Proc. of the 17th Int. Conference on Digital Audio Effects (DAFx-14), Erlangen, Germany, September 1-5, 2014

A METHOD OF MORPHING SPECTRAL ENVELOPES OF THE SINGING VOICE FOR USE WITH BACKING VOCALS

Matthew Roddy
Dept. of Computer Science
and Information Systems,
University of Limerick, Ireland

Jacqueline Walker
Dept. of Electronic
and Computer Engineering,
University of Limerick, Ireland

ABSTRACT

The voice morphing process presented in this paper is based on the observation that, in many styles of music, it is often desirable for a backing vocalist to blend his or her timbre with that of the lead vocalist when the two voices are singing the same phonetic material concurrently. This paper proposes a novel application of recent morphing research for use with a source backing vocal and a target lead vocal. The function of the process is to alter the timbre of the backing vocal using spectral envelope information extracted from both vocal signals to achieve varying degrees of blending. Several original features are proposed for the unique usage context, including the use of LSFs as voice morphing parameters, and an original control algorithm that performs crossfades between synthesized and unsynthesized audio on the basis of voiced/unvoiced decisions.

1. INTRODUCTION

Sound morphing is a term that has been used to describe a wide range of processes and, as of yet, there is no consensus on a standard definition for the term due to variations in usage contexts, goals and methods. Despite the disparities in definitions, Caetano [1] remarks that, in most applications, the aim of morphing can be defined as "obtaining a sound that is perceptually intermediate between two (or more), such that our goal becomes to hybridize perceptually salient features of sounds related to timbre dimensions." The goal of achieving perceptually intermediate timbres is complicated by the multidimensional nature of timbre perception [2]. Classifications of the dimensions associated with timbre [3, 4] usually distinguish between features derived from the temporal envelope of the sound (e.g temporal centroid, log-attack time), and features derived from the spectral envelope of sounds (e.g spectral centroid, spectral tilt).

When attempting to achieve perceptually intermediate spectral features between sounds, many morphing systems adopt sinusoidal models in which the partials of a sound are represented as a sum of sinusoids that, in the case of musical sounds, are often quasi-harmonically related. A common strategy in morphing systems is to establish correspondences between the partials of two sounds and to interpolate the frequency and amplitude values [5, 6]. Methods based on this approach do not account for resonance peaks or formants that are delineated by the contour of the sound's spectral envelope. Consequently, the resulting intermediate spectral envelopes often display undesirable timbral behavior in which formant peaks are smoothed rather than shifted in frequency. Therefore, when hybridizing the non-temporal dimensions of timbre the challenge is finding parameterizations of the spectral envelope that can be interpolated to create perceptually linear shifts in timbre. Some spectral envelope parameterizations that have been proposed are: linear prediction coefficients (LPC) [7], cepstral coefficients (CC) [8], reflection coefficients (RC) [7], and line spectral frequencies (LSF) [9].

Different parameterizations of the spectral envelopes of musical instrument sounds were recently compared at IRCAM [10] using spectral shape features as timbral measures to determine which representations provided the most linear shift in peaks and spectral shape. They found that, of the parameterizations surveyed, LSFs provided the most perceptually linear morphs. This supports previous proposals [9, 11] for the use of LSFs as good parameters for formant modification. In the morphing process introduced below, this research is used in conjunction with research into the formant behavior of singers that has indicated that individual singers will sometimes alter the formant structures of vowels to blend in or stand out in an ensemble situation. Goodwin [12] found that singers in choirs lowered the intensity of their second and third formants, and sometimes shifted the formants down in frequency to blend better. Ternström [13] concluded that singers in barbershop quartets spread out the spacings of their formants to stand out for intonation purposes.

This paper presents a novel voice morphing process that is intended to be used as a studio tool to blend a backing vocal with a lead vocal. The process uses the spectral envelope of a lead vocalist to alter the spectral envelope of the backing vocalist on a frame by frame basis while preserving pitch information. The morphing process is built upon the observation that it is common in many music styles for a backing vocalist to sing the same phonetic material concurrently with the lead vocalist. Given this specific context, the formants of the two signals will be similar, and differences in the spectral envelopes can be attributed to differences in either the singer's pronunciation or the timbral characteristics of the individual's voice. It can be aesthetically desirable in this situation for vocalists to blend their timbre with other vocalists [12, 13]. In this context, if the spectral envelope of the backing vocalist is morphed with that of the lead vocalist, and the morphing method creates a perceptually linear morph, the formants that define phonetic information will remain intelligible and only the envelope information that affects the singer's individual timbre will be altered. Furthermore, since perceptually intermediary timbres between the two can be achieved using LSFs, the process can be used as a subtle effect.

This proposed morphing process could be useful in studio situations where the lead vocalist and a backing vocalist have contrasting timbres. In this scenario, the current common practice to achieve a blended timbre is to multitrack the lead vocalist perform-

ing both the lead and backing parts. In this situation, the timbral results are limited to either being perceptually blended (when the lead vocalist records both parts) or perceptually distinct (when the backing vocalist records their part). The proposed morphing process allows for a larger variety of combined vocal textures by creating gradations in the amount of blending between the two voices. The combined texture created by the two voices can be perceptually blended, perceptually distinct or any gradation in between the two depending on the LSF settings that are used.

The objectives of this voice morphing process differ from those of most morphing processes since the objective is not to achieve the target vocal sound, but rather to use its spectral envelope to modify the timbre of the source vocal, preserving its original harmonic structure and hence its fundamental frequency. The objectives of this morphing process share some similarities with those discussed in [14], in which features from two voices are combined to create a hybrid voice that retains one voice's pitch information.

The proposed morphing process falls within the bounds of some definitions of cross-synthesis in which an "effect takes two sound inputs and generates a third one which is a combination of the two input sounds. The idea is to combine two sounds by spectrally shaping the first sound by the second one and preserving the pitch of the first sound." [15] If this definition is adopted then the proposed process would be defined as cross-synthesis with a preliminary morphing stage in which the spectral envelope of the second sound is altered using envelope features extracted from the first sound.

In the next section the signal model used to morph the envelopes is described and an overview of the structure of an analysis/synthesis system that implements the process is presented. In section 3 the calculation of the LSF spectral envelope parameterization is discussed. In section 4 an original control algorithm that performs crossfades between the synthesized audio and the unsynthesized backing vocal audio is discussed. In section 5 a subjective discussion of the sonic results and the limitations of the process are presented as well as our conclusions.

2. SIGNAL MODEL AND THE STRUCTURE OF THE PROCESS

2.1. Source-filter signal model

This morphing process uses spectral modeling synthesis (SMS), as described by Xavier Serra [16], to synthesize a morphed version of a backing vocal signal. SMS models a sound $x(t)$, by splitting it into two components, a sinusoidal component $x_h(t)$, and a stochastic residual component $x_r(t)$. The sinusoidal component models the quasi-harmonic element of sounds by first detecting spectral peaks according to a quadratic peak-picking algorithm [17], followed by a refinement of these peaks on the basis of harmonic content. This harmonic component of the sound is modeled as a sum of sinusoids using:

$$x_h(t) = \sum_{k=0}^{K(t)} a_k(t) \exp[j\phi_k(t)] \tag{1}$$

where $a_k(t)$ and $\phi_k(t)$ are the amplitude and phase of the k$^{\text{th}}$ harmonic. The residual component is modeled by subtracting the harmonic component from the original signal. The residual is then synthesized using noise passed through a time-varying filter. When using SMS to synthesize the human voice, the residual generally models unvoiced sounds such as consonants and aspiration noise.

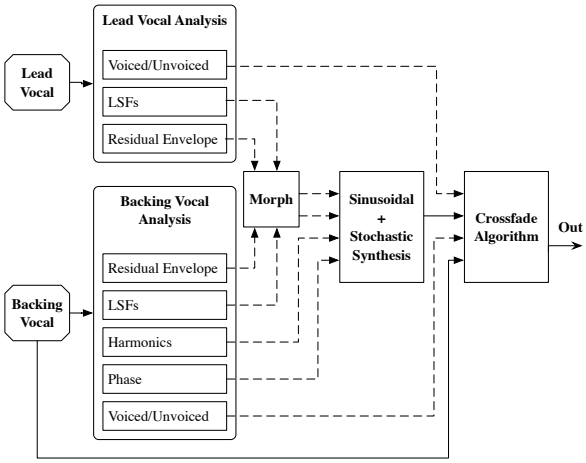

Figure 1: *Flow chart diagram of the morphing process. Dashed lines represent the flow of extracted data. Solid lines represent the flow of audio.*

The synthesis strategy adopted in this morphing process differs from traditional SMS in its use of a source-filter model which considers the amplitudes of the harmonics separately from the harmonics themselves. This model, proposed in [18], divides the harmonic component of a sound into an excitation source, in which the amplitudes of the harmonics are set to unity ($a_k = 1$), and a time-varying filter given by:

$$H(f,t) = |H(f,t)| \exp[j\psi(f,t)] \tag{2}$$

where $|H(f,t)|$ is the amplitude, and $\psi(f,t)$ is the phase of the system. The time-varying filter is derived using spectral envelope estimation methods described in section 3. The model for the representation of the harmonic element is then given by:

$$y_h(t) = \sum_{k=0}^{K(t)} |H[t, f_k(t)]| \exp[j(\phi_k(t) + \psi(f_k(t)))] \tag{3}$$

where $f_k(t) \approx k f_0(t)$, $\phi_k(t)$ is the excitation phase, and $\psi[f_k(t)]$ is the instantaneous phase of the k$^{\text{th}}$ harmonic. As such, the time-varying filter models the curve of the spectral envelope according to the formant structure and individual timbral characteristics of the singer. This approach, which was originally proposed for musical instruments, is adopted for the singing voice instead of traditional source-filter models, such as linear predictive coding, since it offers greater flexibility for timbral manipulation.

2.2. Process Structure

This morphing process belongs to the class of audio effects discussed by Verfaille *et al.* [19] known as external-adaptive audio effects. External-adaptive effects use features extracted from an external secondary input signal as control information to modify a primary input signal. In the case of this morphing process, features used to control the source-filter model described above are

extracted from the lead vocalist's signal (x_{Lv}) to alter the backing vocalist's signal (x_{Bv}) on a frame-by-frame basis. The structure of the process (shown in Fig. 1) can be divided into four stages: an analysis stage, a morphing stage, a synthesis stage, and a control stage.

During the analysis stage the spectral envelopes of the harmonic components of both the lead and backing vocal frames are estimated and parameterized as LSFs using a process described in section 3. The residual envelopes are extracted by subtracting their harmonic components from their respective magnitude spectra. Decimation is then used to create line-segment representations of the residual envelopes. Voiced/unvoiced information is also extracted from the two vocals using a two way mismatch (TWM) algorithm [20]. In addition to the three features listed above that are extracted from both voices, two additional features, the frequencies of harmonics and phase information, are extracted from the backing vocal. These two features are used, unaltered, during the synthesis process. By using the original phase and harmonic structures, the pitch information of the backing vocalist's audio is preserved and only its timbral qualities are altered.

During the morphing stage of the process, the parametric representations of both the harmonic and residual envelopes (LSFs and line segments, respectively) are morphed using:

$$M(\alpha) = \alpha S_{Lv} + [1-\alpha]S_{Bv} \qquad 0 \leq \alpha \leq 1 \qquad (4)$$

where S_{Lv} and S_{Bv} are arrays containing the spectral envelope parameters of the lead and backing vocals respectively. The variable α is the morph factor that controls the amount of timbral blending. The morphed parameters are input into the SMS system during the synthesis stage of the process along with the original harmonic frequencies and phase information of the backing vocalist. The final control stage of the process (described in section 4) uses the voiced/unvoiced information extracted during the analysis stage to perform crossfades between audio produced by the SMS system and the original unvoiced backing vocal audio.

The overall structure of the effect, and the unique control algorithm (discussed in section 4) were designed with the intention of laying the ground-work for a real-time SMS implementation. A possible real-time effect could be implemented using a side-chain to input the lead vocal signal. A similar real-time SMS application has been discussed in [21].

3. MORPHING USING LINE SPECTRAL FREQUENCIES

The chosen method of calculating LSFs begins with the magnitudes of the harmonic component of x_h, which are derived using the peak-picking algorithm. The harmonic component is first squared and then interpolated to create the power spectrum $|X(\omega)|^2$. An inverse-Fourier transform is performed on the power spectrum to calculate autocorrelation coefficients ($r_{xx}(\tau)$) according to the Wiener-Khinchin theorem:

$$r_{xx}(\tau) = \mathcal{F}^{-1}\{|X(\omega)|^2\} \qquad (5)$$

The first p autocorrelation coefficients are used to calculate p linear prediction coefficients using Levinson-Durbin recursion to solve the normal equations:

$$\sum_{k=1}^{p} a_k r_{xx}(i-k) = r_{xx}(i) \quad , i = 1, \ldots, p. \qquad (6)$$

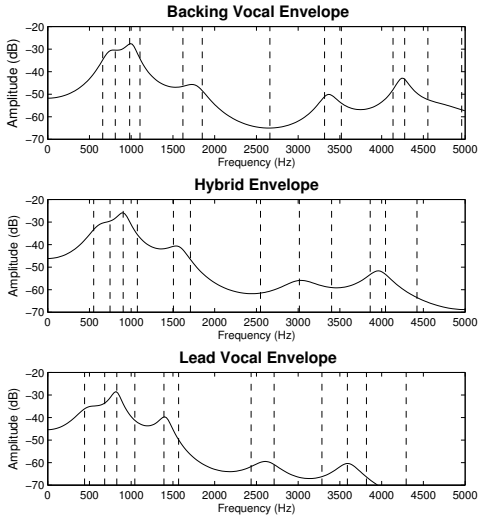

Figure 2: *Spectral envelopes demonstrating the effect of morphing sung [a] vowels using LSFs (overlaid in dashed lines). The hybrid envelope shows the resulting formant shift behavior when a morphing factor (α) of 0.5 is used.*

LSFs are then derived from the linear prediction coefficients (a_k) by considering the coefficients as a filter representing the resonances of the vocal tract. Based on the interconnected tube model of the vocal tract, two polynomials are created that correspond to a complete closure and a complete opening at the source end of the interconnected tubes [22]. The polynomials are generated from the linear prediction coefficients by adding an extra feedback term that is either positive or negative, modeling energy reflection at a completely closed glottis or a completely open glottis respectively. The roots of these polynomials are the LSFs. A thorough explanation of the process of calculating LSFs from linear prediction coefficients, as well as the reverse process, is given in [22].

In the line spectral domain, the LSFs from the backing vocal are morphed with the LSFs from the lead vocals using equation (4). An example of morphed LSFs and the hybrid spectrum created using this process are shown in Fig. 2. The figure shows a clear shift in the amplitudes and central frequencies of the of the third and fourth formants, demonstrating the good interpolation characteristics discussed in [9, 11, 10]. These morphed LSFs are then converted into the linear prediction coefficients that constitute the all-pole filter $H[f_k(t)]$ discussed in section 2.1. Using

$$H[\omega_k] = \frac{1}{1 + \sum_{n=1}^{p} a(n) \exp[-j\omega_k n T_s]} \qquad (7)$$

where $\omega_k = 2\pi f_k$ and T_s is the sampling interval, the linear prediction filter is evaluated at the individual harmonic frequencies.

4. CROSSFADE ALGORITHM

An important feature of this morphing process is a control algorithm that performs crossfades (shown in Fig. 3) between the original unvoiced consonants of the backing vocal and morphed

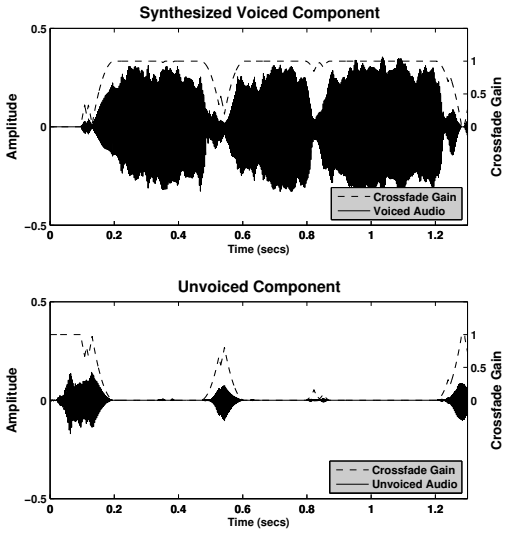

Figure 3: *The synthesized harmonic plus stochastic audio (top figure), the unsynthesized original audio (bottom figure), with their respective crossfade gain values. Crossfades with an exponential value of 2 and a fade length of 2 windows (2048 samples) were used.*

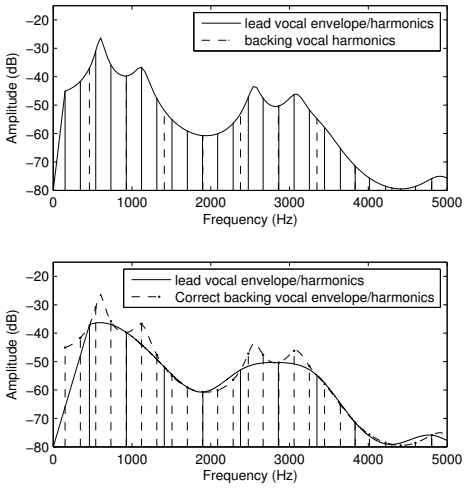

Figure 4: *Demonstration of the vowel spectra of a phoneme ([a]) created when the target lead vocal has either a lower (a) or higher (b) fundamental frequency relative to the backing vocalist. In (a) the lead vocalist has a lower fundamental (f_0 = 147 Hz) and the backing vocalist has a higher fundamental (f_0 = 497 Hz). In (b) the fundamental frequencies are swapped.*

voiced sounds. This reconstruction algorithm for the morphing process uses the voiced/unvoiced classifications for the frame plus a fade position inherited from the previous frame. The crossfades are performed by indexing tables created with user-defined exponential curves. The fades are designed to be at unity gain and the number of samples needed to complete a fade is specified by the user in window lengths. In the experiments discussed below in section 5, the hop size of 256 samples is taken into account when performing the crossfades by applying the indexed gain amount to only 256 samples at a time. The length of the fade was set to 3072 samples with an analysis window-length of 1024 samples and a sampling frequency of 44100 Hz.

The crossfades address a number of issues that are unique to the application context. Firstly, although the morphing process is designed to operate under the condition that both voices are singing the same phonetic material concurrently, there will almost always be discrepancies in the timing of the two voices. To avoid the spectral envelope of a consonant being imposed on the harmonic structure of a vowel, or vice versa, the algorithm checks whether either of the two voices contain unvoiced sounds in their corresponding frames. If so, the algorithm either fades towards the original unsynthesized audio or it remains with the unsynthesized audio at full gain, depending on the initial position of the fade. An equally important reason for using a crossfading system is that the transients of consonants synthesized using the filtered noise of SMS are considered to lack realism due to a loss of sharpness in their attack [17, 23]. A reason for performing a gradual crossfade is to make up for inaccuracies in voiced/unvoiced decisions made by the TWM algorithm during the analysis stage. These inaccuracies can be observed in Fig. 3 by the presence of jagged lines during either steady state voiced sections or during transitions. They represent decisions that change quickly over the course of a small number of frames. They are usually a single voiced frame surrounded by unvoiced frames, or vice versa. The use of gradual transitions masks the overall impact that these isolated voicing classifications have.

5. DISCUSSION

5.1. Informal Testing

The effectiveness of the two principal features of this morphing process (the use of LSFs and the reinsertion of unvoiced consonants using crossfades) were informally tested by comparing the morphing process with a second SMS-based morphing process [24] that uses synthesized unvoiced segments and morphs voiced segments using simple interpolation of the spectral envelopes created by the harmonic components. From a five second recording of a backing vocal, two sets of processed backing vocals were created: one using the morphing process presented here, and another using the second envelope interpolation process used for comparison. In each of the sets, the backing vocal was synthesized using the morphing factors: $\alpha = 0, 0.5, 1.0$. To assess the realism of the resulting audio, the two sets were first played independent of their corresponding lead vocal. Subsequently, the same processed backing vocals were played in conjunction with their corresponding lead vocal to informally assess the level of perceptual blending.

An initial observation was that the realism contributed by the reintroduction of the original unvoiced consonants using the crossfade algorithm was significant when compared with the envelope interpolation process without the reinsertion of consonants. Similar to what was found by [17, 23], the use of SMS to model unvoiced segments was considered to result in consonants that lacked

definition due to being modeled by the noise residual. A drawback of the use of the crossfades was that, as α increased, there were noticeable artifacts that appeared during the transitions between synthesized and unsynthesized audio. These artifacts are due to the differences between the two spectral envelopes that are perceptually highlighted by rapid changes. The effect of these artifacts can be reduced by increasing the length of the crossfade. When considering the realism contributed by the LSFs, as the α value was increased, the resulting voiced sounds of the LSF-based morphing process remained defined and realistic, due to the linear shift in timbral features. In contrast, the voiced segments synthesized using the second SMS morphing process lacked definition at $\alpha = 0.5$, due to the peak smoothing behavior that occurs during the interpolation of envelopes. When the two sets of processed backing vocals were played in conjunction with the lead vocal it was considered that the formant shift behavior due to the use of LSFs increased the level of perceptual blend between the two voices as the α value was increased. With the second SMS morphing process, this was not always the case due to the peak smoothing behavior.

5.2. Limitation

One of the limitations of the morphing process presented here is that it cannot be used to effectively blend backing vocals that have a lower fundamental than their corresponding lead vocals. This is due to the envelope-sampling behavior of harmonics. As shown in Fig. 4, the harmonics sample the vowel envelope at frequencies that are approximately integer multiples of the fundamental. Given the case of a backing vocal with a lower fundamental than the lead vocal, the lead vocal vowel envelope will not be sampled at a high enough rate for the backing vocalist to accurately recreate the formants of the vowel. In addition, the harmonics of the backing vocal that are at lower frequencies than the fundamental of the lead vocal cannot be designated appropriate amplitude values since there is no vowel envelope information at frequencies below the fundamental.

5.3. Conclusion

The voice morphing process presented in this paper uses LSFs to modify the timbral characteristics of a backing vocal, including the frequencies and strengths of formants, to achieve different levels of blending with a target lead vocal. In choral situations, formant modification by singers has been observed in which formant strengths have been lowered and centre frequencies slightly shifted for the purpose of blending [12]. Although the actions of a choral singer and the timbral modifications produced by this process create different results, both are motivated by the objective of producing a homogeneity of timbre through modification of the spectral envelope. For this reason, this process is proposed as a potentially valuable artistic tool for blending two voices.

6. REFERENCES

[1] M. F. Caetano and X. Rodet, "Automatic timbral morphing of musical instrument sounds by high-level descriptors," in *Proceedings of the International Computer Music Conference*, 2010, pp. 254–261.

[2] J. M. Grey, "Multidimensional perceptual scaling of musical timbres," *The Journal of the Acoustical Society of America*, vol. 61, p. 1270, 1977.

[3] K. Jensen, "The timbre model," *Journal of the Acoustical Society of America*, vol. 112, no. 5, p. 2238–2251, 2002.

[4] G. Peeters, "A large set of audio features for sound description (similarity and classification) in the cuidado project," IRCAM, Tech. Rep., 2004.

[5] E. Tellman, L. Haken, and B. Holloway, "Morphing between timbres with different numbers of features," *Journal of the Audio Engineering Society*, vol. 62, no. 2, pp. 678–689, 1995.

[6] K. Fitz and L. Haken, "Sinusoidal modeling and manipulation using lemur," *Computer Music Journal*, vol. 20, no. 4, pp. 44–59, 1996.

[7] J. A. Moorer, "The use of linear prediction of speech in computer music applications," *Journal of the Audio Engineering Society*, vol. 27, no. 3, pp. 134–140, 1979.

[8] M. Slaney, M. Covell, and B. Lassiter, "Automatic audio morphing," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, vol. 2, 1996, pp. 1001–1004.

[9] K. K. Paliwal, "Interpolation properties of linear prediction parametric representations," in *Fourth European Conference on Speech Communication and Technology*, 1995.

[10] M. Caetano and X. Rodet, "Musical instrument sound morphing guided by perceptually motivated features," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 8, pp. 1666–1675, Aug. 2013.

[11] R. W. Morris and M. A. Clements, "Modification of formants in the line spectrum domain," *Signal Processing Letters, IEEE*, vol. 9, no. 1, pp. 19–21, 2002.

[12] A. W. Goodwin, "An acoustical study of individual voices in choral blend," *Journal of Research in Music Education*, vol. 28, no. 2, 1980.

[13] S. Ternstrom and G. Kalin, "Formant frequency adjustment in barbershop quartet singing," in *Intenational Congress on Acoustics*, 2007.

[14] P. Depalle, G. Garcia, and X. Rodet, "The recreation of a castrato voice, Farinelli's voice," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 1995, p. 242–245.

[15] U. Zölzer, Ed., *DAFX: Digital Audio Effects*. John Wiley & Sons, 2011, ch. Glossary, pp. 589–594.

[16] X. Serra and J. Smith, "Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition," *Computer Music Journal*, vol. 14, no. 4, pp. 12–24, 1990.

[17] X. Serra, "A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition," Ph.D. dissertation, Stanford University, 1989.

[18] M. Caetano and X. Rodet, "A source-filter model for musical instrument sound transformation," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, 2012, pp. 137–140.

[19] V. Verfaille, U. Zölzer, and D. Arfib, "Adaptive digital audio effects (a-DAFx): a new class of sound transformations," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1817–1831, Sep. 2006.

[20] R. C. Maher and J. W. Beauchamp, "Fundamental frequency estimation of musical signals using a two-way mismatch procedure," *The Journal of the Acoustical Society of America*, vol. 95, pp. 2254–2263, 1994.

[21] 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, pp. 109–112.

[22] I. V. McLoughlin, "Line spectral pairs," *Signal Processing*, vol. 88, no. 3, pp. 448–467, Mar. 2008.

[23] T. S. Verma and T. H. Meng, "Extending spectral modeling synthesis with transient modeling synthesis," *Computer Music Journal*, vol. 24, no. 2, pp. 47–59, 2000.

[24] J. Bonada, X. Serra, X. Amatriain, and A. Loscos, *DAFX: Digital Audio Effects*. John Wiley & Sons, 2011, ch. Spectral Processing, pp. 393–445.