Journal on Multimodal User Interfaces, Vol. 1, No. 1, March 2007

REALTIME AND ACCURATE MUSICAL CONTROL OF EXPRESSION IN SINGING SYNTHESIS

Nicolas D'Alessandro, Pascale Woodruff, Yohann Fabre, Thierry Dutoit

TCTS Lab, Faculté Polytechnique de Mons, Belgium
{nicolas.dalessandro;pascale.woodruff;yohann.fabre;thierry.dutoit}@fpms.ac.be

Sylvain Le Beux, Boris Doval, Christophe d'Alessandro

LIMSI-CNRS, Université Paris XI, Orsay, France
{sylvain.le.beux;boris.doval;christophe.dalessandro}@limsi.fr

ABSTRACT

In this paper, we describe a full computer-based musical instrument allowing realtime synthesis of expressive singing voice. The expression results from the continuous action of an interpreter through a gestural control interface. In this context, expressive features of voice are discussed. New real-time implementations of a spectral model of glottal flow (CALM) are described. These interactive modules are then used to identify and quantify voice quality dimensions. Experiments are conducted in order to develop a first framework for voice quality control. The representation of vocal tract and the control of several vocal tract movements are explained and a solution is proposed and integrated. Finally, some typical controllers are connected to the system and expressivity is evaluated.

KEYWORDS

Singing voice – Voice synthesis – Voice quality – Glottal flow models – Gestural control – Interfaces.

1. INTRODUCTION

Expressivity is nowadays one of the most challenging topics studied by researchers in speech synthesis. Indeed, recent synthesisers provide acceptable speech in term of intelligibility and naturalness but the need to improve human/computer interactions has brought researchers to develop more human, expressive systems. Some recent realisations have shown that an interesting option was to record multiple databases corresponding to a certain number of labelled expressions (e.g. happy, sad, angry, etc.) [1]. At synthesis time, the expression of the virtual speaker is then set by choosing the units in the corresponding database.

We decided to investigate the opposite option. Indeed, we postulated that emotion in speech was not the result of switches between labelled expressions but a continuous evolution of voice characteristics highly correlated with the context. Thus, we developed a set of flexible voice synthesisers conducted in realtime by an operator [2]. After some tests, it was clear that such a framework was particularly efficient for singing synthesis.

Remarkable achievements have been recently reached in singing voice synthesis. A review of state of the art can be found in [3]. The technology seems mature enough now to allow for the replacement of human vocals with synthetic singing, at least for backing vocals [4] [5]. However, existing singing synthesis systems suffer from two restrictions: they are aiming at mimicking singers rather than creating new instruments, and are generally limited to MIDI controllers.

We think it worthwhile to extend vocal possibilities of voice synthesisers and design new interfaces that will open new musical possibilities. In a first attempt we decided to restrain our survey on voice quality control to the boundaries of natural voice production - in fact, it is always better trying to mimic one particular voice. This process enables us to achieve analysis by synthesis : once we are able to perceive more naturalness in the synthesised voice, then we understood something about the voice production process. It is then easier to diverge from these limits when dealing with a musical application in a more creative way.

2. AIMS OF THIS WORK

Our aims can be summarised in three main axes. First, we target the implementation of intra and inter-dimensional mappings driving low-level parameters of source models (e.g. complex interactions between vocal effort and tenseness, represented by the phonetogram). Then, we investigate the effects of the vocal tract in voice quality variations (e.g. the singer formant, lowering of the larynx). Finally, source/filter coupling effects (e.g. relations between harmonics and formant frequencies) are analysed, and several mechanisms are implemented (e.g. overtone, bulgarian, occidental singing).

3. BACKGROUND IN SINGING SYNTHESIS

Speech and singing both result from the same production system: the vocal apparatus. However, the signal processing techniques developed for their synthesis evolved quite differently. One of the main reasons for this deviation is that the aim for producing voice is different in the two cases. The aim of speech production is to exchange messages. For singing, the main aim is to use the voice organ as a musical instrument. Therefore a singing synthesis system needs to include various tools to control (analyse/synthesise or modify) different dynamics of the acoustic sound produced: duration of the phonemes, vibrato, wide range modifications of the voice quality, the pitch and the intensity, etc., some of which are not needed in most of the speech synthesis systems. A pragmatic reason for that separation is that singing voice synthesisers target almost exclusively musical performances. In this case, playability (flexibility and real-time abilities) is much more important than intelligibility. Discussions about various issues of singing synthesis can be found in [6, 7].

As described in [8], frequency-domain analysis/modification methods are frequently preferred in singing synthesis research due to the need to modify some spectral characteristics of actual

recorded signals. The most popular application of such a technique is the phase vocoder [9], which is a powerful tool used for many years for time compression/expansion, pitch shifting and cross-synthesis.

To increase flexibility, short-time signal frames can be modelled as sums of sinusoids (controlled in frequency, amplitude and phase) plus noise (controlled by the parameters of a filter which is excited by a white noise). HNM (Harmonic plus Noise Model) [10] provides a flexible representation of the signal, which is particularly interesting in the context of unit concatenation. That representation of signals is thus used as a basis in many singing synthesis systems [11, 12, 13, 14].

Another approach is to use the source/filter model. Several models of glottal pulse has been proposed with different quality and flexibility. A complete study and normalisation of the main models can be found in [15]. For example, the R++ model has been used in the famous Voicer [16]. LF [17] and CALM [18] models have been used during eNTERFACE workshops [2]. Other differences appear in the method used to compute the vocal tract transfer function. Some systems [19] compute the formants from the magnitude spectrum: a series of resonant filters (controlled by formants frequencies, amplitudes and bandwidths). Some other systems compute an acoustic representation of the vocal tract, as a cascade of acoustic (variant-shape) tubes. For example, the SPASM synthesiser [20] uses digital waveguides [21] to model acoustic features of oral, nasal cavities and throat radiation (driven by a frequency-domain excitation model). The model was extended to variable length conical sections by Välimäki and Karjalainen [22].

There exist also some particular approaches like FOF (*Formes d'Ondes Formantiques*) synthesis [23], used in CHANT [24], which performs synthesis by convolving a pulse train with parallel formant wave functions (time-domain functions corresponding to individual formants resonance).

4. VOICE PRODUCTION

The vocal apparatus is usually described as a "source/filter" system. Glottal source is a non-linear volume velocity generator where sound is produced by complex movements of vocal folds (larynx) under lung pressure. A complete study of glottal source can be found in [25]. Sounds produced by the larynx are then propagated in oral and nasal cavities which can be seen as time-varying filters. Finally, the flow is converted into radiated pressure waves through the lips and nose openings (cf. Figure 1).

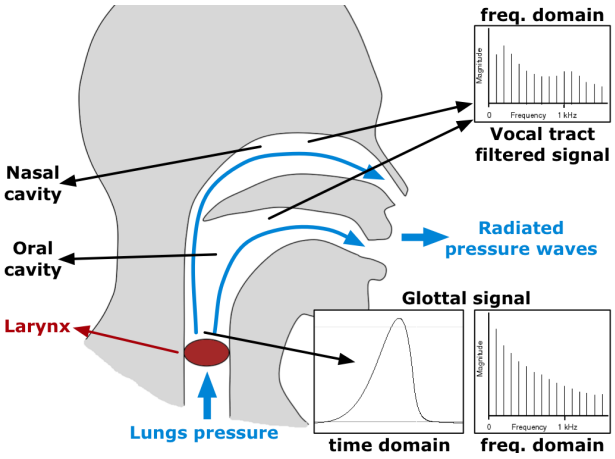

Figure 1: Voice production mechanisms: vocal folds vibrations, vocal tract filtering and lips/nose openings radiation.

In the context of signal processing applications, and particularly in singing synthesis, some simplifications are usually accepted. Firstly, the effect of lip and nose openings can be seen as derivative of the volume velocity signal. It is generally processed by a time-invariant high-pass first order linear filter [26]. Vocal tract effect can be modelled by filtering the glottal signal with multiple (usually 4 or 5) second order resonant linear filters.

Contrary to this "standard" vocal tract implementation, plenty of models have been developed for representation of glottal flow, with differences in accuracy and flexibility. Usual models are KLGLOTT88 [27], R++ [28], Rosenberg-C [29], LF [17, 30] and more recently, CALM [18].

5. THE GLOTTAL SOURCE

In this section, we describe the work related to the realtime generation of the glottal source signal. We first explain our theoretical basics: the modelling of the glottal flow as the response of a causal/anticausal linear system (CALM). Then, we will describe two different implementations achieved: a buffered computation of a causal stable filter (v1.x) and a sample-by-sample computation of a causal unstable filter (v2.x).

5.1. The Causal/Anticausal Linear Model (CALM) [18]

Modelling the human vocal tract in the spectral domain (with resonant filters central frequency, amplitude and bandwidth) is very powerful in term of manipulation because spectral description of sounds is close to auditory perception. Traditionally, glottal flow has been modelled in the time domain. A spectral approach can be seen as equivalent only if both amplitude and phase spectra are considered in the model.

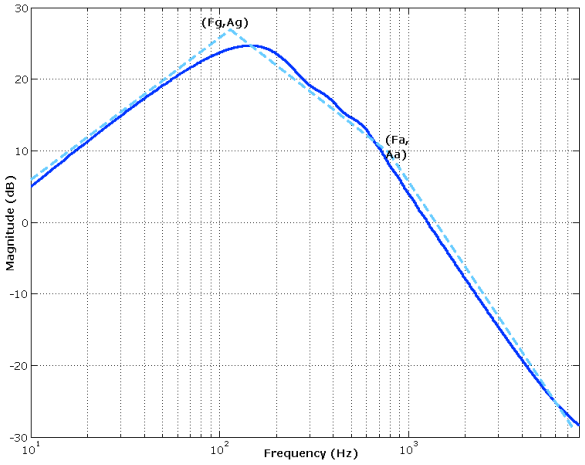

Figure 2: Amplitude spectrum of the glottal flow derivative: illustration of glottal formant (F_g, A_g) and spectral tilt (F_a, A_a).

For amplitude spectrum, two different effects can be isolated (cf. Figure 2). On the one hand, an amount of energy is concentrated in low frequencies (i.e. below 3 kHz). This peak is usually called the "glottal formant". We can see that bandwidth, amplitude and position of the glottal formant can change with voice quality variations. On the other hand, a variation of spectrum slope in higher frequencies (called "spectral tilt") is also related to voice quality modifications.

Considering both "glottal formant" and "spectral tilt" effects, two cascading filters can be used. A second order resonant low-pass filter ($H_1(z)$) for glottal formant, and a first order

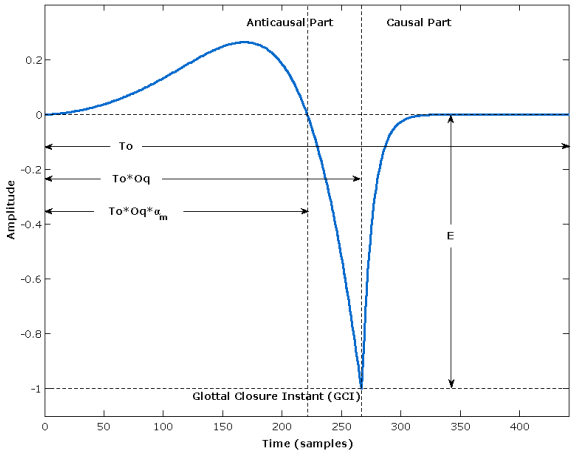

Figure 3: Time-domain representation of derivated glottal pulse: anticausal and causal parts, respectively on the left and right of the glottal closure instant.

low-pass filter ($H_2(z)$) for spectral tilt. But phase information indicates to us that this system is not completely causal. Indeed, as it is illustrated on Figure 3, glottal pulse is a combination of a "increasing" (or active) part and a "decreasing" (or passive) part. The decreasing part, called the return phase, mainly influences the spectral tilt and is causal. And we can also show that the second order low-pass filter has to be anticausal in order to provide a good phase representation. This information is sometimes referred as the mixed-phase representation of voice production [31].

A complete study of spectral features of glottal flow, detailed in [18], gives us equations linking relevant parameters of glottal pulse (F_0: fundamental frequency, O_q: open quotient, α_m: asymetry coefficient and T_l: spectral tilt, in dB at 3000Hz) to $H_1(z)$ and $H_2(z)$ coefficients. Expression of b_1 as been corrected, compared to [18] and [32].

Anticausal second order resonant filter:

$$H_1(z) = \frac{b_1 z}{1 + a_1 z + a_2 z^2}$$
$$a_1 = -2e^{-a_p T_e} cos(b_p T_e)$$
$$a_2 = e^{-2a_p T_e}, b_1 = ET_e$$
$$a_p = -\frac{\pi}{O_q T_0 tan(\pi \alpha_m)}, b_p = \frac{\pi}{O_q T_0}$$

Causal first order filter:

$$H_2(z) = \frac{b_{T_L}}{1 - a_{T_L} z^{-1}}$$
$$a_{T_L} = \nu - \sqrt{\nu^2 - 1}, b_{T_L} = 1 - a_{T_L}$$
$$\nu = 1 - \frac{1}{\eta}, \eta = \frac{\frac{1}{e^{-T_L/10ln(10)}} - 1}{cos(2\pi \frac{3000}{F_e}) - 1}$$

Full anticausal processing is only possible offline, by running algorithms backwards on data buffers. In a realtime context, anticausal response can be processed with two different methods. On the one hand, the response of a causal version of $H_1(z)$ is stored backwards (*v1.x*). On the other hand, $H_1(z)$ is replaced by a unstable causal filter and the "divergent" impulse response is truncated (*v2.x*). We can also note that in order to be useful our implementations have to be able to produce correct glottal flow (GF) and glottal flow derivative (GFD). Indeed, the GFD is the acoustical signal used to synthesise the voiced sounds, but the GF is important in the synthesis of turbulence, involved in unvoiced and breathy sounds.

5.2. RealtimeCALM v1.x Implementation

This implementation is the continuation of the development tasks of eNTERFACE'05 [2] and work presented to NIME'06 [32]. In this algorithm, we generate the impulse response by *period-synchronous anticausal processing*. It means that in order to achieve the requested waveform, the impulse response of a causal version of H_1 (glottal formant) is computed, but stored backwards in a buffer. This waveform is truncated at a length corresponding to instantaneous fundamental frequency (F_0+$Jitter$). This algorithm is now integrated in both Max/MSP [33, 34] and Pure Data [35] external objects (for Mac OS X, Windows and Linux): $almPulse_{\sim}$ $v1.x$. Then the resulting period is filtered by H_2 (spectral tilt). This algorithm is also integrated in both Max/MSP and Pure Data external objects: $stFilter_{\sim}$ $v1.x$. Coefficients of H_1 and H_2 are calculated from equations described in subsection *The Causal/Anticausal Linear Model (CALM)* and [18]. Thus, both time-domain and spectral-domain parameters can be sent.

Actually, we take advantage of physical properties of glottis to propose this real-time algorithm. Indeed, glottal pulse corresponds to opening/closing movements of vocal folds. It means that impulse responses generated by H_1 and H_2 filters can't overlap. Thus, impulse responses can be stored backwards and truncated period-synchronously without excessively changing their spectral properties.

Truncation of the CALM waveform at each period gives quite good synthesis results. Nevertheless, several configurations of parameters (e.g. high value of α_m plus low value of O_q) make the impulse response oscillating inside the period, which gives signals that are no more related to glottal source phenomena and changes voice quality perception. Thus, earlier truncation points and windowing options have been tested (e.g. first zero crossing of the GF, first zero crossing of the GFD). This study has shown us that it is not possible to set a truncation point inside the period which gives simultaneously correct synthesis results on the GF and the GFD (even with a synchronized half-Hanning windowing[1]). This modelization problem and limitations due to the use of period buffer drove us to change the architecture of this synthesis module (*v2.x*). Discontinuity in GFD due to GF truncation is illustrated at the Figure 4.

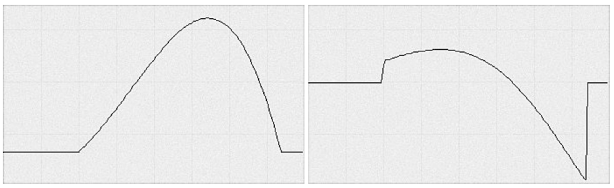

Figure 4: Discontinuity in GFD (right) due to GF truncation at the first zero crossing of the CALM period (left).

5.3. RealtimeCALM v2.x Implementation

This part explains another version of the anticausal filter response computation. It avoids the use of period buffer. The main idea behind this solution was to decrease memory allocations, in

[1] This windowing method multiplies the increasing part of the glottal pulse (flow or derivative) – meaning the part between the zero crossing and the positive maximum – by the left part of a Hanning window.

order to be able to generate simultaneously the glottal flow and the glottal flow derivative, each with their own truncation points and windowings[2].

Instead of computing a causal version of the impulse response off-line and then copying it backwards into a fixed buffer, the computation is here straightforward. The iterative equation corresponds indeed to the unstable anticausal filter. At any rate, the explosion of the filter is avoided by stopping the computation exactly at the Glottal Closure Instant (GCI). We can also note that glottal flow and glottal flow derivative can both be achieved with the same iterative equation, only changing the values of two first samples used as initial conditions in the iteration process.

One other main implementation problem is that the straightforward waveform generation has to be synchronised with the standard Pure Data performing buffer size. This standard size is 64 samples which, at an audio rate of 44100Hz, corresponds to a frequency of approximately 690 Hz. Most of the time, the fundamental frequency of the glottal flow is less than 690 Hz, which means that several buffers are necessary to achieve the complete computation of one period. But whenever a buffer reaches the end, the main performing routine is called and thus the values of a_1 and a_2 have to be frozen as long as the period time has not been reached. A flag related to the opening of the glottis is then introduced, fixed to the value of the period (in samples), and the values of a_1 and a_2 are not changed until this flag is decreased to 0. Once values of T_0, T_e, γ, a_p, and b_p have been calculated at the opening instant, only a_1 and a_2 have to be frozen, as these are the only variables that are taken into account in the equations of the derivative glottal waveform.

We just tested the glottal flow/glottal flow derivative generation alone, without the addition of any vocal tract information. However, extensive tests have been carried out concerning this implementation and revealed that this version is more robust than the previous one. In particular, this implementation is not stuck when exotic values are sent to the algorithm. Finally, we can note that this upgrade only concerns the $almPulse_{\sim}$ module. The spectral tilt filtering module ($stFilter_{\sim}$) was not modified.

5.4. Dimensionnal Issues

The next step in the realisation of our singing tool was to define perceptual dimensions underlying the control of voice quality, and to implement analytic mapping functions with low-level synthesis parameters. Dimensional features of voice were first collected from various research fields (signal processing, acoustics, phonetics, singing), completed, and described in a formalised set [32, 36].

- *Melody (F_0)*: short-term and long-term elements involved in the organisation of temporal structure of fundamental frequency;
- *Vocal Effort (V)*: a representation of the amount of "energy" involved in the creation of the vocal sound. It makes the clear difference between a spoken and a screamed voice for example [37, 38, 39, 40];
- *Tenseness (T)*: a representation of the constriction of the voice source. It makes the difference between a lax and a tensed voice [25];
- *Breathiness (B)*: a representation of the amount of air turbulence passing through the vocal tract, compared to the amount of voiced signal [25, 27];
- *Hoarseness (H)*: a representation of the stability of sound production parameters (especially for fundamental frequency and amplitude of the voice);
- *Mecanisms (M_i)*: voice quality modifications due to type of phonation involved in sound production [41].

[2] We can observe that our method will change the link between those two waveforms. Indeed, if two separated truncation points and windowings are applied, what we call "glottal flow derivative" is no more the derivative of the glottal flow.

5.5. Description of Mapping Functions

Once dimensions are defined, two main tasks can be investigated. First, the implementation of mapping functions between these dimensions and low-level parameters. Then, identification and implementation of inter-dimensional phenomena. In this area, many different theories have been proposed relating to several intra or inter-dimensional aspects of voice production [27, 40, 42, 43, 44, 45]. We decided to focus on some of them, like direct implementation of tenseness and vocal effort, realisation of a phonetogram, etc. and design our synthesis platform in order to be easily extensible (e.g. to correct existing relations and add new mapping functions etc.). All current parameters are defined for a male voice.

Relations between Dimensions and Synthesis Parameters

We focused on several aspects of the dimensionnal process. First, we consider relations between a limited number of dimensions (F_0, V, T and M_i) and synthesis parameters (O_q, α_m and T_l). Then, we decided to achieve our data fusion scheme by considering two different "orthogonal" processes in the dimensionnal control. On the one hand, vocal effort (V) (also related to F_0 variations, cf. next paragraph: *Inter-Dimensionnal Relations*) and mecanisms (M_i) are controlling "offset" values of parameters (O_{q_0}, α_{m_0}, T_{l_0}). On the other hand, tenseness (T) controls "delta" values of O_q and α_m around their offsets (ΔO_q, $\Delta\alpha_m$). Considering this approach, effective values of synthesis parameters can be described as:

$$O_q = O_{q_0} + \Delta O_q$$
$$\alpha_m = \alpha_{m_0} + \Delta\alpha_m$$
$$T_l = T_{l_0}$$

Following equations consider V and T parameters normalized between 0 and 1 and M_i representing the i^{th} phonation mecanism.

- $O_{q_0} = f(V|M_i)$

$$O_{q_0} = 0,8 - 0,4 \times V|M_1$$
$$O_{q_0} = 1 - 0,5 \times V|M_2$$

- $\alpha_{m_0} = f(M_i)$

$$\alpha_{m_0} = 0,8|M_1$$
$$\alpha_{m_0} = 0,6|M_2$$

- $T_{l_0} = f(V)$

$$T_{l_0}(dB) = 55 - 49 \times V$$

- $\Delta O_q = f(T)$

$$\Delta O_q = (1 - 2T)O_{q_0} + 0,8T - 0,4|T \leq 0,5$$
$$\Delta O_q = (2T - 1)O_{q_0} + 2T + 1|T > 0,5$$

- $\Delta\alpha_m = f(T)$

$$\Delta\alpha_m = (0,5T - 1)\alpha_{m_0} - 1,2T + 0,6 | T \geq 0,5$$

$$\Delta\alpha_m = (0,25 - 0,5T)\alpha_{m_0} + 0,4T - 0,2 | T < 0,5$$

Last adaptation on parameters concerns a perceptual distorsion of O_q (square distorsion) and α_m (square root distorsion) between their ranges of variation (O_q: $0,4$ to 1; α_m: $0,6$ to $0,8$) [46].

Inter-Dimensionnal Relations: the Phonetogram

One important characteristic of human voice production is that we are not able to produce any fundamental frequency (F_0) at any vocal effort (V). A strong relationship exists between these two production features. For example, one could not produce a very low pitch (around $80Hz$) at a sound pressure level higher than $80dB$ (for a male speaker) or conversely to produce a high pitch at low intensity. Trying to do so results in a sudden stop of vocal production. This relationship is called the phonetogram, and the evolution of this dependency varies very much from one speaker to an another. Consider, for example, whether the subject is a trained singer or not, male or female, has a pathological voice or not, etc. As a first approach, we decided to implement an average phonetogram, relying on the work of N. Henrich [47]. Figure 5 and Figure 6 represent two average phonetograms for male and female.

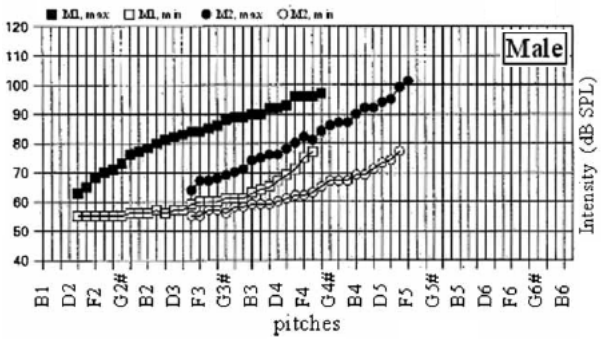

Figure 5: Average voice amplitude range profile (phonetogram) of male singers in mechanisms M_1 and M_2 [47].

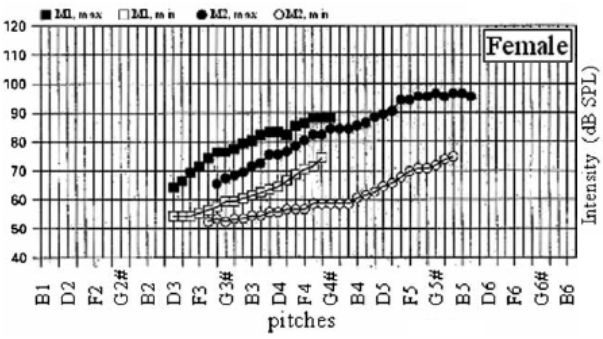

Figure 6: Average voice amplitude range profile (phonetogram) of female singers in mechanisms M_1 and M_2 [47].

Moreover, this phenomenon involves different types of laryngeal configurations. We here dealt with mainly two configurations, first and second mechanisms of the vocal folds (M_1 and M_2). This two laryngeal mechanisms are, in the common singing typology, referred as chest and falsetto registers. Hence, as shown on Figure 5 and Figure 6, it is not possible to produce any frequency in both mechanisms, but the two configurations have an overlapping region in the middle of the phonetogram. This region enables the passing between the two mechanisms. Following the work presented in [48], the frequency range where this passing can occur is about one octave (or 12 semi-tones). The main characteristic of this passing is to provoke a break in the fundamental frequency (F_0). Thus, when producing an increasing glissando from M_1 to M_2, there is an average 8 semi-tones break, whereas it is approximately 12 semi-tones when performing a decreasing glissando. Breaking intervals probabilities are depicted on Figure 7 and Figure 8. In the first one we can actually see that the frequency breaks also depends on the fundamental frequency where it occurs.

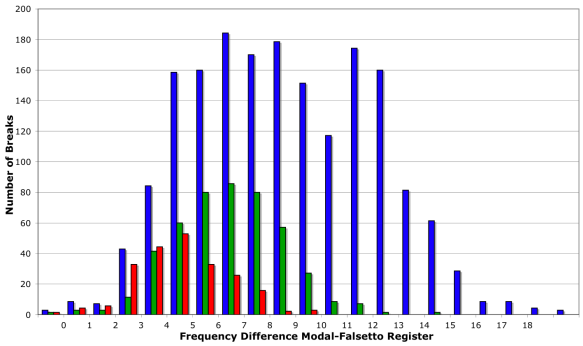

Figure 7: Frequency drops densities in semi-tones from Chest (or Modal) to Falsetto register. In blue, when the break happens at $200Hz$, in green at $300Hz$, in red at $400Hz$ [48].

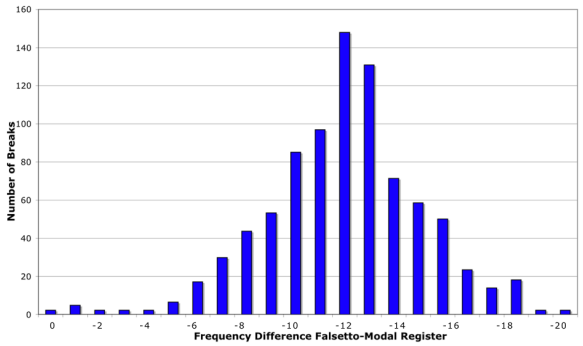

Figure 8: Frequency drops densities in semi-tones from Falsetto to Chest (or Modal) register [48].

In other words, this phenomenon introduces an hysteresis. For most untrained speakers or singers this break is uncontrollable whereas trained singers are able to hide more or less smoothly this break, although they cannot avoid the mechanism switch altogether.

6. THE VOCAL TRACT

In this section, we describe the implementation of a vocal tract model. This module is based on a physical "tube-based" representation of a vocal tract filter, which is simultaneously controllable using geometrical (area) and spectral (formant) parameters.

6.1. The lattice filter, a geometrical approach of vocal tract representation

Linear Predictive Coding [49] is a method for representing the spectral envelope of a digital signal of speech in compressed form, with the information given by a linear predictive model. The order of the filter is related to the complexity of the envelope, and also the number of control parameters. Thus, to represent a five-formant singing vowel, a filter containing five pairs of conjugated poles (for the resonances) is needed, adding up to a total of ten parameters for the vocal tract.

The LPC parameters (commonly named a_i) are non linearly interpolable. This implies that, for two configurations $[a_1 a_2 ... a_n]$ and $[b_1 b_2 ... b_n]$ corresponding to two vowels, a linear interpolation between both of these vectors will not correspond to a linear interpolation between the two spectra, and could even lead to unstable combinations. For these reasons, we will use another implementation of the LPC filter: the *lattice filter*. The control parameters of such a filter are called *reflection* coefficients (commonly named k_i). Such a filter is represented in Figure 9. It is composed of different sections, each characterized by a ki parameter.

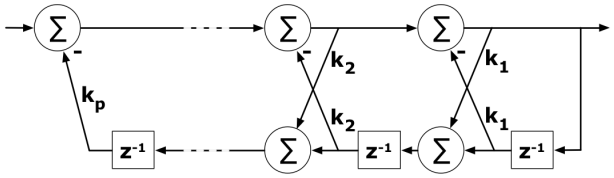

Figure 9: Representation of k_p cells of a lattice filter.

The reflection coefficients correspond to physical characteristics of the vocal tract, which may be represented by a concatenation of cylindrical acoustic resonators, forming a lossless tube. This physical model of the lattice filter is represented in Figure 10. Each filter section represents one section of the tube; the forward wave entering the tube is partially reflected backwards, and the backward wave is partially reflected forwards. The reflection parameter k_i can then be interpreted as the ratio of acoustic reflections in the i^{th} cylindrical cavity, caused by the junction impedance with the adjacent cavity. This value varies from 1 (total reflection) to -1 (total reflection with phase inversion), and is equal to 0 when there is no reflection.

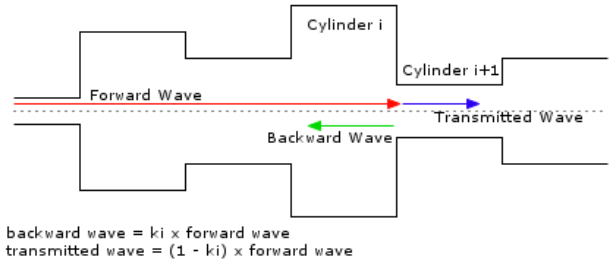

Figure 10: Geometrical interpretation of the lattice filter: transmitted and backwards waves at each cell junction.

The filter will be stable if the k_i parameters are between -1 and 1. However, there is no direct relationship between these parameters and sound: a small modification of k_i does not imply a small modification of the spectrum. Thus, instead of using the reflection coefficients, we will be using the different cylinder areas, named A_i, which can be easily deducted from the reflection coefficients with the following expression:

$$\frac{A_i}{A_{i+1}} = \frac{1+k_i}{1-k_i}$$

By acting on these A_i parameters, the interpreter is directly connected to the physical synthesis instrument. The sound spectrum will then evolve with acoustical coherence, which makes it more natural to use. Moreover, the stability of the filter is guaranteed for all A_i values.

6.2. Coefficients Convertion Framework

In order to use the area parameters of the lattice filter (A_i), a Max/MSP object was created to convert them to k_i values which are used in the lattice filter. Several sets of A_i parameters corresponding to different vowels were calculated. After selecting one of these presets, certain sections of the vocal tract can be modified by a percentage ΔA_i, which has the effect of opening or closing that section of the oral cavity.

A second approach to controlling the lattice filter was considered: a formant-based scheme was used to represent the spectral enveloppe, and the formant features, F_i, were converted to k_i parameters (after conversion to the LPC a_i coefficients), and then to A_i areas to control the lattice filter. This allowed us to easily model certain phenomena that are well known in speech processing, like overtone singing or the singer formant [50, 51], by acting on analytical parameters (the formants) rather than geometrical parameters (the areas). Similarly to the control of the areas, the formants have presets for different vowels and can be modified by a percentage ΔFi.

The parameters conversion framework described above is represented in Figure 11.

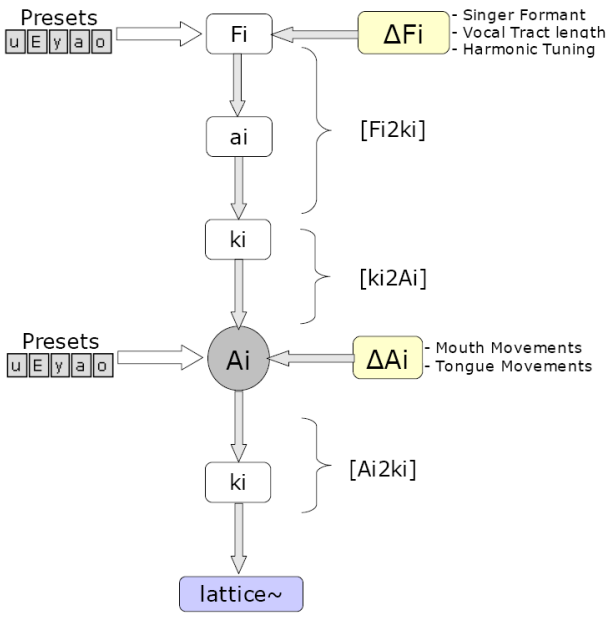

Figure 11: Coefficients convertion and presets/modifications framework, allowing user to modify spectrum-related and shape-related features at the same time.

7. ABOUT THE REAL-TIME CONTROL OF VOICE SYNTHESIS

In this section, we comment some experiments we conducted in order to evaluate expressive and performing abilities of systems

we developed. Modules which were integrated together inside Max/MSP and various control devices (and in various combinations) were dynamically connected using a mapping matrix. This set of tests allowed us to acheive efficient configurations employing several different performing styles (classical singing, overtone singing, etc.).

7.1. Concerning Voice Source

In order to be able to compare expressive skills of this system with the one developed before [2, 32, 36], we decided to keep the same control scheme: a graphic tablet. In that way, we were able to evaluate clearly the improvements achieved using the new mapping functions. Early experimentation demonstrated to us that independent control of tenseness and vocal effort significantly increased performance possibilities. Current mapping equations still provide some unlikely parameters combinations, resulting, for example, in "ultra-tensed" perception or unexpected variations of dynamics.

The implementation of the phonetogram is also a major improvement in term of naturalness. It gives better results in terms of expressivity rather than that without monitored control of loudness which is more linear. Although we have significantly investigated the frequency break phenomenon, we have not yet integrated it into the system, as we did not find a satisfying solution for controlling it. It is not straightforward to translate this frequency break in the control domain, and as our hand gestures are mainly continuous or used as basic switch from one configuration to another is not really satisfying from a musical point of view, and can result in breaks in frequency range and "wrong" notes.

7.2. Concerning Vocal Tract

The vocal tract was controlled using a data glove [52] as shown in Figure 12. The glove was mapped to the area parameters of the lattice filter in four different ways:

- The folding of the fingers control the opening angle of the mouth (represented in Figure 13) (see Figure 14)
- The hand movement along the z-axis controls the position of the "tongue" in the vocal tract (towards the back or the front of the mouth)
- The hand movement along the y-axis controls the vertical position of the tongue (near or far from the palate) (see Figure 14)
- The hand movement along the x-axis changes the vowels (configurable from one preset to another, for example from an */a/* to an */o/*)

This configuration allowed us to achieve vocal tract modification techniques such as overtone singing quite easily. Indeed, as the spectral representation (F_i) is very efficient to configure for some presets (e.g. offset vowel) or to leave running on automatic tasks (e.g. harmonic/formant tuning), the constant access to geometrical "delta" features (ΔS_i) allows the user to refine techniques (e.g. lowering the vocal tract, changing tongue position, etc.) and thus increase expressivity.

7.3. Incidental Remarks

Overall, at this stage of development, the synthesiser allows control of 17 parameters, namely : pitch, vocal effort, tenseness, mechanisms, the first two formants, the singers formant, vocal tract length, gain, transition between vowels, width of the vocal tract, position of the tongue and mouth opening (5 parameters).

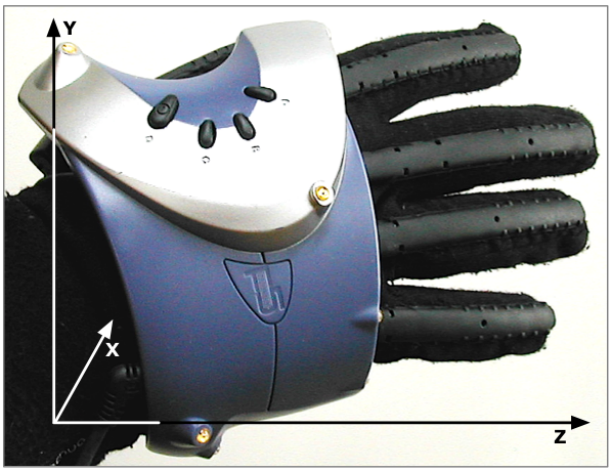

Figure 12: Vocal tract control with a data glove: 5 finger flexion sensors and 3 dimensions (x,y,z) tracking.

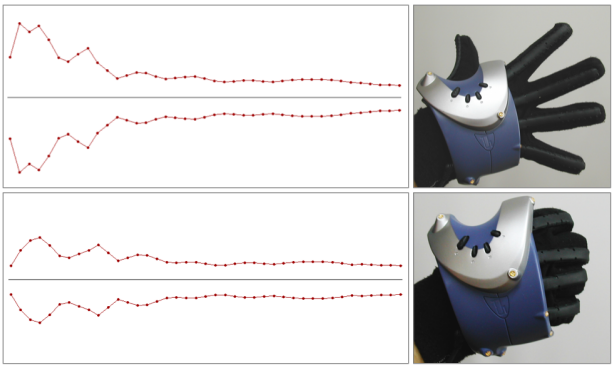

Figure 13: Mouth opening control: finger flexion sensors mapped to variation of 9 first A_i.

Considering all of these parameters, only the actions on mechanisms is not a continuous parameter, thus 16 parameters must be monitored using continuous parameters. From the controller side, we have 17 continuous parameters (out of 33), meaning that we are actually theoretically able to control all needed parameters. However, the problem is that from user's perpective, it is impossible to manipulate three interfaces at the same time. There are actually two solutions: one is to have multiple users (2 or 3) being in control of the interfaces, the other is to use one-to-many mappings, allowing the performer to control several parameters with the same controller.

8. CONCLUSIONS

In this work, our main aim was to build a high performance musical instrument allowing a wide range of expressive singing possibilities. Our actual work resulted in the implementation of new models for voice source and vocal tract, in real-time, which will be strategic tools in order to further this work. Improvements in expressivity of this new system have encouraged us to go forward with this approach. Moreover, our modular architecture inspires us to move towards a highly extensible synthesis platform which will be useful in the integration of other results from existing and forthcoming vocal production techniques.

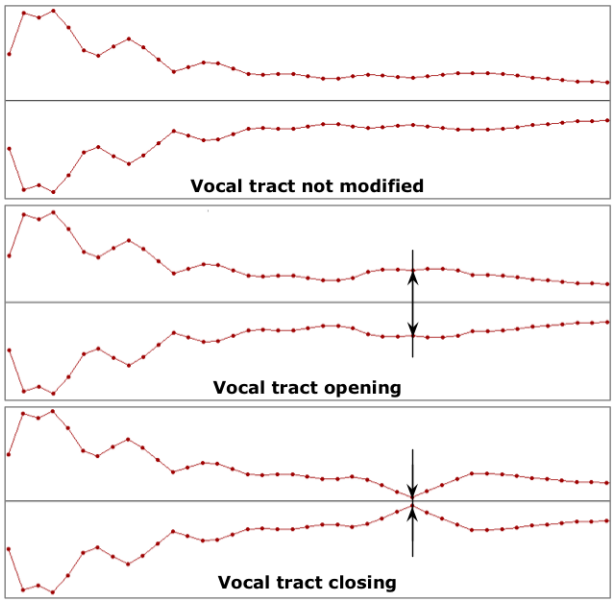

Figure 14: Control of the vertical position of the tongue.

9. ACKNOWLEDGMENTS

The authors would like to thank SIMILAR Network of Excellence (and thus the European Union) which has provided resources to allow researchers from all over Europe to meet, share and work together, thus achieving exciting results. We also would like to thank the Croatian organisation team of eNTERFACE'06, led by Prof. Igor Pandzic, where most of this work were done. Finally, we would like to thank our respective laboratories (TCTS Lab, Mons, Belgium and LIMSI-CNRS, Paris, France) who have adapted their research agendas in order to allow us to collaborate in this project.

10. REFERENCES

[1] http://www.loquendo.com/. 31

[2] C. d'Alessandro, N. D'Alessandro, S. L. Beux, J. Simko, F. Cetin, and H. Pirker, "The Speech Conductor: Gestural Control of Speech Synthesis", in *Proceedings of eNTERFACE'05 Summer Workshop on Multimodal Interfaces*, 2005. 31, 32, 33, 37

[3] M. Kob, "Singing Voice Modelling As We Know It Today", *Acta Acustica United with Acustica*, vol. 90, pp. 649–661, 2004. 31

[4] http://www.virsyn.de/. 31

[5] http://www.vocaloid.com/. 31

[6] X. Rodet and G. Bennet, "Synthesis of the Singing Voice", *Current Directories in Computer Music Research*, 1989. 31

[7] X. Rodet, "Synthesis and Processing of the Singing Voice", in *Proceeding of the First IEEE Benelux Workshop on Model-Based Processing and Coding of Audio (MPCA-2002)*, (Leuven, Belgium), 2002. 31

[8] P. Cook, *Identification of Control Parameters in an Articulatory Vocal Tract Model, with Applications to the Synthesis of Singing*. Ph.d. thesis, Standford University, 1990. 31

[9] J. Moorer, "The Use of the Phase Vocoder in Computer Music Application", *Journal of the Audio Engineering Society*, vol. 26, no. 1-2, pp. 42–45, 1978. 32

[10] J. Laroche, Y. Stylianou, and E. Moulines, "HNS: Speech Modifications Based on a Harmonic plus Noise Model", in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, vol. 2, pp. 550–553, 1993. 32

[11] M. Macon, L. Jensen-Link, J. Oliviero, M. Clements, and E. George, "A Singing Voice Synthesis System Based on Sinusoidal Modeling", in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. 435–438, 1997. 32

[12] K. Lomax, *The Analysis and the Synthesis of the Singing Voice*. Ph.d. thesis, Oxford University, 1997. 32

[13] Y. Meron, *High Quality Singing Synthesis Using the Selection-Based Synthesis Scheme*. Ph.d. thesis, University of Michigan, 2001. 32

[14] P. Cano, A. Loscos, J. Bonada, M. de Boer, and X. Serra, "Voice Morphing System for Impersonating in Karaoke Applications", in *Proceedings of the International Computer Music Conference*, 2000. 32

[15] B. Doval, C. d'Alessandro, and N. Henrich, "The spectrum of glottal flow models", *Acta Acustica*, vol. 92, pp. 1026–1046, 2006. 32

[16] L. Kessous, "A two-handed controller with angular fundamental frequency control and sound color navigation", in *Proceedings of the 2002 Conference on New Interfaces for Musical Expression (NIME-02)*, 2002. 32

[17] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow", *STL-QPSR*, vol. 4, pp. 1–13, 1985. 32

[18] B. Doval and C. d'Alessandro, "The voice source as a causal/anticausal linear filter", in *proc. Voqual'03, Voice Quality: Functions, analysis and synthesis, ISCA workshop*, (Geneva, Switzerland), Aug. 2003. 32, 33

[19] B. Larson, "Music and Singing Synthesis Equipment (MUSSE)", *Speech Transmission Laboratory Quarterly Progress and Statut Report (STL-QPSR)*, pp. (1/1977):38–40, 1977. 32

[20] P. Cook, "SPASM: a Real-Time Vocal Tract Physical Model Editor/Controller and Singer: the Companion Software System", in *Colloque sur les Modèles Physiques dans l'Analyse, la Production et la Création Sonore*, 1990. 32

[21] J. O. Smith, "Waveguide Filter Tutorial", in *Proceedings of the International Computer Music Conference*, pp. 9–16, 1987. 32

[22] V. Välimäki and M. Karjalainen, "Improving the Kelly-Lochbaum Vocal Tract Model Using Conical Tubes Sections and Fractionnal Delay Filtering Techniques", in *Proceedings of the International Conference on Spoken Language Processing*, 1994. 32

[23] X. Rodet, "Time-Domain Formant Wave Function Synthesis", vol. 8, no. 3, pp. 9–14, 1984. 32

[24] X. Rodet and J. Barriere, "The CHANT Project: From the Synthesis of the Singing Voice to Synthesis in General", *Computer Music Journal*, vol. 8, no. 3, pp. 15–31, 1984. 32

[25] N. Henrich, *Etude de la source glottique en voix parlée et chantée*. Ph.d. thesis, Université Paris 6, France, 2001. 32, 34

[26] G. Fant, *Acoustic theory of speech production*. Mouton, La Hague, 1960. 32

[27] D. Klatt and L. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers", *J. Acous. Soc. Am.*, vol. 87, no. 2, pp. 820–857, 1990. 32, 34

[28] R. Veldhuis, "A Computationally Efficient Alternative for the Liljencrants-Fant Model and its Perceptual Evaluation", *J. Acous. Soc. Am.*, vol. 103, pp. 566–571, 1998. 32

[29] A. Rosenberg, "Effect of Glottal Pulse Shape on the Quality of Natural Vowels", *J. Acous. Soc. Am.*, vol. 49, pp. 583–590, 1971. 32

[30] G. Fant, "The LF-Model Revisited. Transformations and Frequency Domain Analysis", *STL–QPSR*, 1995. 32

[31] B. Bozkurt, *Zeros of the Z-Transform (ZZT) Representation and Chirp Group Delay Processing for the Analysis of Source and Filter Characteristics of Speech Signals*. PhD thesis, Faculté Polytechnique de Mons, 2004. 33

[32] N. D'Alessandro, C. d'Alessandro, S. Le Beux, and B. Doval, "Realtime CALM Synthesizer, New Approaches in Hands-Controlled Voice Synthesis", in *NIME'06, 6th international conference on New Interfaces for Musical Expression*, (IRCAM, Paris, France), pp. 266–271, 2006. 33, 34, 37

[33] D. Zicarelli, G. Taylor, J. Clayton, jhno, and R. Dudas, *Max 4.3 Reference Manual*. Cycling'74 / Ircam, 1993-2004. 33

[34] D. Zicarelli, G. Taylor, J. Clayton, jhno, and R. Dudas, *MSP 4.3 Reference Manual*. Cycling'74 / Ircam, 1997-2004. 33

[35] M. Puckette, *Pd Documentation*. 2006. http://puredata.info. 33

[36] C. d'Alessandro, N. D'Alessandro, S. L. Beux, and B. Doval, "Comparing Time-Domain and Spectral-Domain Voice Source Models for Gesture Controlled Vocal Instruments", in *Proc. of the 5th International Conference on Voice Physiology and Biomechanics*, 2006. 34, 37

[37] R. Schulman, "Articulatory dynamics of loud and normal speech", *J. Acous. Soc. Am.*, vol. 85, no. 1, pp. 295–312, 1989. 34

[38] H. M. Hanson, *Glottal characteristics of female speakers*. Ph.d. thesis, Harvard University, 1995. 34

[39] H. M. Hanson, "Glottal characteristics of female speakers : Acoustic correlates", *J. Acous. Soc. Am.*, vol. 101, pp. 466–481, 1997. 34

[40] H. M. Hanson and E. S. Chuang, "Glottal characteristics of male speakers : Acoustic correlates and comparison with female data", *J. Acous. Soc. Am.*, vol. 106, no. 2, pp. 1064–1077, 1999. 34

[41] M. Castellengo, B. Roubeau, and C. Valette, "Study of the acoustical phenomena characteristic of the transition between chest voice and falsetto", in *Proc. SMAC 83, vol. 1*, (Stockholm, Sweden), pp. 113–23, July 1983. 34

[42] P. Alku and E. Vilkman, "A comparison of glottal voice source quantification parameters in breathy, normal and pressed phonation of female and male speakers", *Folia Phoniatr.*, vol. 48, pp. 240–54, 1996. 34

[43] H. Traunmüller and A. Eriksson, "Acoustic effects of variation in vocal effort by men, women, and children", *J. Acous. Soc. Am.*, vol. 107, no. 6, pp. 3438–51, 2000. 34

[44] N. Henrich, C. d'Alessandro, B. Doval, and M. Castellengo, "On the use of the derivative of electroglottographic signals for characterization of non-pathological phonation", *J. Acous. Soc. Am.*, vol. 115, pp. 1321–1332, Mar. 2004. 34

[45] N. Henrich, C. d'Alessandro, M. Castellengo, and B. Doval, "Glottal open quotient in singing: Measurements and correlation with laryngeal mechanisms, vocal intensity, and fundamental frequency", *J. Acous. Soc. Am.*, vol. 117, pp. 1417–1430, Mar. 2005. 34

[46] N. Henrich, G. Sundin, D. Ambroise, C. d'Alessandro, M. Castellengo, and B. Doval, "Just noticeable differences of open quotient and asymmetry coefficient in singing voice", *Journal of Voice*, vol. 17, no. 4, pp. 481–494, 2003. 35

[47] N. Henrich, "Mirroring the voice from garcia to the present day: Some insights into singing voice registers", *Logopedics Phoniatrics Vocology*, vol. 31, pp. 3–14, 2006. 35

[48] G. Bloothooft, M. van Wijck, and P. Pabon, "Relations between Vocal Registers in Voice Breaks", in *Proccedings of Eurospeech*, 2001. 35

[49] J. D. Markel and A. H. Gray, *Linear prediction of speech*. Springer-Verlag, Berlin, 1976. 36

[50] B. Story, "Physical modeling of voice and voice quality", in *proc. Voqual'03, Voice Quality: Functions, analysis and synthesis, ISCA workshop*, (Geneva, Switzerland), Aug. 2003. 36

[51] G. Carlsson and J. Sundberg, "Formant frequency tuning in singing", *J. Voice*, vol. 6, no. 3, pp. 256–60, 1992. 36

[52] http://www.vrealities.com/P5.html. 37