs has been proved to be an effective form of improving the engagement of autistic children (Wood et al. 2017), or to reduce the procedural distress in children with cancer (Jibb et al. 2018), or to augment the positive memories of people with dementia (Khosla et al. 2017), to name a few.

In this sense, it is worth mentioning that early works on speech synthesis already enabled the generation of both speech and singing (e.g. see Cook 1993), as they stood on a source-filter model inspired by the classical acoustic theory of voice production (Taylor 2009). However, the difficulty of defining and adjusting the necessary control parameters to obtain high quality speech led the research towards data-driven approaches (Taylor 2009). Although some approaches used diphone-based TTS systems to generate singing (Flinger 2001; Uneson 2002), most works opted to use databases specifically recorded for singing purposes (Macon et al. 1997; Bonada and Serra 2007; Blaauw and Bonada 2017). Meanwhile, the speech synthesis investigations also moved to corpus-based approaches, deploying TTS systems based on Unit Selection (US), Hidden Markov Models (HMM) or hybrid approaches, and more recently, including Deep Neural Networks (DNN) (e.g. Oord et al. 2016; Wang et al. 2017). Even though these systems can deliver very natural synthetic speech (King 2014), as far as we know, they are not able to speak and sing at the same time.

In order to add singing capabilities to a corpus-based TTS system, the first idea that may come to mind is to incorporate a supplementary singing database. However, occasional singing needs do not justify the cost of building an additional corpus, which may become unfeasible if the original speaker is unavailable or unable to sing properly (Blanco et al. 2016; Freixes et al. 2016). As an alternative, we could take advantage of those approaches which focus on the production of singing from speech following the so-called speech-to-singing (STS) conversion (Röbel and Fineberg 2007; Saitou et al. 2007; Dong et al. 2014). These techniques can be applied to the output of a TTS system to transform speech to singing by maintaining the identity of the speaker (Freixes et al. 2016; Li et al. 2011). However, this straightforward approach has been proved suboptimal in terms of flexibility and computational costs (Freixes et al. 2016).

Building on the preliminary approach presented in (Freixes et al. 2016), this work introduces a Unit Selection based Text-to-Speech-and-Singing (US-TTS&S) synthesis framework that allows the generation of both speech and singing from an input text and a score, respectively, using the same neutral speech corpus. To this end, the framework incorporates Speech-to-Singing (STS) conversion within a TTS system pipeline. The viability of the proposal is evaluated objectively and subjectively through a proof-of-concept implementation of the US-TTS&S framework using a 2.6h Spanish neutral speech corpus.

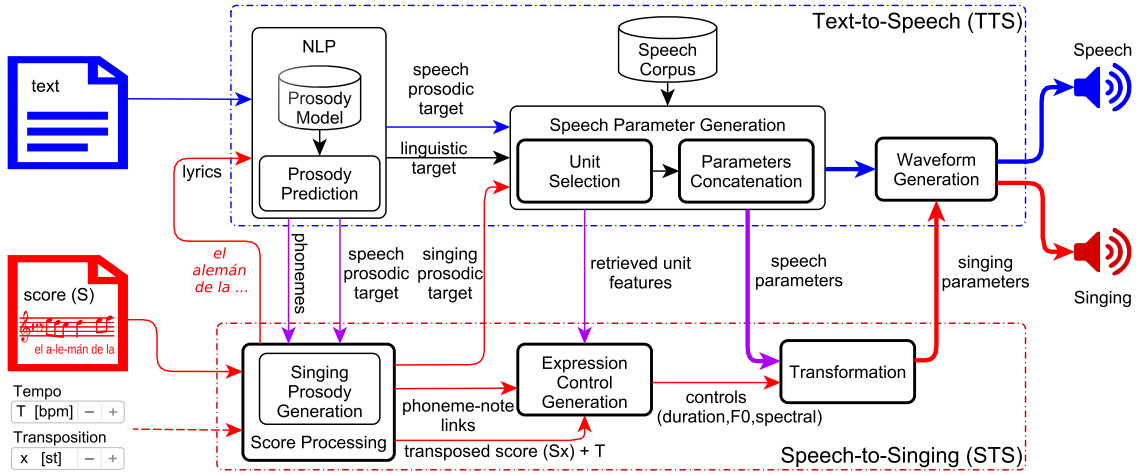


Figure II.1: US-TTS&S framework. Block diagram of the unit-selection text-to-speech and singing (US-TTS&S) synthesis framework from neutral speech. In the speech mode, an input text is converted into synthetic speech by the TTS subsystem (above in the blue box). In the singing mode, the incorporation of the Speech-to-Singing (STS) subsystem (below in the red box) enables the framework to produce synthetic singing from an input score S (containing both the notes and the lyrics), considering optional input values: tempo T in beats per minute and transposition x in semitones.

The paper is structured as follows. Section II.2 reviews the singing and speech-to-singing literature. Then, Section II.3 describes the proposed US-TTS&S framework and the proof-of-concept implementation. The methodology employed for the objective and the subjective evaluation is detailed in Section II.4. Finally, after presenting and discussing the results (Section II.5), the conclusions of this work are drawn in Section II.6.

II.2 Related work

This section includes a review of the singing synthesis approaches which are closely related to speech synthesis and a description of speech-to-singing techniques.

II.2.1 Singing synthesis

Until the late 80s, most of the singing synthesis approaches were closely linked to sound synthesis (Chowning 1980) or to speech synthesis (see Cook 1996 and references therein). The latter correspond to first generation synthesis systems, where according to a synthesis specification (verbal component, pitch values and durations) a rule-based control drives a source-filter model built on the classical acoustic theory of voice production. On the one hand, articulatory speech synthesis (Kelly and Lochbaum 1962) was used to generate one of the first synthetic singing examples[1]. This technology evolved giving rise to systems such as SPAM/Singer (Cook 1993), which could be used for TTS and singing synthesis through control files (Cook et al. 1993). On the other hand, formant speech synthesis inspired the development of singing systems as the MUSSE (MUsic and Singing Synthesis Equipment)

[1] https://ccrma.stanford.edu/~jos/wav/daisy-klm.wav

and the subsequent MUSSE DIG (MUSSE, DIGital version) (Sundberg 2006) or the CHANT project (Rodet et al. 1984). First generation rule-based systems gave way to data-driven approaches mainly due to the difficulty of generating the control parameters to get high quality results (Taylor 2009). However, formant synthesis is still used nowadays in the context of performative singing synthesis (Feugère et al. 2017), where flexibility and real time are the main issues.

In second generation synthesis systems, a unit (typically a diphone) for each unique type was recorded. Pitch and timing of units were modified applying signal processing techniques to match the synthesis specification (Taylor 2009). Some works exploited signal processing capabilities to generate singing from a spoken database. Flinger (Flinger 2001) for instance used residual LPC synthesis and provided several modules in order to enable the Festival TTS system (Festival 2016) to sing. MBROLA was also used to generate both speech and singing from speech units (Uneson 2002; D'Alessandro et al. 2005). Similarly, the Ramcess synthesiser (D'Alessandro et al. 2008) generated singing by convolving vocal tract impulse responses from a database with an interactive model of the glottal source. However, the data-driven paradigm of second generation synthesis systems naturally led to the creation of singing databases.

Finally, it should be noted that there have been some recent attempts to produce singing from speech in a corpus-based TTS system. Some works used the system to get a spoken version of the song and transform it into singing by incorporating a signal processing stage. For instance, in (Li et al. 2011) the synthetic speech was converted into singing according to a MIDI file input, using STRAIGHT to perform the analysis, transformation and synthesis. In (Blanco et al. 2016) an HMM-based TTS synthesiser for Basque was used to generate a singing voice. The parameters provided by the TTS system for the spoken version of the lyrics were modified to adapt them to the requirements of the score.

II.2.2 Speech-to-Singing

Speech-to-singing conversion is the task of transforming the spoken lyrics of a song into singing, while retaining the identity of the speaker and the linguistic content (Vijayan et al. 2019). In (Saitou et al. 2007), the authors proposed a method to transform speech into singing, by modifying the pitch contour, the duration of the phonemes and the spectrum according to the analysis of the features of the singing voice. Phoneme target durations were obtained by applying STS duration conversion rules derived from the analysis of real performances. The pitch contour was derived from a stepwise melody curve by applying a filtering that models the behaviour and dynamics of the fundamental frequency ($F0$) in singing: preparation, overshoot, fine fluctuation and vibrato. Finally, two spectral control models were applied to the envelope to add the singing formant, and to apply a formant amplitude modulation that was synchronised with the vibrato. Analysis, transformation and synthesis were carried out using STRAIGHT (Kawahara et al. 1999).

In order to obtain more natural contours, other approaches have used real singing performances, but they require spoken and sung parallel recordings. In (Röbel and Fineberg

2007), a set of phrases was recorded by a female singer to get a spectral envelope database. The same speech sentences, recorded by an actor, were time-stretched, transposed and aligned with the singing phrases. Finally, the spectral envelope from the singer database was transferred to the speech signal. The transformation was performed by a phase vocoder in this case. However, improved signal models were subsequently proposed (Roebel et al. 2012; Huber and Roebel 2015). In (Dong et al. 2014), I^2R Speech2Singing system was presented. This application was able to convert speech or poor singing into high-quality singing, using a template-based conversion (Cen et al. 2012) with professional singing as a reference model. Parallel singing templates were aligned with the speech input in a 2-step Dynamic Time Warping-based method. Thus, the pitch contour could be derived from actual singing voice and applied to the input speech through STRAIGHT. An improved dual alignment scheme for this system has been recently presented in (Vijayan et al. 2017).

Finally, apart from appropriate timing and F0 contours, spectral transformation is a very important issue in speech-to-singing conversion. Voice conversion and model adaptation techniques were extended to this scenario in (Lee et al. 2014), using a large collection of singing recordings and their corresponding spoken lyrics. The comparison between these methods and the spectral transformation applied in (Saitou et al. 2007) showed that model adaptation outperforms the other approaches in singing quality and similarity provided there is enough data.

II.3 US-TTS&S synthesis framework from neutral speech

This section is organised as follows. Section II.3.1 describes the proposed US-TTS&S synthesis framework. Next, Section II.3.2 details the proof-of-concept implementation of the framework.

II.3.1 Framework

The block diagram of the proposed synthesis framework is depicted in Fig. II.1. It consists of two main subsystems: the text-to-speech subsystem (at the top), which allows the framework to produce neutral synthetic speech for a given input text, and the speech-to-singing subsystem (at the bottom), which provides the framework with singing capabilities.

In the speech mode, the input text is analysed by the Natural Language Processing (NLP) module, which yields a linguistic target (including the phonetic transcription and the linguistic context) and predicts a speech prosodic target (i.e. phrasing and intonation appropriate for the message). The unit selection block searches the corpus for the units that best meet these targets and that can be smoothly joined. Finally, the parametric representations of the selected units are concatenated, thus obtaining a stream of speech parameters that is rendered into synthetic speech through the waveform generation module.

In the singing mode, the input score S, which contains the lyrics as syllables assigned to the notes, is parsed by the score processing module, which extracts the lyrics, integrates score and phonetic information and provides a singing prosodic target to perform the unit selection according to S and the optional tempo and transposition values. Subsequently,

the transformation module converts the retrieved speech parameters into the singing ones, according to the controls computed by the expression control generation module. Finally, the waveform generation module renders the modified parameters into synthetic singing.

The following subsections describe the key modules for the singing mode.

II.3.1.1 Score processing

This module joins the syllables extracted from S in order to get the full lyrics of the song, which are fed into the TTS subsystem (see Fig. II.1). Subsequently, it obtains the links between the phonetic transcription of the lyrics (provided by the NLP module) and the notes. To this end, the phonemes are distributed among the notes according to the assignment of syllables to notes in the score. Furthermore, since a note onset coincides with a vowel (Sundberg and Bauer-Huppmann 2007), the preceding consonants are assigned to the preceding note.

Moreover, this module allows for score transposition according to the input value x, obtaining thereby a transposed score Sx. This score is then used, together with phoneme-note links and tempo T, to compute the singing prosodic target (see Section II.3.1.2) and to generate the expression controls (see Section II.3.1.3).

With regard to tempo, the value of T in beats per minute (bpm) can be extracted from S, or alternatively indicated as an input of the synthesis framework (see Fig. II.1). Tempo is used to compute the duration of each note according to its note value (e.g. quarter note, eighth note, etc.).

Regarding score transposition, this process consists of moving the entire set of notes up or down in pitch by a constant interval in order to fit it within the vocal range of the singer. Accordingly, the notes of S are shifted to get the score Sx, whose pitch range midpoint $F0_m^{\mathrm{Sx}}$ is x semitones above the speech corpus vocal range midpoint $F0_m^{\mathrm{C}}$, which represents an intermediate value within the pitch range covered by the vowels in the corpus C. To this end, the note pitches in S are translated into an integer notation following a twelve-tone equal temperament, which divides the octave into 12 semitone steps equally spaced on a logarithmic scale. Thus, a note number $N^{\mathrm{S}}(i)$ is obtained for each note in S, where $i = \{1..K\}$, being K the total number of notes in the score S. Subsequently, the note numbers for Sx are computed as

$$N^{\mathrm{Sx}}(i) = N^{\mathrm{S}}(i) + x - d(F0_m^{\mathrm{C}}, F0_m^{\mathrm{S}}) \tag{II.1}$$

where

$$d(F0_m^{\mathrm{C}}, F0_m^{\mathrm{S}}) = \left[12 log_2 \left(\frac{F0_m^{\mathrm{S}}}{F0_m^{\mathrm{C}}}\right)\right] \tag{II.2}$$

is the distance in semitones from the speech corpus vocal range midpoint $F0_m^{\mathrm{C}}$ to the input score pitch range midpoint $F0_m^{\mathrm{S}}$, and $[\cdot]$ denotes that the result of the operation is rounded to the nearest integer. Since the perception of pitch is logarithmic, $F0_m^{\mathrm{S}}$ is computed from the lowest and the highest note as the geometric mean of their $F0$ values, i.e.,

$$F0_m^{\mathrm{S}} = \sqrt{F0_{min}^{\mathrm{S}} \cdot F0_{max}^{\mathrm{S}}}. \tag{II.3}$$

II.3.1.2 Singing prosody generation

This block translates the note durations and $F0$s obtained from Sx and T into a prosodic representation of the singing target consisting of phonetic timing and $F0$s. This singing prosodic target enables the US-TTS&S framework to perform the unit selection according to Sx and T. The phonetic timing is obtained by adjusting the duration of the phonemes so that they fit the duration of the notes to which they are linked. Similarly, the $F0$ of each note is assigned to its phonemes considering that the note $F0$ is reached at the vowel onset, so the transition occurs in the precedent phoneme (Sundberg and Bauer-Huppmann 2007).

II.3.1.3 Expression control generation

Expression control in singing synthesis, also known as performance modelling, consists in the manipulation of a set of voice features (e.g. phonetic timing, pitch contour, vibrato, timbre, etc.) that relates to a particular emotion, style or singer (Umbert et al. 2015). Accordingly, the expression control generation module provides the duration, $F0$ and spectral controls required by the transformation module to convert the sequence of speech parameters into singing parameters. To this end, and following the phoneme-note links, this module aligns the units retrieved by the US block with the notes, and generates the controls to transform the spoken features (durations, $F0$ and spectra) into singing ones in accordance with Sx and T. Since obtaining control parameters that are perceived as natural is one of the main issues regarding singing synthesis, several approaches can be found in the literature (see Umbert et al. 2015 and references therein for further details).

II.3.1.4 Speech parameter generation and transformation

In contrast to *pure* unit selection, where an overlap and add (OLA) method is applied to the retrieved units, with the aim of modifying the original waveforms as little as possible (Taylor 2009), the US-TTS&S framework is based on a parametric representation of the speech. This enables the use of more flexible processing techniques to address the highly significant transformations (including spectral ones) involved in the STS conversion.

The framework signal processing pipeline consists of three modules. The speech parameter generation module performs the unit selection (according to the linguistic and prosodic targets) and concatenates the parametric representation of the selected units to obtain a speech parameter sequence. In the speech mode this sequence is directly fed into the waveform generation module to produce synthetic speech. Conversely, in the singing mode the sequence is previously processed by the transformation module, which applies time-scaling, pitch-scaling and spectral transformations to convert the speech parameters into singing ones.

II.3.2 Proof-of-concept implementation

In the following paragraphs, the main elements of the implementation of the US-TTS&S framework are described.

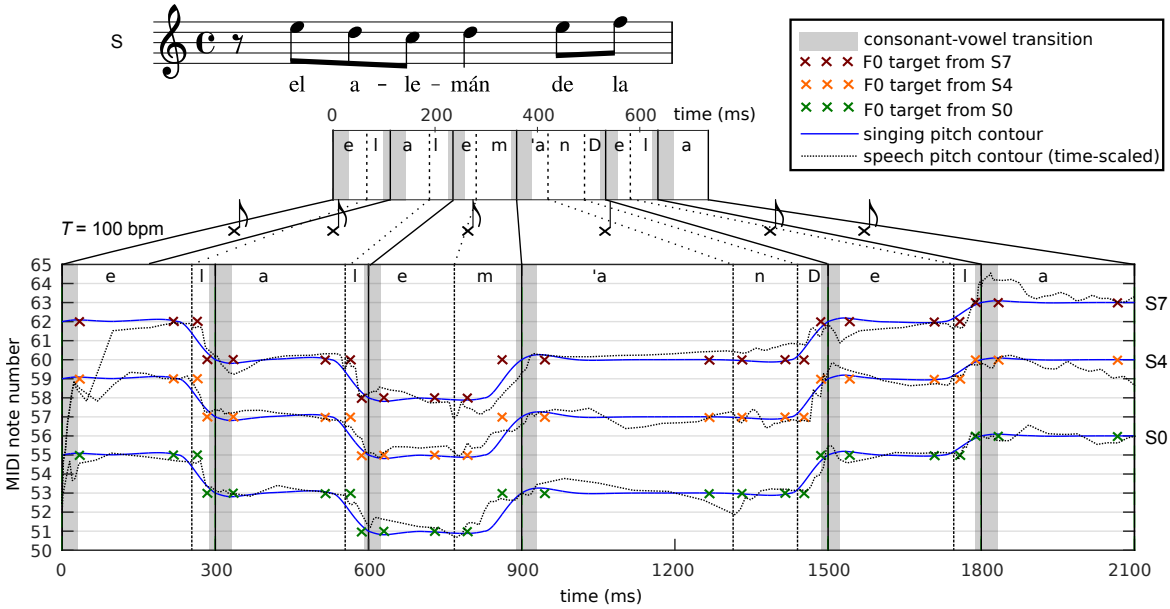


Figure II.2: Example of a song excerpt synthesised with transposed scores S0, S4 and S7. The phonemes from the lyrics phonetic transcription are represented below the input score S, together with their durations, which are: i) predicted from the lyrics by the NLP module when computing the singing prosodic target for the US block (see Fig. II.1), or; ii) those of the retrieved speech units when generating the expression controls. At the bottom, the phoneme durations have been time-scaled to fit the note durations. The crosses represent the $F0$ values of the singing prosodic targets obtained from S0, S4 and S7. The pitch contours (time-scaled) of the retrieved speech units are depicted as dashed grey lines. Finally, the solid blue lines represent the singing pitch contours generated by the expression control generation module. The score-driven US configuration Sxp and $T = 100$ bpm have been used for this example.

II.3.2.1 Text-to-Speech subsystem

The US-TTS system of La Salle - Universitat Ramon Llull (Formiga et al. 2010) has been used as text-to-speech subsystem. This TTS synthesis system includes a Case-Based Reasoning (CBR) prosody prediction block, trained with acoustic prosodic patterns from the speech corpus, and a unit selection block following a classical scheme (Hunt and Black 1996). This block retrieves the units that minimise the prosodic, linguistic and concatenation costs (see Formiga et al. 2010 for more details). The weights for the prosodic target and concatenation subcosts were perceptually tuned by means of active interactive genetic algorithms for speech synthesis purposes (Alías et al. 2011).

The Time-Domain Pitch Synchronous Overlap and Add (TD-PSOLA) waveform generation, used in the original US-TTS system, has been replaced by a Harmonic plus Noise Model (HNM) implementation (Calzada Defez and Socoró 2013). Accordingly, the corpus has been parameterised with HNM representation. The harmonic component (for the voiced frames) is modelled by a sum of sinusoids (each with a certain amplitude and phase) at the multiples of the fundamental frequency up to the 5 kHz maximum voiced frequency (Erro and Moreno 2008). This component is subtracted from the speech signal to get the stochastic (noise) component, which is analysed using an autoregressive model and it is represented with 15-order

Linear Prediction Coefficients (LPC) and the noise variance (Erro and Moreno 2008). The HNM analysis has been performed pitch-synchronously, applying a window around the centre of gravity to avoid phase mismatches when units are concatenated (Stylianou 2001).

II.3.2.2 Score processing

The proof-of-concept implementation of this module has adopted the MusicXML[2] format for the score S. To this end, the scripts from (Nichols et al. 2009) have been considered. In MusicXML, each syllable of the lyrics is assigned to a note with the *lyric* element. This contains a *text* element with the syllable and a *syllabic* element that indicates how the syllable fits into the word. The latter can take the values *single*, *begin*, *end*, or *middle*, and is used to recompose the words and obtain the whole text of the song. The *syllabic* element also provides the syllabic distribution, which is considered to assign the phonemes from each word to their corresponding notes. An example of this alignment is depicted at the top of Fig. II.2.

With regard to the $F0$, each MusicXML note in S is parsed into a MIDI note number $N^{\mathrm{S}}(i)$, whose $F0$ is computed as

$$F0^{\mathrm{S}}(i) = 440 \cdot 2^{(N^{\mathrm{S}}(i)-69)/12}, \tag{II.4}$$

since the MIDI note 69 corresponds to A4 (440 Hz)[3]. If a transposition value of x semitones is introduced into the framework, the shifted MIDI note numbers for Sx are computed following equations (II.1), (II.2) and (II.3).

The speech corpus vocal range is defined from the $F0$ mean values of the vowels within it. According to this, the speech corpus vocal range midpoint is computed in this implementation as

$$F0^{\mathrm{C}}_m = \sqrt{F0^{\mathrm{C}}_5 \cdot F0^{\mathrm{C}}_{95}}, \tag{II.5}$$

where $F0^{\mathrm{C}}_5$ and $F0^{\mathrm{C}}_{95}$ are the 5th and the 95th corpus vowel $F0$ percentiles, respectively, thus avoiding possible outliers.

II.3.2.3 Singing prosody generation

This block generates a singing prosodic target according to the durations and $F0$s obtained from score Sx and tempo T.

On the one hand, the phoneme durations predicted by the prosodic model (represented below the score S in Fig. II.2) are adjusted to fit the note durations by applying the STS conversion rules derived by Saitou from the analysis of real performances (Saitou et al. 2007):

a) When phonemes tied to a note have to be shortened, their original durations are multiplied by the same factor.

b) When phonemes have to be stretched, three parts are differentiated around the boundary between a consonant and a vowel: the consonant, the transition (from 10 ms before to 30 ms after the boundary) and the vowel.

[2] https://www.musicxml.com
[3] https://www.midi.org/specifications

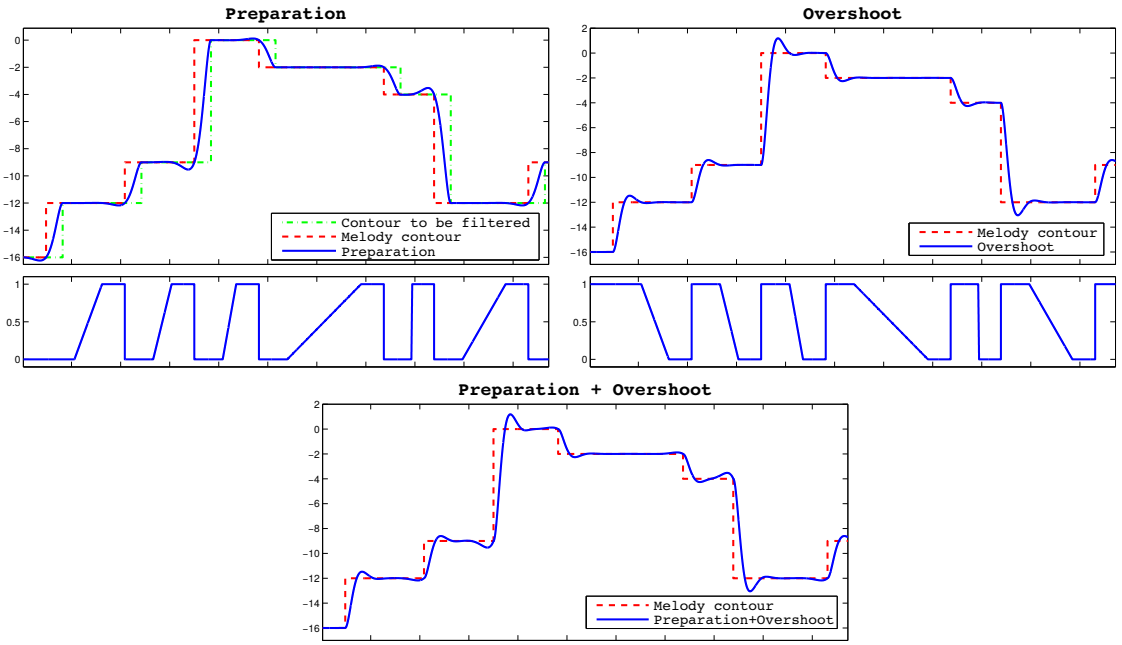


Figure II.3: Singing pitch curve generation. Preparation (upper left), overshoot (upper right), the applied masks (middle) and the resulting mix (bottom).

(i) The consonant part is extended according to fixed rates (1.58 for a fricative, 1.13 for a plosive, 2.07 for a semivowel, 1.77 for a nasal, and 1.13 for a /y/).

(ii) The transition part (depicted as a shadowed area in Fig. II.2) is not extended.

(iii) The vowel part is extended until the phoneme fits the note duration.

In the current implementation the transition length within the vowel (30 ms) has been limited to a maximum of half of its duration, since the corpus contains very short vowels.

On the other hand, the $F0$ target (represented by crosses in Fig. II.2) is assigned at a semiphoneme level. The $F0$ from each note in Sx is assigned to all its corresponding semiphonemes, except in the transitions where the right semiphoneme receives the $F0$ of the following note.

II.3.2.4 Expression control generation

This module computes the duration, $F0$ and spectral controls required to perform the STS conversion, in accordance with Sx and T.

Regarding the duration control, the durations of the phonemes retrieved by the US block are scaled to fit the durations of the notes, by applying the conversion rules detailed in Section II.3.2.3. The correspondence between the original and the scaled durations will drive the time-scaling process performed by the transformation module.

With respect to the $F0$ control, a singing pitch contour (the blue solid lines in Fig. II.2) is obtained following the approach described in (Saitou et al. 2007). According to this, a stepwise pitch contour is built from $F0$s and durations of the notes. Then, this contour is filtered to obtain the singing $F0$ characteristic fluctuations: overshoot, preparation and fine fluctuation. Fig. II.3 depicts an example of a pitch curve generation. Overshoot (upper right)

is obtained by directly filtering the stepwise contour. Alternatively, preparation (upper left) can be obtained by filtering (from the end towards the beginning) a slightly delayed version of the stepwise curve. The mix (bottom) of both fluctuations is obtained by applying the masks (middle), which prioritise the overshoot at the beginning of the note, preparation at the end, and consider a simple cross-fading in between. In this proof of concept, the implementation of vibrato is left for future research.

Finally, the spectral control tries to emulate the singing formant by emphasising the spectral envelope peak around 3kHz within the vowels (Saitou et al. 2007).

II.3.2.5 Speech parameter generation and transformation

The HNM parameters of the retrieved units are concatenated, removing pitch and spectrum mismatches by applying a simple linear interpolation technique around the joins (Stylianou 2001). Transformation and synthesis are performed pitch-synchronously. Thus, when a prosody modification is performed, the HNM parameters in the new epochs are obtained pitch-synchronously through the time-domain linear interpolation of the original parameters. Furthermore, if pitch scaling is done, amplitudes and phases are interpolated in frequency to preserve the original spectral envelope shape (Erro et al. 2007). The new harmonic amplitudes are calculated by the linear interpolation of the spectral envelope in a logarithmic amplitude scale. The phases of the target harmonics are obtained by interpolating the real and the imaginary parts of the harmonic complex amplitudes at the new frequencies. Finally, the amplitudes are scaled to preserve the energy despite the variation in the number of harmonics.

II.4 Methods

This section describes the methods used for the evaluation of the proposed US-TTS&S synthesis framework through the proof-of-concept implementation using a Spanish corpus. The study has been carried out for three vocal ranges and two tempos, and considering a text-driven and three score driven US configurations. The experiments setup is described in Section II.4.1. Then, the objective evaluation (Section II.4.2) analyses the magnitude of the transformations required by the STS process to allow the framework to sing. Finally, the subjective evaluation (Section II.4.3) assesses both the singing capabilities of the framework together with the naturalness of the synthesised singing.

II.4.1 Experiments setup

II.4.1.1 Corpus

The experiments have been performed using a 2.6h Spanish neutral speech corpus recorded by a female professional speaker (Alías et al. 2008). The duration and $F0$ histograms of the corpus vowels are depicted in the Fig. II.4a. Regarding duration, about half of the vowels last 50 ms or less, and there are virtually none beyond 200 ms. The $F0$ histogram has been depicted so each bin coincides with a semitone in an equal temperament. Even though the corpus

contains vowels from 123 until 330, three out of four are between 139 and 196, so only cover 7 semitones. The 5th and the 95th percentiles ($F0_5^{\mathrm{C}}$ and $F0_{95}^{\mathrm{C}}$) are 134.4 Hz and 235.5 Hz, respectively. Therefore, the corpus vocal range midpoint is $F0_m^{\mathrm{C}} = \sqrt{F0_5^{\mathrm{C}} \cdot F0_{95}^{\mathrm{C}}} = 178$ Hz.

II.4.1.2 Vocal ranges and tempos

The first evaluation scenario considered corresponds to singing in the corpus vocal range (S0). However, in order to evaluate the capability of the proposed US-TTS&S system to work in a singer vocal range, a contralto set up has been also examined; this is 7 semitones above the speech corpus pitch range midpoint (S7). Moreover, the study has been completed with an intermediate anchor point (S4). Finally, regarding the tempo, two values have been considered: $T = 100$ bpm corresponding to a moderate speed, and a slow one ($T = 50$ bpm).

II.4.1.3 Unit selection configurations

The evaluation has included a text-driven US configuration, MLC, which considers linguistic (L) and concatenation (C) costs, and the prosodic target predicted from the lyrics by the CBR prosodic model (M). This would correspond to the default US-TTS setting.

Moreover, the study has also considered three score-driven configurations. In this case, the prosodic target is that obtained by the singing prosody generation block according to Sx and T. These configurations are: SxpdLC, which uses the pitch (p) and duration (d) from the score instead of those from the model, SxpdC, which also disables the linguistic cost, and finally Sxp, which only considers the pitch.

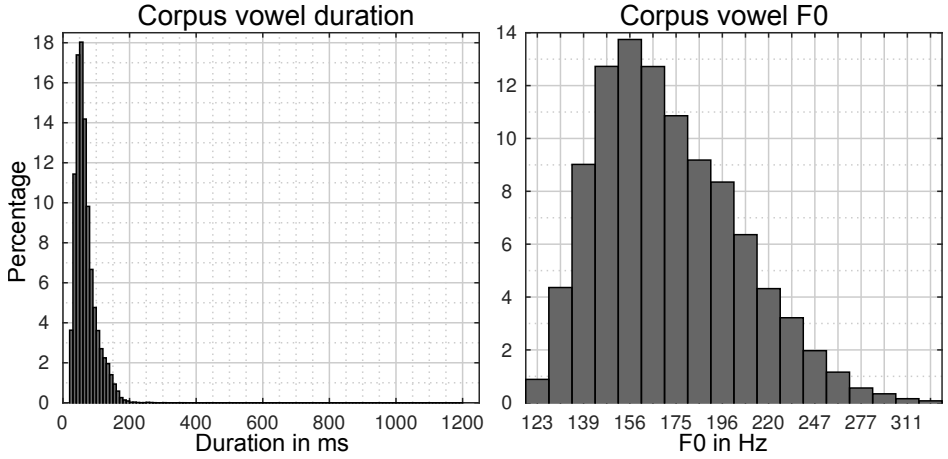
II.4.2 Objective evaluation

The objective analysis has been conducted by feeding a score dataset into the framework to be sung in the aforementioned vocal ranges and tempos with the considered US configurations. Then, the pitch and time-scale factors required to transform the retrieved units into singing have been computed. More specifically, the analysis has been focused on the vowels, where the bulk of the signal processing takes place. Moreover, the approach described in (Karabetsos et al. 2010) has been implemented to get a binary concatenation quality prediction (poor/good) for each join (within the vowels). The subsequent paragraphs describe the details of the experiments.

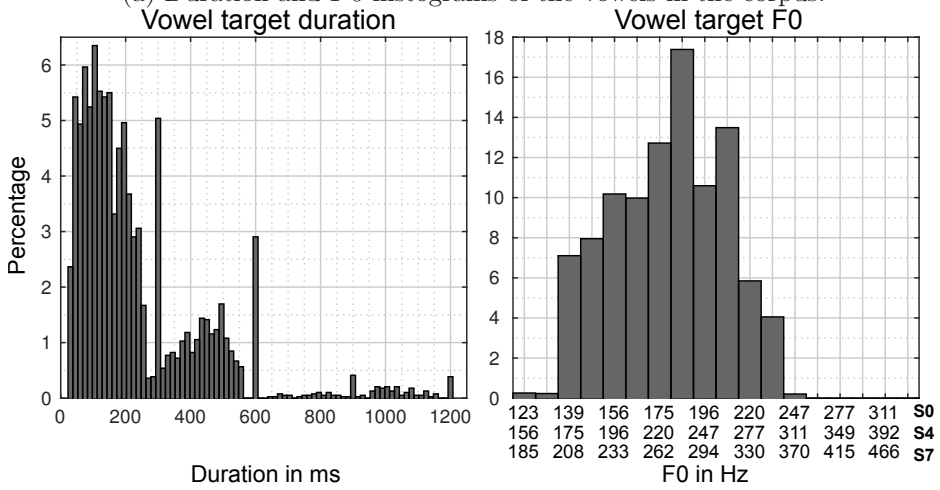
II.4.2.1 Score test dataset

From a score compilation of songs for children (Bensaya 2018), a subset of 279 musical phrases has been selected, by applying a greedy algorithm (Francois and Boëffard 2002) to ensure its phonetic coverage in Spanish. This has resulted in a dataset containing 3899 notes, which spans 29 min and 57 s with $T = 100$ bpm and 59 min and 54 s for $T = 50$ bpm.

Fig. II.4b presents the vowel duration and $F0$ targets generated from the dataset by the singing prosody generation block. The left of Fig. II.4b shows the histogram of the vowel



(a) Duration and $F0$ histograms of the vowels in the corpus.



(b) Vowel target duration and $F0$ histograms predicted from the test score dataset sung with $T = 100$ bpm, and 0, 4 and 7 semitones above the speech corpus vocal range midpoint S0, S4 and S7, respectively).

Figure II.4: Corpus (above) and target (below) vowel duration and $F0$ distributions.

Table II.1: Pitch-scale intervals expressed in absolute number of semitones ($|\alpha_{st}|$) and as multiplying factors (α).

$\lvert\alpha_{st}\rvert$	**[0-4]**	**(4-7]**	**(7-12]**	**>12**
$\alpha < 1$	[0.8-1)	[0.67-0.8)	[0.5-0.67)	<0.5
$\alpha > 1$	(1-1.26]	(1.26-1.50]	(1.50-2]	>2

duration target for the score dataset sung at 100 bpm, while the right section depicts the histogram of the vowel $F0$ target for the dataset performed with S0, S4 and S7.

II.4.2.2 Transformation requirements

A time-scale factor (β) has been calculated for each retrieved vowel as

$$\beta = \frac{Dur_{tgt} - Dur_{trn}}{Dur_{orig} - Dur_{trn}} \tag{II.6}$$

where Dur_{tgt} is the singing target duration and Dur_{orig} the original duration of the retrieved vowel. When the vowel is stretched, Dur_{trn} accounts for the duration of the unscaled transition

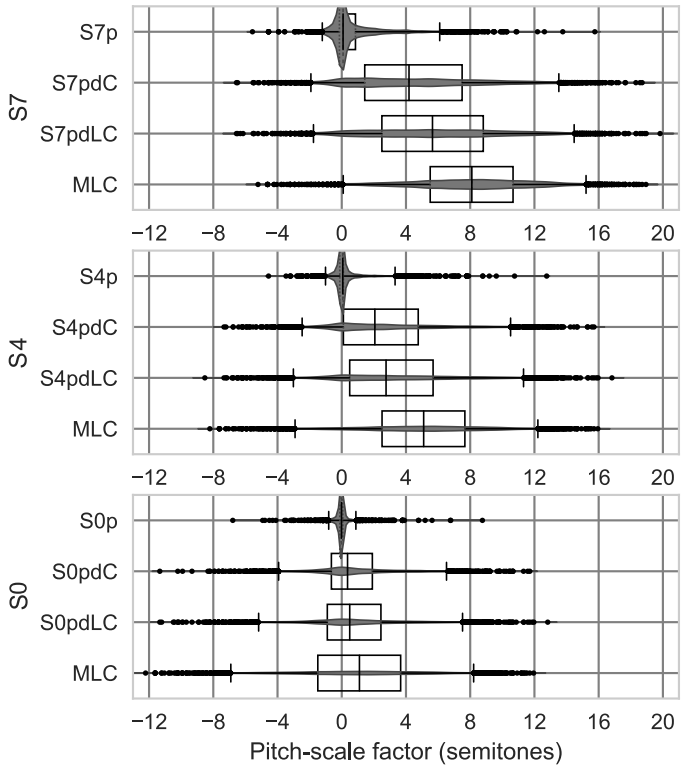


Figure II.5: Pitch-scale factors (α_{st}) for different vocal ranges S0, S4, S7) and unit selection configurations. Whiskers are set to 2nd and 98th percentile. Differences between all configurations are statistically significant ($p<0.01$) except for the pair S0pdC-S0pdLC.

(shadowed areas in Fig. II.2), otherwise $Dur_{trn} = 0$.

Regarding the pitch-scale factors (α), since the core US-TTS works with diphones, we have obtained two values for each vowel, i.e. one for each semiphoneme. The pitch-scale factor has been computed as

$$\alpha = \frac{F0_{tgt}}{mean(F0_{orig})} \tag{II.7}$$

where $F0_{tgt}$ is the target $F0$ assigned from Sx, and $mean(F0_{orig})$ is the mean of the $F0$ values within the retrieved semiphoneme. Pitch-scale factors are expressed in number of semitones as $\alpha_{st} = 12log_2(\alpha)$, since these units are more meaningful from a musical point of view and closer to the logarithmic perception of the pitch.

Moreover, transformation factors have been categorised taking into account reference values in the literature. Regarding time-scale factors, authors in (Moulines and Laroche 1995) considered the values below 4 as moderate, whereas in (Kafentzis et al. 2014) only factors smaller than 2.5 received this consideration. According to this, time-scale factors have been grouped in three categories: low (<2.5), moderate (2.5,4] and high (>4). Similarly, pitch-scale values have also been categorised according to typical values (Kafentzis et al. 2014) (see Table II.1).

Finally, the statistical significance of the differences among the results has been analysed using the Wilcoxon signed-rank test for the transformation factors, and McNemar for the discretised factors.

II.4.3 Subjective evaluation

II.4.3.1 MUSHRA test setup

The subjective evaluation is based on the MUSHRA (MUltiple Stimuli with Hidden Reference and Anchor) test (ITU 2003), and it was done using the Web Audio Evaluation Tool (Jillings et al. 2015). For the evaluation, five sentences were chosen from the speech corpus so that their phonetic stress distribution could coincide with the music stressed beats. These sentences were set to music using eighth notes (the most common note value), thus getting five scores. These songs were synthesised in the 3 vocal ranges (S0, S4 and S7) and the 2 tempos (100 bmp and 50 bmp) considering the 4 US configurations under study. The obtained audios were analysed following the procedure described in Section II.4.2 to check that the transformation factors obtained for the different US configurations fit with those seen in the objective evaluation with the score dataset.

Forty-nine Spanish native speakers took part in the test. From the 30 evaluation sets (5 scores x 3 vocal ranges x 2 tempos), each user evaluated 6 sets corresponding to the 6 case scenarios (3 vocal ranges x 2 tempos). For each set, the participants were told to rate different versions of the same melody compared to a reference on a scale of 0 to 100. Specifically, they were told to evaluate the naturalness and the singing (i.e. how well sung is each stimuli regardless the naturalness). Moreover, they were instructed to give the highest score to the reference. Thus we excluded 14.5% of the sets where participants rated the hidden reference below 70.

Regarding the singing evaluation, the score performed by Vocaloid (Kenmochi 2012) was used as the upper reference and the lyrics synthesised by the TTS (i.e. not sung) as the lower anchor. Since the STS process applied is the same for all the US configurations, only MLC was included together with the hidden reference and the anchor, to minimise the fatigue of the participants. For the naturalness assessment, the upper reference was the original sentence from the corpus, i.e. natural speech, while no lower anchor was available. In this case 7 stimuli were evaluated within each set: MLC, the 3 score-driven configurations (Sxp, SxpdC and SxpdLC), Vocaloid (V) and the hidden reference.

II.5 Results and discussion

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

II.5.1 Objective evaluation

II.5.1.1 Pitch-scale and concatenation analysis

The distributions of the pitch-scale factors (α_{st}) required to convert the retrieved spoken units into singing are depicted in Fig. II.5. Their probability densities are represented by violinplots superposed on the standard boxplots, whose whiskers are set to 2nd and 98th percentiles. The percentages for the categorised pitch-scale factors and for concatenation quality can be seen in

Table II.2: Pitch-scale factor (| α_{st} |) percentages and good concatenation percentages.

			$\mid \alpha_{st} \mid$			Concat.
	Configuration	[0-4]	(4-7]	(7-12]	>12	Good
S7	S7p	94.2	4.5	1.2	0.1	33.1
	S7pdC (100 bpm)	48.0	24.0	22.9	5.1	67.5
	S7pdC (50 bpm)	47.1	24.0	23.7	5.1	68.1
	S7pdLC (100 bpm)	36.2	24.6	30.8	8.3	70.5
	S7pdLC (50 bpm)	36.2	24.1	31.2	8.4	70.4
	MLC	14.3	24.2	46.9	14.6	72.3
S4	S4p	98.6	1.2	0.3	0.0	44.2
	S4pdC (100 bpm)	69.2	19.0	11.1	0.7	70.4
	S4pdC (50 bpm)	68.7	18.9	11.7	0.7	71.2
	S4pdLC (100 bpm)	60.4	22.6	15.7	1.3	72.1*
	S4pdLC (50 bpm)	59.8	22.8	16.2	1.3	71.7
	MLC	37.7	31.4	28.6	2.3	72.3
S0	S0p	99.8	0.2	0.0	0.0	52.9
	S0pdC (100 bpm)	88.1	10.3	1.5	0.0	78.4
	S0pdC (50 bpm)	87.5	10.9	1.6	0.0	77.8
	S0pdLC (100 bpm)	82.5	14.2	3.3	0.0	76.7
	S0pdLC (50 bpm)	82.1	14.6	3.3	0.0	76.5
	MLC	68.1	25.6	6.3	0.0	72.3

Each row shows the percentages corresponding to a particular vocal range (S0, S4 or S7) and US configuration. Differences with respect to MLC are statistically significant ($p<0.01$) for all configurations, except *.

Table II.2. The values obtained at the two tempos have been included for the configurations that consider durations from the score (SxpdC and SxpdLC). However, since the differences due to the tempo are very small, only the distributions obtained with $T = 100$ bpm have been depicted in Fig. II.5.

When singing in the corpus vocal range (look at S0 scenario in Fig. II.5), the distribution of pitch-scale factors is centred around 0 semitones in all the configurations. The interval defined by the 2nd and 98th percentiles ranges from [-6.9,8.2] for MLC to [-0.8,0.9] for S0p. Therefore, the distributions are narrowed when the score is considered. This implies that the percentage of small factors (| α_{st} |< 4) increases from 68.1% in MLC until 99.8% for S0p as can be seen in Table II.2.

When singing beyond the speech corpus vocal range (S4 and S7), the distribution of MLC pitch-scale factors with S0 shifts up 4 semitones for S4 and 7 for S7 as seen in Fig. II.5. Conversely, when the score is taken into account this increase can be mitigated, or even neutralised if only pitch is considered (Sxp). However, Table II.2 shows that 72.3% of good concatenations obtained with MLC drop to 52.9% for S0p, 44.2% for S4p and 33.1% for S7p.

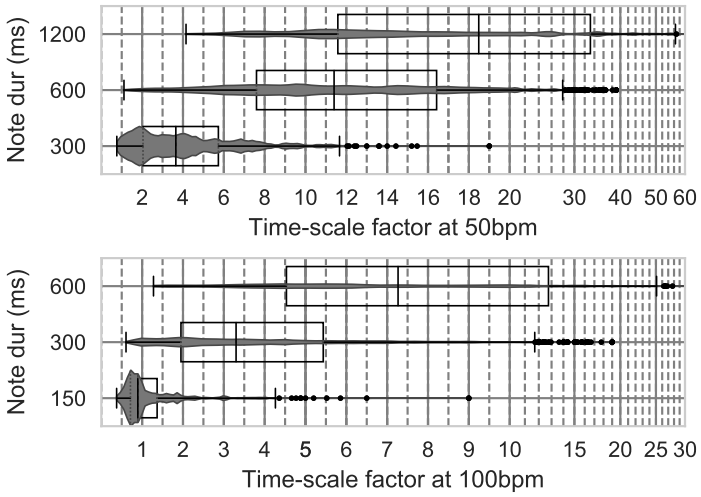


Figure II.6: Time-scale factors (β) obtained with the S4pdLC configuration at 100 bpm and 50 bpm for different note durations. Whiskers are set to minimum and 98th percentile.

By contrast, the intermediate configurations (SxpdC and SxpdLC) still allow for a statistically significant reduction of the pitch-scale factors while minimising the concatenation quality degradation. Finally, it should be noted that in the score-driven configurations the percentage of good concatenations decreases as the distance from the speech corpus vocal range midpoint increases.

II.5.1.2 Time-scale analysis

Regarding the time-scale factors, although the differences between configurations are in some cases statistically significant, they are barely relevant compared to the differences which arise from the tempo and the note values. According to this, and for the sake of clarity, the results of the intermediate configuration S4pdLC are presented for the two tempos under study, breaking them down according to the three most frequent note values: sixteenth note (𝅘𝅥𝅯), eighth note (𝅘𝅥𝅮) and quarter note (𝅘𝅥). These note values respectively account for 14.0%, 59.1% and 21.7% of the notes in the score dataset, and they last 150 ms, 300 and 600 ms for $T = 100$ bpm and double for $T = 50$ bpm.

Fig. II.6 shows the distributions of the time-scale factors (β), with the boxplot whiskers set to the minimum and the 98th percentile. The time-scale factor percentages by category are presented in the Table II.3. We can see in Fig. II.6 that when the tempo goes from 100 bpm to 50 bpm, notes doubled their duration, while time-scale factors more than doubled. This behaviour is also observed between note values within the same tempo. Similarly, Table II.3 shows that while almost all (97.8%) of the shortest notes (150 ms) can be addressed with small time-scale factors ($\beta \leq 2.5$), when moving to medium duration notes (300 ms) 15.2% of high time-scale factors emerge at 100 bpm, and 17.4% at 50 bpm. Finally, as seen in Fig. II.6 time-scale factors up to 28 can be required when singing long notes (600 ms), and even greater than 50 for notes lasting 1200 ms.

Table II.3: Time-scale factor (β) percentages obtained with the S4pdLC configuration at 100 bpm and 50 bpm for different note durations (in ms).

$T = 100$ *bpm*				$T = 50$ *bpm*			
Note	β			Note	β		
dur(ms)	≤2.5	(2.5-4]	>4	dur(ms)	≤2.5	(2.5-4]	>4
600	9.0	23.5	67.5	1200	0.0	0.2	99.8
300	55.1	29.7	15.2	600	3.5	8.4	88.1
150	97.8	1.8	0.3	300	50.1	32.6	17.4

Table II.4: Singing MUSHRA average scores and 95% confidence interval. Best values are in italics.

	Configuration	$T = 100$ bpm	$T = 50$ bpm
S7	MLC	*62 ± 6*	*61 ± 7*
S4	MLC	59 ± 6	60 ± 6
S0	MLC	60 ± 8	58 ± 7

II.5.2 Subjective evaluation

Results from the MUSHRA test are shown in Table II.4 and Table II.5. Regarding the singing assessment (see Table II.4) the US-TTS&S framework has received MUSHRA scores of around 60. Although a slight preference for the contralto vocal range (S7) can be observed (62 at 100 bpm, and 61 at 50 bpm), similar results have been obtained for all the analysed scenarios.

With regard to naturalness (see Table II.5), singing produced by the US-TTS&S framework is far from the Vocaloid (around 40 and 69, respectively). Although the differences between the US configurations are not statistically significant (according to the Wilcoxon signed-rank test), some tendencies can be observed. For instance, looking at the MUSHRA scores in Table II.5 it can be seen that Sxp configurations have received the lowest ratings in all the analysed scenarios except for S4 and $T = 50$ *bpm*. Conversely, when the concatenation cost is enabled (SxpdC and SxpdLC), the naturalness is similar to that of MLC, or in some cases slightly improved, as with S0 and S0pdC at 50 bpm, or with S4 for both configurations and the two tempos.

II.5.3 Discussion

The experiments have been designed to evaluate the proposal through a proof-of-concept implementation. From the objective tests, it can be observed that large time transformation factors arise when dealing with medium duration notes (300 ms), but especially when long and very long notes (600 ms and 1200 ms) are present in the song (see Fig. II.6). This result is in concordance with the corpus characteristics, which contains almost no vowels longer than 200 ms (see Fig. II.4a). As a consequence, we can conclude that score-driven US configurations hardly impact on the time-scaling requirements.

Table II.5: Naturalness MUSHRA average scores and 95% confidence interval. Best values achieved by the proposed system in each scenario are in italics.

	Configuration	T = 100 bpm	$T = 50$ bpm
S7	V	74 ± 6	70 ± 6
	S7pdLC	41 ± 5	*44 ± 6*
	S7pdC	39 ± 6	43 ± 6
	S7p	36 ± 5	40 ± 6
	MLC	*42 ± 5*	*44 ± 5*
S4	V	69 ± 7	67 ± 7
	S4pdLC	*42 ± 6*	38 ± 6
	S4pdC	39 ± 6	*41 ± 6*
	S4p	35 ± 6	38 ± 6
	MLC	38 ± 6	37 ± 6
S0	V	66 ± 7	70 ± 6
	S0pdLC	*44 ± 5*	38 ± 4
	S0pdC	*44 ± 5*	*42 ± 6*
	S0p	41 ± 6	35 ± 6
	MLC	*44 ± 6*	39 ± 5

Regarding pitch-scaling, the obtained moderate transformation factors required to sing in the speech corpus vocal range (S0) are consistent with the overlap between the $F0$ distribution from the score dataset and that from the corpus vowels (see Fig. II.4). Conversely, when moving towards a contralto singer vocal range (S7), the overlap between $F0$ distributions is significantly reduced as it can be seen in Fig. II.4. Thus, even though Sxp configurations are able to find almost all the vowels close to the desired pitch, it becomes harder to find units that also join adequately and meet the other target specifications (SxpdC and SxpdLC). This can be observed in the last column of Table II.2, in the decreasing percentage of good concatenations when moving away from the corpus vocal range. Hence, although the score-driven US strategies have been proved helpful to reduce the pitch-scaling requirements, their effectiveness could be higher if a larger speech corpus with a greater coverage was available.

From the perceptual tests, a slight preference for singing in an actual singer vocal range (S7) has been observed (see Table II.4). However, this preference is not significant with respect to the other vocal ranges under study (with MUSHRA scores of around 60). With regard to naturalness (see Table II.5), the ratings achieved by the proof-of-concept with respect to natural speech are significantly different to those obtained by Vocaloid (with MUSHRA ratings around 40 and 69, respectively). Nevertheless, this is not surprising since Vocaloid is a commercial high-quality singing synthesiser exclusively designed for this purpose, which uses databases including diphones, sustained vowels and optionally triphones, sung by professional singers in several pitches to cover their vocal range (Kenmochi 2012). Conversely, the proposal has to deal with the spoken units available in the corpus, which are low-pitched and very short compared to what could be found in a singing database. Therefore, converting them into singing involves high demanding transformations factors as seen in the objective evaluation.

In this context, it has also been observed that the substantial pitch-scale factors reduction achieved by the score-driven US configurations has had a small impact on the naturalness, obtaining scores similar to those received by the text-driven US configuration. Besides the aforementioned restrictions due to the corpus size,this could be explained by the impossibility of relaxing the time-scale requirements. This may be important, considering that the ability to reproduce the behaviour of sustained vowels is known to be essential in singing synthesis (Kenmochi 2012).

Finally, it is worth mentioning that the validation of the proposal has been carried out with a specific speech corpus on a US-TTS system, since this approach enabled the study of the STS transformation factors required to produce singing from speech. Nevertheless, other corpus and adjustments of the cost function weights could be considered, and even other corpus-based approaches, such as statistical parametric speech synthesis using HMM or DNN.

II.6 Conclusions

This work has proposed a synthesis framework that provides singing capabilities to a US-TTS system from neutral speech, through the integration of Speech-To-Singing (STS) conversion. The proposal has been evaluated by means of a proof-of-concept implementation on a 2.6h Spanish neutral speech corpus, considering different vocal ranges and tempos and studying diverse text-driven and score-driven US configurations.

Results show that high demanding STS transformation factors are required to sing beyond the corpus vocal range and/or when notes longer than 150 ms are present. However, the pitch-scale factors can be reduced by considering score-driven US configurations. Conversely, the time-scale requirements can not be reduced because of the short length of the vowels available in the corpus.

Regarding the subjective evaluation, text-driven and score-driven US configurations have obtained a similar naturalness in all the analysed scenarios, with MUSHRA scores around 40. Although these values are far from those of Vocaloid (around 69), the obtained singing ratings around 60 validate the capability of the framework to address eventual singing needs.

The obtained results encourage us to continue working on the proposal to improve the performance of the system. To this aim, the focus will be placed on the generation of long sustained vowels, exploring advanced time-scale and spectral transformation strategies, and incorporating vibrato to the singing expression control generation module. Furthermore, other signal processing techniques could be considered for the transformation module to better cope with the challenge of generating singing from neutral speech.

Abbreviations

TTS: Text-to-speech; STS: Speech-To-Singing; US: Unit Selection; HMM: Hidden Markov Model; DNN: Deep Neural Network; US-TTS&S: Unit Selection based Text-to-Speech-and-Singing; NLP: Natural Language Processing; OLA: overlap and add; CBR: Case-Based

Reasoning; TD-PSOLA: Time-Domain Pitch Synchronous Overlap and Add; MUSHRA: MUltiple Stimuli with Hidden Reference and Anchor

Availability of data and materials

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

Competing interests

The authors declare that they have no competing interests.

Author's contributions

All authors participated in the design of both the framework and its evaluation. MF and JCS carried out the proof-of-concept implementation, and MF conducted the experiments. The paper was mostly written by MF and FA. All authors have read and approved the final manuscript.

Acknowledgements

We want to thank Yannis Stylianou for his advice regarding the HNM model, Jordi Bonada and Merlijn Blaauw for the Vocaloid samples, Lisa Kinnear for English proofreading and the people that took the perceptual test.

Funding

This work has been partially supported by the Agencia Estatal de Investigación (AEI) and FEDER, EU, through project GENIOVOX TEC2016-81107-P. Marc Freixes received the support of the European Social Fund (ESF) and the Catalan Government (SUR/DEC) with the FI grant No. 2016FI_B2 00094. Francesc Alías acknowledges the support from the Obra Social "La Caixa" under grants ref. 2018-URL-IR1rQ-021 and 2018-URL-IR2nQ-029.

References

Alías, Francesc, Formiga, Lluís, and Llorá, Xavier (May 2011). "Efficient and reliable perceptual weight tuning for unit-selection text-to-speech synthesis based on active interactive genetic algorithms: A proof-of-concept". In: *Speech Communication* vol. 53, no. 5, pp. 786–800.

Alías, Francesc, Sevillano, Xavier, Socoró, Joan Claudi, and Gonzalvo, Xavier (2008). "Towards high quality next-generation Text-to-Speech synthesis: a Multidomain approach by automatic domain classification". In: *IEEE Transactions on Audio, Speech and Language Processing* vol. 16, no. 7, pp. 1340–1354.

Blaauw, Merlijn and Bonada, Jordi (2017). "A Neural Parametric Singing Synthesizer Modeling Timbre and Expression from Natural Songs". In: *Applied Sciences* vol. 7, no. 12, p. 1313.

Blanco, Eder del, Hernaez, Inma, Navas, Eva, Sarasola, Xabier, and Erro, Daniel (Sept. 2016). "Bertsokantari: a TTS Based Singing Synthesis System". In: *Interspeech*. San Francisco, CA, USA: ISCA, pp. 1240–1244.

Bonada, Jordi and Serra, Xavier (Mar. 2007). "Synthesis of the Singing Voice by Performance Sampling and Spectral Models". In: *IEEE Signal Processing Magazine* vol. 24, no. 2, pp. 67–79.

Calzada Defez, Àngel and Socoró, Joan Claudi (Dec. 2013). "Voice Quality Modification Using a Harmonics Plus Noise Model". In: *Cognitive Computation* vol. 5, no. 4, pp. 473–482.

Cen, Ling, Dong, Minghui, and Chan, Paul (Mar. 2012). "Template-based personalized singing voice synthesis". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Kyoto, Japan: IEEE, pp. 4509–4512.

Chowning, John M. (1980). "Computer synthesis of the singing voice". In: *Sound Generation in Winds, Strings, Computers*. Stockholm, Sweden: Royal Swedish Academy of Music, pp. 4–13.

Cook, Perry R. (1993). "SPASM, a Real-Time Vocal Tract Physical Model Controller; and Singer, the Companion Software Synthesis System". In: *Computer Music Journal* vol. 17, no. 1, pp. 30–44.

Cook, Perry R. (1996). "Singing voice synthesis: History, current work, and future directions". In: *Computer Music Journal* vol. 20, no. 3, pp. 38–46.

Cook, Perry R., Kamarotos, Dimitris, Diamantopoulos, Taxiarchis, and Philippis, Giorgos (1993). "IGDIS (Instrument for Greek Diction and Singing): A Modern Greek Text to Speech/Singing Program for the SPASM/Singer Instrument". In: *International Computer Music Conference (ICMC)*. Tokyo, Japan: ICMA, pp. 387–389.

D'Alessandro, Nicolas, Babacan, Onur, Bozkurt, Baris, Dubuisson, Thomas, Holzapfel, Andre, Kessous, Loic, Moinet, Alexis, and Vlieghe, Maxime (2008). "RAMCESS 2.X framework—expressive voice analysis for realtime and accurate synthesis of singing". In: *Journal on Multimodal User Interfaces* vol. 2, pp. 133–144.

D'Alessandro, Nicolas, Sebbe, Raphael, Bozkurt, Baris, and Dutoit, Thierry (Sept. 2005). "MaxMBROLA: A Max/MSP MBROLA-Based Tool for Real-Time Voice Synthesis". In: *13th European Signal Processing Conference (EUSIPCO)*. Antalya, Turkey: EURASIP.

Dong, Minghui, Lee, S. W., Li, Haizhou, Chan, Paul, Peng, Xuejian, Ehnes, Jochen Walter, and Huang, Dongyan (Sept. 2014). "I2R Speech2Singing Perfects Everyone's Singing". In: *Interspeech*. Singapore: ISCA, pp. 2148–2149.

Erro, Daniel and Moreno, Asunción (2008). "Intra-lingual and cross-lingual voice conversion using harmonic plus stochastic models". PhD thesis. Universitat Politecnica de Catalunya.

Erro, Daniel, Moreno, Asunción, and Bonafonte, Antonio (Aug. 2007). "Flexible Harmonic/Stochastic Speech Synthesis". In: *6th ISCA Workshop on Speech Synthesis (SSW6)*. Bonn, Germany: ISCA, pp. 194–199.

Festival (2016). *The Festival speech synthesis system*.

Feugère, Lionel, D'Alessandro, Christophe, Doval, Boris, and Perrotin, Olivier (Dec. 2017). "Cantor Digitalis: chironomic parametric synthesis of singing". In: *EURASIP Journal on Audio, Speech, and Music Processing* vol. 2017, p. 2.

Flinger (2001). *Flinger: Festival Singer*.

Formiga, Lluís, Trilla, Alexandre, Alías, Francesc, Iriondo, Ignasi, and Socoró, Joan Claudi (Nov. 2010). "Adaptation of the URL-TTS system to the 2010 Albayzin evaluation campaign". In: *FALA 2010, Jornadas en Tecnología del Habla and Iberian SLTech Workshop*. Vigo, Spain: ISCA IL-SIG, pp. 363–370.

Francois, Hélene and Boëffard, Olivier (May 2002). "The greedy algorithm and its application to the construction of a continuous speech database". In: *3rd International Conference on Language Resources and Evaluation (LREC)*. Las Palmas de Gran Canaria, Spain: ELRA, pp. 1420–1426.

Freixes, Marc, Socoró, Joan Claudi, and Alías, Francesc (2016). "Adding Singing Capabilities to Unit Selection TTS Through HNM-Based Conversion". In: *Advances in Speech and Language Technologies for Iberian Languages: Third International Conference, IberSPEECH*. Vol. 10077 LNAI. Lisbon, Portugal: Springer International Publishing, pp. 33–43.

Fridin, Marina (Jan. 2014). "Storytelling by a kindergarten social assistive robot: A tool for constructive learning in preschool education". In: *Computers & Education* vol. 70, pp. 53–64.

Huber, Stefan and Roebel, Axel (Sept. 2015). "On glottal source shape parameter transformation using a novel deterministic and stochastic speech analysis and synthesis system". In: *Interspeech*. Dresden, Germany: ISCA, pp. 289–293.

Hunt, A.J. and Black, Alan W. (May 1996). "Unit selection in a concatenative speech synthesis system using a large speech database". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Vol. 1. Atlanta, GA, USA: IEEE, pp. 373–376.

ITU, Recommendation (2003). "ITU-R BS.1534-1: Method for the subjective assessment of intermediate quality level of coding systems". In: *International Telecommunication Union*.

Jibb, Lindsay A., Birnie, Kathryn A., Nathan, Paul C., Beran, Tanya N., Hum, Vanessa, Victor, J. Charles, and Stinson, Jennifer N. (Sept. 2018). "Using the MEDiPORT humanoid robot to reduce procedural pain and distress in children with cancer: A pilot randomized controlled trial". In: *Pediatric Blood & Cancer* vol. 65, no. 9, e27242.

Jillings, Nicholas, De Man, Brecht, Moffat, David, and Reiss, Joshua D. (July 2015). "Web audio evaluation tool: A browser-based listening test environment". In: *12th International Conference in Sound and Music Computing (SMC)*. Maynooth, Ireland: SMC network, pp. 147–152.

Kafentzis, George P., Degottex, Gilles, Rosec, Olivier, and Stylianou, Yannis (May 2014). "Pitch modifications of speech based on an adaptive Harmonic Model". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Florence, Italy: IEEE, pp. 7924–7928.

Karabetsos, Sotiris, Tsiakoulis, Pirros, Chalamandaris, Aimilios, and Raptis, Spyros (2010). "One-class classification for spectral join cost calculation in unit selection speech synthesis". In: *IEEE Signal Processing Letters* vol. 17, no. 8, pp. 746–749.

Kawahara, Hideki, Masuda-Katsuse, Ikuyo, and Cheveigné, Alain de (Apr. 1999). "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds". In: *Speech Communication* vol. 27, no. 3-4, pp. 187–207.

Kelly, John L. and Lochbaum, Carol C. (Aug. 1962). "Speech synthesis". In: *4th International Congress on Acoustics (ICA)*. Copenhagen, Denmark: ICA, G42.

Kenmochi, Hideki (Mar. 2012). "Singing synthesis as a new musical instrument". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Kyoto, Japan: IEEE, pp. 5385–5388.

Khosla, Rajiv, Nguyen, Khanh, and Chu, Mei-Tai (June 2017). "Human Robot Engagement and Acceptability in Residential Aged Care". In: *International Journal of Human–Computer Interaction* vol. 33, no. 6, pp. 510–522.

King, Simon (June 2014). "Measuring a decade of progress in Text-to-Speech". In: *Loquens* vol. 1, no. 1, e006.

Lee, S. W., Wu, Zhizheng, Dong, Minghui, Tian, Xiaohai, and Li, Haizhou (Sept. 2014). "A Comparative Study of Spectral Transformation Techniques for Singing Voice Synthesis". In: *Interspeech*. Singapore: ISCA, pp. 2499–2503.

Li, Jinlong, Yang, Hongwu, Zhang, Weizhao, and Cai, Lianhong (2011). "A lyrics to singing voice synthesis system with variable timbre". In: *Applied Informatics and Communication. ICAIC 2011. Communications in Computer and Information Science*. Vol. 225, pp. 186–193.

Macon, Michael W., Jensen-Link, Leslie, Oliverio, James, Clements, Mark A., and George, E. Bryan (Apr. 1997). "A singing voice synthesis system based on sinusoidal modeling". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Vol. 1. Munich, Germany: IEEE, pp. 435–438.

Montaño, Raúl and Alías, Francesc (Dec. 2016). "The role of prosody and voice quality in indirect storytelling speech: Annotation methodology and expressive categories". In: *Speech Communication* vol. 85, pp. 8–18.

Moulines, Eric and Laroche, Jean (Feb. 1995). "Non-parametric techniques for pitch-scale and time-scale modification of speech". In: *Speech Communication* vol. 16, no. 2, pp. 175–205.

Nichols, Eric, Morris, Dan, Basu, Sumit, and Raphael, Christopher (Oct. 2009). "Relationships Between Lyrics and Melody in Popular Music". In: *International Symposium on Music Information Retrieval (ISMIR)*. Kobe, Japan: ISMIR, pp. 471–476.

Oord, Aaron van den, Dieleman, Sander, Zen, Heiga, Simonyan, Karen, Vinyals, Oriol, Graves, Alex, Kalchbrenner, Nal, Senior, Andrew, and Kavukcuoglu, Koray (Sept. 2016). "WaveNet: A Generative Model for Raw Audio". In: pp. 1–15.

Röbel, Axel and Fineberg, Joshua (Aug. 2007). "Speech to chant transformation with the phase vocoder". In: *Interspeech*. Antwerp, Belgium: ISCA, pp. 4007–4008.

Rodet, Xavier, Potard, Yves, and Barriere, Jean-baptiste (1984). “The CHANT Project: From the Synthesis of the Singing Voice to Synthesis in General”. In: *Computer Music Journal* vol. 8, no. 3, pp. 15–31.

Roebel, Axel, Huber, Stefan, Rodet, Xavier, and Degottex, Gilles (Mar. 2012). “Analysis and modification of excitation source characteristics for singing voice synthesis”. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Kyoto, Japan: IEEE, pp. 5381–5384.

Saitou, Takeshi, Goto, Masataka, Unoki, Masashi, and Akagi, Masato (Oct. 2007). “Speech-to-Singing Synthesis: Converting Speaking Voices to Singing Voices by Controlling Acoustic Features Unique to Singing Voices”. In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. New Paltz, NY, USA: IEEE, pp. 215–218.

Bensaya, Pablo (2018). *Cantos Infantiles Educativos, de Pablo Bensaya.*

Stylianou, Yannis (2001). “Applying the harmonic plus noise model in concatenative speech synthesis”. In: *IEEE Transactions on Speech and Audio Processing* vol. 9, no. 1, pp. 21–29.

Sundberg, Johan (2006). “The KTH synthesis of singing”. In: *Advances in Cognitive Psychology* vol. 2, no. 2-3, pp. 131–143.

Sundberg, Johan and Bauer-Huppmann, Julia (May 2007). “When Does a Sung Tone Start?” In: *Journal of Voice* vol. 21, no. 3, pp. 285–293.

Taylor, Paul (2009). *Text-to-Speech Synthesis*. Cambridge, UK: Cambridge University Press, p. 626.

Umbert, Marti, Bonada, Jordi, Goto, Masataka, Nakano, Tomoyasu, and Sundberg, Johan (Nov. 2015). “Expression Control in Singing Voice Synthesis: Features, approaches, evaluation, and challenges”. In: *IEEE Signal Processing Magazine* vol. 32, no. 6, pp. 55–73.

Uneson, Marcus (2002). “Outlines of Burcas - A simple MIDI-to-singing voice synthesis system”. In: *Fonetik*. Vol. 44. 1, pp. 133–136.

Vijayan, Karthika, Dong, Minghui, and Li, Haizhou (Dec. 2017). “A dual alignment scheme for improved speech-to-singing voice conversion”. In: *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. Kuala Lumpur, Malaysia: IEEE, pp. 1547–1555.

Vijayan, Karthika, Li, Haizhou, and Toda, Tomoki (Jan. 2019). “Speech-to-Singing Voice Conversion: The Challenges and Strategies for Improving Vocal Conversion Processes”. In: *IEEE Signal Processing Magazine* vol. 36, no. 1, pp. 95–102.

Wang, Yuxuan et al. (Aug. 2017). “Tacotron: Towards End-to-End Speech Synthesis”. In: *Interspeech*. Stockholm, Sweden: ISCA, pp. 4006–4010.

Wood, Luke, Dautenhahn, Kerstin, Robins, Ben, and Zaraki, Abolfazl (Aug. 2017). “Developing child-robot interaction scenarios with a humanoid robot to assist children with autism in developing visual perspective taking skills”. In: *26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. Lisbon, Portugal: IEEE, pp. 1055–1060.

Yamagishi, Junichi, Veaux, Christophe, King, Simon, and Renals, Steve (2012). "Speech synthesis technologies for individuals with vocal disabilities: Voice banking and reconstruction". In: *Acoustical Science and Technology* vol. 33, no. 1, pp. 1–5.

Authors' addresses

Marc Freixes GTM – Grup de Recerca en Tecnologies Mèdia, La Salle - Universitat Ramon Llull
Quatre Camins, 30, 08022 Barcelona, Spain
marc.freixes@salle.url.edu

Paper III

Glottal Source Contribution to Higher Order Modes in the Finite Element Synthesis of Vowels

Marc Freixes, Marc Arnela, Joan Claudi Socoró, Francesc Alías, Oriol Guasch

Published in *Applied Sciences - Special Issue "IberSPEECH 2018: Speech and Language Technologies for Iberian Languages"*, October 2019, volume 9(21), pp. 4535. DOI: 10.3390/app9214535.

III

Abstract

Articulatory speech synthesis has long been based on one-dimensional (1D) approaches. They assume plane wave propagation within the vocal tract and disregard higher order modes that typically appear above 5 kHz. However, such modes may be relevant in obtaining a more natural voice, especially for phonation types with significant high frequency energy (HFE) content. This work studies the contribution of the glottal source at high frequencies in the 3D numerical synthesis of vowels. The spoken vocal range is explored using an LF (Liljencrants–Fant) model enhanced with aspiration noise and controlled by the R_d glottal shape parameter. The vowels [ɑ], [i], and [u] are generated with a finite element method (FEM) using realistic 3D vocal tract geometries obtained from magnetic resonance imaging (MRI), as well as simplified straight vocal tracts of a circular cross-sectional area. The symmetry of the latter prevents the onset of higher order modes. Thus, the comparison between realistic and simplified geometries enables us to analyse the influence of such modes. The simulations indicate that higher order modes may be perceptually relevant, particularly for tense phonations (lower R_d values) and/or high fundamental frequency values, $F0$s. Conversely, vowels with a lax phonation and/or low F0s may result in inaudible HFE levels, especially if aspiration noise is not considered in the glottal source model.

III.1 Introduction

Voice can be generated simulating acoustic wave propagation within the vocal tract. For years only plane waves were considered, which allowed the use of 1D vocal tract models to produce a voice of fairly good quality (see e.g., Story 2013; Birkholz 2013). Nonetheless,

the accuracy of 1D models is limited up to about 4–5 kHz, depending on the generated sound (see e.g., Arnela et al. 2019 which compares vowels and diphthongs generated in 1D and 3D). Beyond this frequency, higher order modes also exist, resulting in resonances and anti-resonances that cannot be predicted in 1D and which strongly modify the high frequency energy (HFE) content of the spectrum (Blandin et al. 2015; Arnela et al. 2016b). Until now, however, little attention has been paid to the high frequency range. An exception is found in some recent works that point out that HFE may be important for voice quality, speech localisation, speaker recognition, and intelligibility (see Monson et al. 2014 and references therein).

Plane wave propagation along the vocal tract midline is not a constraint for 3D acoustic models. Some examples of the latter can be found, for instance, in (Vampola et al. 2008) where the finite element method (FEM) was used to study the production of Czech vowels using 3D vocal tracts, reconstructed from magnetic resonance imaging (MRI) data. In (Takemoto et al. 2010), a finite-difference time-domain method was adopted to analyse the MRI-based vocal tracts of Japanese vowels. Moreover, the results of the simulations were validated through experiments performed in physical models constructed from the same MRI data. Similarly, in (Arnela et al. 2016a), measurements on 3D-printed mechanical replicas presented very close results to those from 3D FEM acoustic simulations on MRI-based vocal tracts.

Nevertheless, the use of a 3D acoustic model does not necessarily entail the propagation of higher-order modes. For instance, these will rarely appear in a straight, axisymmetric vocal tract excited at the glottis, as observed in (Arnela et al. 2016b). In fact, several geometric simplifications were analysed in (Arnela et al. 2016b) which preserved the cross-sectional areas of the vocal tract but introduced modifications in their cross-sectional shapes and midline curvature. Results showed a similar behaviour for the analysed configurations in frequencies below 4–5 kHz but very large deviations beyond that value. This highlights the limits of the plane wave assumption and also shows that changes in the vocal tract shape modify the HFE content. Nonetheless, there are other important factors that must be considered to determine the HFE content of a voice such as phonation type. In (Monson et al. 2011) for instance, loud and soft phonations of sustained vowels showed significant differences in HFE content. Moreover, results showed that modifications of HFE levels are more easily detected by listeners in a loud phonation case.

In this work, we study the contribution of the glottal source excitation in the 3D numerical synthesis of vowels, paying special attention to HFE content. 3D realistic vocal tracts for vowels [ɑ], [i], and [u] were considered for this purpose, as well as their simplified counterparts consisting of straight ducts of varying circular cross-section (Arnela et al. 2016b). The latter allowed us to mitigate the onset of higher order modes and thus examine their influence on HFE by comparison with the 3D realistic outputs. Vocal tract impulse responses have been computed from FEM simulations in the time domain (Arnela and Guasch 2013). Vowels have finally been synthesised by convolving impulse responses with the desired glottal source excitations. An LF (Liljencrants–Fant) model (Fant et al. 1985) enhanced with aspiration noise has been employed to generate the latter. Although this model does not take into

account the interaction between the vocal tract and the vocal folds (Murtola et al. 2018; Erath et al. 2013), it has proven useful to explore the phonatory tense-lax continuum (Murphy et al. 2017) by controlling the R_d glottal shape parameter (Fant 1995). The R_d parameter has thus been incorporated in the LF model and used to examine different phonation types, ranging from a lax to a tense phonation. Moreover, the influence of the fundamental frequency $F0$ on HFE content has also been examined. Several plausible combinations of R_d and $F0$ were considered, thus covering to a large extent the phonation range for male speech. Finally, aspiration noise was also evaluated to study its impact on HFE levels. A preliminary version of this work was presented in (Freixes et al. 2018).

The paper is structured as follows. Section III.2 details the methodology we propose to study the production of vowels [ɑ], [i], and [u] for different phonation types and for both, the realistic and simplified vocal tract geometries. Computations are carried out and the results are analysed and discussed in Section III.3. Finally, conclusions and future work close the paper in Section III.4.

III.2 Methodology

Figure III.1 represents the process used to synthesise the different versions of the vowels [ɑ], [i], and [u]. These were obtained by convolving the glottal source signals with the impulse responses of the vocal tract geometries. As explained in the Introduction, the realistic vocal tract geometries from (Arnela et al. 2016b) were used as well as their simplified counterparts with straight mid-line and circular cross-sections (see Section III.2.1). With regard to the impulse responses $h(t)$, those were computed using the 3D FEM acoustic model

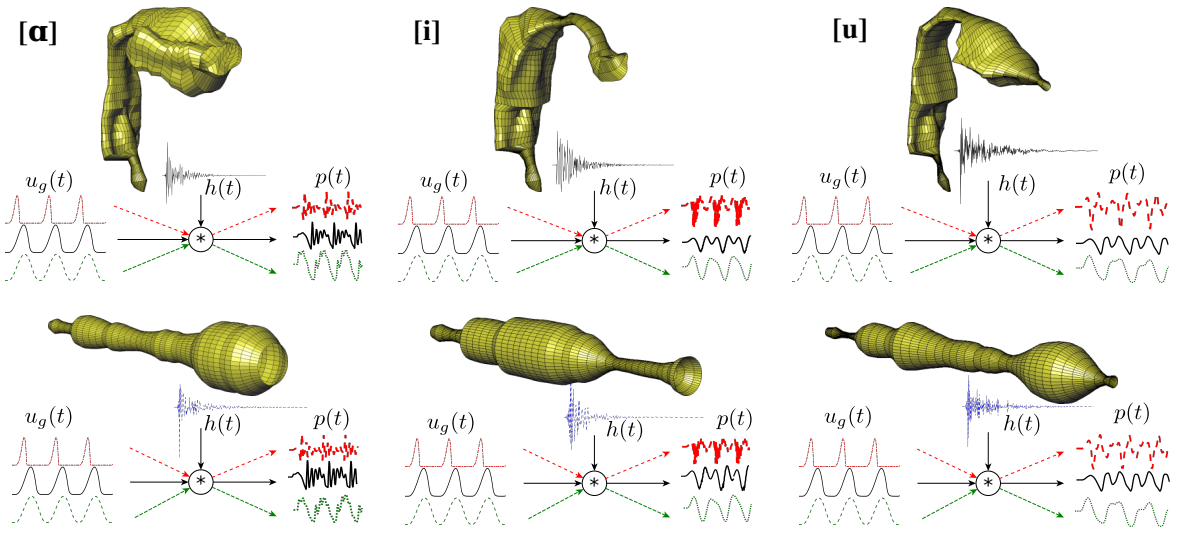


Figure III.1: Synthesis of vowels [ɑ], [i], and [u] with realistic vocal tract geometries (above) and their simplified counterparts of circular cross-sections set in a straightened midline (below). The output pressure signal $p(t)$ is computed as the convolution of the glottal source $u_g(t)$ with the vocal tract impulse response $h(t)$ obtained from a 3D FEM (finite element method) simulation. Three phonation type examples are represented in the figure: Tense (dashed red line), modal (solid black line), and lax (dotted green line).

detailed in Section III.2.2. Besides, the glottal source signals $u_g(t)$ were generated by means of a R_d controlled LF model enhanced with aspiration noise, described in Section III.2.3. The synthesised acoustic pressure $p(t)$ for each vowel was finally analysed according to the methodology in Section III.2.4.

III.2.1 Vocal Tract Geometries

The two vocal tract representations (realistic and simplified) of vowels [ɑ], [i], and [u] used in this work are depicted in Figure III.1. They were generated in (Arnela et al. 2016b) from adapted versions of the 3D complex vocal tract geometries reconstructed from MRI data in (Aalto et al. 2014). Neither the realistic nor the simplified vocal tracts include the subglottal tube, lips, and face (see Arnela et al. 2013 and Arnela et al. 2016a for the influence of the head and lips on simulations) or the side branches, such as the piriform fossae and valleculae (see e.g., Takemoto et al. 2010; Takemoto et al. 2013 for their acoustic effects).

In fact, the realistic representations consist of cross-sections extracted from the adapted MRI-based vocal tract geometries. An adaptive grid approach, which considers the cross-sections as being perpendicular to the vocal tract midline, was used for that purpose. The cross-sections were then linearly interpolated to reconstruct a 3D vocal tract geometry. It was shown in (Arnela et al. 2016b) that such types of geometries correctly mimic the behaviour of MRI-based vocal tracts without side branches.

The simplified representations involve strong additional modifications of realistic vocal tracts. First, the shape of each cross-section was replaced with a circle of an equivalent area. Second, the resulting cross-sections were set in a straightened vocal tract midline, obtained by computing the Euclidean distance between the centers of the cross-sections (sagittal variations of the cross-section centers were excluded from the computations to avoid an artificial lengthening of the vocal tract Story et al. 1996). Linear interpolation was then applied to obtain the 3D vocal tract geometry (see Arnela et al. 2016b).

III.2.2 Vocal Tract Impulse Response

The impulse response of each vocal tract was obtained from the time-domain FEM simulations (Arnela and Guasch 2013). The propagation of acoustic waves within the 3D vocal tracts was provided by the FEM solution to the acoustic wave equation:

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

where $p(\mathbf{x}, t)$ stands for the acoustic pressure, ∂_{tt}^2 for the second order time derivative, and c_0 for the speed of sound. c_0 was set to the usual value of 350 m/s. An exterior domain was included to let waves emanate from the mouth and account in this way for radiation losses. A perfectly matched layer (PML) was imposed on the computational domain boundaries to prevent wave reflections. Wall losses were considered by prescribing a boundary admittance coefficient μ on the vocal tract walls which was set to $\mu = 0.005$. Sound waves were generated within the vocal tract imposing a volume velocity $u_g(t)$ on the glottal cross-sectional area.

Specifically, the following Gaussian pulse was used:

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

with $T_{gp} = 0.646/f_c$ and $f_c = 10$ kHz.

Numerical simulations were performed with a sampling frequency of $f_s = 8000$ kHz. This unusually large value was selected to ensure the stability of the explicit discrete time scheme used to solve the wave Equation (III.1). Time events of 20 ms were simulated, tracking the acoustic pressure, $p_0(t)$ at a mesh node located 4 cm away from the mouth exit. The vocal tract transfer function $H(f)$ was then obtained as:

$$H(f) = \frac{P_o(f)}{U_g(f)}, \tag{III.3}$$

with $P_o(f)$ and $U_g(f)$ respectively being the Fourier transforms of $p_o(t)$ and $u_g(t)$. Note that this compensates for the slight spectral decay introduced by the Gaussian pulse. $H(f)$ was computed up to 12 kHz in order to generate speech at 24 kHz. This sampling frequency allowed us to cover the whole 8 kHz octave band, in which the HFE levels would be computed.

Figure III.2 shows the computed vocal tract transfer functions $H(f)$ of [ɑ], [i], and [u] for the realistic and simplified geometries. Observe that below 5 kHz the two representations behaved very similarly, whereas above that frequency strong differences emerged. This is consistent with the results presented in (Arnela et al. 2016b), where it was observed that plane wave propagation, which dominates below 5 kHz, is barely affected by the cross-sectional shape and vocal tract bending. Beyond that limit, however, higher order modes also propagate and play a significant role in the realistic configurations. In contrast, radial symmetry prevents the onset of most higher order modes in the simplified vocal tracts for the examined frequency range (Blandin et al. 2015; Arnela et al. 2016b). As observed in Figure III.2, the depicted vocal tract transfer functions show an almost flat global trend in contrast to the spectral characteristics of speech because they do not include the effect of the glottal source.

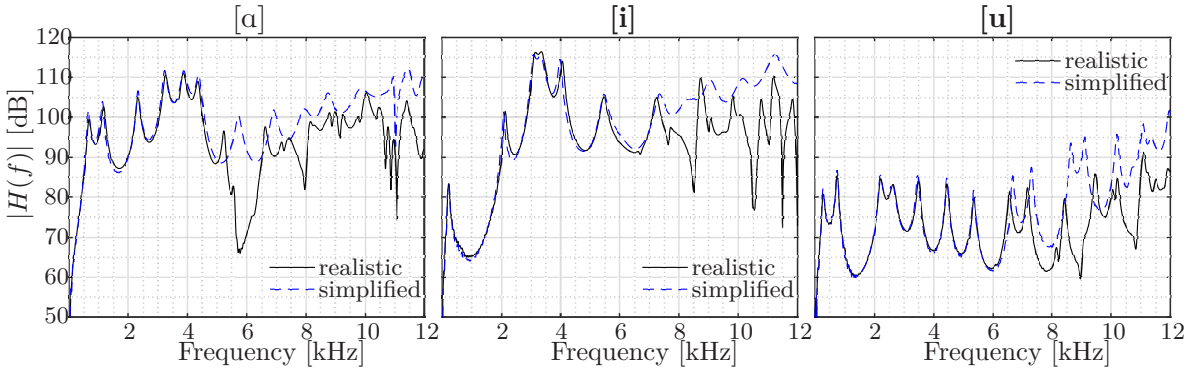


Figure III.2: Vocal tract transfer function magnitude $|H(f)|$ of vowels [ɑ], [i], and [u] for realistic and simplified vocal tract geometries.

The vocal tract impulse responses $h(t)$ were finally obtained from the inverse Fourier transform of the vocal tract transfer functions $H(f)$ (see Figure III.1).

III.2.3 Voice Source Signal

Voice source signals were generated according to the LF model (Fant et al. 1985). Specifically, Kawahara's implementation (Kawahara et al. 2017) was chosen to obtain aliasing-free glottal flow derivative waveforms $u'_g(t)$. The shape of a glottal pulse is controlled by the parameters T_p, T_e, T_a, T_c, and T_0 (see Figure III.3). However, this control can be simplified as described in the transformed LF model (Fant 1995). The latter reduces parameter redundancy in the glottal pulse description. To this end, a global waveshape parameter, R_d, is introduced as:

$$R_d = \frac{T_d}{T_0}\frac{1}{110} = \frac{U_0}{E_e}\frac{F0}{110} \quad , \tag{III.4}$$

where T_d is the declination time, T_0 the period, and $F0$ the fundamental frequency. The declination time T_d corresponds to the quotient between the glottal flow peak U_0 and the negative amplitude of the differentiated glottal flow E_e. The scale factor was chosen so as to make the numerical value of R_d the same as the declination time in seconds for $F0 = 110$ Hz (Fant 1995). The glottal shape parameter R_d was integrated into Kawahara's implementation, thereby allowing us to simulate from a tense, very adducted phonation ($R_d = 0.3$) to a lax, very abducted phonation ($R_d = 2.7$). T_p, T_e, and T_a are derived from R_d according to the equations in (Fant 1995) and T_c is set to T_0. The glottal flow $u_g(t)$ is computed by performing the cumulative integration of $u'_g(t)$ using the composite trapezoidal rule (Davis and Rabinowitz 2007).

Furthermore, the voice source model was extended to incorporate aspiration noise, $S_{AH}(t)$, which is added to the glottal flow $u_g(t)$. To this end, the method presented in (Gobl 2006) was implemented. This consists of automatically generating the temporal dynamics of $S_{AH}(t)$ according to the voice source parameters as follows:

$$S_{AH}(t) = AH\ E_e^{1.35} F0^{1.05}\ n(t)\sqrt{\frac{U_{ac}}{U_0}u_g(t) + U_{dc}}, \tag{III.5}$$

where $U_{ac} = (379/R_d) - 91$, $U_{dc} = 83T_d + 34$ and $T_d = 110\ T_0 R_d$, according to (Gobl 2006). The noise amplitude factor AH was perceptually adjusted to 3×10^{-14}. The noise signal $n(t)$ was generated by filtering white Gaussian noise with a 2nd order Butterworth bandpass

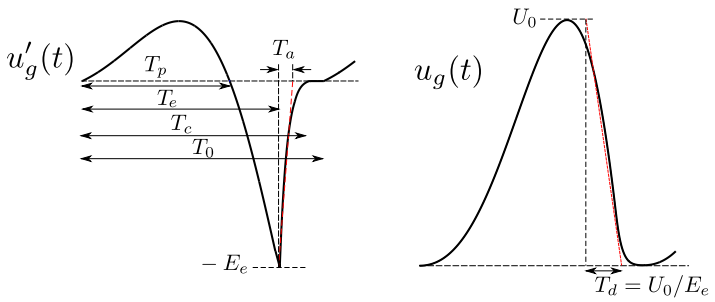


Figure III.3: Glottal flow $u_g(t)$ and its time derivative $u'_g(t)$ according to the LF (Liljencrants–Fant) model (Fant et al. 1985). T_p is the rise time, T_e is the duration of the open phase, T_a corresponds to the effective duration of the return phase, T_c is the location of the complete closure, T_d is the declination time, and T_0 is the period. U_0 is the peak of the glottal flow and E_e corresponds to the negative amplitude of the differentiated glottal flow.

filter with cutoff frequencies of 300 and 3000 Hz as in (Story 2013). Finally, a SoX resampling (http://sox.sourceforge.net/SoX/Resampling) was incorporated to adapt the glottal flow signals originally generated at 44,100 Hz to the sampling rate at which the speech signals were synthesised (24 kHz).

The glottal flow signals generated for this work cover the R_d range $[0.3, 2.7]$ (Fant 1995), considering 49 logarithmically spaced values of R_d (24 steps from 0.3 to 1 and 24 more steps from 1 to 2.7). Regarding $F0$, a pitch contour was extracted from a real sustained vowel lasting for 2 s. This curve was successively pitch-shifted from a $F0$ mean value of 75.6 Hz to 240 Hz in 81 steps of 0.25 semitones, thereby covering the male speech range (Pabon and Ternström 2018). For each possible combination of R_d and $F0$ values, two glottal flow versions were obtained with and without aspiration noise. The pulse amplitude, U_0, was selected to have 70 $\mathrm{dB_{SPL}}$ with the realistic geometry, $F0 = 120$ Hz and $R_d = 1$, which resulted in values $U_0 = 6.296 \times 10^{-5}$ $\mathrm{m^3/s}$ for vowel [ɑ], $U_0 = 3.455 \times 10^{-5}$ $\mathrm{m^3/s}$ for [i] and $U_0 = 6.657 \times 10^{-5}$ $\mathrm{m^3/s}$ for [u].

III.2.4 Acoustic Analysis

The Welch's power spectral density (PSD) estimate of each synthesised vowel was computed using a 2048-point FFT, with a 15 ms Hanning window and 50% overlap. The PSD was scaled by the equivalent noise bandwidth of the window to get the long-term average spectrum (LTAS). Moreover, HFE levels were computed as the integral of the PSD estimate within the 8 kHz octave band, as in (Monson et al. 2012; Monson et al. 2011) and in the three 1/3 octaves conforming that band, i.e., 6.3 kHz, 8 kHz, and 10 kHz. In the same way, the overall energy levels were obtained by considering the full bandwidth from 0 Hz to 12 kHz. The 16 kHz octave band was not considered in this study because its HFE variations have been shown to be almost perceptually irrelevant (see Monson et al. 2011).

III.3 Results

The vowels [ɑ], [i], and [u] were synthesised modifying the glottal source model in the whole phonation range, defined in this work as the space comprising fundamental frequencies $F0 \in [75.6, 240]$ Hz for $R_d \in [0.3, 2.7]$. Amplitude variations of the glottal pulses, U_0, could have also been incorporated in the study. However, they were not considered because they simply produce a level increment proportional to U_0. For instance, doubling the amplitude of U_0 simply generates a constant level offset of +6 dB at all frequencies.

In the following subsections we will start examining tense, modal, and lax phonations with $R_d = \{0.3, 1, 2.7\}$, respectively, for an intermediate $F0$ value of 120 Hz. The analysis will be then extended over the whole simulated phonation range, namely for $(F0, R_d) \in [75.6, 240] \times [0.3, 2.7]$.

III.3.1 Analysis of Tense, Modal, and Lax Phonations for Fixed $F0$ and R_d Values

Figure III.4 shows the LTAS of vowels [ɑ], [i], and [u] for tense, modal, and lax phonations ($R_d = \{0.3, 1, 2.7\}$) with $F0 = 120$ Hz. The figure thus contains nine subplots covering all possible combinations. In turn, each subplot presents four curves. Those correspond to the results with the realistic and simplified vocal tract geometries for activated and deactivated aspiration noise. The overall and HFE levels of the LTAS curves are shown in Table III.1 for the 8 kHz octave frequency band and also for its corresponding 1/3 octave bands, 6.3 kHz, 8 kHz, and 10 kHz. Values in parentheses indicate the level rise produced by including aspiration noise in the glottal source model.

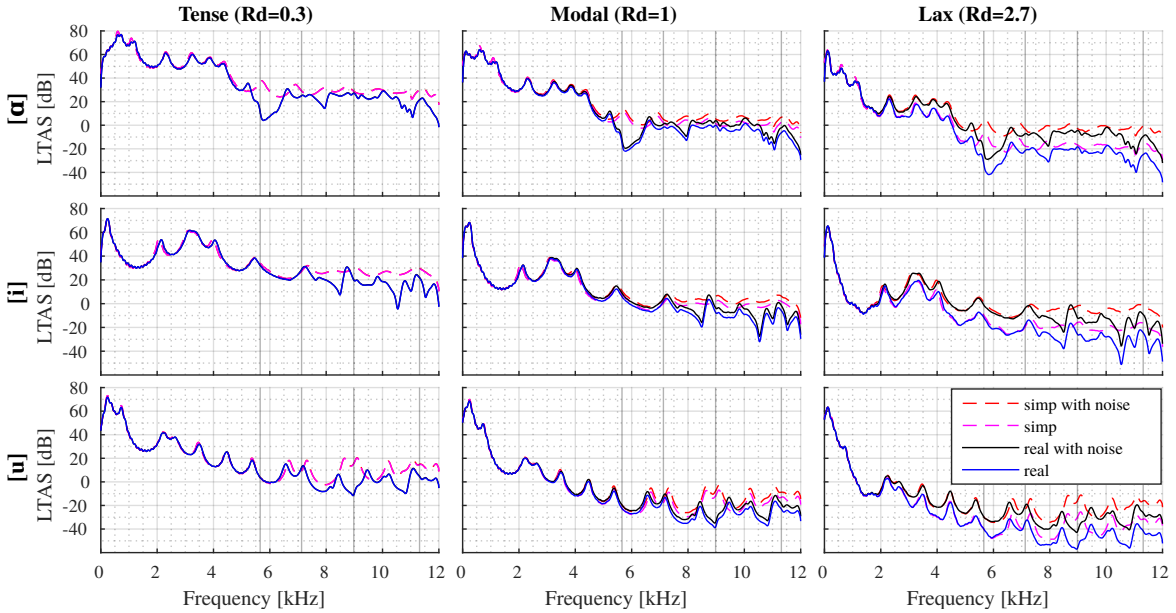


Figure III.4: Long-term average spectra (LTAS) of the FEM synthesised vowels [ɑ], [i], and [u] using the realistic and simplified vocal tract geometries with and without aspiration noise. Vowels were generated with a tense ($R_d = 0.3$), a modal ($R_d = 1$), and a lax ($R_d = 2.7$) phonation with $F0 = 120$ Hz. Vertical lines depict the boundaries of the 1/3 octave bands 6.3 kHz, 8 kHz, and 10 kHz.

Let us first focus on the comparison between the realistic and simplified vocal tract geometries. Looking at Figure III.4, we can observe that the vocal tract geometry did not have a significant effect on frequencies below ∼5 kHz, as already mentioned before, in contrast to the high frequency range. As explained, this is because planar modes mainly propagate at lower frequencies, whereas the higher order ones mostly appear in the high frequency range. This was clearly the case for the realistic geometry. Note, for instance, that a large valley is produced close to 6 kHz in the realistic configuration of [ɑ] (generated by a transverse mode, see Arnela et al. 2016b), whereas a resonance appears instead in the simplified configuration. The lack of higher order modes in the latter due to radial symmetry will allow us to determine their influence by comparing the results from the two configurations.

In general, higher order modes diminished the HFE levels, regardless of the phonation type and examined vowel. Note that in the 8 kHz octave band of Table III.1 the level of

Table III.1: Overall and high frequency energy (HFE) levels (in dB) obtained in the realistic and simplified vocal tract configurations of vowels [ɑ], [i], and [u]. Values correspond to vowels with a tense ($R_d = 0.3$), a modal ($R_d = 1$), and a lax ($R_d = 2.7$) phonation without considering aspiration noise. The values in parentheses denote the increment in dB obtained due to adding aspiration noise.

Vowel	Geometry	R_d	Overall	1/1 Octave Band	1/3 Octave Band		
				8 kHz	6.3 kHz	8 kHz	10 kHz
[ɑ]	realistic	0.3	82.3 (+0.0)	41.5 (+0.2)	35.3 (+0.1)	37.6 (+0.2)	37.2 (+0.2)
		1.0	70.0 (+0.0)	14.5 (+3.8)	8.7 (+3.0)	10.6 (+3.7)	9.8 (+4.4)
		2.7	63.5 (+0.0)	−4.5 (+14.4)	−10.3 (+13.1)	−8.5 (+14.3)	−9.1 (+15.2)
	simplified	0.3	83.5 (+0.0)	47.4 (+0.1)	43.4 (+0.1)	42.4 (+0.2)	41.9 (+0.2)
		1.0	71.4 (+0.0)	20.5 (+3.5)	17.0 (+2.8)	15.3 (+3.6)	14.5 (+4.4)
		2.7	64.9 (+0.0)	1.6 (+13.8)	−1.8 (+12.4)	−3.7 (+14.3)	−4.4 (+15.3)
[i]	realistic	0.3	73.6 (+0.0)	41.0 (+0.2)	37.0 (+0.1)	37.9 (+0.2)	31.8 (+0.2)
		1.0	70.0 (+0.0)	14.2 (+3.3)	10.6 (+2.7)	10.9 (+3.5)	4.4 (+4.4)
		2.7	65.6 (+0.0)	−4.7 (+13.4)	−8.2 (+12.3)	−8.0 (+13.8)	−14.5 (+15.2)
	simplified	0.3	73.4 (+0.0)	44.8 (+0.2)	38.5 (+0.1)	40.7 (+0.2)	40.5 (+0.2)
		1.0	70.0 (+0.0)	17.8 (+3.7)	12.2 (+2.7)	13.7 (+3.6)	13.1 (+4.4)
		2.7	65.6 (+0.0)	−1.0 (+14.0)	−6.6 (+12.2)	−5.0 (+13.8)	−5.7 (+15.2)
[u]	realistic	0.3	74.0 (+0.0)	22.8 (+0.2)	19.6 (+0.1)	16.3 (+0.2)	18.0 (+0.3)
		1.0	70.0 (+0.0)	−4.0 (+3.6)	−6.9 (+3.0)	−10.6 (+3.3)	−9.4 (+4.5)
		2.7	64.2 (+0.0)	−23.2 (+14.1)	−26.4 (+13.5)	−29.3 (+13.3)	−28.3 (+15.3)
	simplified	0.3	75.0 (+0.0)	29.9 (+0.2)	21.5 (+0.1)	25.4 (+0.2)	27.0 (+0.2)
		1.0	71.1 (+0.0)	2.7 (+4.0)	−5.2 (+3.2)	−1.7 (+3.7)	−0.4 (+4.4)
		2.7	64.2 (+0.0)	−16.0 (+14.4)	−23.6 (+12.6)	−20.5 (+14.2)	−19.2 (+15.1)

the realistic geometry became between 5.6 dB and 6.1 dB lower than that of the simplified geometry for vowel [ɑ], between 3.6 and 4.3 dB for vowel [i], and between 6.7 and 7.5 dB for [u]. More details can be obtained from the 1/3 octave bands levels at 6.3 kHz, 8 kHz, and 10 kHz. The first one at 6.3 kHz only presents small spectral differences for [u] and [i], which ranged from 1.8 dB to 2.8 dB for [u] and were of the order of 1.5 dB for [i]. Actually, this 1/3 octave band did not contain higher order modes for these vowels (see Figure III.4). Conversely, the dip around 6 kHz in the realistic [ɑ] vowel caused level decreases of between 7.8 dB and 8.5 dB. For this vowel, the onset of higher order modes took place at a lower frequency since vowel [ɑ] has a bigger oral cavity than [u] and [i]. In the second 1/3 octave band, centred at 8 kHz, the largest differences were found for [u]. In this case, the realistic configuration presented a valley close to 9 kHz, whereas resonances appeared for the simplified geometry. This results in variations ranged from 8.8 dB to 9.6 dB. Finally, [i] and [u] exhibited the largest deviations for the third 1/3 octave band at 10 kHz, which varied from 8.8 dB to 9.0 dB for the realistic configuration. According to (Monson et al. 2011), minimum difference limen scores of about 1 dB were obtained for normal-hearing listeners in the 1/1 octave band of 8 kHz. Therefore, one could hypothesise taking into account the aforementioned differences, that higher order modes could be perceptually relevant. However, this relevance may also depend on the HFE levels, which in turn greatly depend on the glottal source.

The glottal source not only modified the overall energy level but also introduced a spectral decay that can be appreciated by comparing the LTAS in Figure III.4 with the $H(f)$ of Figure III.2. This decay, also known as the spectral tilt, is strongly dependent on the

phonation type. In Figure III.4, it can be observed that the laxer the phonation (i.e., for growing R_d) the stronger the spectral tilt, especially at higher frequencies. For instance, moving from the modal phonation to the lax one produces an overall energy decay between 4.4 dB and 6.9 dB, considering all values in Table III.1. However, this reduction is much larger in the high frequency range. It can reach ∼19 dB if no aspiration noise is considered. When aspiration noise is present, the decrease was not so prominent and only ranged from 8.0 dB to 9.4 dB. On the other hand, going from a modal to a tense phonation resulted in the opposite behaviour. The spectral tilt was reduced, which increased the overall levels between 3.5 dB and 12.2 dB. Again, the HFE levels were more sensitive and increased from 23.1 dB to 27.4 dB. It is worthwhile observing that in this case the aspiration noise did not play a determinant role, since as the LTAS of Figure III.4 shows, the tense phonation remained unaltered.

Let us then analyse the influence of aspiration noise in more detail. As seen from Table III.1, the aspiration noise had no effect at all on the overall levels of any of the analysed configurations. It only affected the HFE content, resulting in significant energy increments for laxer phonations but in negligible differences for the tense ones. Level increments in the 8 kHz octave band of Table III.1 were less than 0.2 dB for the latter. In the case of modal phonation, the energy slightly increased beyond 4 kHz (see Figure III.4), which resulted in a level rise from 3.3 dB to 4.0 dB in the 8 kHz octave band. As expected, the most sensitive phonation type was the lax one, which was strongly influenced by aspiration starting from ∼2.5 kHz. The 8 kHz octave band levels increased from 13.4 dB to 14.4 dB in this case.

To summarise, higher order modes diminished HFE levels between 3.6 dB and 7.5 dB in the 8 kHz octave band when considering all tested configurations. These level reductions were comparable to those in (Monson et al. 2011) and could therefore be perceptually relevant. Nevertheless, the differences induced by the higher order modes would only be perceivable if the energy input at the high frequency range was substantial. This seems to be the case of the tense phonation, whose levels in the aforementioned frequency band were higher than 41 dB for [ɑ] and [i], and above 22 dB for [u] (see Table III.1). When the phonation was modal, higher order modes might still have been relevant for [ɑ] and [i], which presented HFE levels above 14 dB, in contrast with [u], where the levels remained between −4.0 dB and 6.7 dB. Finally, in the case of a lax phonation, perceptually significant HFE values could only be achieved for [ɑ] and [i] if aspiration noise was considered (with variations between 8.7 dB and 15.4 dB). Higher order modes became irrelevant for [u].

III.3.2 Analysis for the Whole Phonation Range

The analysis for the whole phonation range comprises of the overall and HFE levels in the 8 kHz octave band for the realistic geometry of the three vowels, with and without aspiration noise in the glottal source model. That results in the nine contour subplots are shown in Figure III.5. A rainbow colour scale is used for all of them, with red and blue respectively representing the highest and lowest energy levels. Note that the realistic cases analysed in the previous Section III.3.1 correspond to the vertical lines in the subplots, which have been

indicated with a diamond symbol.

The colour maps in Figure III.5 exhibit a pattern of diagonal contours with a general tendency to increase the overall and HFE levels from bottom left to top right. That is to say the minimum levels were obtained for the lowest $F0$ and laxest phonation, $R_d = 2.7$, and gradually increased when moving to higher $F0$ and smaller R_d values. This means that the obtained energy levels not only depended on R_d, as already observed in the previous Section III.3.1, but also on the $F0$ of the excitation. In regards to the overall levels (first column in Figure III.5), the lowest values were similar for all vowels ranging between 57.6 dB and 59.7 dB. In contrast, the highest levels depended on each vowel and reached 91.6 dB, 80.4 dB, and 78.8 dB for [ɑ], [i], and [u], respectively. The vocal tract of vowel [ɑ] produced the highest overall levels thanks to its first three formants, which are the most prominent ones as seen from the VTTFs in Figure III.2. Moreover, these resonances took place below 2.5 kHz, where the energy decay of the tense voice source was still moderate (see the top-left subplot in Figure III.4). On the other hand, the contours for the overall levels in Figure III.4 present some deviations with respect to the aforementioned diagonal pattern, especially for [i] and [u]. These deviations occurred at those $F0$s that were sub-multiples of the frequency of each vowel first formant F_1. For instance, vowel [ɑ] presented level increases at $F0 = 168.3$ Hz and $F0 = 224.3$ Hz, which correspond to $F_1/4$ and $F_1/3$, respectively. This effect was even more exaggerated for vowels [i] and [u], since they had a lower F_1 frequency. Note that the levels of these two vowels significantly increased at 110 Hz and 144.5 Hz, i.e., at $F_1/2$.

An even more interesting analysis is the examination of HFE content at 8 kHz with and without aspiration noise. When comparing the second and third columns in Figure III.5, we can appreciate how the inclusion of aspiration curves the iso-contours in the bottom left corner of the subplots. Hence, the impact of aspiration noise increased as the phonation became laxer, whereas its effect was negligible for tense phonations ($R_d < 0.74$) as quoted in Section III.3.1. The contour maps also revealed the effect of aspiration noise increased for decreasing $F0$. When the aspiration noise was not considered, the HFE levels were similar for [ɑ] and [i], ranging between -14.6 dB and 54.5 dB, while they were much lower for vowel [u] with variations ranging from -32 dB to 33 dB. Vowel [u] produced lower HFE values despite having similar overall levels to [i]. On the other hand, when the aspiration noise was added, the minimum HFE levels for [ɑ] and [i] were 3.3 and 2.1 dB, respectively. Thus, the levels for these two vowels were above the theoretically audible threshold of 0 dB, in the analysed phonation range. Conversely, the region of [u] with higher R_d values and lower $F0$s remained below 0 dB even if aspiration noise was incorporated. The minimum HFE level for this vowel was -15.7 dB. Therefore, depending on the phonation type and also on the vowel, the HFE levels may have been too small to perceive differences in simulations with the realistic or simplified vocal tract geometries. This may occur for high R_d values and/or low $F0$s, especially if there was no aspiration noise. In this respect, the presence of the latter seemed to suffice to obtain audible HFE levels for vowels [ɑ] and [i], but not for [u], in the very lax region.

Let us next examine the influence of the geometry and consequently that of higher order

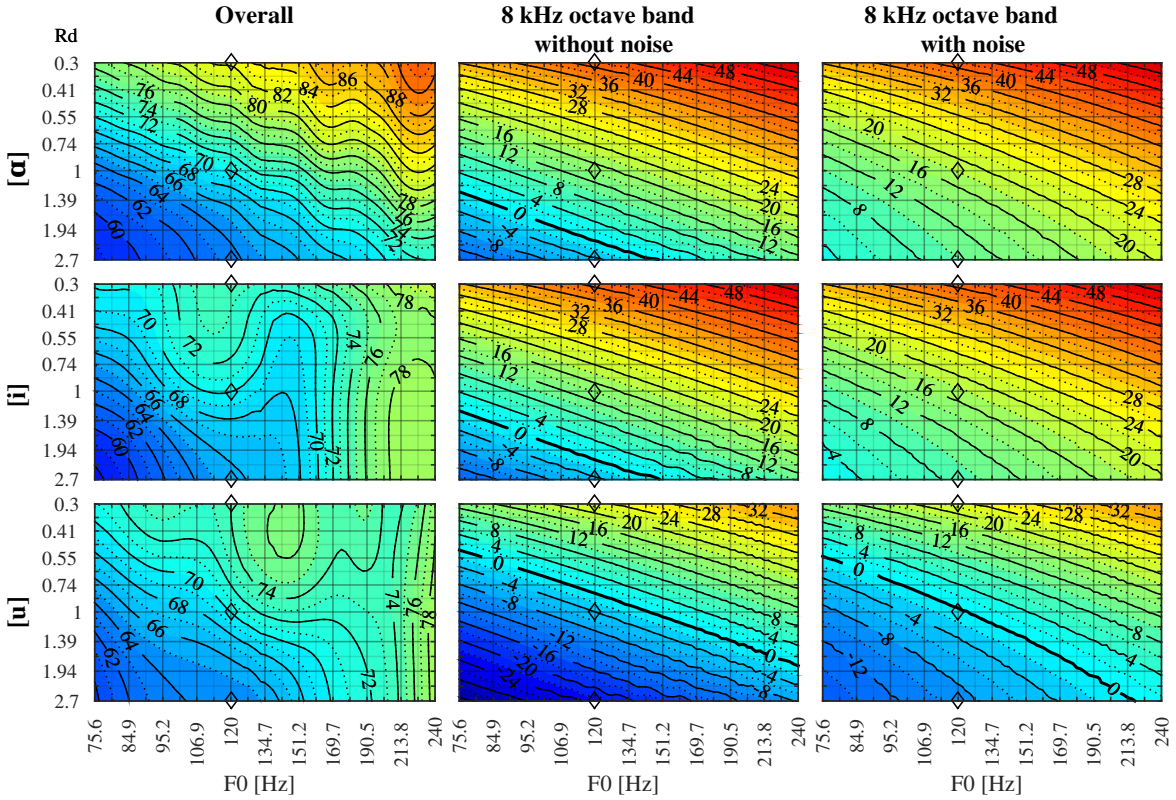


Figure III.5: Contour plots showing the overall and HFE levels (dB) in the 8 kHz 1/1 octave band for the realistic vocal tract geometry of vowels [ɑ], [i], and [u]. HFE levels are computed with and without introducing aspiration noise in the glottal source model. Each plot depicts the equal level contours for the whole phonation range, representing the $F0$ in the abscissas and the R_d value in the ordinates. Diamonds represent the points analysed in Section III.3.1.

modes in the considered phonation range. To do so, HFE level differences between the vowels generated with the realistic and simplified geometries were computed for each combination of $F0$-R_d-AspirationNoise. Table III.2 depicts the mean increments of the simplified configuration over the realistic ones in 1/1 and 1/3 octave frequency bands. It is worth mentioning that the standard deviation of these increments was less than 0.3 dB for all vowels and bands, since the LF model does not consider the interaction between the vocal tract and the vocal folds. The differences obtained in the 8 kHz 1/1 octave band were similar for [ɑ] and [u], with mean increments of 6.0 dB and 6.8 dB, respectively. Nevertheless, vowel [ɑ] primarily concentrated the differences in the first 1/3 octave band, while changes for [u] mainly occurred in the other two bands. In turn, differences for [i] basically manifested in the 10 kHz band, the mean increment in the 8 kHz 1/1 octave band being only ∼3.6 dB . All these values could slightly vary when aspiration noise is included (see the increments in parentheses in Table III.2). Note that the above observations were in line with the analysis derived in Section III.3.1 for the selected three pairs of $(F0, R_d)$.

Table III.2: HFE level mean increments (in dB) obtained for the simplified geometries with respect to the realistic ones. The values have been computed for the 8 kHz octave band and its corresponding 1/3 octave bands. The values in parentheses denote the additional increment in dB due to aspiration noise.

Vowel	**1/1 Octave Band**	**1/3 Octave Band**		
	8 kHz	**6.3 kHz**	**8 kHz**	**10 kHz**
[ɑ]	6.0 (−0.2)	8.2 (−0.2)	4.8 (+0.0)	4.7 (+0.0)
[i]	3.6 (+0.3)	1.5 (+0.0)	2.8 (+0.1)	8.8 (+0.0)
[u]	6.8 (+0.4)	1.8 (+0.0)	9.2 (+0.3)	9.0 (+0.0)

III.4 Conclusions

In this work, we analysed the relevance of higher order modes in the 3D finite element synthesis of vowels [ɑ], [i], and [u], considering different glottal source excitations. It was shown that higher order modes induced a reduction of between 3.6 dB and 7.2 dB in the HFE levels of the 8 kHz octave band which, according to previous works in literature, may be perceptually relevant. However, the influence of higher order modes strongly depended on the phonation type and fundamental frequency $F0$. Influence was greater for phonations with high HFE levels, such as the tense ones (small R_d), and/or for high $F0$s. On the other hand, HFE levels dropped rapidly for lax phonations and/or low $F0$s. The presence of aspiration noise could partially alleviate such decreases for [ɑ] and [i] vowels. Conversely, the levels obtained for [u] suggested that differences between realistic and simplified geometries may not be perceptually relevant for this vowel when the phonation was lax, even if aspiration noise is included. Future work will focus on the perceptual validation of the results presented herein. To this end, we will generate pseudowords containing vowels and consonants to have a broader assessable phonetic context, instead of only considering sustained vowels. However, the synthesis of such utterances is still being developed in FEM-based approaches for voice simulation.

Finally, we would like to point out that the outcomes for the realistic vocal tracts in this work correspond to those of a specific individual. Analysis for other speakers may result in some differences, yet we believe that the reported general tendencies will still be valid for them. In the future, though, it would be interesting to extend the investigation to further MRI-based geometries and using other glottal source models as well.

Author Contributions

Conceptualisation, writing—reviewing, and editing, M.F., M.A., J.C.S., F.A., and O.G.; methodology, software, formal analysis, investigation, and writing—original draft preparation, M.F. and M.A.; validation, data curation, and visualisation, M.F.; supervision, J.C.S., F.A., and O.G.

Funding

This research was funded by the Agencia Estatal de Investigación (AEI) and FEDER, EU, through project GENIOVOX TEC2016-81107-P. The fourth and fifth authors also acknowledge the support from the Obra Social "La Caixa" under respective grants ref. 2018-URL-IR2nQ-029 and 2018-URL-IR2nQ-031.

Acknowledgments

The authors are grateful to Saeed Dabbaghchian for the design of the vocal tract geometry simplifications and Lisa Kinnear for the English proofreading.

Conflicts of Interest

The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

HFE High Frequency Energy
FEM Finite Element Method
LF Liljencrants–Fant
MRI Magnetic Resonance Imaging
PML Perfectly Matched Layer
PSD Power Spectral Density
LTAS Long-Term Average Spectrum

References

Aalto, Daniel et al. (2014). "Large scale data acquisition of simultaneous MRI and speech". In: *Applied Acoustics* vol. 83, pp. 64–75.

Arnela, Marc, Blandin, Rémi, Dabbaghchian, Saeed, Guasch, Oriol, Alías, Francesc, Pelorson, Xavier, Van Hirtum, Annemie, and Engwall, Olov (2016a). "Influence of lips on the production of vowels based on finite element simulations and experiments". In: *The Journal of the Acoustical Society of America* vol. 139, no. 5, pp. 2852–2859.

Arnela, Marc, Dabbaghchian, Saeed, Blandin, Rémi, Guasch, Oriol, Engwall, Olov, Van Hirtum, Annemie, and Pelorson, Xavier (2016b). "Influence of vocal tract geometry simplifications on the numerical simulation of vowel sounds". In: *The Journal of the Acoustical Society of America* vol. 140, no. 3, pp. 1707–1718.

Arnela, Marc, Dabbaghchian, Saeed, Guasch, Oriol, and Engwall, Olov (2019). "MRI-based vocal tract representations for the three-dimensional finite element synthesis of diphthongs". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* vol. 27, pp. 2173–2182.

Arnela, Marc and Guasch, Oriol (2013). "Finite element computation of elliptical vocal tract impedances using the two-microphone transfer function method". In: *The Journal of the Acoustical Society of America* vol. 133, no. 6, pp. 4197–4209.

Arnela, Marc, Guasch, Oriol, and Alías, Francesc (2013). "Effects of head geometry simplifications on acoustic radiation of vowel sounds based on time-domain finite-element simulations". In: *The Journal of the Acoustical Society of America* vol. 134, no. 4, pp. 2946–2954.

Birkholz, Peter (2013). "Modeling Consonant-Vowel Coarticulation for Articulatory Speech Synthesis". In: *PLoS ONE* vol. 8, no. 4, e60603.

Blandin, Rémi, Arnela, Marc, Laboissière, Rafael, Pelorson, Xavier, Guasch, Oriol, Van Hirtum, Annemie, and Laval, Xavier (2015). "Effects of higher order propagation modes in vocal tract like geometries". In: *The Journal of the Acoustical Society of America* vol. 137, no. 2, pp. 832–8.

Davis, Philip J. and Rabinowitz, Philip (2007). *Methods of numerical integration*. Courier Corporation.

Erath, Byron D., Zañartu, Matías, Stewart, Kelley C., Plesniak, Michael W., Sommer, David E., and Peterson, Sean D. (2013). "A review of lumped-element models of voiced speech". In: *Speech Communication*, pp. 667–690.

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

Fant, Gunnar, Liljencrants, Johan, and Lin, Qi-guang (1985). "A four-parameter model of glottal flow". In: *Speech Transmission Laboratory Quarterly Progress and Status Report (STL-QPSR)* vol. 26, no. 4, pp. 1–13.

Freixes, Marc, Arnela, Marc, Socoró, Joan Claudi, Alías, Francesc, and Guasch, Oriol (2018). "Influence of tense, modal and lax phonation on the three-dimensional finite element synthesis of vowel [A]". In: *Proc. IberSPEECH 2018*, pp. 132–136.

Gobl, Christer (2006). "Modelling aspiration noise during phonation using the LF voice source model". In: *Interspeech*, pp. 965–968.

Kawahara, Hideki, Sakakibara, Ken-Ichi, Banno, Hideki, Morise, Masanori, Toda, Tomoki, and Irino, Toshio (2017). "A new cosine series antialiasing function and its application to aliasing-free glottal source models for speech and singing synthesis". In: *Interspeech*, pp. 1358–1362.

Monson, Brian B., Hunter, Eric J., Lotto, Andrew J., and Story, Brad H. (2014). "The perceptual significance of high-frequency energy in the human voice". In: *Frontiers in Psychology*.

Monson, Brian B., Lotto, Andrew J., and Story, Brad H. (2012). "Analysis of high-frequency energy in long-term average spectra of singing, speech, and voiceless fricatives". In: *The Journal of the Acoustical Society of America* vol. 132, no. 3, pp. 1754–1764.

Monson, Brian B., Lotto, Andrew J., and Ternström, Sten (2011). "Detection of high-frequency energy changes in sustained vowels produced by singers". In: *The Journal of the Acoustical Society of America* vol. 129, no. 4, pp. 2263–2268.

Murphy, Andy, Yanushevskaya, Irena, Chasaide, Ailbhe Ní, and Gobl, Christer (Aug. 2017). "Rd as a Control Parameter to Explore Affective Correlates of the Tense-Lax Continuum". In: *Interspeech*. Stockholm, Sweden, pp. 3916–3920.

Murtola, Tiina, Alku, Paavo, Malinen, Jarmo, and Geneid, Ahmed (2018). "Parameterization of a computational physical model for glottal flow using inverse filtering and high-speed videoendoscopy". In: *Speech Communication* vol. 96, pp. 67–80.

Pabon, Peter and Ternström, Sten (2018). "Feature Maps of the Acoustic Spectrum of the Voice". In: *Journal of Voice*.

Story, Brad H. (2013). "Phrase-level speech simulation with an airway modulation model of speech production". In: *Computer Speech & Language* vol. 27, no. 4, pp. 989–1010.

Story, Brad H., Titze, Ingo R., and Hoffman, E. A. (1996). "Vocal tract area functions from magnetic resonance imaging". In: *The Journal of the Acoustical Society of America* vol. 100, no. 1, pp. 537–554.

Takemoto, Hironori, Adachi, Seiji, Mokhtari, Parham, and Kitamura, Tatsuya (2013). "Acoustic interaction between the right and left piriform fossae in generating spectral dips". In: *The Journal of the Acoustical Society of America* vol. 134, no. 4, pp. 2955–2964.

Takemoto, Hironori, Mokhtari, Parham, and Kitamura, Tatsuya (2010). "Acoustic analysis of the vocal tract during vowel production by finite-difference time-domain method". In: *The Journal of the Acoustical Society of America* vol. 128, no. 6, pp. 3724–3738.

Vampola, Tomáš, Horáček, Jaromír, and Švec, Jan G. (2008). "FE Modeling of Human Vocal Tract Acoustics. Part I: Production of Czech vowels". In: *Acta acustica united with Acustica* vol. 94, no. 5, pp. 433–447.

Authors' addresses

Marc Freixes GTM – Grup de Recerca en Tecnologies Mèdia, La Salle - Universitat Ramon Llull
Quatre Camins, 30, 08022 Barcelona, Spain
marc.freixes@salle.url.edu

Chapter 2

Conclusions and future work

The aim of this thesis has been mainly focused on incorporating expressiveness in the generation of synthetic voice without recording or collecting specific expressive data from the original speaker. The conclusions derived from the research conducted for this purpose are presented in this chapter together with some future research directions. Section 2.1 describes the main outcomes obtained after adding expressiveness to a unit-selection TTS system to generate singing and storytelling increasing suspense, in accordance with objective O1. Section 2.2 details the main findings regarding the first steps taken to incorporate expressiveness in the numerical voice production (objective O2). Finally, results are discussed and future perspectives are pointed out in section 2.3 .

2.1 Adding expressiveness to a unit-selection TTS system

In order to add expressiveness to a US-TTS system in the context of storytelling without the need of recording new corpora from the original speaker, we have proposed to incorporate neutral-to-expressive speech transformation modules into the classic TTS system pipeline. Regarding the expressive styles, the work has focused on singing and increasing suspense both in the storytelling domain.

In an initial stage, we proposed two approaches to add singing capabilities to a US-TTS synthesis system by: i) adding HNM-based STS transformation modules to the output of the TTS system, and; ii) integrating the STS modules within the TTS pipeline, thereby building a hybrid US+HNM TTS system. The conducted perceptual tests show that it is viable to generate singing of reasonable quality with a US-TTS system fed with neutral speech, without the need of explicitly recording a singing database. Moreover, although the two considered approaches received similar rates in terms of naturalness and singing resemblance, the hybrid strategy reduced the computation cost by almost a 250% with respect to the chaining-based approach.

Building on the findings of this first work, we have designed a Unit Selection based Text-to-Speech-and-Singing (US-TTS&S) synthesis framework that allows the generation of both speech and singing from an input text and a score, respectively, using a small corpus of neutral speech. Besides the integration of the STS transformation modules as in the hybrid approach, a score prosody generation module has been incorporated to enable a score-driven US strategy beyond the default text-driven US mode. The proposal has been evaluated by means of a proof-of-concept implementation using a 2,6 h Spanish neutral speech corpus, and considering different vocal ranges and tempos as well as several text-driven and score-driven US configurations. Results have revealed that high demanding STS transformation factors are required to sing beyond the corpus vocal range and/or when notes longer than 150 ms

are present in the input score. Score-driven US configurations have been proved effective in reducing the required pitch-scale factors but not the time-scale ones, mainly due to the short length of the vowels available in the neutral corpus as it was originally designed for general purpose TTS synthesis purposes. With respect to the subjective evaluation, a similar naturalness has been achieved by text-driven and score-driven US configurations in all the analysed scenarios. Compared to the first work, the perceptual test has incorporated reference stimuli generated with the professional singing synthesiser Vocaloid[1]. Although the obtained naturalness is obviously far from that achieved by Vocaloid, the received singing ratings validate the capability of the framework to satisfactorily address eventual singing needs.

The hybrid approach has been also applied to generate storytelling suspense speech using an adaptive Harmonic Model (aHM) and an expert system based on prosodic rules derived from the analysis of few but representative utterances of increasing suspense. The results obtained from the conducted perceptual test show that our approach has achieved a better naturalness and storytelling quality than those obtained with the rules from (Theune et al. 2006), which have been used as a baseline. Nevertheless, that work provided evidences that there is room for improvement regarding suspense arousal. In this respect, some participants commented that a warmer and more whispery voice could improve this aspect. These commentaries suggest that VoQ plays an important role in storytelling suspense and it should be therefore included in further investigations.

2.2 Adding expressiveness to numerical voice production

As a first step to incorporate expressiveness in the numerical simulation of voice, a decoupled source-filter based approach has been considered and the study has focused on the source component, thereby avoiding the need to acquire additional vocal tract MRI-based geometries. Our proposal has been built on: i) the generation and manipulation of glottal source signals using a LF glottal flow model; and ii) the extraction of glottal source characteristics from the analysis of speech signals by means of glottal source processing techniques. In particular, we have analysed the contribution of the glottal source to higher order modes in the FEM synthesis of vowels considering realistic vocal tract geometries. Moreover, a GlottDNN based analysis of emotional vowels from a parallel corpus has been proposed with the aim to characterise and incorporate their expressive particularities into the numerical production of vowels.

The first work done in this line analysed the influence of phonation on the 3D FEM synthesis of vowel [ɑ], comparing the results of the simulations when considering a realistic and a simplified vocal tract geometry. Specifically, the study has focused on three predefined R_d values corresponding to tense, modal and lax phonation, and considering a standard $F0$ value of 120 Hz. As expected, LTAS for both geometries are very similar below 5 kHz. Nevertheless, significant differences appear beyond this frequency due to the propagation of higher order modes, which only happens in the realistic geometry. These modes imply a reduction of the HFE levels at the 8 kHz band between 5.6 and 5.9 dB, depending on the

[1] http://www.vocaloid.com/en/

phonation type. Although these differences are expected to be perceptually relevant according to previous studies in the literature, such relevance is conditional on the phonation type. The lax phonation yields to small HFE levels, which suggests the influence of higher order modes may be imperceptible so a simpler 1D simulation would suffice for this phonation type. Conversely, HFE levels given by a modal or a tense voice indicate that a realistic 3D vocal tract geometry would be needed for the accurate FEM simulation of vowel [ɑ].

The contribution of the glottal source into the 3D finite element simulation of vowels has been further analysed considering not only vowel [ɑ] but also vowels [i] and [u], besides covering the complete R_d range from lax to tense as well as the spoken vocal range of $F0$, beyond the three R_d values and the single $F0$ value considered in the first work. Moreover, aspiration noise has been incorporated into the LF model to analyse its contribution to higher order modes depending on the phonation types. The performed simulations show that higher order modes lead to decreases in the HFE levels at the 8 kHz octave band between 3.6 dB and 7.2 dB, mainly depending on the vowel, hence a wider range of values compared with those obtained when only vowel [ɑ] was analysed. These decreases may still be perceptually relevant according to the literature, conditioned on the HFE levels. In this respect, the relevance of higher order modes is greater for phonations with high HFE levels, namely the tense ones (small R_d) and/or for high $F0$s. Conversely, lax phonations and/or low $F0$s present lower HFE levels, especially if aspiration noise is not incorporated. The results suggest that differences between realistic and simplified geometries may be perceptually relevant for tense and modal phonations. This seems also to be the case for lax phonations if aspiration noise is incorporated, except for the vowel [u], whose HFE levels are significantly lower due to the frequency response of this vowel, which presents the lowest gain at high frequencies.

In the aforementioned works, the R_d values have been modified to explore the phonation tense-lax continuum. Nevertheless, we also wanted to know which R_d values should be considered to introduce certain expressive styles into the numerical simulations of vowels. To this end, we have explored the glottal source differences between happy and aggressive emotional styles with respect to neutral speech through the analysis-by-synthesis of vowel [a]. The analysis has focused on features that could be translated to a LF model controlled by the R_d parameter, which has been used to generate the excitation for the 3D FEM-based simulation of [a] vowels. More specifically, 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 been subsequently translated into LF parameters in order to add expressiveness to the LF-FEM based synthesis of vowels [a] and ['a]. The conducted experiments suggest that the synthesis of aggressive and happy vowels requires greater $F0$ and lower R_d values than those used for neutral speech. The differences are greater for happy than for aggressive especially for the stressed ['a]. Finally, it is worth mentioning that the R_d values obtained from the analysis of expressive speech are not as extreme as the theoretical values considered in our previous studies, where the aim was to study the phonation tense-lax continuum.

2.3 Discussion and future perspectives

This thesis has addressed the generation of expressive speech on two neutral voice synthesis systems, specifically: i) a corpus-based US-TTS system and; ii) FEM-based numerical simulation of voice.

Expressiveness in corpus-based synthesis systems is restricted to that available in the corpora. In this sense, the overall trend in current speech synthesis approaches is to build ever larger corpora, especially in E2E systems. However, collecting data to cover *any* expressive style remains time-consuming and costly and besides, it is not always possible to record additional samples from the original speaker. In this respect, we found interesting to explore alternative approaches to augment the expressive capabilities of a TTS fed with a small neutral corpus so as to extend its basic capabilities of synthesising neutral utterances. To this end, we have proposed a framework which integrates several modules into a US-TTS system to enable the synthesis of singing through: i) the selection of speech units from the corpora according to that expressive target and; ii) the transformation of the selected speech signals to singing.

Classical US-TTS systems build on the idea of *choosing the best to modify the least* (Balestri et al. 1999). Inspired by this strategy, an expressive prosody generation module has been incorporated to choose the units from the neutral corpus that are closer to the desired target prosody, which is determined by the score in the case of singing. While this approach seems useful to reduce the pitch modification requirements, it has entailed a small impact on time-scaling in the performed experiments. Time-scale factors remain very high, especially those needed to generate long notes. This may justify, in part, why in the conducted perceptual tests the score-driven US strategy has yield similar results to that obtained with the default text-driven US. The very different nature of speech and singing may also explain such results. Units that are closer to the musical prosody can be seen as outliers from the speech synthesis point of view, especially when singing beyond the corpus vocal range.

It would be interesting to adapt the expressive prosody generation module to other storytelling speaking styles. It seems reasonable to expect better results of the proposed US-based strategy for such styles because their prosody would not be as extreme as that of singing, so it may be easier to find units in the corpora closer to the desired expressive prosody. On the other hand, it could be explored the use of expressive speech corpora in addition to the neutral corpus. Expressive corpora, besides covering a wider range of prosodic patterns, may contain speech conveying different VoQ. The selection of units from these corpora could therefore be explored as a means of achieving expressive VoQ effects without requiring of timbre transformations. It would be interesting to study the combination of different expressive styles and corpora, starting for example with singing and subsequently extending the analysis to other expressive speaking styles of the storytelling.

Another fundamental aspect of the proposal is the transformation of speech signals from the neutral corpora into expressive speech or singing. This goal sets very high transformation requirements, especially for the latter. In order to address those requirements, harmonic modelling of speech has been used. This model allow for an almost transparent resynthesis

of speech and it has already been integrated in US-TTS systems as it can also smooth the joints between the selected speech units. In the present work, this model has been applied in a much more demanding scenario: the transformation of neutral speech to singing. One of the major challenges for such transformations is the generation of long sung vowels from short spoken vowels given the large time-scale factors involved. Further work should therefore be undertaken in this aspect, exploring for instance other time-scale and spectral transformation approaches, besides incorporating a vibrato model.

Work done on singing has mainly focused on prosodic transformations. Nevertheless, the results obtained from the experiments conducted on storytelling increasing suspense show the need for VoQ transformations in addition to the prosodic ones. Specifically, the responses of the participants suggest that this style requires a more whispered voice. Such work considered an aHM model, which outperforms the HNM model regarding the resynthesis accuracy. Conversely, the HNM model appears to be more flexible to handle potential VoQ transformations because it separately parameterises the harmonic and the stochastic component of speech instead of modelling the whole speech signal as a sum of sinusoids. In this way, a more whispered voice could be obtained by increasing the amplitude of the stochastic component. Moreover, the aspiration noise should be incorporated pitch-synchronously to ensure it sounds integrated with the harmonic component. Accordingly, and with a view towards further developments, a pitch synchronous HNM model has been implemented for the US-TTS&S framework, replacing a previously considered constant frame-rate model. This new HNM implementation also incorporates some of the improvements introduced by the aHM model, such as the adaptive iterative refinement of the $F0$, which allows for a more accurate modelling of the voiced speech component.

Informal tests suggest that the incorporation of aspiration noise should be done together with changes in the tension of the phonation to achieve convincing results. This observation is consistent with the aspiration noise model implemented for the numerical simulation of vowels, where the noise amplitude depends on the R_d parameter, which correlates with the phonation tension. In this sense, we could profit from the experience in the numerical simulation of voice both exploring the tense-lax continuum of phonation and analysing the glottal source characteristics of expressive vowels. It would also be interesting to consider alternatives closer to the source-filter model than the HNM model, like the one provided by the GlottDNN vocoder. These alternatives could facilitate the independent modification of the glottal source and the vocal tract characteristics to achieve a more lax or even a breathy voice, which is necessary to resemble specific expressive speaking styles such as the aforementioned increasing suspense.

Numerical 3D simulations are a very powerful tool to study the physics involved in human voice production. Up to now, only few phonemes, diphthongs and vowel-consonant-vowel sequences have been generated so far. While research is being done to broaden the range of simulated utterances, the work presented in this thesis has focused on adding expressiveness to the FEM-based numerical simulation of vowels following the source-filter paradigm. To pursue this aim, we have focused on the modelling and modification of the excitation component,

leaving the study of the relevance of the vocal tract response for future works. To this end, we have considered the glottal flow LF model to generate the excitation signal for the numerical simulations of expressive voice. Despite the simplifications made by this approach, such as the assumption of independence between glottal source and vocal tract, it has allowed us to generate different phonation types and study their effect on the numerical simulations of expressive vowels, considering different vocal tract geometries. Nevertheless, in order to complete the investigations other glottal source models could be considered in the future, like the two-mass model for instance. The latter can account for the dependency between glottal source and vocal tract, though its control is more complicated. It is worth mentioning that the results obtained in these works correspond to vocal tract geometries from a specific individual. Although the evaluation has been done through relative comparisons between simplified and realistic geometries, analysis for other speakers may result in some differences. Therefore, in the future it would be interesting to extend the study to other MRI-based 3D geometries to evaluate to what extent the obtained results can be generalised to other speakers.

The analysis of expressive speech through inverse filtering has allowed us to study the characteristics of the glottal source of different expressive styles with tense voice. Nevertheless, the accuracy of inverse filtering algorithms is still an open issue, especially for female voices and for expressive speech styles which include irregular phonation patterns such as those of a creaky voice, for instance. Furthermore, it has been observed that a voice with a lax phonation affects the estimation of the glottal source spectral tilt because of the presence of aspiration noise. Regarding the studied vowels, inverse filtering performs better for [a] than for other vowels as reported in the literature. Therefore, further work should be done on studying and developing inverse filtering approaches to extend the analysis to other vowels and expressive styles with the goal of integrating the characteristics of real voice into the numerical production model of voice. The work done in this thesis has shown the relevance of the glottal source in the generation of expressive vowels. Future work will also analyse the contribution of vocal tract to the production of expressive voice. Moreover, in order to incorporate expressive effects associated with the vocal tract into the numerical simulations, modifications of the vocal tract geometry should be performed. Dynamic vocal tract geometries could be considered to simulate expressive utterances like diphthongs. The synthesis of such utterances longer than static vowels could also facilitate the perceptual evaluation, which would complement the investigations presented herein.

Our proposal builds on the idea of controllability and modularity. The control allows for a gradual transition from a style to another. This aspect is of great importance because of the theoretically infinite expressive possibilities of the human voice. This is especially relevant in storytelling, which is an expressive speaking style with a great variability and expressive subtle nuances. Control has been also very useful to study the influence of the glottal source characteristics in the numerical simulation of vowels. It has allowed to explore the tense lax-continuum of phonation, and the effect of $F0$ and aspiration noise in these simulations. On the other hand, modularity opens the door to evolve the proposed frameworks according to the aforementioned directions by upgrading their modules or by incorporating

additional ones. In this respect, the work done on singing has focused on the integration of STS modules within the US-TTS pipeline, analysing the transformation requirements and how to reduce these requirements by means of a score-driven US strategy. However, recent advances in STS transformations including voice conversion could be incorporated by upgrading the modules of US-TTS&S framework responsible for the transformation and the expressive control. The transformation module could be also adapted to work with other models of speech, as well as to handle VoQ transformations. Similarly, the expressive control module could be extended to other expressive categories of the indirect storytelling speech or even to direct speech (i.e., characters interventions). The US-TTS&S framework has been built on top of a US-TTS system. Nevertheless, US-based synthesis systems have been recently outperformed by approaches based on deep learning. It would therefore be interesting to study how our proposal, where the availability of data is very limited, could benefit from those approaches, which typically rely on large amounts of speech data. In this respect, although the US-TTS&S framework is far from a E2E pipeline, other options such as the use of neural vocoders could be considered in future works.

With regard to the numerical simulation of vowels, a module to deal with vocal tract modifications could be envisioned to account for the expressive effects associated with the filter component of the source-filter model. In this respect, although first steps have been taken to incorporate expressiveness into the FEM-based numerical production of voice, there is still a long road ahead to achieve an approach able of generating expressive voice considering excitation and a realistic vocal tract in a unified manner.

Finally, possible applications of the developed investigations could be envisaged. The US-TTS&S framework could be incorporated in applications where the synthesis of speech plus singing could be of interest. For instance, in storytelling when one of the characters sings, or in assistive technologies such as voice output communication aid devices, which could allow people with special needs to talk and sing. It could also be used in the production of videogames or animated series, or as a support tool for children studying music, choir singers, composers, etc. The 3D numerical simulation of voice can help understanding and modelling of the physics of the vocal apparatus, which could be useful in the study of VoQ effects and pathological voices, and it could also be potentially applicable in the field of speech therapy.

References

Balestri, Marcello, Pacchiotti, Alberto, Quazza, Silvia, Salza, Pier Luigi, and Sandri, Stefano (Sept. 1999). “Choose the best to modify the least: a new generation concatenative synthesis system”. In: *Eurospeech*. Budapest, Hungary: ISCA.

Theune, Mariët, Meijs, Koen, Heylen, Dirk, and Ordelman, Roeland (2006). “Generating expressive speech for storytelling applications”. In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* vol. 14, pp. 1137–1144.

Other articles

Paper IV

Generating Storytelling Suspense from Neutral Speech using a Hybrid TTS Synthesis framework driven by a Rule-based Prosodic Model

Raul Montaño, Marc Freixes, Francesc Alías, Joan Claudi Socoró

Published in *Proceedings of IberSPEECH 2016*, November 2016, pp. 129–138.

Abstract

There is a growing interest in the analysis and synthesis of expressive speech containing particular speaking styles. However, collecting enough representative speech data for each and every specific expressive style is a very daunting task, becoming almost unfeasible for those styles sporadically present in the speech. This is of special relevance for storytelling speech, where many subtle speech nuances and characters impersonations may take place. In this paper, we describe a hybrid Unit Selection-adaptive Harmonic Model text-to-speech synthesis framework that integrates a prosodic rule-based model derived from a small but representative set of utterances to convey suspense from neutral speech. The perceptual tests conducted on increasing suspense show that the introduced synthesis framework achieves better naturalness and storytelling resemblance than previous approaches, and similar suspense arousal.

IV.1 Introduction

Until the beginning of the 21st century, the main focus of the research community working on the analysis and synthesis of expressive speech was placed on emotions (see Scherer 2003; Schröder 2001, and references therein). From then on, a growing number of studies have coped with other expressive speaking styles mainly following corpus-based approaches (cf., Schröder 2009).

In order to bridge the daunting task of building ad-hoc corpus for each and every expressive speaking style when possible (e.g., Iriondo et al. 2007; Alías et al. 2008), some works have tackled the generation of synthetic expressive speech following quite diverse approaches. In (Theune et al. 2006; Zovato et al. 2004; Montaño et al. 2013), basic fixed acoustic rules were

applied to transform neutral to expressive synthetic speech. Differently, adaptation techniques have been considered in Hidden Markov Model (HMM)-based synthesizers to interpolate between statistical models trained on different expressive databases (Yamagishi and Kobayashi 2007). Hybrid approaches have also been introduced with the same aim. An Unit Selection (US)-based conversion system using Harmonic plus Noise Model (HNM) was developed to generate emotions from neutral speech in (Erro et al. 2010). Later, an emotion transplantation approach consisting of adaptation functions as pseudo-rules for modifying the HMM-based models was presented in (Lorenzo-Trueba et al. 2015). Although both approaches are based on rather small corpora, they still need non-negligible speech data for each expressive style (e.g., 6–30 min. Lorenzo-Trueba et al. 2015 and around 10 min. in Erro et al. 2010 per style), besides presenting other limitations such as the need of parallel corpora (Erro et al. 2010) or yielding over-smoothed speech quality characteristic of statistical approaches (Barra-Chicote et al. 2010). In this respect, adaptive Harmonic Model (aHM) (Degottex and Stylianou 2013) has been proved to provide better synthesis quality than HNM (Kafentzis et al. 2014) and other vocoders (Hu et al. 2013). However, as far as we know there are still no expressive speech synthesis works applying aHM.

One recently studied speaking style with rich expressive content is storytelling. While some studies have directly used audiobooks containing stories to generate corpus-based expressive synthetic speech (Jauk et al. 2015; Charfuelan and Steiner 2013; Prahallad and Black 2011), others have been focused on the detailed analysis of specific prosodic aspects of oral storytelling (Theune et al. 2006; Doukhan et al. 2011; Montaño et al. 2013). Even though using audiobooks can be used to generate expressive speech with good quality in average, there are several subtle expressive nuances within the storytelling speaking style that need further analysis to fully accomplish the requirements of storytelling applications (see e.g., Leite et al. 2015; Alofs et al. 2015).

As initial steps to this aim, some works have analysed and modelled specific types of storytelling speech to synthesize them from neutral speech. In (Theune et al. 2006) a set of *fixed* prosodic rules (including mathematical functions) was defined and applied in a diphone-based text-to-speech (TTS) synthesizer, reaching a significant improvement of storytelling and suspense perception. However, the prosodic rules for suspense were derived from very few sentences (e.g., only *one* sentence for increasing suspense and *two* sentences for sudden suspense) as they are rarely found in stories. Conversely, in (Montaño et al. 2013) speaking rate, mean pitch, pitch standard deviation and mean intensity of several sentences were analysed for different storytelling categories. A hybrid US-HNM framework was considered to prosodically transform neutral speech to the different expressive categories according to mean values of each category. As a consequence of using simple constant conversion factors, subtle expressive nuances were not captured accurately.

In this paper, we focus on the analysis and synthesis of increasing suspense as a key expressive style in storytelling speech, but with the added difficulty that is present in very specific instants of the story (i.e., very few sentences can be found). A hybrid US-aHM TTS synthesis framework is introduced to generate suspenseful storytelling speech from neutral

speech. To that effect, the TTS system is driven by a rule-based prosodic model that captures the subtle nuances of increasing suspense from a reduced set of representative utterances.

This paper is structured as follows. Section IV.2 reviews the main works dealing with suspense in storytelling speech. Next, Section IV.3 explains the proposed US-aHM synthesis framework. Then, Section IV.4 describes the development of the approach on increasing suspense as a proof of concept. After that, Section IV.5 describes the conducted perceptual evaluation and the results obtained after comparing our approach to the prosodic rules from (Theune et al. 2006). Finally, Section IV.6 end this paper with the conclusions.

IV.2 Related work

Suspense is the feeling of excitement or anxiety that the audience (listeners or readers) feels because of waiting for something to happen, i.e., the outcome is uncertain (Lehne and Koelsch 2015). Up to our knowledge, only two works have shed some light in how suspense can be evoked in the audience by means of modifying speech prosody. In (Doukhan et al. 2011), the authors suggested that a low intensity may induce suspense, but no further analyses were applied. On the contrary, (Theune et al. 2006) observed and defined two kinds of suspense found within their speech material: the *sudden* suspense and the *increasing* suspense. The former corresponds to an unexpected dramatic moment in the story, such as a startling revelation or a sudden momentous event. In the latter, the dramatic event is expected in advance and the suspense is built up until a pause, which is followed by the revelation of the important information. In this paper, we focus on the increasing suspense, whereas the sudden suspense is left for future works.

In (Theune et al. 2006), the authors defined a set of fixed prosodic rules for the increasing suspense based on the analysis of *one* sentence uttered by a professional actor. The acoustic characteristics observed in that utterance were a gradual increase in pitch and intensity, accompanied by a decrease in tempo. Then, a pause was present before the description of the actual dramatic event. Thus, this type of suspense was divided into two zones (Theune et al. 2006): before (zone 1) and after (zone 2) the pause. From this analysis the following prosodic modifications were applied to a neutral synthetic utterance generated with the Fluency Dutch TTS system. In the first zone, a sinusoidal function applied to stressed syllables was proposed to model the gradual increase of pitch (from +25 to +60Hz), whereas a constant increase up to +10 dB (on the whole signal) and +150% (on stressed vowels) was considered for intensity and duration transformations, respectively. In the second zone, pitch and durations gradually decreased to their normal values, whereas for intensity an increase of +6 dB was applied to the first word with no further modifications afterwards.

IV.3 Hybrid US-aHM synthesis framework

The US-aHM TTS synthesis system depicted in Fig. IV.1 builds on the idea of enabling US-TTS synthesis to manage different expressive styles within the same synthesis framework (Alías et al. 2008). The process starts by building the rule-based prosodic model from utterances

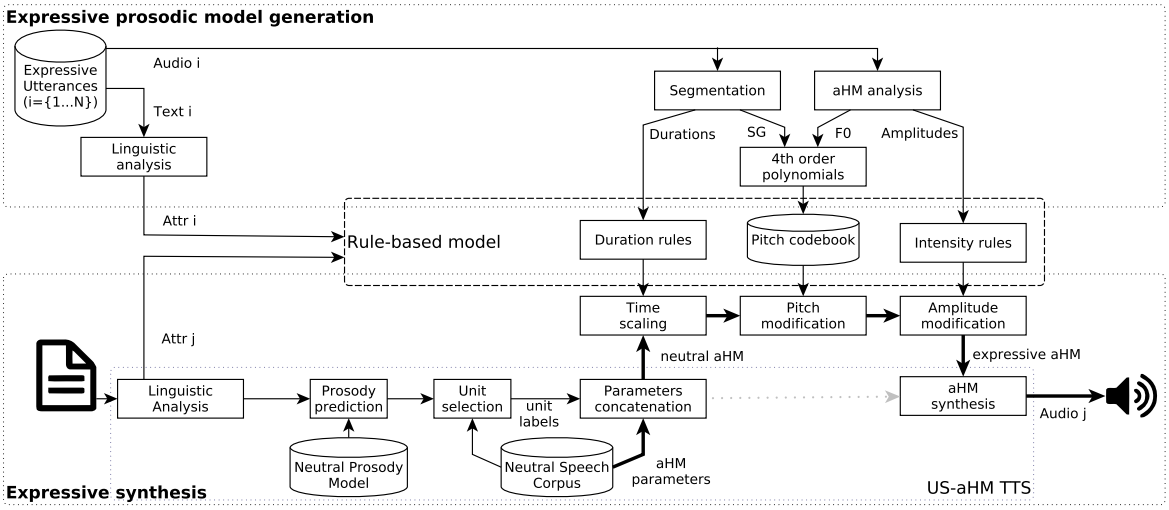


Figure IV.1: Hybrid US-aHM TTS expressive synthesis framework based on a rule-based prosodic model.

containing the desired expressive speaking style. During the synthesis stage, the TTS system converts any input text to the target expressive speaking style from a neutral Spanish female voice.

IV.3.1 Expressive prosodic model generation

Firstly, it is worth remarking that the basic intonation unit considered in our synthesis framework is the stress group (henceforth SG) defined as a stressed syllable plus all succeeding unstressed syllables within the same compound sentence (Erro et al. 2010; Iriondo et al. 2007). As it can be observed in Fig. IV.1, each selected expressive utterance is linguistically analysed and segmented, obtaining the following SG-level attributes (see Fig. IV.2):

- **Intonational Phrase (IP)**: This attribute identifies to which IP within the utterance the SG belongs to.
- **nSGs**: Refers to the total number of SGs within each IP.
- **SGpos**: The SGpos indicates the position of the SG within each IP, differentiating PRE (unstressed SG of initial position), BEG (Beginning), MID (Middle), PEN (Penultimate), and END (Final SG).

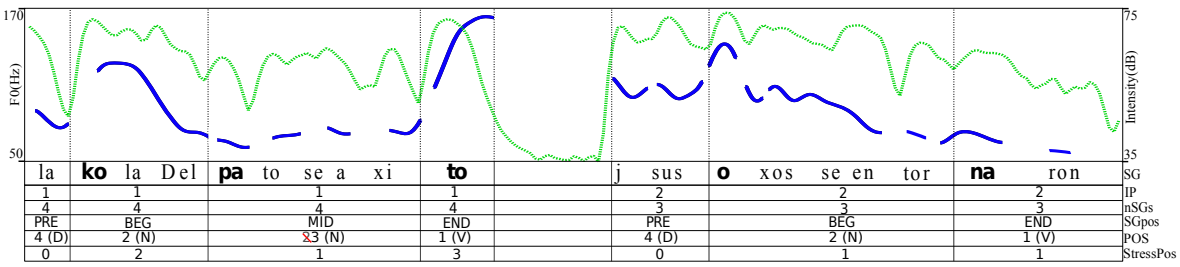


Figure IV.2: Increasing suspense example: *La cola del pato se agitó, y sus ojos se entornaron* ("The duck's tail twitched, and its eyes narrowed"). Stressed syllables are in bold. The phonetic transcription of the SG tier is in SAMPA for Spanish. Blue solid line: F0. Green dotted line: Intensity.

- **Part Of Speech (POS)**: Freeling POS labels for Spanish are used (Lloberes et al. 2010). A relevance score is assigned to each POS label: verbs (1), nouns and adjectives (2), adverbs (3), and rest (4). If a SG complements another SG, its relevance score is degraded (except in verbs).
- **StressPos**: is a numerical value that represents the position of the stressed vowel end within the SG, i.e., first (1), second (2), and third (3) SG part. Unstressed SG before the first stressed syllable is assigned a 0.

The expressive utterances considered to derive the prosodic model are also analysed by means of the aHM technique implemented in the COVAREP (version 1.4.1) algorithms (Degottex et al. 2014) to extract the F0 and amplitude parameters. A pitch contour is obtained for each SG by considering both the aHM F0 parameters and the SG segmentation. The SG-level attributes together with the 4th-order coefficients obtained from the polynomial fitting of the pitch contour (Iriondo et al. 2007) are used to define each SG codeword (i.e., a vector containing attributes and polynomial coefficients) that is stored in the pitch codebook (CB). Regarding intensity and durations, a set of rules is also derived from a detailed analysis of the utterances.

IV.3.2 Expressive synthesis stage

At run time, the input text to be synthesized is fed into the US-aHM TTS system. The TTS system extracts the aforementioned linguistic attributes and accesses the rule-based prosodic model to get the corresponding expressive prosodic conversions (see Fig. IV.1). After retrieving the selected units from the neutral speech database, the corresponding aHM parameters are converted according to the target expressive style. Finally, the aHM-based synthesis generates the synthetic expressive speech.

Linguistic attributes combined with some rules are used to retrieve from the pitch codebook the possible pitch contours for each SG. Then, a simple yet effective combination cost is defined to assess which combinations are more suitable to be concatenated. Concretely, when two consecutive SG pitch contours come from different utterances, the cost is increased by 1. If several combinations contain the minimum cost, the final sequence is randomly chosen in order to increase synthesis variability (Alías et al. 2005). Following a similar approach to (Alías et al. 2005), an interpolation technique is applied to avoid discontinuities between consecutive SGs pitch contours. Thereupon, since we deal with two different speakers, the obtained pitch contour must be scaled, shifting it from the source f0 reference value (f0 mean of the expressive utterances) to the target f0 reference value (f0 mean of neutral corpus used in the synthesis). Finally, SG-level 4th order polynomial fitting is applied to the original f0 curve and the resulting pitch contour is replaced by the pitch contour obtained from the codebook (see Fig IV.3.).

IV.4 Developing a rule-based prosodic model of increasing suspense

IV.4.1 Material

The increasing suspense speech was obtained from an audiobook interpreted by a Spanish professional male storyteller. The storyteller interpreted a story that belongs to the fantasy and adventures genres (with children and pre-teenagers as its main target audience). The audiobook contains around 4 hours of storytelling speech. However, only eight utterances that fully fit the expressive profile of increasing suspense have been found. All the utterances were manually segmented to allow reliable subsequent analyses. Fig. IV.2 depicts an example of the complete labelling at the SG-level of an increasing suspense utterance.

IV.4.2 Analysis oriented to synthesis

In this section the rule-based prosodic model specifically conceived for our US-aHM Neutral TTS synthesis framework is described.

IV.4.2.1 Duration.

Theune *et al.* observed a pause of 1.04 s between both zones in their utterance. However in our set of utterances, such pause duration is much lower (mean duration of 0.4 s±0.1 s). Furthermore, Theune *et al.* observed a progressive increase of stressed vowels durations in the first zone. This pattern was detected in one of the eight increasing suspense utterances. Nevertheless, as 7 out of the 8 sentences did not present that pattern, we opted for not including this Theune *et al.* observation in our rules. Despite further detailed analyses of rhythm patterns and changes of speech tempo between both zones, no clear patterns whatsoever were found. Therefore, in this work, the only duration rule included in the synthesis framework is to apply a value of 0.4 s to the pause between both zones.

IV.4.2.2 Fundamental Frequency.

Similarly to Theune *et al.* we have observed a tendency consisting of a F0 increase along zone 1 and a gradual decrease in zone 2 in all the utterances. However, not all the utterances show

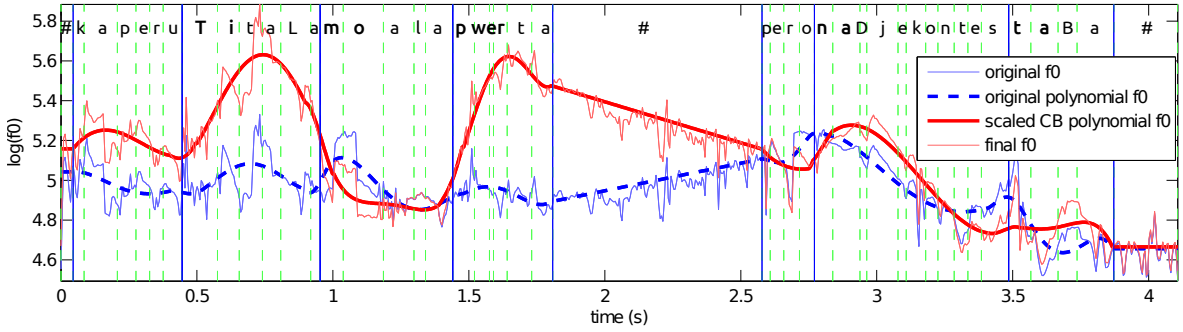


Figure IV.3: Pitch modification example. *"Caperucita llamó a la puerta, pero nadie contestaba"* ("Little Red Cap knocked on the door, but no one answered").

a gradual F0 increase in all the stressed syllables of the first zone. For instance, in Fig. IV.2 it can be observed that the word *"pato"* ("duck") is not F0-accented in the stressed syllable. On the contrary, the F0 curve drops as if the storyteller wanted to emphasize even more the last SG "*agitó*" ("twitched"). This phenomenon also manifests in the rest of utterances without a gradual increase, being related to the POS of the SG. Other examples can also be an adjective complementing a verb, e.g., "***era*** *evidente*" ("it **was** clear"), or an adjective complementing a noun, e.g., "***hombre*** *alto*" ("tall **man**"). Another clear pattern observed in all the utterances is a substantial rise of F0 in the last SG of zone 1. This rise is preceded in all cases by a downfall except if the penultimate SG of zone 1 is a verb, e.g., *"**inundó** la habitación"* ("**flooded** the room"), where two F0 rises are present (reaching a higher point in the last SG). Finally, within zone 2 the only clear pattern observed is a F0 boost in the first SG whose POS corresponds to a verb, a noun, an adjective, or an adverb, accompanied with a gradual decrease until the end of the utterance.

From this analysis the rules to access the pitch codebook are derived, i.e. which linguistic attributes are used and in what order. Thus, codeword candidates are obtained through a selection based first on the first attribute, a subsequent selection within the previous subset which meet the second attribute, and so on. When in a selection step none of the codewords meet the attribute, codewords nearest to this attribute are chosen and the process is finished.

For the first zone:

- Pitch contours for each SG are retrieved according to its position within the zone (note that the IP is equivalent to the zone in increasing suspense) and its stress position, in that order.
- In case of having more than one MID SG, the POS is also considered (before the stress position) in order to establish which SG should be F0-accented.

For the second zone:

- The SG pitch contours are retrieved according to the number of SGs, the SG position, and the stress position, in that order.

IV.4.2.3 Intensity

Similarly to what was observed in the analysis of F0, the gradual intensity increase reported by Theune *et al.* was not observed either within the analysed material. Therefore, we opted for modifying energy coherently with the F0 curve following (Sorin et al. 2015), which is based on the fundamental relationship between the instantaneous F0 and instantaneous energy of a speech signal. In order to validate this approach, we performed a correlation analysis between F0 and intensity curves in our speech corpus obtaining a value of $r = 0.654$ and a linear regression slope of 9.8 dB/octave. These values confirm the viability of the considered approach as they are very similar to the $r = 0.670$ and 9 dB/octave obtained in (Sorin et al. 2015).

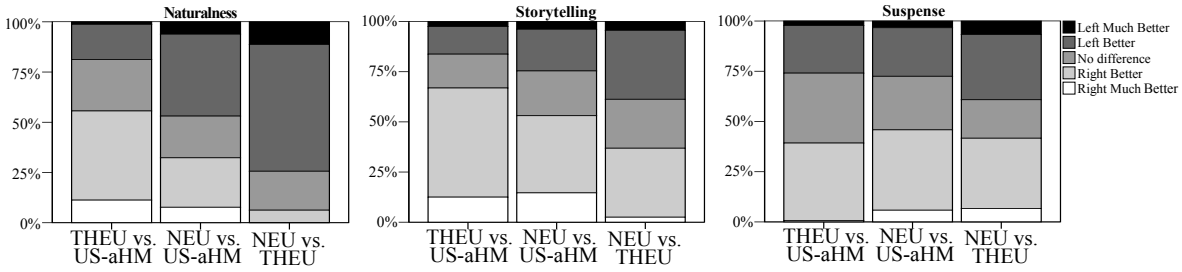


Figure IV.4: Percentage bars representing the answers of the subjects for each evaluation. NEU: Neutral; THEU: Theune *et al.*

IV.5 Perceptual evaluation

The perceptual evaluation was conducted by means of a 5-point scale ($[-2, +2]$) Comparative Mean Opinion Score (CMOS) on 5 synthetic utterances using the TRUE online platform (Planet et al. 2008). Such utterances, were generated from made-up sentences with a semantic content related to stories (see for example Fig. IV.3). In each comparison, two utterances synthesized through the aHM-US TTS framework were presented to the evaluator (randomly ordered in each comparison), using either the introduced rule-based prosodic model, the fixed rules of Theune et al., or the neutral synthetic speech as baseline (5 utterances x 3 methods = 15 comparisons).

All subjects were asked to relatively grade both speech fragments in terms of naturalness, storytelling resemblance, and expression of suspense. As no specific target was available, no reference audio was included to avoid biasing the CMOS towards our method if some of the prosodic patterns of the analysed utterances were included. It is worth noting that three control points were added to remove unreliable evaluators from subsequent analyses (18 comparisons plus a final survey in total). From the total of 32 subjects (mean age 34±10), 4 were discarded for the aforementioned reliability criterion. The results from the subjective test were analysed in terms of percentage scores (see Fig. IV.4) and differences in the CMOS median (Mdn) values. The latter, were analysed by means of a one-sample Wilcoxon signed-rank test with significance level $p < 0.05$.

Regarding naturalness, our approach significantly outperforms Theune *et al.* (Mdn = 1; 55% US-aHM better/much better) and it is perceived equal to the neutral synthetic counterpart (Mdn = 0; 53% US-aHM no difference/better/much better). On the contrary, the method of Theune *et al.* obtains significantly lower results than the neutral synthetic speech (Mdn = -1; 74% neutral better/much better). Moreover, storytelling quality results indicates that the proposed method outperforms both Theune *et al.* (Mdn = 1; US-aHM 63% better/much better) and the neutral synthetic speech (Mdn = 1; US-aHM 53% better/much better). Differently, Theune *et al.* is perceived similar to neutral in this evaluation (Mdn = 0; neutral 63% no difference/better/much better). Finally, results regarding the expression of suspense show that all methods are perceived similarly, even though the proposed method is perceived as slightly better with respect to Theune *et al.* (26% preferred Theune *et al.* and 40% preferred the US-aHM method) together with a significant preference in front of the neutral synthesis (Mdn = 1; US-aHM 48% better/much better).

IV.6 Conclusions

In this paper, a hybrid text-to-speech synthesis framework based on unit selection and adaptive Harmonic Model has been adapted to generate storytelling suspense speech using a rule-based prosodic model derived from the analysis of few but representative utterances of increasing suspense (less than 1 min of speech). The US-aHM approach has been evaluated on a subjective test comparing it to the fixed prosodic rules introduced in (Theune et al. 2006), using the neutral synthetic speech as baseline. Our proposed approach obtains better naturalness and storytelling resemblance, although it is similar to the baseline in terms of suspense arousal.

In this respect, some evaluators commented that a warmer and more whispery voice could improve the suspenseful perception. From these results, we reckon that voice quality should be included in future works as a means to fully resemble suspense. Moreover, we will keep working to gather more data to improve the robustness of the model. Finally, since comparable acoustic patterns among storytellers of similar linguistic communities have been observed (Montaño and Alías 2015), we plan to study to what extent the current results obtained for Spanish are generalizable.

IV.7 Acknowledgements

Raúl Montaño and Marc Freixes thank the support of the European Social Fund (ESF) and the Catalan Government (SUR/DEC) for the pre-doctoral FI grants No. 2015FI_B2 00110 and 2016FI_B2 00094, respectively. This work has been partially funded by SUR/DEC (grant ref. 2014-SGR-0590). We also want to thank the people that took the perceptual test.

References

Alías, Francesc, Iriondo, Ignasi, Formiga, Lluís, Gonzalvo, Xavier, Monzo, Carlos, and Sevillano, Xavier (2005). "High quality Spanish restricted-domain TTS oriented to a weather forecast application". In: *Interspeech*. Lisbon, Portugal, pp. 2573–2576.

Alías, Francesc, Sevillano, Xavier, Socoró, Joan Claudi, and Gonzalvo, Xavier (2008). "Towards high quality next-generation Text-to-Speech synthesis: a Multidomain approach by automatic domain classification". In: *IEEE Transactions on Audio, Speech and Language Processing* vol. 16, no. 7, pp. 1340–1354.

Alofs, Thijs, Theune, Mariët, and Swartjes, Ivo (2015). "A tabletop interactive storytelling system: Designing for social interaction". In: *International Journal of Arts and Technology* vol. 8, no. 3, pp. 188–211.

Barra-Chicote, Roberto, Yamagishi, Junichi, King, Simon, Montero, Juan M., and Macias-Guarasa, Javier (2010). "Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech". In: *Speech Communication* vol. 52, no. 5, pp. 394–404.

Charfuelan, Marcela and Steiner, Ingmar (2013). "Expressive speech synthesis in MARY TTS using audiobook data and EmotionML". In: *Interspeech*, pp. 1564–1568.

Degottex, Gilles, Kane, John, Drugman, Thomas, Raitio, Tuomo, and Scherer, Stefan (May 2014). "COVAREP - A collaborative voice analysis repository for speech technologies". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 960–964.

Degottex, Gilles and Stylianou, Yannis (2013). "Analysis and Synthesis of Speech Using an Adaptive Full-Band Harmonic Model". In: *IEEE Transactions on Audio, Speech, and Language Processing* vol. 21, no. 10, pp. 2085–2095.

Doukhan, David, Rilliard, Albert, Rosset, Sophie, Adda-Decker, Martine, and d'Alessandro, Christophe (2011). "Prosodic analysis of a corpus of tales". In: *Interspeech*, pp. 3129–3132.

Erro, Daniel, Navas, Eva, Hernáez, Inma, and Saratxaga, Ibon (2010). "Emotion conversion based on prosodic unit selection". In: *IEEE Transactions on Audio, Speech, and Language Processing* vol. 18, pp. 974–983.

Hu, Qiong, Richmond, Korin, Yamagishi, Junichi, and Latorre, Javier (Aug. 2013). "An experimental comparison of multiple vocoder types". In: *8th ISCA Workshop on Speech Synthesis (SSW)*, pp. 135–140.

Iriondo, Ignasi, Socoró, Joan Claudi, and Alías, Francesc (Apr. 2007). "Prosody Modelling of Spanish for Expressive Speech Synthesis". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Vol. 4, pp. 821–824.

Jauk, Igor, Bonafonte, Antonio, Lopez-Otero, Paula, and Docio-Fernandez, Laura (2015). "Creating Expressive Synthetic Voices by Unsupervised Clustering of Audiobooks". In: *Interspeech*, pp. 3380–3384.

Kafentzis, George P., Degottex, Gilles, Rosec, Olivier, and Stylianou, Yannis (May 2014). "Pitch modifications of speech based on an adaptive Harmonic Model". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Florence, Italy: IEEE, pp. 7924–7928.

Lehne, Moritz and Koelsch, Stefan (2015). "Toward a general psychological model of tension and suspense". In: *Frontiers in Psychology* vol. 6, pp. 1–11.

Leite, Iolanda, McCoy, Marissa, Lohani, Monika, Ullman, Daniel, Salomons, Nicole, Stokes, Charlene K., Rivers, Susan, and Scassellati, Brian (2015). "Emotional Storytelling in the Classroom: Individual versus Group Interaction between Children and Robots". In: *ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 75–82.

Lloberes, Marina, Castellón, Irene, and Padró, Lluís (2010). "Spanish FreeLing Dependency Grammar". In: *International Conference on Language Resources and Evaluation (LREC)*. La Valletta, Malta.

Lorenzo-Trueba, Jaime, Barra-Chicote, Roberto, San-Segundo, Rubén, Ferreiros, Javier, Yamagishi, Junichi, and Montero, Juan M. (2015). "Emotion transplantation through adaptation in HMM-based speech synthesis". In: *Computer Speech & Language* vol. 34, no. 1, pp. 292–307.

Montaño, Raúl and Alías, Francesc (2015). "The role of prosody and voice quality in text-dependent categories of storytelling across languages". In: *Interspeech*. Dresden, Germany, pp. 1186–1190.

Montaño, Raúl, Alías, Francesc, and Ferrer, Josep (2013). "Prosodic analysis of storytelling discourse modes and narrative situations oriented to Text-to-Speech synthesis". In: *8th ISCA Workshop on Speech Synthesis (SSW)*, pp. 171–176.

Planet, Santiago, Iriondo, Ignasi, Martínez, Elisa, and Montero, José Antonio (2008). "TRUE: an online testing platform for multimedia evaluation". In: *Workshop on Corpora for Research on Emotion and Affect*. Marrakech, Morocco.

Prahallad, Kishore and Black, Alan W. (2011). "Segmentation of Monologues in Audio Books for Building Synthetic Voices". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* vol. 19, no. 5, pp. 1444–1449.

Scherer, Klaus R. (2003). "Vocal communication of emotion: A review of research paradigms". In: *Speech Communication* vol. 40, no. 1-2, pp. 227–256.

Schröder, Marc (2001). "Emotional Speech Synthesis: A review". In: *Interspeech*. Aalborg, Denmark, pp. 561–564.

Schröder, Marc (2009). "Expressive speech synthesis: Past, present, and possible futures". In: *Affective information processing*, pp. 111–126.

Sorin, Alexander, Shechtman, Slava, and Pollet, Vincent (2015). "Coherent modification of pitch and energy for expressive prosody implantation". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 4914–4918.

Theune, Mariët, Meijs, Koen, Heylen, Dirk, and Ordelman, Roeland (2006). "Generating expressive speech for storytelling applications". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* vol. 14, pp. 1137–1144.

Yamagishi, Junichi and Kobayashi, Takao (Feb. 2007). "Average-Voice-based Speech Synthesis using HSMM-based Speaker Adaptation and Adaptive Training". In: *IEICE Transactions on Information and Systems* vol. E90–D, no. 2, pp. 533–543.

Zovato, Enrico, Pacchiotti, Alberto, Quazza, Silvia, and Sandri, Stefano (2004). "Towards Emotional Speech Synthesis: A rule based approach". In: *5th ISCA Workshop Speech Synthesis (SSW)*, pp. 219–220.

Authors' addresses

Raúl Montaño GTM – Grup de Recerca en Tecnologies Mèdia, La Salle - Universitat Ramon Llull
Quatre Camins, 30, 08022 Barcelona, Spain
raulma@salle.url.edu

Marc Freixes GTM – Grup de Recerca en Tecnologies Mèdia, La Salle - Universitat Ramon Llull
Quatre Camins, 30, 08022 Barcelona, Spain
marc.freixes@salle.url.edu

Paper V

Influence of tense, modal and lax phonation on the three-dimensional finite element synthesis of vowel [ɑ]

Marc Freixes, Marc Arnela, Joan Claudi Socoró, Francesc Alías, Oriol Guasch

Published in *Proceedings of IberSPEECH 2018.*, November 2018, pp. 132–136. DOI: 10.21437/IberSPEECH.2018-28.

Abstract

One-dimensional articulatory speech models have long been used to generate synthetic voice. These models assume plane wave propagation within the vocal tract, which holds for frequencies up to ∼5 kHz. However, higher order modes also propagate beyond this limit, which may be relevant to produce a more natural voice. Such modes could be especially important for phonation types with significant high frequency energy (HFE) content. In this work, we study the influence of tense, modal and lax phonation on the synthesis of vowel [ɑ] through 3D finite element modelling (FEM). The three phonation types are reproduced with an LF (Liljencrants-Fant) model controlled by the R_d glottal shape parameter. The onset of the higher order modes essentially depends on the vocal tract geometry. Two of them are considered, a realistic vocal tract obtained from MRI and a simplified straight duct with varying circular cross-sections. Long-term average spectra are computed from the FEM synthesised [ɑ] vowels, extracting the overall sound pressure level and the HFE level in the 8 kHz octave band. Results indicate that higher order modes may be perceptually relevant for the tense and modal voice qualities, but not for the lax phonation.

V.1 Introduction

For many years, works on articulatory speech synthesis have considered a simplified one-dimensional (1D) representation of the vocal tract. This is built from the so-called vocal tract area functions, which describe the cross-sectional area variations along the vocal tract center midline (see e.g., Story et al. 1996). Voice is then synthesised by simulating the propagation of acoustic waves within this 1D representation of the vocal tract (see e.g., Story 2013; Birkholz 2013; Stone et al. 2018). However, 1D approaches assume plane wave propagation,

so they can only correctly approximate the acoustics of the vocal tract in the frequency range below 4-5 kHz. Beyond this limit, not only planar modes get excited but also higher order propagation modes appear, which strongly change the high frequency energy (HFE) content of the spectrum (Blandin et al. 2015; Arnela et al. 2016b) compared to that from a 1D model. Although the high frequency range has not received much attention in the literature, some recent studies point out that the HFE may be relevant for voice quality, speech localisation, speaker recognition and intelligibility (see Monson et al. 2014 and references therein).

On the other hand, three-dimensional (3D) models do not need to assume plane wave propagation, since they can directly deal with 3D vocal tract geometries to emulate the complex acoustic field generated during voice production (Vampola et al. 2008; Takemoto et al. 2010; Arnela et al. 2016a). However, higher order modes do not always appear even if a 3D acoustic model is used. As shown in (Arnela et al. 2016b), a straightened vocal tract based on circular cross-sections prevents the onset of such modes due to radial symmetry, in contrast to what occurs for realistic vocal tract geometries based on MRI data. Other vocal tract geometries simplifications were studied in that work, all of them showing large variations in the HFE while keeping a similar behaviour for low frequencies. One can then assert that the vocal tract shape is determinant for the HFE content of the generated sound. However, the vocal tract shape is not the only factor affecting the HFE. The type of phonation can also modify it, as shown for instance in (Monson et al. 2011) for sustained vowels with loud and soft phonation.

In this work we study the effect of tense, modal and lax phonation on the synthesis of vowel [ɑ], paying special attention to the HFE content. These three phonation types are reproduced using an LF (Liljencrants-Fant) model (Fant et al. 1985). Although this model cannot consider the interaction between the vocal tract and the vocal folds (Murtola et al. 2018; Erath et al. 2013), it has proved to be useful to explore the phonatory tense-lax continuum (Murphy et al. 2017) by controlling the R_d glottal shape parameter (Fant 1995). Regarding the vocal tract, we consider an MRI-based realistic geometry, and its simplified counterpart considering circular cross-sections in a straightened midline (Arnela et al. 2016b). This allows us to somewhat "activate" and "deactivate" the higher order modes. Different versions of vowel [ɑ] are generated by convolving the LF glottal source signals with the vocal tract impulses responses obtained using a 3D acoustic model based on the Finite Element Method (FEM) (Arnela and Guasch 2013). In order to analyse the relevance of the higher order propagation modes for the lax, modal and tense phonation, the long-term average spectra (LTAS) and the HFE levels of the synthesised vowels are computed and compared.

The paper is structured as follows. The methodology used to study the production of vowel [ɑ] with the three phonation types and the two vocal tract geometries is explained in Section V.2. Next, the obtained results are discussed in Section V.3. Finally, conclusions and future work are presented in Section V.4.

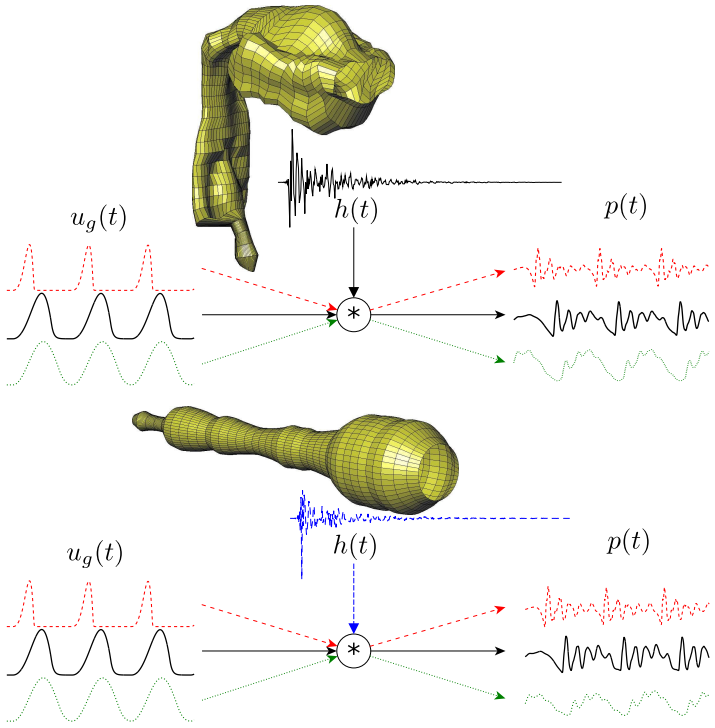


Figure V.1: Synthesis of vowel [ɑ] with a realistic vocal tract geometry (above) and its simplified counterpart of circular cross-sections in a straightened midline (below). Three phonation types are considered to reproduce a tense (dashed red line), a modal (solid black line) and a lax (dotted green line) voice production. The output pressure signal $p(t)$ is computed as the convolution of the glottal source $u_g(t)$ with the vocal tract impulse response $h(t)$ obtained from 3D FEM simulations.

V.2 Methodology

Figure V.1 depicts the process followed to synthesise six versions of vowel [ɑ]. These were obtained by convolving three glottal source signals with the FEM impulse responses of two vocal tract geometries that produce this vowel. In particular, and as mentioned before, we used the realistic vocal tract and the simplified straightened simplification with circular cross-sections from (Arnela et al. 2016b) (see Section V.2.1), and computed their impulse responses $h(t)$ using the FEM (see Section V.2.2). The glottal source signals $u_g(t)$ were generated by means of an R_d controlled LF model. The values $R_d = 0.3, 1$ and 2.7 were selected from the R_d range $[0.3, 2.7]$ (see Fant 1995) to reproduce a tense, a modal, and a lax phonation, respectively (see Section V.2.3).

For each vowel, the LTAS was computed as the Welch's power spectral density estimate, with a 15 ms hamming window, 50% overlap and a 2048-point FFT. The overall energy levels and the HFE levels in the 8 kHz octave band were also extracted as in (Monson et al. 2011). The 16 kHz octave band was not considered, since HFE changes in this frequency range were found almost perceptually irrelevant in (Monson et al. 2011).

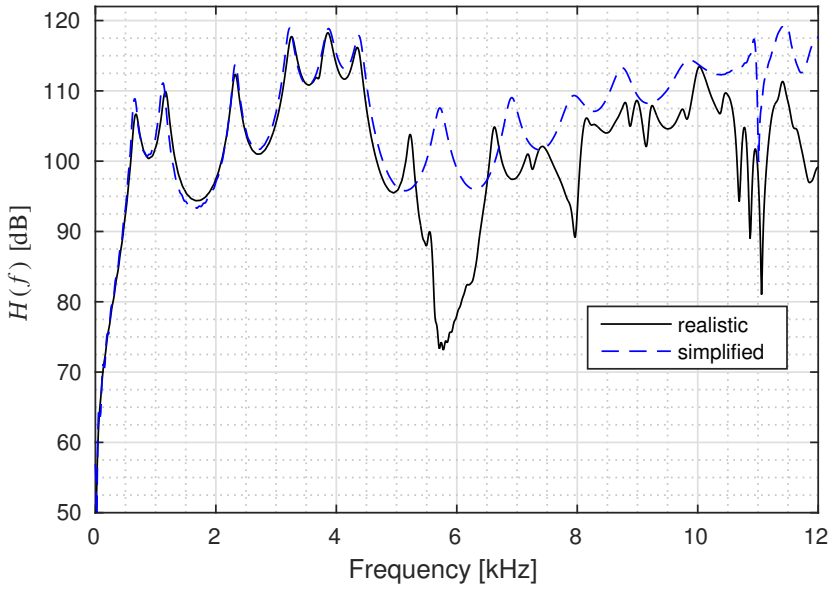


Figure V.2: Vocal tract transfer function $H(f)$ for the vowel [ɑ] with the realistic and simplified vocal tract geometries.

V.2.1 Vocal Tract Geometries

Two vocal tract geometry simplifications of vowel [ɑ] have been employed in this work, namely, the realistic configuration and the simplified straight vocal tract with circular shape (see Fig. V.1). These geometries were obtained in (Arnela et al. 2016b) by simplifying the MRI-based vocal tract geometries in (Aalto et al. 2014). In a nutshell, the procedure consisted in the following. First, the subglottal tube, the face and the lips were removed from the original geometry (see Arnela et al. 2016a for a detailed analysis of the lips influence on simulations). Moreover, side branches such as the piriform fossae and valleculae were occluded (see e.g. Takemoto et al. 2010; Takemoto et al. 2013 for their acoustic effects). Cross-sections were next extracted as typically done to generate 1D area functions, but preserving their shapes and locations in the vocal tract midline. The realistic configuration was generated by linearly interpolating the resulting cross-sections. As shown in (Arnela et al. 2016b), this simplification provides very similar results to the original MRI-based vocal tract geometry without branches.

In the simplified straight vocal tract configuration, the cross-sectional shapes were modified to be that of a circle, preserving the same area. These circular cross-sections were located in a straightened version of the vocal tract midline and then linearly interpolated. The two configurations are hereafter referred as the realistic and the simplified vocal tracts.

V.2.2 Vocal Tract impulse response

The impulse response of each vocal tract geometry was computed using a custom finite element code that numerically solves the acoustic wave equation,

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

combined with a Perfectly Matched Layer (PML) to account for free-field propagation (Arnela and Guasch 2013). In Eq. (V.1) $p(\mathbf{x}, t)$ is the acoustic pressure, ∂_{tt}^2 stands for the second

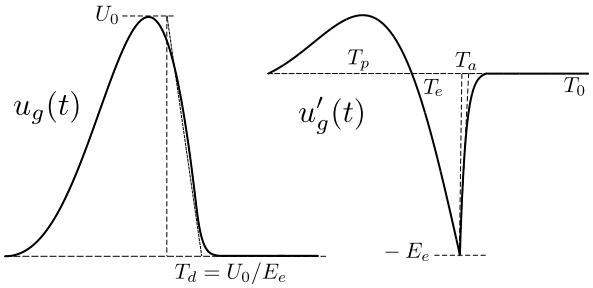


Figure V.3: Glottal flow $u_g(t)$ and its time derivative $u'_g(t)$ according to the LF model (Fant et al. 1985).

order time derivative, and c_0 is the speed of sound which is set to the usual value of 350 m/s. A Gaussian pulse was introduced on the glottal cross-sectional area as an input volume velocity $u_g(t)$. This pulse is of the type

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

with $T_{gp} = 0.646/f_c$ and $f_c = 10$ kHz. Wall losses were considered by imposing a boundary admittance coefficient of $\mu = 0.005$ on the vocal tract walls. A 20 ms simulation was then performed capturing the acoustic pressure $p_0(t)$ at a node located outside of the vocal tract, 4 cm away from the mouth aperture center. The sampling frequency was set to $f_s = 8000$ kHz, which ensures a restrictive stability condition of the Courant-Friedrich-Levy type required by explicit numerical schemes (see Arnela and Guasch 2013 for details on the numerical scheme).

A vocal tract transfer function $H(f)$ was computed from each simulation to compensate for the slight energy decay in frequency of the Gaussian pulse. This is defined as

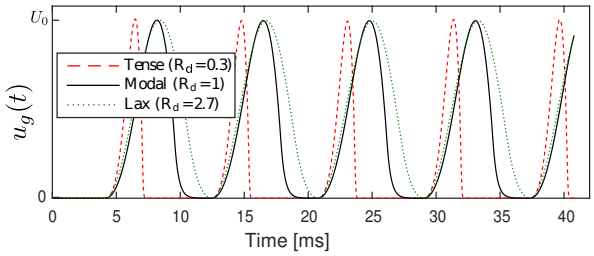
$$H(f) = \frac{P_o(f)}{U_g(f)}, \tag{V.3}$$

with $P_o(f)$ and $U_g(f)$ being the Fourier Transform of $p_o(t)$ and $u_g(t)$, respectively. $H(f)$ was computed up to 12 kHz, to allow the calculation of HFE level in the 8 kHz octave band (Monson et al. 2011). The vocal tract transfer functions $H(f)$ for the realistic and the simplified geometries of vowel [a] are shown in Fig. V.2 (also reported in Arnela et al. 2016b, but only up to 10 kHz). As can be observed, planar modes are mainly produced below 5 kHz giving place to the first vowel formants. Beyond this value, higher order modes can also propagate, resulting in the more complex spectrum of the realistic geometry. Note, however, that these modes do not appear in the spectrum of the simplified configuration. The radial symmetry of this geometry prevents their onset (Blandin et al. 2015; Arnela et al. 2016b).

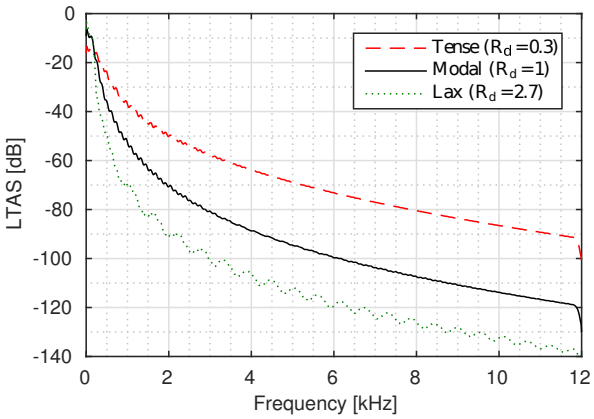
Finally, the inverse Discrete Fourier Transform was applied to the vocal tract transfer functions $H(f)$ to obtain the vocal tract impulse responses $h(t)$ of the two geometries (see Fig. V.1).

V.2.3 Voice Source Signal

An LF model (Fant et al. 1985) was used to produce the voice source signal. This model approximates the glottal flow $u_g(t)$ and its time derivative $u'_g(t)$ in terms of four parameters



(a) Waveform



(b) Long-term average spectrum

Figure V.4: Glottal source for a tense ($R_d = 0.3$), a modal ($R_d = 1$), and a lax ($R_d = 2.7$) phonation.

(T_p, T_e, T_a, E_e) that describe its time-domain properties (see Fig. V.3). The control of this model can be simplified with the single glottal shape parameter R_d (Fant 1995). This is defined as

$$R_d = \frac{T_d}{T_0}\frac{1}{110} = \frac{U_0}{E_e}\frac{F_0}{110}, \tag{V.4}$$

where T_d is the declination time, T_0 the period, and F_0 the fundamental frequency. The declination time T_d corresponds to the quotient between the glottal flow peak U_0 and the negative amplitude of the differentiated glottal flow E_e.

In this work, we used the Kawahara's implementation of the LF model (Kawahara et al. 2017), which generates a free-aliasing excitation source signal. We adapted this model to our purposes, modifying the sampling frequency from its original value of 44100 Hz to 24 kHz. Moreover, we introduced the R_d glottal shape parameter. This allows one to easily control the voice source with a single parameter, which runs from $R_d = 0.3$ for a very adducted phonation, to $R_d = 2.7$ for a very abducted phonation (see Fant 1995). From the R_d range $[0.3, 2.7]$ two extreme values plus a middle one were chosen. We used $R_d = 0.3$ to generate a tense phonation, $R_d = 2.7$ for a lax production, and $R_d = 1$ for a normal (modal) voice quality. With regard to F0, a pitch curve was obtained from a real sustained vowel lasting 4.4 seconds. This pitch contour was placed around 120 Hz to generate all the source signals. Figure V.4a shows four periods of the three simulated voice source waveforms. Moreover, the LTAS of the glottal source signals are represented in Fig. V.4b. As observed, the phonation type obviously changes the glottal pulse shape, thus modifying the spectral energy distribution of the source signal.

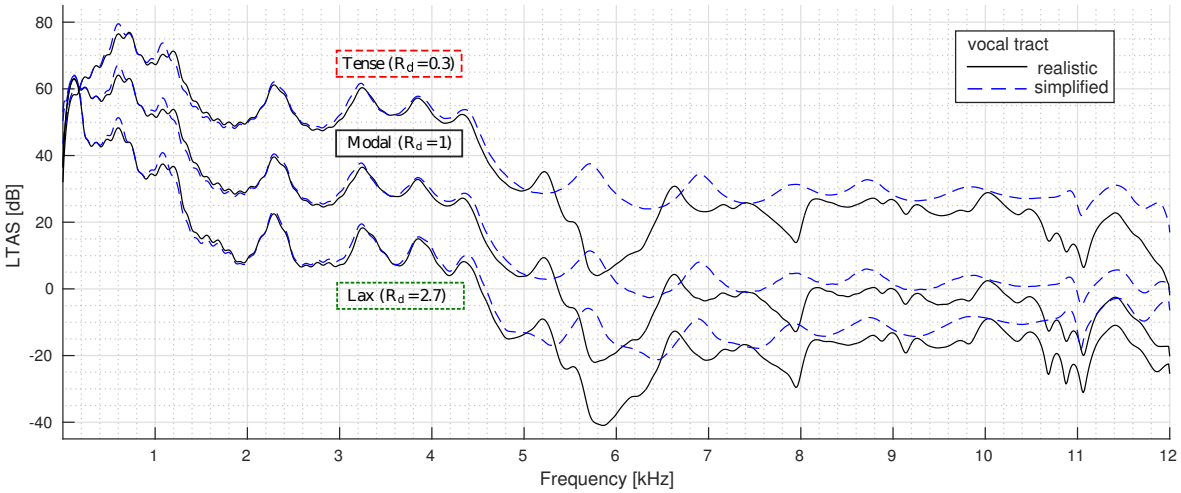


Figure V.5: Long-term average spectra (LTAS) of the FEM synthesised vowel [ɑ] using the realistic and simplified vocal tract geometries with a tense ($R_d = 0.3$), a modal ($R_d = 1$), and a lax ($R_d = 2.7$) phonation.

V.3 Results

Six versions of vowel [ɑ] (see Fig. V.1) have been generated using the three glottal source signals corresponding to a tense, a modal and a lax phonation, and the two impulse responses obtained from the 3D FEM simulations of the realistic and simplified vocal tract geometries. The six synthesised vowels are normalised with the same scaling factor to obtain reasonable sound pressure levels. This factor has been selected so as to produce 70 $\mathrm{dB_{SPL}}$ in the realistic geometry with a modal phonation ($R_d = 1$). The LTAS have then been computed for each audio.

Figure V.5 shows the obtained LTAS for the six generated vowels. As also appreciated in the vocal tract transfer functions (see Fig. V.2), small differences between geometries are produced for frequencies below 5 kHz, whereas beyond this range higher order modes propagate in the realistic case, thus inducing larger deviations. This behaviour can be observed for the three phonation types. Essentially the glottal source modifies the overall energy level and also introduces an energy decay in frequency (compare Fig. V.2 with Fig. V.5). This decay, known as the spectral tilt, strongly depends on the phonation type. The laxer the phonation the larger the spectral tilt (Fant 1995). Furthermore, the voice source also affects the energy balance of the first harmonics (below $\sim$ 500 Hz). For instance, the lax phonation has the lowest overall energy values among all phonation types. However, one can see that the first harmonic (close to 120 Hz) has larger amplitude levels than the rest of the spectrum, in contrast to what occurs for the other phonations.

HFE levels have been computed by integrating the power spectral density in the 8 kHz octave band, as in (Monson et al. 2011). In addition, the overall energy levels have been calculated following the same procedure but for the whole examined frequency range.

The obtained results are listed in Table V.1. Note first that in the realistic case with a modal phonation ($R_d = 1$) the overall level is 70 $\mathrm{dB_{SPL}}$. Remember that this value was fixed to compute the scaling factor used to normalise the audio files. The overall level variations

Table V.1: Overall and High-Frequency Energy (HFE) levels (in dB) obtained in the realistic and simplified vocal tract configurations of vowel [ɑ] with a tense ($R_d = 0.3$), a modal ($R_d = 1$), and a lax ($R_d = 2.7$) phonation. Δ denotes the difference between the two vocal tract geometries.

R_d	Geometry	Overall	ΔOverall	HFE	ΔHFE
0.3	realistic	82.2	1.2	41.4	5.8
	simplified	83.4		47.3	
1	realistic	70.0	1.3	14.9	5.9
	simplified	71.3		20.8	
2.7	realistic	63.5	1.4	1.0	5.6
	simplified	64.9		6.6	

for the other configurations will thus correspond to modifications either introduced by the vocal tract geometry or by the glottal source. As expected, the larger the R_d value (laxer phonation) the smaller the overall levels.

Far more interesting is to compare the results between geometries. The HFE levels decay between 5.6 dB and 5.9 dB for the realistic vocal tract depending on the phonation type, which only manifests as an overall level difference of 1.2 dB and 1.4 dB. The higher order modes tend to reduce the levels in the HFE content. According to (Monson et al. 2011), minimum difference limen scores of about 1 dB are given for normal-hearing listeners in the 8 kHz octave band, so one may hypothesise that the higher order modes may be perceptually relevant. However, depending on the phonation type the HFE could be too small to notice any difference. This seems to be the case of the lax phonation ($R_d = 2.7$), which gives HFE levels of 1.0 dB and 6.6 dB, depending on the geometry. We may then conjecture, that for this phonation type no differences in the outputs from the two geometries will be perceived. In other words, we would not notice the influence of higher order modes.

V.4 Conclusions

In this work we have studied the influence of tense, modal and lax phonation on the 3D finite element synthesis of vowel [ɑ], considering a realistic and a simplified vocal tract geometry. The 3D simulations behave very similarly for both geometries below 5 kHz, but significant differences appear beyond this frequency because of the rising of higher order propagation modes. It is worth mentioning that these modes only appear when using the realistic vocal tract. They induce a reduction of the HFE levels at the 8 kHz octave band from 5.6 to 5.9 dB, depending on the phonation type. These differences may be perceptually relevant, according to previous works in the literature. Specifically, a realistic 3D vocal tract geometry would be required for an accurate synthesis of vowel [ɑ] through 3D FEM, when trying to simulate a modal and a tense voice production. Conversely, when a lax phonation is considered, the influence of higher order propagation may be imperceptible, since the HFE levels are very small. Therefore, a simpler 1D simulation would suffice in this case.

Future work will consider other R_d values and geometry simplifications as well as other vowels to complete the study. Finally, we also plan to include aspiration noise in the LF

model to evaluate its impact on the HFE content of the numerical simulations.

V.5 Acknowledgements

The authors are grateful to Saeed Dabbaghchian for the design of the vocal tract geometry simplifications. This research has been supported by the Agencia Estatal de Investigación (AEI) and FEDER, EU, through project GENIOVOX TEC2016-81107-P. The fourth author acknowledges the support from the Obra Social "La Caixa" under grant ref. 2018-URL-IR1rQ-021.

References

Aalto, Daniel et al. (2014). "Large scale data acquisition of simultaneous MRI and speech". In: *Applied Acoustics* vol. 83, pp. 64–75.

Arnela, Marc, Blandin, Rémi, Dabbaghchian, Saeed, Guasch, Oriol, Alías, Francesc, Pelorson, Xavier, Van Hirtum, Annemie, and Engwall, Olov (2016a). "Influence of lips on the production of vowels based on finite element simulations and experiments". In: *The Journal of the Acoustical Society of America* vol. 139, no. 5, pp. 2852–2859.

Arnela, Marc, Dabbaghchian, Saeed, Blandin, Rémi, Guasch, Oriol, Engwall, Olov, Van Hirtum, Annemie, and Pelorson, Xavier (2016b). "Influence of vocal tract geometry simplifications on the numerical simulation of vowel sounds". In: *The Journal of the Acoustical Society of America* vol. 140, no. 3, pp. 1707–1718.

Arnela, Marc and Guasch, Oriol (2013). "Finite element computation of elliptical vocal tract impedances using the two-microphone transfer function method". In: *The Journal of the Acoustical Society of America* vol. 133, no. 6, pp. 4197–4209.

Birkholz, Peter (2013). "Modeling Consonant-Vowel Coarticulation for Articulatory Speech Synthesis". In: *PLoS ONE* vol. 8, no. 4, e60603.

Blandin, Rémi, Arnela, Marc, Laboissière, Rafael, Pelorson, Xavier, Guasch, Oriol, Van Hirtum, Annemie, and Laval, Xavier (2015). "Effects of higher order propagation modes in vocal tract like geometries". In: *The Journal of the Acoustical Society of America* vol. 137, no. 2, pp. 832–8.

Erath, Byron D., Zañartu, Matías, Stewart, Kelley C., Plesniak, Michael W., Sommer, David E., and Peterson, Sean D. (2013). "A review of lumped-element models of voiced speech". In: *Speech Communication*, pp. 667–690.

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

Fant, Gunnar, Liljencrants, Johan, and Lin, Qi-guang (1985). "A four-parameter model of glottal flow". In: *Speech Transmission Laboratory Quarterly Progress and Status Report (STL-QPSR)* vol. 26, no. 4, pp. 1–13.

Kawahara, Hideki, Sakakibara, Ken-Ichi, Banno, Hideki, Morise, Masanori, Toda, Tomoki, and Irino, Toshio (2017). "A new cosine series antialiasing function and its application to aliasing-free glottal source models for speech and singing synthesis". In: *Interspeech*, pp. 1358–1362.

Monson, Brian B., Hunter, Eric J., Lotto, Andrew J., and Story, Brad H. (2014). "The perceptual significance of high-frequency energy in the human voice". In: *Frontiers in Psychology*.

Monson, Brian B., Lotto, Andrew J., and Ternström, Sten (2011). "Detection of high-frequency energy changes in sustained vowels produced by singers". In: *The Journal of the Acoustical Society of America* vol. 129, no. 4, pp. 2263–2268.

Murphy, Andy, Yanushevskaya, Irena, Chasaide, Ailbhe Ní, and Gobl, Christer (Aug. 2017). "Rd as a Control Parameter to Explore Affective Correlates of the Tense-Lax Continuum". In: *Interspeech*. Stockholm, Sweden, pp. 3916–3920.

Murtola, Tiina, Alku, Paavo, Malinen, Jarmo, and Geneid, Ahmed (2018). "Parameterization of a computational physical model for glottal flow using inverse filtering and high-speed videoendoscopy". In: *Speech Communication* vol. 96, pp. 67–80.

Stone, Simon, Marxen, Michael, and Birkholz, Peter (2018). "Construction and evaluation of a parametric one-dimensional vocal tract model". In: *IEEE/ACM Transactions on Audio Speech and Language Processing* vol. 26, no. 8, pp. 1381–1392.

Story, Brad H. (2013). "Phrase-level speech simulation with an airway modulation model of speech production". In: *Computer Speech & Language* vol. 27, no. 4, pp. 989–1010.

Story, Brad H., Titze, Ingo R., and Hoffman, E. A. (1996). "Vocal tract area functions from magnetic resonance imaging". In: *The Journal of the Acoustical Society of America* vol. 100, no. 1, pp. 537–554.

Takemoto, Hironori, Adachi, Seiji, Mokhtari, Parham, and Kitamura, Tatsuya (2013). "Acoustic interaction between the right and left piriform fossae in generating spectral dips". In: *The Journal of the Acoustical Society of America* vol. 134, no. 4, pp. 2955–2964.

Takemoto, Hironori, Mokhtari, Parham, and Kitamura, Tatsuya (2010). "Acoustic analysis of the vocal tract during vowel production by finite-difference time-domain method". In: *The Journal of the Acoustical Society of America* vol. 128, no. 6, pp. 3724–3738.

Vampola, Tomáš, Horáček, Jaromír, and Švec, Jan G. (2008). "FE Modeling of Human Vocal Tract Acoustics. Part I: Production of Czech vowels". In: *Acta acustica united with Acustica* vol. 94, no. 5, pp. 433–447.

Authors' addresses

Marc Freixes GTM – Grup de Recerca en Tecnologies Mèdia, La Salle - Universitat Ramon Llull
Quatre Camins, 30, 08022 Barcelona, Spain
marc.freixes@salle.url.edu

Paper VI

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, Joan Claudi Socoró

Published in *Proceedings of 10th ISCA Speech Synthesis Workshop (SSW10)*, September 2019, pp. 132–136. DOI: 10.21437/SSW.2019-24.

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.

VI.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., Aalto et al. 2014). 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

(Blandin et al. 2015; Arnela et al. 2016). Many 3D approaches can be found in literature. The most extended ones are those based on the Finite Element Method (FEM) (Arnela and Guasch 2013), but also on finite differences (Takemoto et al. 2010), digital waveguides (Speed et al. 2014), and multimodal approaches (Blandin et al. 2015). 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 (Birkholz et al. 2017). Given that the latter are difficult to obtain from speech signals (Birkholz et al. 2015), 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 (Yanushevskaya et al. 2018), 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 (Birkholz et al. 2015) 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 (Fant 1995) was used in (A. Murphy et al. 2017) to simulate the tense-lax continuum and explore its affective correlates. Likewise an Auto-Regressive eXogenous variant of the LF model was proposed in (Li et al. 2018) 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 (Chien et al. 2017) have allowed for competitive glottal vocoders (Airaksinen et al. 2018). 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 (Lorenzo-Trueba et al. 2012). In this context, it has been recently proposed a GlottDNN-based speaking style conversion from natural to Lombard speech (Seshadri et al. 2019).

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 (Freixes et al. 2018), 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 VI.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

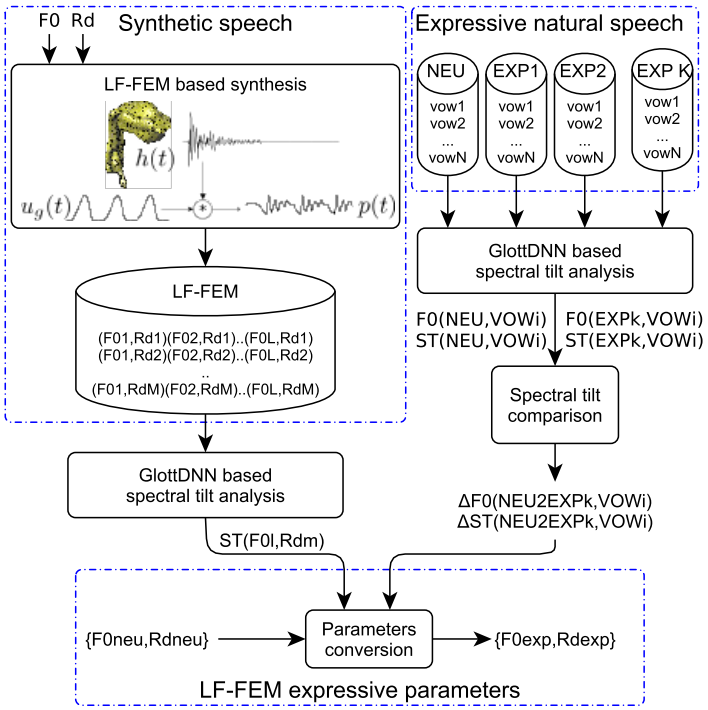


Figure VI.1: Workflow diagram used for the analysis and comparison of expressive natural speech respect with synthetic speech generated with 3D FEM-based acoustic model that uses an LF model as glottal excitation.

are described and discussed in Section VI.3. Finally, Section VI.4 closes the paper with the conclusions.

VI.2 Methodology

Figure VI.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 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 (Airaksinen et al. 2016), 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 VI.1.

VI.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.

VI.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 (Arnela and Guasch 2013). In particular, it numerically solves the acoustic wave equation for the acoustic pressure $p(\mathbf{x}, t)$,

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

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 (Arnela and Guasch 2013).

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 $p_o(t)$,

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

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

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

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

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

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 $H(f)$.

VI.2.1.2 Glottal Source Model

An LF model (Fant et al. 1985) is used to generate the train of glottal pulses $u_g(t)$ needed in (VI.2) to synthesise a vowel sound. In particular, the Kawahara's implementation (Kawahara et al. 2017) 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 VI.2). The original code [1] has been adapted

[1] https://github.com/HidekiKawahara/SparkNG

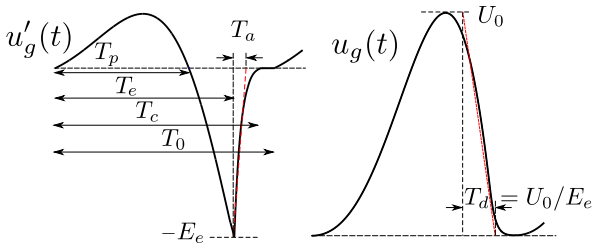


Figure VI.2: Glottal flow $u_g(t)$ and its time derivative $u'_g(t)$ according to the LF model (Fant et al. 1985). T_p is the rise time, T_e is the open phase duration, T_c corresponds to the complete closure, T_0 is the period, T_a is the effective duration of the return phase and T_d is the declination time. The maximum amplitudes of the glottal flow and its derivative are respectively U_0 and E_e.

to introduce the glottal shape parameter R_d (Fant 1995), defined as

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

In Eq. (VI.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 (Fant 1995). 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 (Gobl 2017). 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 $h(t)$.

VI.2.2 Spectral tilt analysis

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

VI.2.2.1 GlottDNN-based spectral tilt extraction

The GlottDNN vocoder (Airaksinen et al. 2016) 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 (Seshadri et al. 2019). 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.

VI.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 (P. Murphy et al. 2008; Kakouros et al. 2017) a

[2] http://sox.sourceforge.net/SoX/Resampling

scalar-based measure of the spectral tilt, ST, has been computed as

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

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.

VI.2.3 Expressive LF-FEM based synthesis

VI.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

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

and the increment of spectral tilt (in dB) as

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

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.

VI.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. VI.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.

VI.3 Experiments and results

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

VI.3.1 Experiments setup

VI.3.1.1 Expressive natural speech

This work has used an emotional Spanish speech corpus, which was explicitly designed to elicit expressive speech (see Iriondo et al. 2009 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.

VI.3.1.2 LF-FEM synthesis

Synthesis of vowel [a] has been done using the LF-FEM based model described in Section VI.2.1. For this purpose, we have used the 3D vocal tract geometry originally generated from MRI in (Aalto et al. 2014) and latter adjusted in (Arnela et al. 2016), 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. (VI.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 (Gobl 2017) 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. (VI.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 (Freixes et al. 2018).

VI.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 VI.3 depicts the results, where each circle represents the $\Delta F0$ and ΔST from a neutral vowel to its

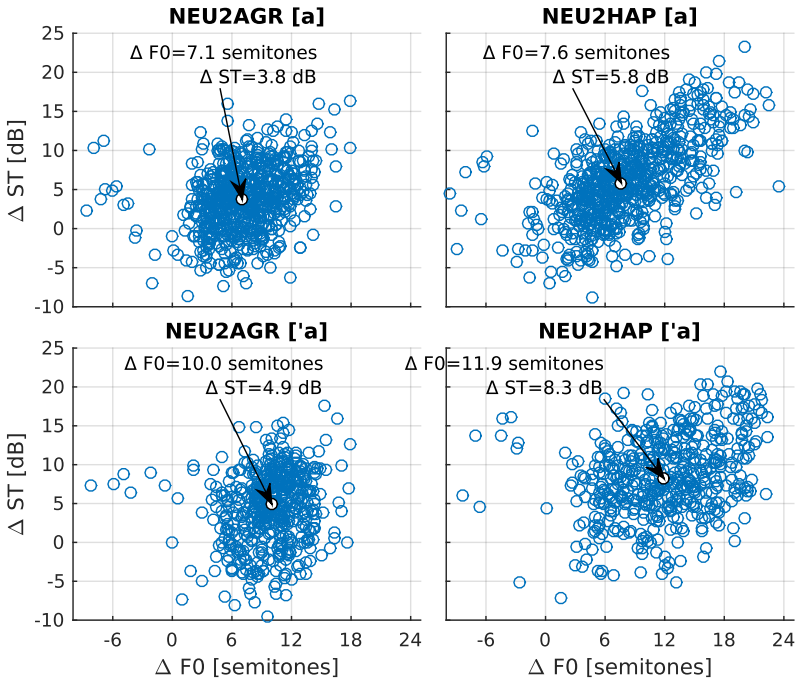


Figure VI.3: Distributions of $\Delta F0$ and ΔST for stressed and unstressed [a] vowels from neutral to aggressive (NEU2AGR), and from neutral to happy (NEU2HAP). The centroid of each distribution is represented as a white dot indicated with an arrow.

Table VI.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

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 VI.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

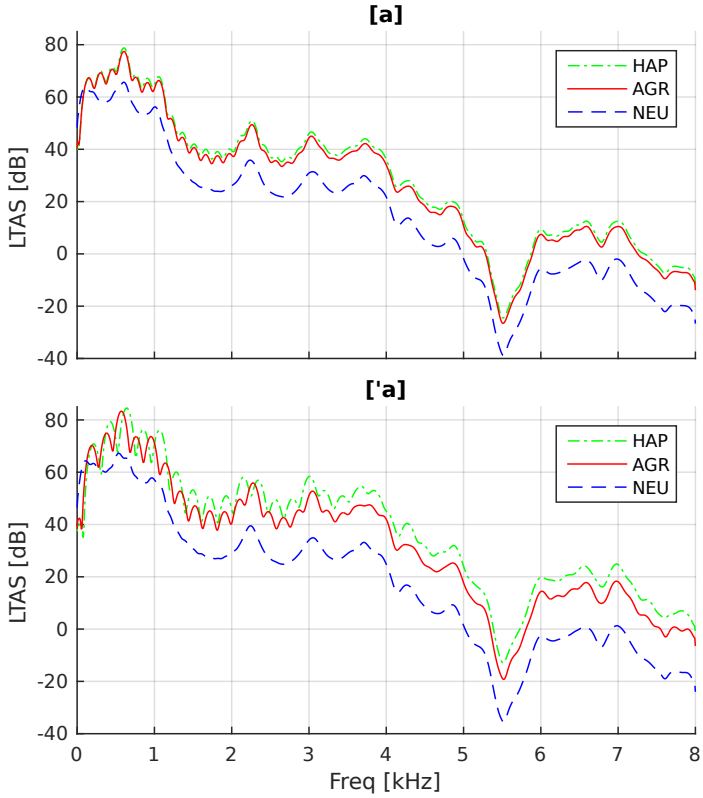


Figure VI.4: Long Term Average Spectra of the unstressed (top) and stressed (bottom) [a] vowels synthesised with the LF-FEM model for the neutral, aggressive and happy styles.

corpus, and as Table VI.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 VI.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. VI.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 VI.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 (Arnela et al. 2016). 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 VI.3, the happy style entails a higher ST than the aggressive one, thus producing the observed increment in the high frequency range.

VI.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.

VI.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.

References

Aalto, Daniel et al. (2014). “Large scale data acquisition of simultaneous MRI and speech”. In: *Applied Acoustics* vol. 83, pp. 64–75.

Airaksinen, Manu, Bollepalli, Bajibabu, Juvela, Lauri, Wu, Zhizheng, King, Simon, and Alku, Paavo (Sept. 2016). “GlottDNN - A full-band glottal vocoder for statistical parametric speech synthesis”. English. In: *Interspeech*, pp. 2473–2477.

Airaksinen, Manu, Juvela, Lauri, Bollepalli, Bajibabu, Yamagishi, Junichi, and Alku, Paavo (Sept. 2018). “A Comparison Between STRAIGHT, Glottal, and Sinusoidal Vocoding in Statistical Parametric Speech Synthesis”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* vol. 26, no. 9, pp. 1658–1670.

Arnela, Marc, Dabbaghchian, Saeed, Blandin, Rémi, Guasch, Oriol, Engwall, Olov, Van Hirtum, Annemie, and Pelorson, Xavier (2016). "Influence of vocal tract geometry simplifications on the numerical simulation of vowel sounds". In: *The Journal of the Acoustical Society of America* vol. 140, no. 3, pp. 1707–1718.

Arnela, Marc and Guasch, Oriol (2013). "Finite element computation of elliptical vocal tract impedances using the two-microphone transfer function method". In: *The Journal of the Acoustical Society of America* vol. 133, no. 6, pp. 4197–4209.

Birkholz, Peter, Martin, Lucia, Willmes, Klaus, Kröger, Bernd J., and Neuschaefer-Rube, Christiane (2015). "The contribution of phonation type to the perception of vocal emotions in German: An articulatory synthesis study". In: *The Journal of the Acoustical Society of America* vol. 137, no. 3, pp. 1503–1512.

Birkholz, Peter, Martin, Lucia, Xu, Yi, Scherbaum, Stefan, and Neuschaefer-Rube, Christiane (2017). "Manipulation of the prosodic features of vocal tract length, nasality and articulatory precision using articulatory synthesis". In: *Computer Speech and Language* vol. 41, pp. 116–127.

Blandin, Rémi, Arnela, Marc, Laboissière, Rafael, Pelorson, Xavier, Guasch, Oriol, Van Hirtum, Annemie, and Laval, Xavier (2015). "Effects of higher order propagation modes in vocal tract like geometries". In: *The Journal of the Acoustical Society of America* vol. 137, no. 2, pp. 832–8.

Chien, Yu-Ren, Mehta, Daryush D., Guðnason, Jón, Zañartu, Matías, and Quatieri, Thomas F. (2017). "Evaluation of glottal inverse filtering algorithms using a physiologically based articulatory speech synthesizer". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* vol. 25, no. 8, pp. 1718–1730.

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

Fant, Gunnar, Liljencrants, Johan, and Lin, Qi-guang (1985). "A four-parameter model of glottal flow". In: *Speech Transmission Laboratory Quarterly Progress and Status Report (STL-QPSR)* vol. 26, no. 4, pp. 1–13.

Freixes, Marc, Arnela, Marc, Socoró, Joan Claudi, Alías, Francesc, and Guasch, Oriol (2018). "Influence of tense, modal and lax phonation on the three-dimensional finite element synthesis of vowel [A]". In: *Proc. IberSPEECH 2018*, pp. 132–136.

Gobl, Christer (2017). "Reshaping the Transformed LF Model : Generating the Glottal Source from the Waveshape Parameter R d". In: *Interspeech*. 1, pp. 3008–3012.

Iriondo, Ignasi, Planet, Santiago, Socoró, Joan Claudi, Martínez, Elisa, Alías, Francesc, and Monzo, Carlos (2009). "Automatic refinement of an expressive speech corpus assembling subjective perception and automatic classification". In: *Speech Communication* vol. 51, no. 9, pp. 744–758.

Kakouros, Sofoklis, Räsänen, Okko, and Alku, Paavo (2017). "Evaluation of Spectral Tilt Measures for Sentence Prominence Under Different Noise Conditions". In: *Interspeech*, pp. 3211–3215.

Kawahara, Hideki, Sakakibara, Ken-Ichi, Banno, Hideki, Morise, Masanori, Toda, Tomoki, and Irino, Toshio (2017). "A new cosine series antialiasing function and its application to aliasing-free glottal source models for speech and singing synthesis". In: *Interspeech*, pp. 1358–1362.

Li, Yongwei, Li, Junfeng, and Akagi, Masato (2018). "Contributions of the glottal source and vocal tract cues to emotional vowel perception in the valence-arousal space." In: *The Journal of the Acoustical Society of America* vol. 144, no. 2, p. 908.

Lorenzo-Trueba, Jaime, Barra-Chicote, Roberto, Raitio, Tuomo, Obin, Nicolas, Alku, Paavo, Yamagishi, Junichi, and Montero, Juan M. (Sept. 2012). "Towards Glottal Source Controllability in Expressive Speech Synthesis". In: *Interspeech*. Portland, USA, pp. 2–5.

Murphy, Andy, Yanushevskaya, Irena, Chasaide, Ailbhe Ní, and Gobl, Christer (Aug. 2017). "Rd as a Control Parameter to Explore Affective Correlates of the Tense-Lax Continuum". In: *Interspeech*. Stockholm, Sweden, pp. 3916–3920.

Murphy, Peter, Mcguigan, Kevin, Walsh, Michael, and Colreavy, Michael (Apr. 2008). "Investigation of a glottal related harmonics-to-noise ratio and spectral tilt as indicators of glottal noise in synthesized and human voice signals". In: *The Journal of the Acoustical Society of America* vol. 123, pp. 1642–52.

Seshadri, Shreyas, Juvela, Lauri, Räsänen, Okko, and Alku, Paavo (2019). "Vocal Effort based Speaking Style Conversion using Vocoder Features and Parallel Learning". In: *IEEE Access*.

Speed, Matt, Murphy, Damian, and Howard, David (2014). "Modeling the Vocal Tract Transfer Function Using a 3D Digital Waveguide Mesh". In: *IEEE Transactions on Audio, Speech, and Language Processing* vol. 22 (2), pp. 453–464.

Takemoto, Hironori, Mokhtari, Parham, and Kitamura, Tatsuya (2010). "Acoustic analysis of the vocal tract during vowel production by finite-difference time-domain method". In: *The Journal of the Acoustical Society of America* vol. 128, no. 6, pp. 3724–3738.

Yanushevskaya, Irena, Gobl, Christer, and Chasaide, Ailbhe Ní (2018). "Cross-language differences in how voice quality and f_0 contours map to affect". In: *The Journal of the Acoustical Society of America* vol. 144, no. 5, p. 2730.

Authors' addresses

Marc Freixes GTM – Grup de Recerca en Tecnologies Mèdia, La Salle - Universitat Ramon Llull
Quatre Camins, 30, 08022 Barcelona, Spain
marc.freixes@salle.url.edu