A Mouth Opening Effect Based on Pole Modification for Expressive Singing Voice Transformation

Luc Ardaillon, Axel Roebel

A mouth opening effect based on pole modification for expressive singing voice transformation

Luc Ardaillon, Axel Roebel

IRCAM - UMR STMS (IRCAM - CNRS - Sorbonne Universits) Paris, France

luc.ardaillon@ircam.fr, axel.roebel@ircam.fr

Abstract

Improving expressiveness in singing voice synthesis systems requires to perform realistic timbre transformations, e.g. for varying voice intensity. In order to sing louder, singers tend to open their mouth more widely, which changes the vocal tract's shape and resonances. This study shows, by means of signal analysis and simulations, that the main effect of mouth opening is an increase of the 1^{st} formant's frequency (F_1) and a decrease of its bandwidth (BW_1). From these observations, we then propose a rule for producing a mouth opening effect, by modifying F_1 and BW_1, and an approach to apply this effect on real voice sounds. This approach is based on pole modification, by changing the AR coefficients of an estimated all-pole model of the spectral envelope. Finally, listening tests have been conducted to evaluate the effectiveness of the proposed effect.

Index Terms: mouth opening, intensity, speech transformation, singing synthesis, spectral modelling

1. Introduction

If singing voice synthesis systems have nowadays reached a rather satisfying quality, generated voices still lack of expressiveness compared to real singers. One of the major current challenges to improve expressiveness in such systems is to allow for realistic voice timbre changes for varying pitch, intensity, and voice quality. For statistical-based and concatenative approaches, all those timbre variations should ideally be included in the system's database in order to be reproduced accurately. However, due to the very wide variety of timbres and parameters to be considered, this task is hardly achievable. It is therefore necessary to design perceptually relevant rules and algorithms to allow realistic and flexible timbre transformations.

Past studies have already permitted to gather knowledge about how various voice spectral features may change during singing, which could be exploited in synthesis systems by defining a set of rules, as proposed in [1] with formant tuning for pitch change. But if such rules are easy to implement in formant synthesizers like [2], where all source and formants' parameters can be explicitly controlled, this task is much more complicated for systems based on signal transformations like [3] and [4].

Our current research is focused on producing vocal intensity variations in the context of concatenative singing voice synthesis. For this purpose, various aspects have to be considered. As pointed out in several sources [5, 6, 7], a vocal effort increase is physiologically linked to an increase in subglottal pressure, an increase in vocal-fold tension, and a wider mouth opening. From the view point of the source-filter model, the well-known effect of increasing the vocal folds' tension on the source's spectrum is an increase of the glottal formant frequency and a decrease of the spectral tilt, while the main effect related to a wider mouth opening is an increase of the 1^{st} formant frequency (F_1), as observed in many studies [5, 6, 7, 8, 9]. Additional aspects to be considered may be the level of aspiration noise, or the singing formant's prominency [10]. For a fully realistic effect, all those aspects should ideally be considered.

A first possible approach to create such transformations of vocal intensity is spectral morphing, that uses target templates recorded at different levels of low and high vocal effort, as proposed in [11, 12] and [13] for diphone speech synthesis. We also explored the potential of such a morphing approach for singing in a previous study [14]. The advantage of this method is that it includes both the effects on the source and the vocal tract. However, the target envelopes are taken from stable parts of phonemes only and thus don't allow to correctly reproduce the timbre variations occurring on coarticulation parts, which may create unnatural transitions at vowels' boundaries. Moreover, recordings at different vocal intensities are not always available. It thus seems pertinent to use a parametric approach to produce such effect without the requirement of additional recordings.

In that direction, [15] and [16] focus on the source-related spectral characteristics, filtering in the spectral domain to modify spectral tilt and glottal formant. The perceived tenseness of the source can also be modified by changing the R_d parameter of the transformed LF model [17] in parametric voice resynthesis, as done in [18]. However, modifying only the source spectrum is not sufficient for producing a convincing effect, and the Vocal Tract Filter (VTF) should also change accordingly.

[5] proposes a mean to add missing high-frequencies harmonics in the spectrum for weak-to-loud transformations, and compares morphing and spectral modelling approaches for modifying the spectral tilt and F_1, thus including an effect of the VTF. In [19], authors also employ a spectral modelling approach to apply intensity transformations of singing voice, using a parametric model of spectral envelope based on 4-pole resonators to modify gain, spectral tilt, and formants' frequencies and bandwidths, based on regressions computed from 60 recorded vowels. However, it requires to properly extract the parameters of all formants, which is not straightforward without manual correction. Another drawback in [19] is that the regressions are computed from all vowels, without considering the gender of the singer or the type of vowel. This probably oversmoothes the variations and doesn't allow to observe the typical move of F1, as mouth opening can't physiologically be as prominent for closed vowels like /u/ than for open vowels like /a/.

Thus, it would be advantageous to decompose the effect of intensity variations in several physiologically meaningful components that can then be adapted to the context of vowel, gender, or singing style. This work presents a first step towards a complete parametric vocal intensity transformation, focusing on effects induced by mouth opening. First, real signals analysis and simulations allowed us to get more insight on the variation of

Work supported by the ANR project ChaNTeR (ANR-13-CORD-011)

F_1 and BW_1 with intensity and mouth opening, and to quantify it to infer a simple rule to be used for synthesis. Then, an approach based on pole modification from an all-pole envelope model is proposed in order to apply this rule on real sounds. Finally, an evaluation has been conducted to validate the effectiveness of the proposed rule and approach.

2. Real signals analysis

As a starting point for our study, we analyzed spectral envelopes on the 15 French vowels recorded by a professional male singer, sustained on 1 pitch ($135 H_z$) and 5 levels of intensity, from pianissimo (*pp*) to fortissimo (*ff*). In order to observe the change in the vocal tract's resonances, we employed the following procedure for estimating the VTF:

- First, the DAP algorithm [20] with order 50 was used to estimate an all-pole model of the spectral envelope;
- Then, the R_d parameter of LF source model was estimated based on algorithm described in [21, 22], and the contribution of the source was removed by spectral division of the DAP envelope with the spectral shape associated with the estimated R_d value.

Informal observations of these analysis confirmed an increase of F_1 with intensity in many cases, especially for open vowels like /a/, and sometimes a decrease of BW_1. However, these observations present an important variability, depending on the vowel, but also maybe related to some limitations of the analysis algorithms, or to the consistency of the recordings across the various intensity levels. This variability makes it difficult to infer a precise rule to be used in a synthesis system, based on those observations only. For this purpose, we thus employed a simulation approach, as described in the next section.

3. Simulations

As seen in sections 1 and 2, it is rather clear that F_1 should increase with mouth opening. But it is not clear to what extent, and most studies don't evoke a possible change of BW_1. Another possible approach to make assumptions on the voice characteristics behaviours is to use simulations.

As explained in [23, 24], there is a direct equivalence between the simple acoustic tube model of the vocal tract and linear prediction. We can thus use this relation to estimate a Vocal Tract Shape (VTS) from an all-pole model, modify this VTS to simulate mouth opening, and convert it back to the all-pole model to observe the effect on formants parameters. For this purpose, reflection coefficients can be computed from the AR coefficients using the step-down procedure (and conversely the step-up procedure to recover back AR coefficients), and the area ratios of the acoustic tube model (and from there the vocal tract area function) are deduced from the reflection coefficients, as explained in [24]. In the following experiment, we used the open-source sparkNG software [25, 26], that implements those procedures to convert back and forth between the VTS and its equivalent all-pole model, while providing a convenient GUI to manipulate, display, and store both.

3.1. Simulation procedure

Figure 1 shows the interface of the software. In the top left panel, the VTS is displayed from the lips to the glottis as areas of the acoustic tube model; the top right panel shows the corresponding all-pole spectral envelope model. The frequency and bandwidth of the formants can be displayed just below the plot.

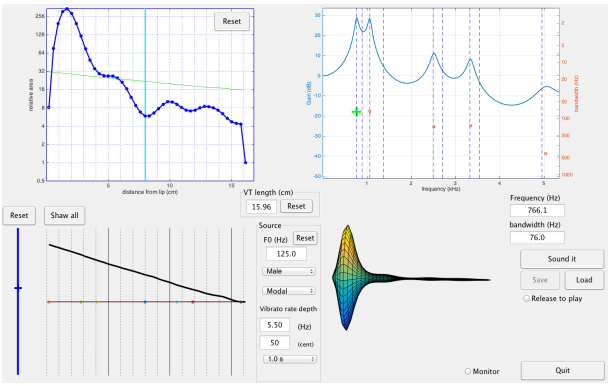

Figure 1: *Interface of the sparkNG software*

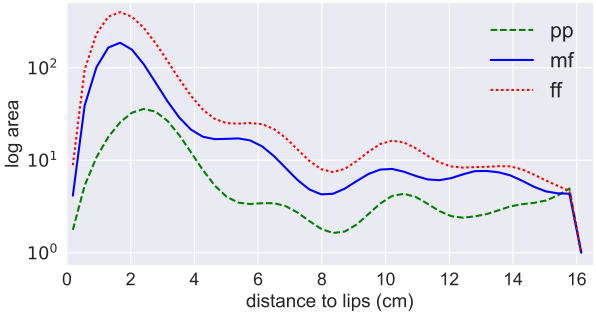

Figure 2: *Estimated VTS for vowel /a/ at intensities pp, mf, ff*

The bottom left panel allows to draw a modification curve to be added to the original VTS, and the all-pole model is automatically re-computed. The slider on the left allows to set the degree of this modification. A button also allows to synthesize the corresponding sound, giving an idea of the perceptual effects.

Based on the analysis exposed in section 2, we used the estimated all-pole models to get the corresponding VTS. Figure 2 shows as an example the estimated VTS for the vowel /a/ sung at 3 different intensity levels (*pp*, *mf* and *ff*). As could be expected, it clearly exhibits an increasing mouth opening from pp to ff. Similar results can be obtained for other open vowels, but not for closed vowels like /u/.

For the rest of our experiment, we thus only used the 4 most open vowels: /a/, /E/, /9/ and /O/ (in SAMPA notation). As a first approximation of the VTS change induced by mouth opening, we used a linear slope from the glottis to the lips, as shown in figure 1, and applied it to the estimated VTS of the 4 vowels sung at medium intensity level (*mf*). By scaling this shape modification curve using the slider, such that the opening at the lips was multiplied by factor $\gamma \in [0.25, 0.5, 1., 2, 4]$ (on a linear scale), we could then measure the variations of the formants parameters induced according to the degree of mouth opening.

3.2. Results

Figure 3 shows the ratios of the estimated formants frequencies (R_{F_i}) after modification of the shape over the original values, for the 4 vowels (with γ displayed on a log scale). Similarly, figure 4 shows the ratios for the estimated bandwidths (R_{BW_i}). As one can see, the main effect of the simulated mouth opening is a linear increase of F_1 and linear decrease for BW_1 (according to $log(\gamma)$), similarly for the 4 vowels. Comparatively, the effect on the 3 other formants is negligible. Again, this increase of F_1 is coherent with previous studies and observations, but

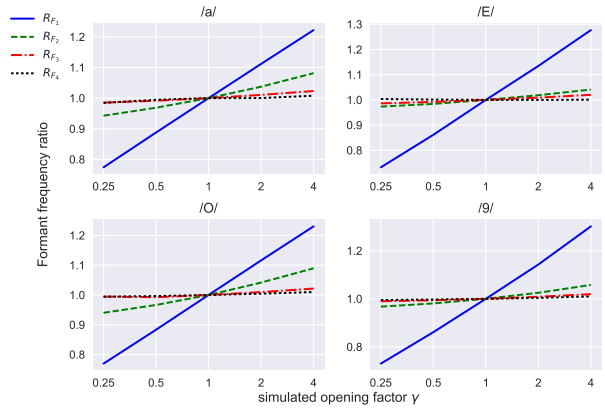

Figure 3: *Ratios of new and original formants frequencies as a function of γ (in log scale) for vowels /a/, /E/, /O/ and /9/*

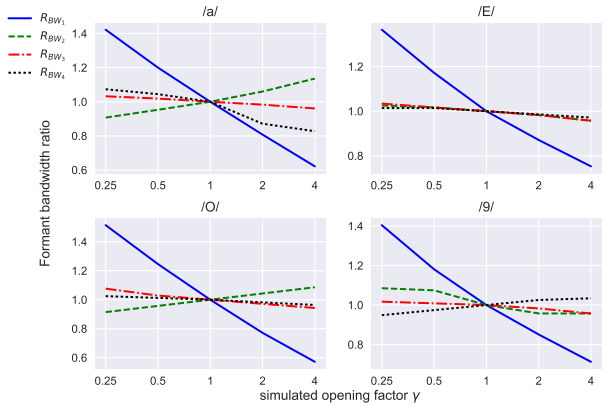

Figure 4: *Ratios of new and original formants bandwidths as a function of γ (in log scale) for vowels /a/, /E/, /O/ and /9/*

such a change in bandwidth is rarely reported in the literature. However, a decrease of all formants bandwidths with intensity was reported in [19], and [27] also evokes an increase of the 1^{st} formant bandwidth in soft breathy voices, which is thus coherent with the results of the present simulation.

4. Mouth opening effect

4.1. Rule for formant's modification

From these simulations, a simple rule has been inferred, based on the mean slope for F_1 and BW_1 over the 4 simulated vowels. This rule is given by the following 2 equations:

$$F1_{new} = F_1 \cdot (1 + 0.25\alpha) \quad (1)$$

$$BW1_{new} = BW_1 \cdot (1 - 0.4\alpha) \quad (2)$$

where $\alpha \in [-1; 1]$ is the opening factor, and $F1_{new}$ and $BW1_{new}$ are the new values of F_1 and BW_1 to be applied for transforming the original sound. The gain of the formant was not included in this rule, as it is not independent from its bandwidth and from interaction with other close formants.

4.2. Proposed transformation approach

In order to apply the defined rule, the spectral envelope has to be modified appropriately to include the change in F_1 and BW_1. In [28], frequency warping is used in order to transform the spectral envelope to create natural pitch changes in singing voice. However, it is not easy to precisely change the formants parameters using frequency warping. Especially, the natural increase of amplitude when bandwidth decreases is not reproduced, and formants can't be merged together when getting closer as they should.

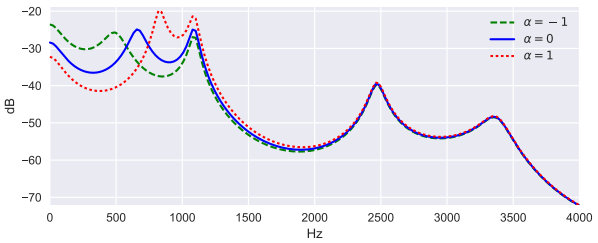

Figure 5: *Transformation of vowel /a/ with $\alpha \in [-1, 0, 1]$*

Another possible approach that we propose to use here is pole modification, based on an all-pole model of spectral envelope. This possibility is mentioned in [29], but has been discarded for being too complicated. A similar approach was used in [30] and [31] to modify formants, focusing on controlling pole interaction. These works are also mentioned in [32], but author evokes as a limit that the amplitude and bandwidth of the formants can't be controlled independently. However, we assume that this is not necessary, as amplitude should naturally change when bandwidth is modified, as stated in [27]. The proposed method is described below.

An all-pole model can be defined by its transfer function:

$$H(z) = \frac{G}{1 + \sum_{k=1}^{P} a_k z^{-k}} \quad (3)$$

where the a_k are the AR coefficients of the model, P is the model order, and G is a fixed gain coefficient. With such a representation, assuming the model is estimated with an appropriate order, the most important formants of the voice should be associated with a conjugate poles pair.

From the AR coefficients, the roots r_i of the polynomial can be computed. Then, the frequencies and -3dB bandwidths of the poles, in H_z, are given by the following formulas [32]:

$$F_i = \frac{f_s}{2\pi} \angle r_i \quad (4)$$

$$BW_i = -\frac{f_s}{\pi} ln(|r_i|) \quad (5)$$

Once the frequency and bandwidth of each pole are known, the values corresponding to the 1^{st} formant can be selected and modified according to equations 1 and 2. From there, functions defined in equations 4 and 5 can be inverted in order to get the new roots of the model. Finally, these modified roots are converted back to the AR coefficients of the new transfer function, using Leja ordering to limit effects of rounding errors in computation [33]. Figure 5 shows an example of using this pole modification approach to apply the given rule on a spectral envelope of the vowel /a/, for $\alpha \in [-1, 0, 1]$.

The transformation is then applied by inverse filtering the original sound by the estimated spectral envelope, and filtering it back with the newly modified envelope.

5. Evaluation

In order to evaluate the proposed transformation, 2 online listening tests have been run for evaluating both the quality of the transformed sounds in term of naturalness, and the proper perception of the degree of mouth opening produced by the effect. A demo page with the sounds used in the tests can be found at url [34]. All sounds were normalized at the same level, and listeners had to use headphones or earphones to do the tests.

5.1. Original sounds

For the 2 tests, recordings of the vowels /a/, /E/, /9/ and /O/, sung by both a male and a female professional singer, at a fixed pitch (135Hz for the male voice, and 250Hz for the female one) and a medium intensity (*mf*), were selected. On each of these sounds, the transformation was applied with $\alpha \in [-1; -0.5; 0; 0.5; 1]$, 0 meaning no transformation. A total of 40 sounds (8 original sounds $\times$ 5 α values) were thus used in the tests.

5.2. AR model estimation

For the transformations, the spectral envelope was first analyzed for each sound, using an implementation of the DAP algorithm [20] in the SuperVP software [35]. An important parameter to be considered for analysis is the order of the model. From the results of [36, 37] and informal tests on our database, we used an order of 50 which seemed appropriate in our case.

5.3. Estimation of the 1^{st} formant

Ideally, the pair of conjugate poles with the lowest absolute frequency should correspond to the 1^{st} formant, but depending on the analysis, this may not always be the case. For a more robust estimation of the 1^{st} formant, we thus imposed as a constraint that $F_1 \in [350 - 1000]H_z$. However, this ranges could potentially be adapted according to the analyzed voice and vowel. Then, the pair of conjugate poles with the lowest frequency following this constraints is selected.

5.4. Stable frames selection

For transforming a sound, the pole modification has to be applied on each frame. Neighbouring frames should have similar formants and therefore pole parameters. However, the analysis may sometimes exhibit some jumps in the estimated poles' frequencies from one frame to another, which could create artifacts. To avoid this, the median value of the F_1 over the central stable part of the vowel is computed. Then, only the 25% of the frames for which the estimated F_1 are the closest to the median are kept. The other frames are discarded and the remaining ones are interpolated after transformation, in order to fill the gaps.

5.5. Test 1: Quality of the transformation

In the first test, 15 of the 40 sounds where randomly selected and presented in random order, with no repetition. The sounds then had to be rated on a 1 to 5 scale by listeners following a standard MOS test procedure [38], according to the perceived naturalness of the sounds.

48 listeners answered this test. Figure 6 shows the results for the different α values. As one can see, the answers highly overlap and span the whole [1-5] range for all sounds categories, with no difference between α values -0.5, 0, and 0.5. The mean rating is however a bit lower for higher transformations levels -1 and 1. The results were slightly better for the male voice. As a result, we can thus assume that the proposed approach can apply the transformation appropriately without important degradation of the naturalness and sound quality.

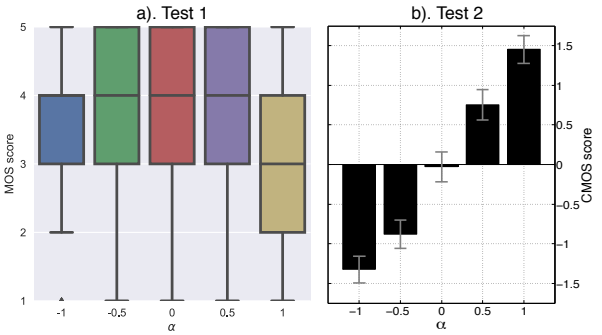

Figure 6: *Results of listening tests*

5.6. Test 2: Perception of transformed mouth opening

The second test presented 15 pairs of sounds to the listener. For each pair, one voice (male or female) and one vowel where first randomly selected, and 2 different sounds were randomly chosen among the 5 possible transformation levels to form the pair. Then, for each pair, a grade from 0 to 3 had to be given according to which sound seemed to be related to a more widely open mouth and how big was the perceived difference between the 2 sounds, following a standard CMOS procedure [39].

Before starting the test, listeners were given, as a reference, sounds of a vowel /a/ recorded with 5 degrees of mouth opening (from maximally closed to maximally open), in order to give an idea of the expected perceptual effect. Also, as the transformation may sometimes change the perception of the vowel identity (an /a/ pronounced with a closed mouth may sound almost like an /o/ or /2/), the vowel was given above each pair of sounds.
36 persons answered this test. As one can see on figure 6, those results perfectly reflect the expected effect of the transformation. None of the confidence intervals overlap, which means that the difference between each degree of the applied effect is clearly perceived by listeners, and the results were very similar for both voices. As a result, one can assume that the proposed rule and transformation method are effective for simulating mouth opening.

6. Conclusions

From signal analysis and simulations, we proposed a rule for modifying the frequency and bandwidth of the 1^{st} formant in order to produce a mouth opening effect. An approach to apply it on real sounds was proposed, using pole modification based on an all-pole model of the spectral envelope. 2 listening tests were conducted in order to validate the effectiveness of the transformation in terms of naturalness of the transformed sounds and of perception of the degree of mouth opening induced. The results of these tests revealed that the transformations were perceived as expected, and that the proposed rule and method were thus very effective for simulating mouth opening on real voice sounds, with very few degradation of the naturalness and sound quality.

In future work, this mouth opening effect should be included as part of a more general intensity transformation, accompanying a modification of the source characteristics (glottal formant, spectral tilt, noise level, and singing formant). However, it should be noted that the effect should be adapted to the nature of the transformed vowel, as the degree of mouth opening is physiologically limited for some closed vowels like /i/ or /u/.

The same approach could also be used for applying other rules, such as proposed in [1] or [29] for tuning formants frequencies to the f_0 for more realistic pitch transformations.

7. References

[1] N. Henrich, J. Smith, and J. Wolfe, "Vocal tract resonances in singing : Strategies used by sopranos , altos , tenors , and baritones," 2010.

[2] L. Feugère, C. D'Alessandro, B. Doval, and O. Perrotin, "Cantor Digitalis: chironomic parametric synthesis of singing," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2017, no. 1, p. 2, 2017.

[3] H. Kenmochi and H. Ohshita, "VOCALOID Commercial singing synthesizer based on sample concatenation," no. August, pp. 3–4, 2007.

[4] L. Ardaillon, G. Degottex, and A. Roebel, "A multi-layer F0 model for singing voice synthesis using a B-spline representation with intuitive controls," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2015-Janua, no. April 2016, pp. 3375–3379, 2015.

[5] O. Perrotin and C. D'Alessandro, "Vocal effort modification for singing synthesis," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 08-12-Sept, 2016, pp. 1235–1239.

[6] H. Traunmü Ller and A. Eriksson, "Acoustic effects of variation in vocal effort by men, women, and children," *Journal of the Acoustical Society of America*, vol. 107, no. 6, pp. 3438–3451, 2000.

[7] J. S. Liénard and C. Barras, "Fine-grain voice strength estimation from vowel spectral cues," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, no. August, 2013, pp. 128–132.

[8] J. S. Liénard and M. G. Di Benedetto, "Effect of vocal effort on spectral properties of vowels." *The Journal of the Acoustical Society of America*, vol. 106, no. 1, pp. 411–22, 1999.

[9] J. E. Huber, E. T. Stathopoulos, G. M. Curione, and T. A. Ash, "Formants of children , women , and men : The effects of vocal intensity variation," no. October 1999, 1999.

[10] J. Sundberg, "Level and center frequency of the singer's formant," *Journal of Voice*, vol. 15, no. 2, pp. 176–186, 2001.

[11] M. Schröder and M. Grice, "Expressing vocal effort in concatenative synthesis," 2003, pp. 2589–2592.

[12] O. Turk, M. Schroder, B. Bozkurt, and L. M. Arslan, "Voice Quality Interpolation for Emotional Text-To-Speech Synthesis," no. 2, 2005.

[13] C. Defez, J. C. Socor, and R. A. J. Clark, "Parametric model for vocal effort interpolation with Harmonics Plus Noise Models," 2013, pp. 25–30.

[14] G. Degottex, L. Ardaillon, and A. Roebel, "Multi-Frame Amplitude Envelope Estimation for Modification of Singing Voice," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 24, no. 7, pp. 1242–1254, 2016.

[15] C. Alessandro and B. Doval, "Voice quality modification for emotional speech synthesis." no. January 2003, 2003.

[16] C. Alessandro, B. Doval, and O. Cedex, "Experiments in voice quality modification of natural speech signals : The spectral approach," 1998.

[17] G. Fant, "The LF-model revisited . Transformations and frequency domain analysis," 1995.

[18] S. Huber, "Voice Conversion by modelling and transformation of extended voice characteristics," Ph.D. dissertation, 2015.

[19] E. Molina, I. Barbancho, A. M. Barbancho, and L. J. Tard, "Parametric model of spectral envelope to synthesize realistic intensity variations in singing voice," pp. 634–638, 2014.

[20] A. El-Jaroudi and J. Makhoul, "Discrete All-Pole Modeling," *IEEE transactions on signal processing*, vol. 39, 1991.

[21] G. Degottex, A. Roebel, and X. Rodet, "Phase Minimization for Glottal Model Estimation," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1080–1090, 2011.

[22] S. Huber and A. Roebel, "On the use of voice descriptors for glottal source shape parameter estimation," *Computer Speech and Language*, vol. 28, no. 5, pp. 1170–1194, 2014.

[23] H. Wakita, "Direct Estimation of the Vocal Tract Shape by Inverse Filtering of Acoustic Speech Waveforms," *IEEE Transactions on Audio and Electroacoustics*, vol. 21, no. 5, pp. 417–427, 1973.

[24] J. D. Markel and A. H. J. Gray, *Linear prediction of speech*. Springer Science & Business Media, 2013, vol. 12.

[25] H. Kawahara, "SparkNG: Interactive MATLAB tools for introduction to speech production, perception and processing fundamentals and application of the aliasing-free L-F model component," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 08-12-Sept, pp. 1180–1181, 2016.

[26] "sparkNG software download page," 2017. [Online]. Available: http://www.wakayama-u.ac.jp/ kawahara/MatlabRealtimeSpeechTools/

[27] H. Park, "Time Course of the First Formant Bandwidth," no. 1, pp. 213–224, 2002.

[28] M. Dong, P. Chan, L. Cen, and S. W. Lee, "Spectral Transformation of Singing Vowels by Dynamic Frequency Warping," pp. 3–6, 2011.

[29] J. L. Santacruz, L. J. Tardón, Isabel Barbancho, and A. M. Barbancho, "Spectral Envelope Transformation in Singing Voice for Advanced Pitch Shifting," 2011.

[30] H. Mizuno, M. Abe, and T. Hirokawa, "Waveform-based speech synthesis approach with a formant frequency modification," pp. 195–198, 1993.

[31] Y. Hsiao and D. Childers, "A new approach to formant estimation and modification based on pole interaction," *Signals, Systems and Computers, ...*, pp. 783–787, 1996. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=601162

[32] M. E. Lee, "Acoustic Models for the Analysis and Synthesis of the Singing Voice," Ph.D. dissertation, 2005.

[33] C. F. Pedersen, "Leja ordering LSFs for accurate estimation of predictor coefficients," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 2545–2548, 2011.

[34] "demo page," 2017. [Online]. Available: http://recherche.ircam.fr/anasyn/ardaillon/mouthOpening2017/demo.php

[35] "SuperVP software," 2017. [Online]. Available: http://anasynth.ircam.fr/home/english/software/SuperVP

[36] F. Villavicencio, A. Röbel, and X. Rodet, "All-pole spectral envelope modelling with order selection for harmonic signals," 2007.

[37] A. Röbel, F. Villavicencio, and X. Rodet, "On cepstral and all-pole based spectral envelope modeling with unknown model order," 2007.

[38] ITU-T, "Recommendation ITU-T-P.800.2 : Mean opinion score interpretation and reporting," Tech. Rep., 2016.

[39] ITU, "BS.1284-1 General methods for the subjective assessment of sound quality," pp. 1–13, 2003.