The Role of Glottal Source Parameters for High-Quality Transformation of Perceptual Age

Xavier Favory, Nicolas Obin, Gilles Degottex, Axel Roebel

THE ROLE OF GLOTTAL SOURCE PARAMETERS FOR HIGH-QUALITY TRANSFORMATION OF PERCEPTUAL AGE

Xavier Favory, *Nicolas Obin*, *Gilles Degottex*, *Axel Roebel*

IRCAM - UMR STMS IRCAM-CNRS-UPMC
Paris, France

ABSTRACT

The intuitive control of voice transformation (e.g., age/sex, emotions) is useful to extend the expressive repertoire of a voice. This paper explores the role of glottal source parameters for the control of voice transformation. First, the SVLN speech synthesizer (Separation of the Vocal-tract with the Liljencrants-fant model plus Noise) is used to represent the glottal source parameters (and thus, voice quality) during speech analysis and synthesis. Then, a simple statistical method is presented to control speech parameters during voice transformation : a GMM is used to model the speech parameters of a voice, and regressions are then used to adapt the GMMs statistics (mean and variance) to a control parameter (e.g., age/sex, emotions). A subjective experiment conducted on the control of perceptual age proves the importance of the glottal source parameters for the control of voice transformation, and shows the efficiency of the statistical model to control voice parameters while preserving a high-quality of the voice transformation.

Index Terms : voice transformation, statistical modelling, glottal source and vocal tract, perceptual age.

1. INTRODUCTION

In a time where speech synthesizers can perform high-quality transformation of a speech signal (e.g., STRAIGHT [1], SVLN [2]), the intuitive control of the voice transformation is highly desirable for extending the expressive repertoire of a voice [3] (typically, to create avatars in video-games and movies). To do so, statistical methods are used to model the speech parameters in order to control the voice identity [4], emotion [5], and "voice quality" [6, 7] of a speech recording. Among the number of linguistic expressions that are used to describe a voice [8] : sex ($\pm$ male/female) and age ($\pm$ young/old) are the most common expressions. In particular, recent research has presented statistical methods to control the age of a voice : from speech [9, 6, 10] to singing voice [11, 12]. This paper focuses on the control of perceptual age in voice transformation.

The evolution of speech through age affects all of the speech characteristics : from timbre, to voice quality, and prosody. First, this affects the vocal tract characteristics : the position of the first formant F_1 tends to decrease with age [13]. Second, this also affects the glottal source characteristics : the fundamental frequency (F_0) [14, 15, 16] and the signal-to-noise ratio (SNR) [17] decrease with age. Also, the glottal source become more irregular with age : the period-to-period variations in frequency (F_0 jitter [18, 19]) and amplitude (F_0 shimmer [18, 19, 20]) of the speech signal tends to increase. In other words, the voice quality changes through age : the voice becomes more noisy (breathiness), and more irregular (creakiness) with age. Finally, speech prosody (F_0 dynamics and speech rate) is also affected through age [21].

Statistical methods based on regressions and one-to-many voice conversion (VC) techniques have been proposed to control the age of a voice. One-to-many voice conversion is used to construct joint acoustical models (GMMs) of source/target speakers from a large database of speakers [22, 6]. Then, multiple-regressive GMM (MR-GMM) is used to adapt the mean vectors of a speaker to a target control parameter [6, 7]. VC systems have been successfully used to control the age of a voice [10, 11, 12]. However, VC systems suffer from several limitations : first, the degradation of the transformed speech signal is a main issue of VC systems [23, 22] ; second, speaker's identity is not preserved during transformation [21] ; third, the glottal source remains largely ignored in voice transformation. Last but not least, VC systems generally require parallel speech databases, which constitutes a serious constraint for real-world applications.

This paper presents a simple statistical method for the high-quality control of voice transformation, and explores the role of glottal source characteristics for the control of voice transformation. The statistical modelling is based on the GMM modelling of the acoustic parameters of a voice, and regressions are used to adapt the GMMs statistics to a control parameter (e.g., age/sex, emotions). The main contributions of this paper are : the transformation is only based on the adaptation of global GMM statistics (mean and variance) of a voice, which guarantees high-quality voice transformation. The statistical modelling is not based on joint source/target speakers, which preserves the source speaker's identity during transformation, and does not require parallel speech databases. Also, a representation of the glottal source which covers tension, breathiness, and creakiness, is used for statistical modelling and transformation of the voice. The role of glottal source and the efficiency of the statistical modelling are investigated in a subjective experiment on the control of perceptual age in voice transformation.

2. SPEECH ANALYSIS AND SYNTHESIS

STRAIGHT [1] is the most popular speech synthesizer widely used for voice conversion [22]. Recent generations of speech synthesizers include an explicit representation of the glottal source (e.g., LF model [24, 2]) in the speech synthesizer, which substantially improves the control of the voice quality during speech synthesis. In particular, the SVLN speech synthesizer (Separation of the Vocal-tract with the Liljencrants-fant model plus Noise) [2] allows the intuitive control of the glottal source (and thus, voice quality) during speech synthesis [25], with a limited number of parameters. This section summarizes the main principles of the SVLN speech synthesizer.

2.1. Analysis of SVLN parameters

The voiced segments of the speech signal are assumed to be stationary and periodic in a short analysis window (of 3.5 periods). Using the source-filter model in the frequency domain, the voice production model of an observed speech spectrum $S(\omega)$ can be described as follows (Figure 1) :

$$S(\omega) = \left[H^{F_0}(\omega) \cdot G^{Rd}(\omega) + N^{\sigma_g}(\omega)\right] \cdot C^{\bar{c}c}(\omega) \cdot L(\omega) \quad (1)$$

where :

- $H^{F_0}(\omega)$ is the harmonic structure modeling a periodic impulse train of fundamental frequency F_0 : $H^{F_0}(\omega) = \sum_{k\in\mathbb{Z}} e^{j\omega k/F_0}$
- $G^{Rd}(\omega)$ is the shape of the deterministic component of the glottal source, the LF model parametrized by Rd and E_e, its shape and amplitude parameters respectively [26].
- $N^{\sigma_g}(\omega)$ is the stochastic component of the glottal source. This noise is assumed to obey a Gaussian distribution of standard-deviation σ_g.
- $C^{\bar{c}c}(\omega)$ is the Vocal-Tract Filter (VTF) representing the resonances and anti-resonances of the vocal-tract. This filter is parametrized by a vector of cepstral coefficients $\bar{c}$.
- $L(\omega)$ is the filter corresponding to the radiation at the lips and nostrils level. Here, we assume $L(\omega) = j\omega$.

The estimation of the SVLN parameters is described in details in [2]. Additionally, glottal closure instants (GCI) are estimated [27], and used to measure the period-to-period regularity of the glottal pulse :

$$\sigma_{GCI}(n) = \frac{\Delta GCI(n)}{\frac{1}{2K+1}\sum_{k=n-K}^{n+K} \Delta GCI(k)} - 1 \quad (2)$$

where : $\Delta GCI(n) = |GCI(n+1) - GCI(n)|$.

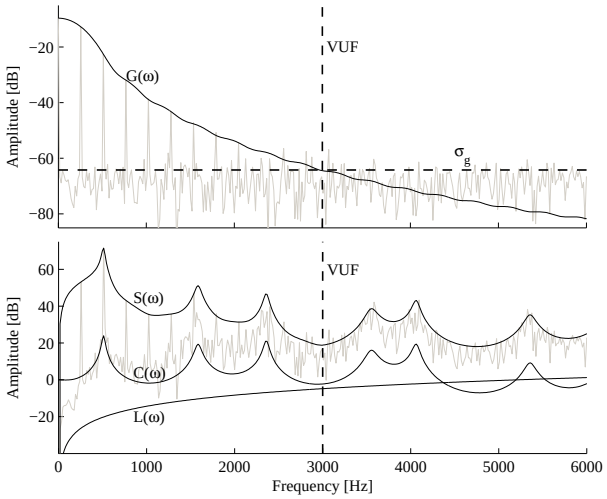

Fig. 1. On top : the glottal source model. On bottom : the voice production model. The observed spectrum is plotted in gray line (3.5 periods), and the estimated parameters are in bold lines (one period). The Voiced/Unvoiced Frequency (VUF) is defined as the frequency above which the signal is considered as unvoiced.

2.2. Synthesis from SVLN parameters

A speech utterance can be synthesized from the SVLN parameters : short segments of stationary signals are first synthesized, and then overlap-added to form the speech signal.

2.2.1. Positioning of speech segments

Temporal marks m_k of the k^{th} speech segment are first placed at intervals according to F_0 (Figure 2). Then the starting time t_k of the k^{th}-segment is defined as the opening instant of the LF model and the ending time of this segment is the starting time of the next.

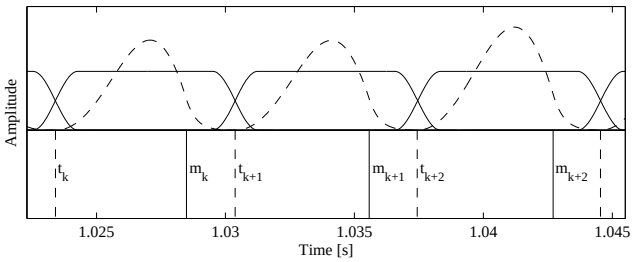

Fig. 2. Three segments : LF models are in dashed lines, and synthesis windows are in solid lines.

2.2.2. Synthesis of speech segments

For all speech segments, the synthetic speech spectrum $S_k(\omega)$ is :

$$S_k(\omega) = \left(e^{-j(\omega m_k + \varphi_k)} \cdot G^{Rd_k}(\omega) + N_k(\omega)\right) \cdot C^{\bar{c}_k}(\omega) \cdot L(\omega) \quad (3)$$

where : $e^{-j\omega m_k}$ is a delay placing the instant t_e of the LF model at the mark m_k, and $e^{-j\omega\varphi_k}$ an additional delay displacing the mark m_k according to the period-to-period regularity of the GCI.

The stochastic positioning of the glottal pulse φ_k is used to introduce period-to-period irregularities (creakiness) in the speech signal. The random delay φ_k is generated according to the Gaussian distribution σ_{GCI}, and then weighted by an activation function which depends on the F_0. A sigmoid function is used as the non-linear activation function : this activation function is centred on the mean F_0 of the speech utterance, 1 for the minimal F_0 of the speech utterance, and 0 for the maximal F_0 of the speech utterance. The use of this activation function is motivated by the fact that period-to-period irregularities generally occur in the low to extreme-low F_0 range of a speaker.

The stochastic component $N_k(\omega)$ of the glottal source is synthesized according to the Gaussian distribution $N^{\sigma_g}(\omega)$ as estimated from the SVLN analysis. This noise is then modulated synchronously with the F_0, and coloured (according to the Voiced/Unvoiced Frequency - VUF) in order to guarantee the coherence of the stochastic and the deterministic components of the speech signal. Then, the deterministic component $G^{Rd_k}(\omega)$ of the glottal pulse is synthesized, and is added to the noise segment. Finally, the vocal-tract filter $C^{\bar{c}_k}(\omega)$ and radiation filters $L(\omega)$ are applied in order to synthesize the speech segment $S_k(\omega)$ [2].

Finally, the speech signal corresponding to each speech segment is retrieved by inverse Fourier transform of $S_k(\omega)$. Then, the entire speech signal is constructed by overlap-adding all speech segments.

3. STATISTICAL MODELLING : MULTIPLE REGRESSION GMM

This section presents the statistical method used to control the evolution of speech parameters with the perceptual age of a speaker.

3.1. Training : Multiple Regression GMMs

3.1.1. GMM statistics

First, a Gaussian mixture model (GMM) is used to represent the acoustic distribution of each speech recording S of each speaker :

$$p(\boldsymbol{x}) = \sum_{i=1}^{M} \alpha_i^{(S)} \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu_i}^{(S)}, \boldsymbol{\Sigma_i}^{(S)}) \quad (4)$$

where : $\boldsymbol{x}$ is a vector of speech parameters, $\alpha_i^{(S)}$ is the weight, $\boldsymbol{\mu_i}^{(S)}$ the mean vector, and $\boldsymbol{\Sigma_i}^{(S)}$ the covariance matrix of the i-th component of the GMM for speech recording S, and M the number of GMM components.

3.1.2. Multiple regression on GMM parameters

Then, a regression is performed to control the GMM statistics with a control parameter $\boldsymbol{w_c}^{(S)}$ (here, the age of the speaker) [6] :

$$\boldsymbol{\mu}^{(S)} = \boldsymbol{w_c}^{(S)} \boldsymbol{\beta} + \boldsymbol{\epsilon} \quad (5)$$

where : $\boldsymbol{\mu}^{(S)}$ is the GMM mean vector, $\boldsymbol{\beta}$ denotes the regression vector, $\boldsymbol{w_c}^{(S)}$ is the Vandermonde matrix of the control parameter, and $\boldsymbol{\epsilon}$ is the bias vector which follows a centred normal distribution $\mathcal{N}(\boldsymbol{\mu_\epsilon} = \mathbf{0}, \boldsymbol{\sigma_\epsilon})$.

The regression parameters $(\boldsymbol{\beta}, \boldsymbol{\epsilon})$ are estimated through the maximum likelihood estimate (MLE) of the normal distribution $\mathcal{N}(\boldsymbol{w_c}\ \boldsymbol{\beta}, \boldsymbol{\sigma}_\varepsilon)$:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{W_c}^\top \boldsymbol{W_c})^{-1} \boldsymbol{W_c}^\top \boldsymbol{\mu} \quad (6)$$

$$\hat{\boldsymbol{\sigma}}_\varepsilon = \frac{1}{N_S} (\boldsymbol{\mu} - \boldsymbol{W_c}\ \hat{\boldsymbol{\beta}})^\top (\boldsymbol{\mu} - \boldsymbol{W_c}\ \hat{\boldsymbol{\beta}}) \quad (7)$$

where : $\boldsymbol{W_c} = [\boldsymbol{w_c}^{(S_1)^\top}, ..., \boldsymbol{w_c}^{(S_{N_S})^\top}]^\top$ is the matrix of control parameters, and $\boldsymbol{\mu} = [\boldsymbol{\mu}^{(S_1)^\top}, ..., \boldsymbol{\mu}^{(S_{N_S})^\top}]^\top$ is the vector of GMM mean vectors, and N_S the number of speech recordings.

Regression parameters are estimated separately for the mean vector $\boldsymbol{\mu}^{(S)}$, and the covariance matrix $\boldsymbol{\Sigma}^{(S)}$ from annotated speech database. In this paper, the chronological age of the speakers is used as the control parameter $\boldsymbol{w_c}^{(S)}$ for regression.

3.2. Generation : Adaptation of Speech Parameters

During transformation, source speech parameters $\boldsymbol{x}_t^{src}$ at time t are modified into target speech parameters $\boldsymbol{x}_t^{tgt}$ according to the control parameter $\boldsymbol{w}^{tgt}$:

$$\boldsymbol{x}_t^{tgt} = \sum_{i=1}^{M} p_i(\boldsymbol{x}_t^{src})(\boldsymbol{\mu_i}^{tgt} + \boldsymbol{\Sigma}_i^{tgt} \boldsymbol{\Sigma}_i^{src-1} (\boldsymbol{x}_t^{src} - \boldsymbol{\mu}_i^{src})) \quad (8)$$

where : $(\boldsymbol{\mu_i^{src}}, \boldsymbol{\Sigma_i^{src}})$ denotes the GMM statistics of source speech parameters, and $(\boldsymbol{\mu_i^{tgt}}, \boldsymbol{\Sigma_i^{tgt}})$ denotes the target GMM statistics obtained after regression on the control parameter $\boldsymbol{w}^{tgt}$.

Then, with $M = 1$:

$$\boldsymbol{x}_t^{tgt} = \boldsymbol{\mu}^{tgt} + \boldsymbol{\Sigma}^{tgt} \boldsymbol{\Sigma}^{src-1} (\boldsymbol{x}_t^{src} - \boldsymbol{\mu}^{src}) \quad (9)$$

This is finally a simple update of the source speech parameters statistics $(\boldsymbol{\mu}^{src}, \boldsymbol{\Sigma}^{src})$ to the target statistics $(\boldsymbol{\mu}^{tgt}, \boldsymbol{\Sigma}^{tgt})$ as obtained from the regression.

Conversely to VC systems [6], the proposed statistical modelling (from training to generation of speech parameters) does not assume any constraint about the linguistic content of speech recordings, and thus can be used with any non-parallel speech databases.

4. EXPERIMENT

4.1. Speech Database

The speech database used for the experiment consists of speech recordings of French celebrities collected from multi-media archives (radio, TV). The speech database is composed of 88 speech recordings of 13 French celebrities (6 males, 7 females) collected in the context of interviews (i.e., non-acted speech). For each speaker, an average of 7 speech recordings is used to represent the evolution of the voice through age. The chronological age of the speakers ranges from 14 to 80 years, and dates of the speech recording range from 1958 to 2012. Speech recordings are encoded into 12 kHz / 16 bits audio format - where 12 kHz is the minimum sampling rate of all speech recordings.

4.2. Experimental Setups

The short-term SVLN parameters are extracted for each speech recording of the speech database. This comprises : the fundamental frequency F_0, the spectral envelope of the vocal-tract filter $\bar{cc}$, the shape of the glottal source Rd, the noise of the glottal source N^{σ_g}, and the period-to-period regularity of the glottal closure instants σ_{GCI}. Statistical models are estimated separately for each speech parameter : a single Gaussian is used with diagonal covariance matrix, and regression is performed on Gaussian statistics (mean vector and diagonal covariance matrix) with 1-st to 3-rd order polynoms.

4.3. Subjective Experiment : Estimation of Perceptual Age

A subjective experiment was conducted to investigate the role of glottal source parameters on the identification of the perceptual age of a voice. For this purpose, 10 speakers (5 males / 5 females) were selected from the TIMIT read speech database of American-English [28], each reading the same sentence ("*She had your dark suit in greasy wash water all year.*"). For each speaker, the chronological age was directly retrieved from the TIMIT meta-information, and then the speech recording was transformed to 7 target ages : 15, 25, 35, 45, 55, 65, and 75 - which ranges from teenager (15) to very old (75). The speech recording was transformed according to 5 combinations of speech parameters : 1) baseline vocal tract / glottal source parameters : F_0 and spectral envelope (F_0, $\bar{cc}$); 2) the baseline parameters plus addition of Rd (tension), 3) N^{σ_g} (breathiness), 4) σ_{GCI} (creakiness) glottal source parameters, and the complete set of speech parameters (F_0, $\bar{cc}$, Rd, N^{σ_g}, σ_{GCI}). This constitutes a total of 36 speech recordings for each speaker (1 original speech recording, and 35 transformed speech recordings).

For the subjective experiment, a speaker was randomly selected and the 36 speech recordings of a speaker, and then randomly presented to the participant. For each speech recording, the participant

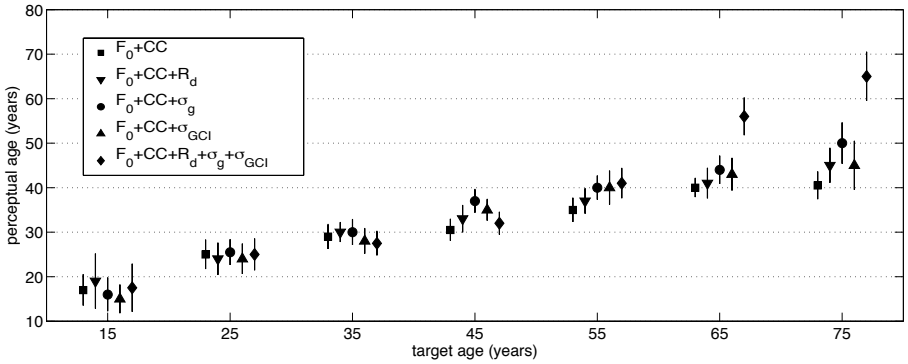

Fig. 3. Perceptual age versus target age : mean and 95% confidence interval of the perceptual age.

was asked : 1) to identify the perceptual age of the speaker (from 0 to 100 years), and 2) to rate the quality of the speech recording on a 5-degrees mean opinion score scale (MOS) (from 1 = bad, to 5 = perfect) [29]. The experiment was conducted on-line with headphones, with 40 participants.

4.4. Results

Figure 3 presents the statistics of perceptual age (mean and 95% confidence interval) for each target age, and for each combination of speech parameters. This figure confirms the role of glottal source parameters to control of perceptual age of a voice. The individual contribution of N^{σ_g} and σ_{GCI} slightly improves the control of the perceptual age of old voices (target age = 55-75 years). Moreover, the complete set of vocal tract and glottal source parameters significantly improves the control of the perceptual age of old voices (target age = 65-75 years). This indicates that the coherence of the speech parameters is required to control the perceptual age of a voice. These observations are confirmed by the Pearson correlation calculated between the target age and the perceptual age (Table 1) : the correlation increases with the addition of glottal source parameters (N^{σ_g}, σ_{GCI}), and the highest correlation is obtained for the complete set of speech parameters. The correlation score obtained (r=0.72) is comparable to those obtained in the literature for the age estimation of adult speakers (from r=0.68 to r=0.88, see [30] for a review).

PARAMETERS	PEARSON CORRELATION
(F0, CC)	0.63
(F0, CC) + Rd	0.62
(F0, CC) + σ_g	0.70
(F0, CC) + σ_{GCI}	0.69
(F0, CC) + Rd + σ_g + σ_{GCI}	0.72

Table 1. Pearson correlation of target age and perceptual age.

Figure 4 presents the MOS statistics (mean and 95% confidence interval) obtained for each combination of speech parameters. The original speech recordings used as a reference obtained a 4.60 score in average. The transformation of the standard speech parameters (F_0 and $\bar{cc}$) reaches a good sound quality (MOS = 3.7). Then, the sound quality drops slightly with the transformation of the glottal source parameters (MOS=3.2-3.6), which is due to the degradation caused by the modification of glottal source parameters (Rd, N^{σ_g}, σ_{GCI}) in the SVLN speech synthesizer. Nevertheless, the sound quality obtained for the complete set of speech parameters remains comparable to the sound quality one that can be obtained with current speech synthesizers (MOS= 3.5-4, see [2] for analysis/resynthesis with current speech synthesizers), and significantly surpasses the sound quality obtained with standard VC systems (MOS = 2-3, see [23, 22]). This constitutes promising results towards high-quality voice transformation.

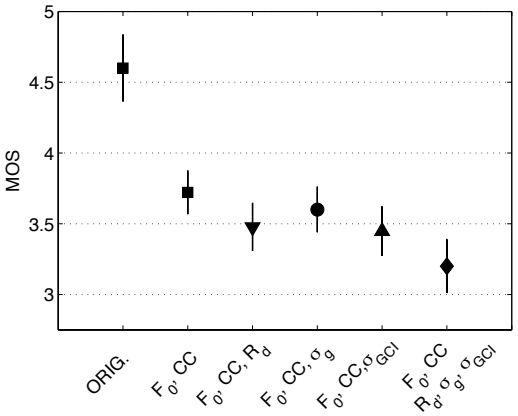

Fig. 4. Mean Opinion Score : mean and 95% confidence interval.

5. CONCLUSION

This paper explored the role of glottal source parameters for the control of perceptual age in voice transformation. A subjective experiment conducted on the estimation of perceptual age proved the importance of the glottal source for the control of voice transformation, and showed the efficiency of the statistical modelling to control the voice parameters while preserving a high-quality of the voice transformation. Further research will face the main limitations of voice transformation : the accurate control of the voice transformation, and the high-quality transformation of the voice quality. The first limitation is the accurate control of the voice transformation : advanced statistical modelling and the integration of speech prosody constitute the main directions for the control of voice transformation. The second limitation is the modification of the voice quality (e.g., tension, breathiness, creakiness) in speech synthesizers, in order to render the sound quality of synthetic speech comparable to that of natural speech.

6. REFERENCES

[1] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigné, "Restructuring Speech Representations using a Pitch-Adaptative Time-Frequency Smoothing and an Instantaneous-Frequency-based F0 Extraction : Possible Role of a Repetitive Structure in Sounds," *Speech Communication*, vol. 27, pp. 187–207, 1999.

[2] G. Degottex, P. Lanchantin, A. Roebel, and X. Rodet, "Mixed Source Model and its Adapted Vocal Tract Filter Estimate for Voice Transformation and Synthesis," *Speech Communication*, vol. 55, no. 2, pp. 278–294, 2013.

[3] S. Farner, A. Roebel, and X. Rodet, "Natural Transformation of Type and Nature of the Voice for Extending Vocal Repertoire in High-Fidelity Applications," in *International Acoustic Engineering Society Conference (AES)*, London, UK, 2009.

[4] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous Probabilistic Transform for Voice Conversion," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, 1998.

[5] C. Veaux and X. Rodet, "Intonation Conversion from Neutral to Expressive Speech," in *Interspeech*, Florence, Italy, 2011, pp. 2765–2768.

[6] K. Ohta, Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Regression Approaches to Voice Quality Control Based on One-to-Many Eigenvoice Conversion," in *ISCA Workshop on Speech Synthesis (SSW)*, Bonn, Germany, 2007, pp. 101–106.

[7] K. Ohta, T. Toda, Y. Ohtani, H. Saruwatari, and K. Shikano, "Adaptive Voice-Quality Control based on One-to-Many Eigenvoice Conversion," in *Interspeech*, Makuhari, Japan, 2010, p. 2158–2161.

[8] H. Kido and H. Kasuya, "Everyday Expressions associated with Voice Quality of Normal Utterance — Extraction by Perceptual Evaluation," *Journal of the Acoustic Society of Japan*, vol. 57, no. 5, pp. 337–344, 2001.

[9] M. Tachibana, T. Nose, J. Yamagishi, , and T. Kobayashi, "A Technique for Controlling Voice Quality of Synthetic Speech using Multiple Regression HSMM," in *Interspeech*, Pittsburgh, USA, 2006, pp. 2438–2441.

[10] K. Yamamoto, T. Toda, H. Doi, H. Saruwatari, and K. Shikano, "Statistical Approach to Voice Quality Control in Esophageal Speech Enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, 2012, pp. 4497–4500.

[11] K. Kobayashi, H. Doi, T. Toda, T. Nakano, M. Goto, G. Neubig, S. Sakti, and S. Nakamura, "An Investigation of Acoustic Features for Singing Voice Conversion based on Perceptual Age," in *Interspeech*, Lyon, France, 2013, pp. 1057–1061.

[12] K. Kobayashi, T. Toda, T. Nakano, M. Goto, G. Neubig, S. Sakti, and S. Nakamura, "Regression Approaches to Perceptual Age Control in Singing Voice Conversion," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, 2014, pp. 7954–7958.

[13] S. E. Linville and J. Rens, "Vocal Tract Resonance Analysis of Aging Voice using Long-Term Average Spectra," *Journal of Voice*, vol. 15, no. 3, pp. 323–330, 2001.

[14] H. Hollien and T. Shipp, "Speaking Fundamental Frequency and Chronological Age in Males," *Journal of Speech and Hearing Research*, vol. 15, pp. 155–159, 1972.

[15] A. Russell, L. Penny, and C. Pemberton, "Speaking Fundamental Frequency Changes over Time in Women : A Longitudinal Study," *Journal of Speech and Hearing Research*, vol. 38, pp. 101–109, 1995.

[16] R. J. Baken, "The Aged Voice : a New Hypothesis," *Journal of Voice*, vol. 19, no. 3, pp. 317–325, 2005.

[17] C. T. Ferrand, "Harmonics-to-Noise Ratio : An Index of Vocal Aging," *Journal of Voice*, vol. 16, pp. 480–487, 2002.

[18] L. Ramig and R. Ringel, "Effects of Physiological Aging on Selected Acoustic Characteristics of Voice," *Journal of Speech and Hearing Research*, vol. 26, pp. 22–30, 1983.

[19] S. Linville, *Vocal Aging*. Singular Thomson Learning, 2001.

[20] S. A. Xue and D. Deliyski, "Effects of Aging on Selected Acoustic Voice Parameters : Preliminary Normative Data and Educational Implications," *Educational Gerontology*, vol. 27, pp. 159–168, 2001.

[21] M. Pettorino, E. Pellegrino, and M. Maffia, ""Young" and "Old" Voice : the Prosodic Auto-Transplantation Technique for Speaker's Age Recognition," in *Speech Prosody*, Dublin, Ireland, 2014, pp. 135–139.

[22] T. Toda, Y. Ohtani, and K. Shikano, "One-to-Many and Many-to-One Voice Conversion Based on Eigenvoices," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Honolulu, Hawai, 2007, pp. 1249–1252.

[23] T. Toda, A. W. Black, and K. Tokuda, "Voice Conversion based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[24] J. Cabral, S. Renals, J. Yamagishi, and K. Richmond, "HMM-based Speech Synthesiser using the LF-model of the Glottal Source," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, 2011, p. 4704–4707.

[25] G. Degottex, A. Roebel, and X. Rodet, " Pitch Transposition and Breathiness Modification using a Glottal Source Model and its Adapted Vocal-Tract Filter ," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Prague, Czeck Republic, May 2011, pp. 5128–5131.

[26] G. Fant, "The LF-Model Revisited. Transformations and Frequency Domain Analysis," K.T.H. Quarterly Progress Report and Status Progress. Departement for Speech, Music and Hearing, Tech. Rep. 2-3, 1995.

[27] G. Degottex, A. Roebel, and X. Rodet, "Glottal Closure Instant Detection from a Glottal Shape Estimate," in *13th International Conference on Speech and Computer*, St-Petersburg, Russia, 2009, pp. 226–231.

[28] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgrena, , and V. Zue., "TIMIT Acoustic-Phonetic Continuous Speech Corpus," *Linguistic Data Consortium*, 1993.

[29] ITU-T P.800, *Methods for Subjective Determination of Transmission Quality - Series P : Telephone Transmission Quality ; Methods for Objective and Subjective Assessment of Quality*, 1996.

[30] L. Cerrato, M. Falcone, and A. Paoloni, "Subjective Age Estimation of Telephonic Voices," *Speech Communication*, vol. 31, no. 2-3, pp. 107–112, 2000.