Vocal effort modification for singing synthesis

Olivier Perrotin, Christophe d'Alessandro

INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Vocal effort modification for singing synthesis

Olivier Perrotin, Christophe d'Alessandro

LIMSI, CNRS, Université Paris-Saclay, France
olivier.perrotin@limsi.fr cda@limsi.fr

Abstract

Vocal effort modification of natural speech is an asset to various applications, in particular, for adding flexibility to concatenative voice synthesis systems. Although decreasing vocal effort is not particularly difficult, increasing vocal effort is a challenging issue. It requires the generation of artificial harmonics in the voice spectrum, along with transformation of the spectral envelope. After a raw source-filter decomposition, harmonic enrichment is achieved by 1/ increasing the source signal impulsiveness using time distortion, 2/ mixing the distorted and natural signals' spectra. Two types of spectral envelope transformations are used: spectral morphing and spectral modeling. Spectral morphing is the transplantation of natural spectral envelopes. Spectral modeling focuses on spectral tilt, formant amplitudes and first formant position modifications. The effectiveness of source enrichment, spectrum morphing, and spectrum modeling for vocal effort modification of sung vowels was evaluated with the help of a perceptive experiment. Results showed a significant positive influence of harmonic enrichment on vocal effort perception with both spectral envelope transformations. Spectral envelope morphing and harmonic enrichment applied on soft voices were perceptively close to natural loud voices. Automatic spectral envelope modeling did not match the results of spectral envelope morphing, but it significantly increased the perception of vocal effort.
Index Terms: vocal effort, speech transformation, singing synthesis, spectral model

1. Introduction

Vocal effort, or voice perceived power, corresponds to changes of loudness and timbre in the voice. In singing, it is employed for aesthetic purposes as it contributes to the dynamics of musical pieces. The vocal effort dimension is as decisive as pitch and rhythm control for expressive singing performances. However, replication of vocal effort variations remains a challenge in concatenative singing synthesis. Since the latter aims at selecting and combining singing units extracted from a database, only vocal efforts levels that were recorded can be synthesized and perceived [1]. To avoid the tedious recording of numerous vocal effort levels, signal processing techniques are often employed to modify the perceived vocal effort level of recorded singing units.

While a large number of studies have been dedicated to the analysis of spectral properties of vocal effort [2], [3], [4], [5], few have dealt with synthesis. Among them, one can identify two types of synthesis techniques: spectral morphing and spectral modeling. Spectral morphing consists in extracting spectral envelopes from units with low and high vocal effort levels, and applies a weighted average spectral envelope to the low or high effort signal to synthesize intermediate vocal effort levels [6], [7]. With spectral modeling, spectral transformations based on the analysis of spectral properties of vocal effort are applied to single units to change their vocal effort from one level to another [8], [9]. Moreover, it has been pointed out in the latter studies that while decreasing the vocal effort of natural speech is easily achieved by the attenuation of high-frequency parts of the loud voice spectrum, increasing vocal effort is a more challenging issue since it requires the generation of frequency components not found in the soft voice spectrum. Yet, most of the previous methods mainly focused on spectral envelope transformation.

This study focuses on the question of increasing vocal effort only, in the context of singing. The aim is to transform "soft" voice utterances into "loud" utterances. A new method for harmonic enrichment of the voice spectrum and a model for spectral envelope transformation are proposed. The system is detailed in section 2 and evaluated in section 3. Discussions and conclusions are given in the last section.

2. Vocal effort modification

2.1. Signal model

Linear acoustic theory describes the speech signal s according to a source-filter model, where the glottal air flow, its resonances in the vocal tract, and the sound radiation at the lips are independent linear filters of frequency responses G, V and L, respectively. The source is the sum of an impulse train of frequency F_0 for voice sounds, and a noise component R for unvoiced sounds. Different filters for the glottal flow model are applied on the voiced and unvoiced components (G_u and G_r, respectively). A spectral description of acoustic properties of vocal effort is adopted in this paper as it is tightly linked to human perception:

$$\begin{aligned} S(f) &= \left[\sum_{k=-\infty}^{\infty} \delta(f - kF_0)\right] G_u(f)V(f)L(f) \\ &+ R(f)G_r(f)V(f)L(f) \end{aligned} \quad (1)$$

Each term of this decomposition contributes to the perception of vocal effort, and is addressed in our system.

2.2. Source modification

An increase of vocal effort is mainly caused by a more abrupt closure of the vocal folds, leading to sharper peaks of minimum amplitude in the glottal flow derivative [2]. Sharper peaks in the time domain correspond to more high harmonics in the spectrum. Therefore, the source periodic component, which reflects the amount of vocal fold vibration, is more prominent than the noise component for high vocal effort levels. Ratios between periodic and aperiodic contributions have been proved significant for vocal effort classification [10]. Increasing vocal effort requires generating higher harmonics. We propose a method for harmonic enrichment by using signal time distortion.

http://dx.doi.org/10.21437/Interspeech.2016-1096

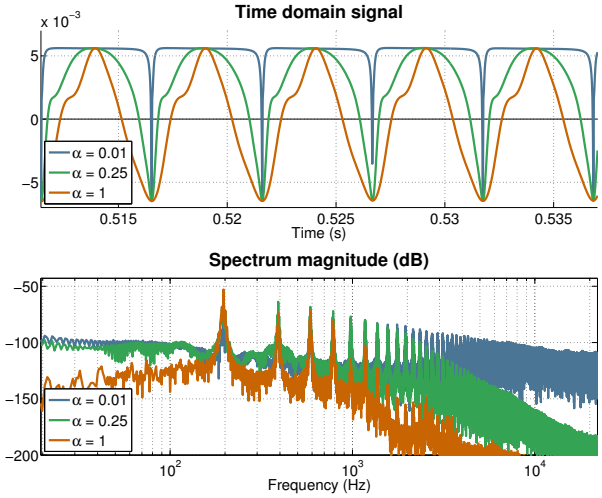

Figure 1: *Effect of distortion on a sung /a/ vowel with three coefficients: $\alpha = 1$ (no distortion), $\alpha = 0.25$, and $\alpha = 0.01$.*

2.2.1. Source estimation

Harmonic enrichment consists in giving more weight to the periodic source component, i.e., first term of equation 1. For this sake, a rough estimation of the source is carried out by filtering the initial low effort singing signal s_{lVE} with an IIR 2nd order bandpass filter h_{BP} with cutting frequencies of $0.5F_0$ and $1.2F_0$, to keep mainly its first harmonic.

$$s_{source}(t) = s_{lVE}(t) * h_{BP}(t) \quad (2)$$

This process strongly attenuates both the noise component and the filters contributions, while keeping at the same time the characteristics of a voice signal.

2.2.2. Distortion

To simulate the rising abruptness of vocal folds closure, the estimated source signal is contracted around each period's peak of minimum amplitude. For this sake, a time warping procedure is employed with the "square" warping function, commonly used in music to create a distortion effect like that of an overdriven guitar amplifier [11], and defined on the interval $[t_i, t_f]$ as:

$$g_{dist}(t,\alpha) = \left(\frac{\frac{t-t_i}{t_f-t_i}}{\alpha + (1-\alpha)|\frac{t-t_i}{t_f-t_i}|} \right)(t_f - t_i) + t_i \quad (3)$$

α is the distortion coefficient. No distortion is obtained for $\alpha = 1$ whereas maximum distortion is achieved for $\alpha = 0$. In this case, the output signal is the scaled sign of the input signal.

A time warping distortion with $\alpha = 0.1$ is chosen, giving a good compromise between generation of high frequency harmonics without amplifying too much background noise. A pitch-synchronous peak detection is implemented to find the time instants t_n of each period's peak of minimum amplitude. The distorted signal s_{dist} calculated for the n^{th} period of s_{source} is expressed as:

$$s_{dist}(t) = s_{source}\left[g_{dist}(t,\alpha)\right], t \in [t_i, t_f] \quad (4)$$

where $[t_i, t_f] = \left[\frac{t_{n-1}+t_n}{2}; \frac{t_n+t_{n+1}}{2}\right]$ for each n. Figure 1 displays examples of distortion of the estimated source of a sung /a/ with different coefficients, and their corresponding spectra.

2.2.3. Source-filter reconstruction

The signal obtained after time distortion contains new harmonics but its spectral envelope does not match with the initial soft signal. Therefore, to reintroduce the filter contribution, the original signal's spectral envelope is extracted and applied on the distorted signal. For this sake, the spectrum is decomposed in periodic and aperiodic components. A periodic component is defined as a frequency band with a width of $F_0/2$ located around a multiple of F_0. The RMS value of each periodic component is computed and interpolated for each frequency to give a spectral envelope. The equalized spectrum S_{EQ} of the distorted signal is then expressed as:

$$S_{EQ}(f) = S_{dist}(f)\frac{E_{lVE}(f)}{E_{dist}(f)} \quad (5)$$

where S_{lVE} and S_{dist} are the Fourier transforms of s_{lVE} and s_{dist}, and E_{lVE} and E_{dist} are the spectral envelopes of S_{lVE} and S_{dist}, respectively.

2.2.4. Harmonic enrichment

To minimize artifacts that might be caused by distortion, the new generated harmonics are introduced into the original signal only where they are missing, by mixing the original and the distorted signals' spectra. For this sake, a mixing window is designed, whose values equal to one in a frequency band $[f_{min}, f_{max}]$ and zero elsewhere. Transients at f_{min} and f_{max} are half Hanning windows with a length of 1000 Hz.

Harmonics are detected in the initial signal S_{lVE} if the RMS ratio between periodic and its adjacent aperiodic components are higher than 12 dB. f_{min} is defined as the frequency after which harmonics are no longer detected. f_{max} is set to 10 kHz. Then, the spectra of the original and distorted signal are mixed within this band:

$$S_{mix}(f) = \beta W(f) S_{EQ}(f) + [1 - \beta W(f)] S_{lVE}(f) \quad (6)$$

$\beta \in [0, 1]$ is the mixing coefficient and allows to choose the periodic / aperiodic ratio of the mixed signal.

2.3. Filter modification

2.3.1. Spectral tilt

The combined spectral contributions of the source and sound radiation at the lips can simply be modeled by a second order bandpass filtered called glottal formant, approximately located between F_0 and $2F_0$, and a first or second order low-pass filter with a cutting frequency beyond 1-2 kHz, leading to a spectral tilt of -40 dB/decade in high frequencies [2].

Changes of spectral tilt are considered here, as they significantly contribute to vocal effort perception: a higher vocal effort leads to a decrease of spectral tilt, allowing higher frequencies in the signal. For this sake, a γ coefficient in dB/decade is chosen to compute a gain in dB to be added for each frequency as

$$\begin{cases} G_{slope}(f) = & \gamma \log_{10}(f/F_0) & \text{for } f \in [F_0, f_{maxslope}] \\ G_{slope}(f) = & 0 & \text{elsewhere} \end{cases} \quad (7)$$

To avoid the amplification of high frequency background noise, a maximum frequency $f_{maxslope}$ is set, beyond which the spectral tilt variation is not applied. This limit is calculated by default as 3 kHz after the position of the 5th vocal tract resonance. Finally, the spectral slope is modified in the signal by:

$$S_{slope}(f) = S_{mix}(f) + G_{slope}(f) \quad (8)$$

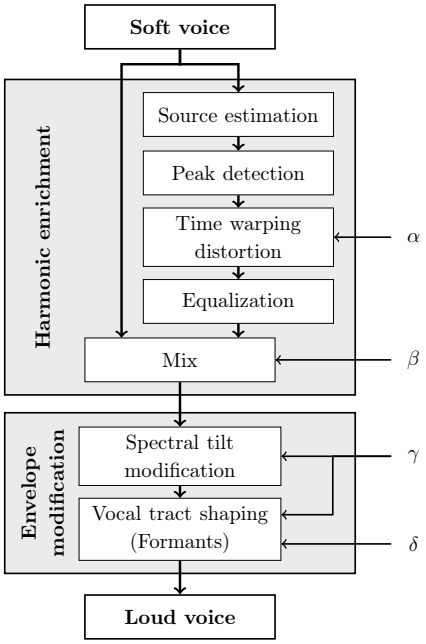

Figure 2: *Algorithm for vocal effort modification.*

2.3.2. Formants

With the increase of spectral tilt for higher vocal effort levels, the vocalic formant amplitudes naturally increase [12]. Nevertheless, if the initial soft voice has few harmonics, vocalic formants can be little prominent, or even nonexistent. In this case, the decrease of spectral tilt alone would amplify high frequency harmonics that are not formant-filtered, and vowel intelligibility would be degraded. To preserve vowel perception, 5 formants are added to the synthesized signal. Their positions $F_i, i \in [1, 5]$ are extracted from the initial signal S_{lVE} after source-filter decomposition with the Iterative Adaptive Inverse Filtering (IAIF) method [13]. Their amplitudes $A_i, i \in [1, 5]$ are defined as the gain provided by the new spectral slope at the formant positions. An additional gain δ in dB can be added if necessary:

$$A_i = G_{slope}(F_i) + \delta = \gamma \log_{10}(F_i/F_0) + \delta \qquad (9)$$

Moreover, vocal effort increase is physiologically linked to a wider mouth opening, strongly correlated to the position of the first vocalic formant. An increase of the first formant position with vocal effort has been demonstrated in several studies, from approximately 3.5 Hz/dB [12] to 10 Hz/dB [14]. Additionally, increases of vocal effort also augment the frequency of the glottal formant [15]. Therefore, both increases of glottal and first vocalic formants are modeled by the addition of 10 Hz/dB to the position of the first formant H_1.

Finally, the signal with decreased spectral slope is passed through 5 parallel formants filters, modeled as 2-poles 2-zeros digital resonator filters of transfer function $H_i, i \in [1, 5]$, and all the filtered signal are summed:

$$S_{final}(f) = \left[1 + \sum_{i=1}^{5} H_i(f)\right] S_{slope}(f) \qquad (10)$$

To conclude, Figure 2 summaries the system's algorithm.

3. Experiment

To assess the performance of our system, we seek to evaluate the contribution of, on one hand, harmonic enrichment, and on the other hand, spectral envelope modification, on vocal effort perception.

3.1. Corpus

3.1.1. Natural voice

Voice transformations were realized on a corpus recorded by two professional singers (male - baritone, and female - soprano) for the design of a concatenative singing synthesis system (http://chanter.limsi.fr). Sounds were recorded with a sample rate of 44100 Hz and a quantification of 32 bits. Three vowels were selected for this experiment: /a/, /i/ and /u/. Each vowel was sung at three pitch levels by the female singer: B3 ($F_0 = 247$ Hz), F4 ($F_0 = 349$ Hz) and C5 ($F_0 = 523$ Hz), and was sung twice at one pitch level by the male singer: G3 ($F_0 = 196$ Hz). Two vocal effort levels were selected for each vowel and note: *pianissimo* and *fortissimo*, which are the musical terms used for extreme low and extreme high vocal effort in singing, and given as instructions during the database recording. In total, 15 pairs of vocal stimuli (with low and high vocal effort) were used with a vowel factor (3 levels), and a note factor (4 levels).

3.1.2. Vocal effort modification

For each low/high vocal effort pair of our corpus, we aimed at increasing the vocal effort of the low effort stimuli. Four transformations were conducted: by spectral envelope modeling with and without harmonic enrichment; by spectral envelope morphing with and without harmonic enrichment.

- Harmonic enrichment followed the method presented in section 2.2. The distortion coefficient was kept constant: $\alpha = 0.1$. Then, the mixing coefficient was $\beta = 1$ for conditions with harmonic enrichment and $\beta = 0$ for conditions without.
- Spectral envelope modeling was made with an increase of the spectral slope, an amplification of the formants and a translation of the first formant, as presented above. We systematically chose a spectral slope coefficient $\gamma = 10$ dB/decade and no additional gain for formant amplification ($\delta = 0$ dB).
- For spectral envelope morphing, the high and low effort signal's spectral envelopes E_{hVE} and E_{lVE} were extracted with the procedure presented in section 2.2.3. Then, the high effort envelope was applied to the mixed signal by:

$$S_{morph}(f) = S_{mix}(f) \frac{E_{hVE}(f)}{E_{lVE}(f)} \qquad (11)$$

Overall, four synthesized stimuli were generated for each pair of natural signals, giving a total of 90 stimuli. Finally, all stimuli were RMS normalized to have the same level of loudness. Then, the stimuli only differed in timbre, i.e. spectral characteristics.

3.2. Protocol

A mean opinion score (MOS) paradigm was adopted to assess the overall perception of vocal effort of our stimuli. The subject's task consisted in listening to audio recordings of the 90

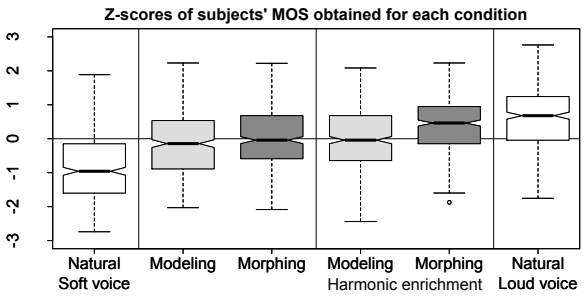

Figure 3: *Z-scores of subject's perception for each condition.*

stimuli presented above, and rating their perceived vocal effort on a scale from 1 (soft) to 5 (loud). The stimuli were presented in random order through a Beyerdynamic DTX900 headset. The experiment took place in an acoustically insulated and treated room designed for perceptual experiments.

In total, 25 subjects (17 males, 8 females, average 21 years old) participated in the experiment. All had musical experience (average 11 years). Before beginning the experiment, all subjects were instructed of the task and listened to six low/high effort pairs of natural voice extracted from the database. These stimuli presented different vowels and pitch levels than the one used for the experiment. Each subject required approximately 10 min to complete the test.

3.3. Results

Z-scores were computed for each subject's MOS to remove their influence on the results. The latter were analyzed through an analysis of variance with the Type of stimuli (6 levels: 2 natural and 4 synthesized signals), the Vowel (3 levels), and the Pitch level (4 levels) as fixed factors. Table 1 gives the analysis results. Each factor has a significant influence on subject's Z-scores. Nevertheless, the Type of stimuli and the Pitch level have major explicative powers (η^2 = 0.29 and η^2 = 0.24, respectively). An interaction between Vowel and Pitch level is also observed. Each factor influence was tested under a post-hoc HSD-Tukey test.

Table 1: *Analysis of the variance explained by each significant factor and their two-ways interactions on the subjects' Z-scores. Results report the F-statistics for the factor's degrees of freedom (df), the associated p level and the effect size (η^2).*

Factor	df	F	p	η^2
Type	5	175.6	$< 10^{-3}$	0.29
Vowel	2	64.8	$< 10^{-3}$	0.06
Pitch	3	231.6	$< 10^{-3}$	0.24
Vowel:Type	10	3.68	$< 10^{-3}$	0.02
Vowel:Pitch	6	26.9	$< 10^{-3}$	0.07

Effects of Type on subjects' Z-scores are depicted in Figure 3 for the natural signals (left: soft voice; right: loud voice) and the four transformations (second and third boxes: modeling and morphing of spectral envelope without harmonic enrichment; fourth and fifth boxes: modeling and morphing of spectral envelope with harmonic enrichment). Each box contains the second and third quartiles of the values and the thick lines represent the medians. Firstly, natural soft (resp. loud) voice signals were judged with lower (resp. higher) vocal effort than every other signal. Then, significant influence of the spectral envelope modification emerges, as the morphing method gives stimuli perceived with higher effort than the modeling method. Finally, the influence of harmonic enrichment is significant, as stimuli with harmonic enrichment are perceived with higher effort than stimuli without.

Secondly, results indicate a significant perception of higher vocal effort for the highest pitch (C5) and perception of lower vocal effort for the lowest pitch (G3). Additionally, stimuli with /i/ vowel were perceived with higher effort, mainly caused by the presence of higher frequencies in /i/ than /a/ or /u/. This leads to the Type and Vowel interaction, where the influence of Type was less pronounced for /i/ vowels than others. Finally, the Vowel and Pitch interaction is explained by the absence of vowel influence for the highest pitch level, as soprano singers tend to adjust their formants around harmonics for higher pitch for a better sound production, at the expense of vowel intelligibility [16].

4. Discussion and conclusions

We implemented and evaluated a method for vocal effort increase with two aspects: harmonic enrichment and modification of the spectral envelope.

Spectral envelope modeling was proved efficient, as it perceptively increased the vocal effort of soft voice signals. However, the effort was not perceived as high as with spectral envelope morphing for two main reasons. First, we chose not to adapt the model to the target signal (high effort signal), and similar gains were added to the spectral tilt and formants for every stimulus. Therefore, the rate of vocal effort increase might have been underestimated for some stimuli. Second, while the full spectral envelope of the high effort signal was applied to the low effort voice with the morphing method, our model focused on the spectral slope, the formant gains and the first formant position. This proves that our model does not explain all spectral features of vocal effort modification. For instance, in the particular case of lyric singing, it has been shown that singers tend to cluster their 3rd to 5th formants to produce what is called the singer's formant [17]. This resonance is typically located around 3 kHz but is strongly singer dependent. An alternative to the previous 5 first formants amplification is the addition of a single singer's formant. Moreover, a reinforcement of higher frequency formants should be considered.

Harmonic enrichment was shown significant with both spectral envelope transformations. The addition of harmonics in the signal with morphed envelope was perceived with an effort close to the natural loud voice. As the spectral envelopes are similar in both signals, this means the generation of harmonics, i.e., the periodic/aperiodic ratio is essential in the perception of vocal effort.

To conclude, the combination of harmonic enrichment and spectral envelope modification of a soft voice signal leads to a high quality transformation of vocal effort. Future developments will focus on the model, to quantify the influence of each spectral feature on vocal effort perception.

5. Acknowledgements

This work was supported by the ANR-ChaNTeR project, under grant ANR-13-CORD-0011.

6. References

[1] M. Schroder and M. Grice, "Expressing vocal effort in concatenative synthesis," in *Proceedings of the International Conferenct of Phonetic Sciences (ICPhS)*, Barcelona, Spain, 2003, pp. 2589–2592.

[2] B. Doval, C. d'Alessandro, and N. Henrich, "The spectrum of glottal flow models," *Acta Acustica*, vol. 92, no. 6, pp. 1026–1046, 2006. [Online]. Available: https://hal.archives-ouvertes.fr/hal-00368131

[3] G. Seshardi and B. Yegnanarayna, "Perceived loudness of speech based on the characteristics of glottal excitation source," *Acoustical Society of America*, vol. 4, pp. 2061–2071, 2009.

[4] C. Harwardt, "Comparing the impact of raised vocal effort on various spectral parameters," in *Proc. of Interspeech*, Florence, Italy, August 28-31 2011, pp. 2941–2944.

[5] J.-S. Liénard and C. Barras, "Fine-grain voice strength estimation from vowel spectral cues," in *Proceedings of Interspeech*, Lyon, France, August 25-29 2013.

[6] O. Turk, M. Schroder, B. Bozjurt, and L. M. Arslan, "Voice quality interpolation for emotional text-to-speech synthesis," in *Proc. of Interspeech*, Lisbon, Portugal, September 4-8 2005, pp. 797–800.

[7] À. C. Defez, J. Claudi, S. Carrié, and R. A. J. Clark, "Parametric model for vocal effort interpolation with harmonics plus noise models," in *ISCA Speech Synthesis Workshop*, Barcelona, Spain, Augist 31 - September 2 2013.

[8] C. d'Alessandro and B. Doval, "Experiments in voice quality modification of natural speech signals: the spectral approach," in *3rd ESCA International Workshop on Speech Synthesis*, Australia, November 26-29 1998, pp. 277–282.

[9] ——, "Voice quality modification for emotional speech synthesis," in *Proceedings of Eurospeech*, Geneva, Switzerland, 2003, pp. 1653–1656.

[10] N. Obin, "Cries and whispers: Classification of vocal effort in expressive speech," in *Proc. of Interspeech*, Portland, Oregon, USA, September 9-13 2012, pp. 2234–2237.

[11] "Music-dsp source code archive." [Online]. Available: http://www.musicdsp.org/

[12] 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. 411–422, 1999.

[13] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Communication*, vol. 11, pp. 109–118, 1992.

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

[15] N. Henrich, C. d'Alessandro, and B. Doval, "Spectral correlates of voice open quotient and glottal flow asymmetry: Theory, limits and experimental data," in *Proceedings of Eurospeech*, Aalborg, Denmark, September 3-7 2001.

[16] N. Henrich, J. Smith, and J. Wolfe, "Vocal tract resonances in singing: Strategies used by sopranos, altos, tenors, and baritones," *Acoustical Society of America*, vol. 129, no. 2, pp. 1024–1035, 2011.

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