Vocal tract length estimation based on vowels using a database consisting of 385 speakers and a database with MRI-based vocal tract shape information

Hideki Kawahara[1], *Tatsuya Kitamura*[2], *Hironori Takemoto*[3], *Ryuichi Nisimura*[1], *Toshio Irino*[1]

[1]Faculty of Systems Engineering, Wakayama University, Wakayama, Japan
[2]Faculty of Intelligence and Informatics, Konan University, Kobe, Japan
[3]National Institute of Information and Communications Technology, Kyoto, Japan

kawahara@sys.wakayama-u.ac.jp, t-kitamu@konan-u.ac.jp, takemoto@nict.go.jp
nisimura@sys.wakayama-u.ac.jp, irino@sys.wakayama-u.ac.jp

10.21437/Interspeech.2014-219

Abstract

A highly-reproducible estimation method of vocal tract length (VTL) and text independent VTL estimation method are proposed based on a Japanese vowel database spoken by 385 male and female speakers ranging from age 6 to 56 and other vowel database with MRI-based vocal tract shape information. Proposed methods are based on interference-free power spectral representation and systematic suppression of biasing factors. MRI data is used to calibrate VTL estimation result to be represented in terms of physically meaningful unit. These databases are normalized based on the estimated VTL information to provide a reference template, which is used to implement a text independent VTL estimation method. A prototype system for text independent estimation of VTL is implemented using Matlab and runs faster than realtime on a PC.

Index Terms: Speech analysis, speech synthesis, speech processing, vowels, diverse density

1. Introduction

Vocal tract length is one of the major sources of speaker variability [1, 2], which hinders automatic speech recognition (ASR). Nowadays, VTL normalization [3, 4, 5, 2] is a common practice in ASR. In speech processing, introducing VTL normalization improves processed speech quality when applied to voice morphing [6]. In addition to these engineering interest, it is important to note that it plays an important role in human speech perception [7]. VTL also provides clues for gender perception and has essential roles in non and para-linguistic aspect of speech communications [8]. Acoustic model of speech production led to VTL estimation methods based on formant frequencies [9, 10]. Unfortunately, in practice, formant frequencies are prone to biased estimation due to F0 and glottal waveform. Sometimes formants are difficult to identify.

Recent introduction of two sources of information opened an interesting possibility which was not available in the related works mentioned in the previous paragraph. The first one is a large scale Japanese vowel database with relevant physical data of speakers [11]. The other is a MRI database of vocal tract shape and length [12]. VTL information given in the database is derived based on the method proposed in the reference [13]. These lead to our proposal of a new VTL estimation method outlined in the next section. The proposed methods do not rely on formant estimation.

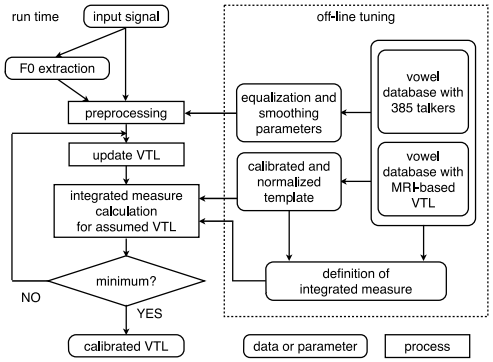

Figure 1: Schematic diagram of the proposed text independent VTL estimation system. The highly-reproducible VTL estimation is used in the off-line tuning.

2. Outline of the proposed methods

Figure 1 shows a schematic diagram of the proposed text independent VTL estimation method. Left side shows the run time system. Constituent procedures (shown as rectangular boxes) in the run time system are designed by the off-line process shown in the right side using databases.

The box named "preprocessing" is the first key-component of the proposed method. It is designed to suppress biasing factors in VTL estimation described in the next section. The preprocessing consists of power spectrum calculation, F0-adaptive spectral smoothing, equalization and additional smoothing.

The box named "integrated measure calculation for assumed VTL" is the second key-component. The integrated measure is based on *Diverse Density* [14] usually used in machine learning. It also uses a set of calibrated and normalized vowel templates based on a vowel database with 385 speakers [11]. The calibration uses the MRI-based VTL data [12].

Right side shows off-line tuning of system parameters used in these key-component procedures. Parameters in "preprocessing" are tuned using a vowel database with 385 speakers. The vowel database is also used to design a set of calibrated and normalized vowel templates used as the reference of VTL estimation. This template is the primary component in the procedure "integrated measure calculation for assumed VTL." The

integrated measure is also designed and tuned using sentence database consisting of 28 (14 male and 14 female) speakers speaking 30 sentences each [15] and the two vowel databases introduced before. For this tuning process and design process, a highly-reproducible VTL estimation method is introduced.

3. Biasing factors

Power spectra of speech sounds not solely reflect VTL information. Many other factors are contributing on the final shape of the power spectra. This section lists those factors and outlines the necessary pre-preprocessing for reducing their effects.

- **Periodicity:** Periodic excitation in voiced sounds effectively samples underlying smooth spectral shape at harmonic frequencies. This results in significant mismatch when fundamental frequencies (F0) of two samples are different. Band-unlimited nature of vocal tract transfer function [16] when sampled in the frequency domain makes this situation worse.
- **Glottal waveform and radiation:** In source filter model of speech production, the output speech spectrum of voiced sound is a product of the glottal source spectrum, vocal tract transfer function and the transfer function from lip to observation point [17]. Shape of glottal waveform significantly differs depending on speaker, effort, speaking style, emotion and so on. Also combination of glottal source and radiation yields a spectral peak around lower harmonic frequency region (glottal formant), which interferes with the first formant of some vowels. It also introduces complex zeros [18].
- **Nonlinear interaction between source and filter:** Acoustic impedance at glottis varies within one pitch period due to periodic variation of the area of the glottal opening. This modifies boundary condition of the vocal tract and modulates its transfer function. This variation of mechanical impedance also modulates vocal fold vibration itself [19].
- **Three dimensional propagation:** Branching of pyriform fossa introduces deep dips into the transfer function [20]. Transversal modes of wave propagation due to three dimensional vocal tract shape have significant impact on its transfer function in higher frequency region [21].

4. Preprocessing

The goal of spectral distance design is to make minimization of the distance measure equates relative VTL of two speech samples. This section introduces procedures for reducing effects of biasing factors described in the previous section.

4.1. Interference free power spectrum

Temporal as well as spectral variations caused by periodic excitation are effectively suppressed by two procedures introduced in TANDEM-STRAIGHT [22]. In the current application, only F0-adaptive spectral smoothing is used because VTL normalization does not need fine temporal resolution and allows the use of relatively long time windowing function. The following F0 adaptive smoothing is a relevant simple approximation of the original smooth spectrum since it is equivalent to piecewise linear interpolation when applied to harmonic line spectra.

$$P_S(f) = \frac{1}{\omega_0}\int_{-f_0}^{f_0} P(f-\lambda)h(\lambda)d\lambda, \quad (1)$$

$$\text{where } h(\lambda) = \begin{cases} \left|\frac{f_0-\lambda}{f_0}\right| & |\lambda| \le f_0 \\ 0 & \text{otherwise} \end{cases},$$

where f_0 represents the fundamental frequency and $P(f)$ represents the power spectrum as a function of frequency f.

4.2. Equalization and smoothing

Average formant frequency spacing on the frequency axis is 1000 Hz when length of the vocal tract is 17 cm. Removing spectral variation coarser than this spacing removes global effects of glottal waveform and radiation characteristics. Smoothing spectral details much finer than this spacing effectively suppresses effects of spectral zeros and formant band width differences. Two functions $h_W(\lambda)$ and $h_N(\lambda)$ are used for equalization and smoothing respectively and yield equalized power spectrum $P_E(f)$ and standardized power spectrum $P_C(f)$.

$$P_E(f) = \frac{P_S(f)}{\int_{-f_W}^{f_W} P_S(f-\lambda)h_W(\lambda)d\lambda}, \quad (2)$$

$$P_C(f) = \int_{-f_N}^{f_N} P_E(f-\lambda)h_N(\lambda)d\lambda, \quad (3)$$

where $h_X(\lambda) = 0$ for $|\lambda| > f_X$ represents positive finite support functions. (Substitute W or N for X.) In the current implementation the following form is used.

$$h_X(\lambda) = c_0\left(1+\cos\left(\frac{\pi\lambda}{f_X}\right)\right), \quad (4)$$

where c_0 is a normalization constant to make $\int h_x(\lambda)d\lambda = 1$.

4.3. Trimming frequency range and weighting

Spectral variations in lower frequency region are mainly caused by F0 differences and glottal source waveform differences. Spectral variations in higher frequency region are dominated by detailed three dimensional individual shape differences. Effects of VTL difference are dominant only in the middle frequency region. Trimming frequency range and using the middle region makes spectral distance minimization to provide the desired distance behavior.

Amount of spectral deformation by VTL modification is proportional to frequency. To equalize this effect, spectral difference between two canonical spectrum is calculated as a Euclid norm of two mean normalized log power spectra sampled at discrete frequencies located equidistance on the logarithmic frequency axis.

5. Spectral distance and parameter tuning

The spectral distance $d_F(P_C^{(A)}, P_C^{(B)}; r_{AB})$ which takes all these preprocessing into account is defined by the following equations with assumed VTL ratio $r_{AB} = l_B/l_A$ where l_X represents VTL of X.

$$d_F(P_C^{(A)}, P_C^{(B)}; r_{AB}) = \frac{1}{\#(\boldsymbol{f}_d)}\left\| L^{(A)}(\boldsymbol{f}_d) - L^{(B)}(r_{AB}\boldsymbol{f}_d)\right\|_2 \quad (5)$$

where

$$L^{(X)}(\boldsymbol{f}_d) = 10\log_{10}\left(P_C^{(X)}(\boldsymbol{f}_d)\right) - \mu^{(X)}(\boldsymbol{f}_d) \quad (6)$$

$$\mu^{(X)}(\boldsymbol{f}_d) = \frac{1}{\#(\boldsymbol{f}_d)}\sum_{f\in\boldsymbol{f}_d} 10\log_{10}\left(P_C^{(X)}(f)\right) \quad (7)$$

$$\boldsymbol{f}_d = [f_L, f_2, \ldots, f_k, \ldots, f_H]^T, \quad (8)$$

where $\boldsymbol{f}_d$ represents a vector consisting of the set of frequencies for discretization and the superscript T represents transposition. The function $\#(\boldsymbol{f}_d)$ returns the number of elements in the vector $\boldsymbol{f}_d$ and $L^{(X)}$ represents dB power spectrum with mean normalization. (Substitute A or B for X.)

5.1. Tuning by a vowel database with 385 speakers

The spectral distance with preprocessing $d_F(P_C^{(A)}, P_C^{(B)}; r_{AB})$ consists of a set of performance-determining parameters $\Theta = \{f_W, f_N, f_L, f_H\}$. [1] The set of performance-determining parameters Θ is tuned based on consistency of the estimated VTL ratios using the vowel database with 385 speakers.

Let $L_v^{(n)}(\boldsymbol{f}_d)$ represent an averaged dB power spectrum with mean normalization of the speaker ID $= n$ and vowel $v \in V$ ($V = \{$/a/, /i/, /u/, /e/, /o/$\}$). Then, the estimated VTL ratio r_{nm} of speakers ID $= n$ and ID $= m$ is defined by the following equation.

$$r_{nm} = \underset{r}{\operatorname{argmin}} \sum_{v\in V}\left|L_v^{(n)}(\boldsymbol{f}_d) - L_v^{(m)}(r\boldsymbol{f}_d)\right|^2. \quad (9)$$

Note that the sum in the right hand side is proportional to squared sum of d_F defined by Eq.5.

Assume that estimation error of $\ln(r_{nm})$ obeys normal distribution. Then, the following least square solution (Eq.10) provides relative estimates of $\ln(l_k)$.

$$\hat{\boldsymbol{l}}_{\log} = (H^TH)^{-1}H^T\boldsymbol{r}_{\log}, \quad (10)$$

where $\hat{\boldsymbol{l}}_{\log}$, $\boldsymbol{r}_{\log}$ and H are defined below.

$$\hat{\boldsymbol{l}}_{\log} = [\ln(\hat{l}_1), \ldots, \ln(\hat{l}_k), \ldots, \ln(\hat{l}_K)]^T \quad (11)$$

$$\boldsymbol{r}_{\log} = [\ln(r_{1,2}), \ldots, \ln(r_{nm}), \ldots, \ln(r_{K,K-1}), 0]^T \quad (12)$$

$$H = [\boldsymbol{h}_1, \ldots, \boldsymbol{h}_p, \ldots, \boldsymbol{h}_{K(K-1),\boldsymbol{1}}]^T \quad (13)$$

$$\{\boldsymbol{h}_p\}_n = -1,\ \{\boldsymbol{h}_p\}_m = 1,\ \text{for}\ \{\boldsymbol{r}_{\log}\}_p = \ln(r_{nm}), \quad (14)$$

The last row of H is a constraint for setting geometric mean of all speakers' VTL equal to one.

The solution $\hat{\boldsymbol{l}}_{\log}$ is a function of the set of performance-determining parameters Θ and is explicitly represented as $\hat{\boldsymbol{l}}_{\log}(\Theta)$. Relative consistency $\eta(\Theta)$ of the estimates is used to tune the set Θ.

$$\eta(\Theta) = \frac{\sum_{m=1}^{m\neq n,K}\sum_{n=1}^{K}|\hat{r}_{nm}(\Theta) - \bar{\hat{r}}(\Theta) - r_{nm}(\Theta) + \bar{r}(\Theta)|^2}{\sum_{m=1}^{m\neq n,K}\sum_{n=1}^{K}|r_{nm}(\Theta) - \bar{r}(\Theta)|^2}, \quad (15)$$

where $\bar{x}$ represents average of variable x and $\hat{r}_{nm} = \hat{l}_m/\hat{l}_n$.

[1] Density of frequency discretization of $\boldsymbol{f}_d$ is set to 24 frequency points in one octave. This density is fine enough and has negligible impact on performance.

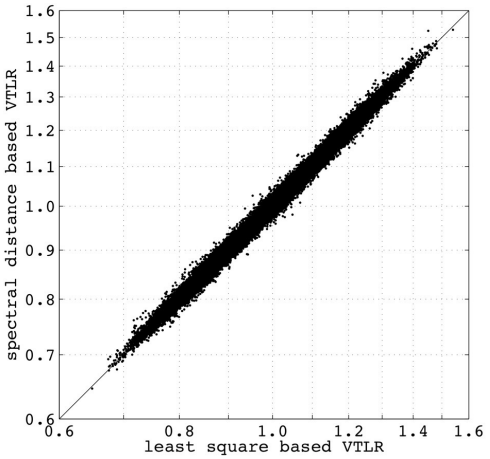

Figure 2: Scatter plot of estimated VTL ratios based on least square method and estimated VTL ratios based on spectral distance minimization

Figure 2 shows the scatter plot of the VTL ratios of all ($385 \times 384 = 147,840$) speaker pairs. The horizontal axis represents the VTL ratio $\hat{r}_{nm}$, which is calculated from the least square estimates and the vertical axis represents the spectrally estimated VTL ratio r_{nm} using the tuned Θ. The standard deviation of $\log(r_{nm}(\Theta)/\hat{r}_{nm}(\Theta))$ for the best parameter set ($f_N = 600$ Hz, $f_W = 2000$ Hz, $f_L = 400$ Hz and $f_H = 3500$ Hz) was 0.86%. This reproducibility is higher than the recent formant based method [10] using acoustic sensitivity functions [23, 24] and our report [25]. However, it is important to note that the relative VTL obtained here does not represent the physical VTL directly. It is a highly reproducible parameter to align two power spectra of different speakers'. Relation between this parameter and the physically defined VTL is tested.

5.2. Calibration by MRI-based VTL

A vowel database with MRI-based shape and VTL information [12] and a similar data set for female speakers were used for calibrating the relative VTL obtained in the previous section. The former database consists of data of 15 male speakers[2] and the latter data set consists of 6 female speakers. The VTL information in these sets was derived from the shape data based on the procedure described in the reference [13]. Since MRI measurements produce strong acoustic noise, vowel sounds recorded in an office environment by the same speakers were used. Isolated and sustained vowel sounds were produced in the same manner as in the MRI recoding.

The sounds were converted to averaged dB spectrum $L_{vMRI}^{(n)}(\boldsymbol{f}_d)$ using the similar procedure described in the previous section. One difference is the F0 extractor used. The SWIPE' F0 extractor [26] was used in this calibration, because it is tolerant to the background noise found in the MRI data sets. Each speaker's averaged dB spectrum was compared with those of the 385 speakers' and yielded 385 relative VTL values. The median of 385 relative VTL values was used as the relative VTL estimates $\hat{l}_{rel}$.

Regression analysis of the relative VTL estimates $\hat{l}_{rel}$ and VTL measured from MRI $\hat{l}_{MRI}$ was conducted using logarithmic conversion. Both regression coefficient and intercept were

[2] The data by three speakers were eliminated because of missing information. In the following analyses, data of 12 speakers were used.

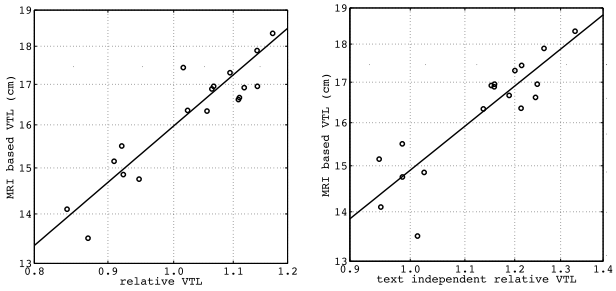

Figure 3: (Left plot) Scatter plot of the relative VTL and the MRI based VTL. Regression line is also shown. (Right plot) Scatter plot of the text independent relative VTL and the MRI based VTL. MRI based VTL is the average of five vowels.

highly significant ($p = 5 \times 10^{-5}$, $p < 10^{-15}$ respectively) and yielded the equation:

$$\hat{l}_{MRI} = \exp\Big(0.8025 \times \log(\hat{l}_{rel}) + 2.771\Big) \ . \quad (16)$$

Left plot of Fig. 3 shows the scatter plot of the least square based relative VTL and the MRI based ones. The R-square is 0.82 and the standard error around this regression line is about 6 mm at $\hat{l}_{rel} = 0$.

6. Text independent relative VTL

By scaling the standardized dB power spectrum with mean normalization $L_v^{(n)}(\boldsymbol{f}_d)$ using the estimated relative VTL $\hat{l}_n$ yields VTL normalized version of the standardized dB power spectrum $L_{vN}^{(n)}(\boldsymbol{f}_d)$.

$$L_{vN}^{(n)}(\boldsymbol{f}_d) = L_v^{(n)}(\hat{l}_n \boldsymbol{f}_d). \quad (17)$$

The set of VTL normalized version of the standardized dB power spectrum $L_{vN}^{(n)}(\boldsymbol{f}_d)$ for all combination of speakers (speaker ID is n) and vowels (vowel ID is v) is used as the reference template $\boldsymbol{V}_N$ of the proposed text independent relative VTL estimation.

Assume $d_v^{(n)}(x, m; l)$ to represent a distance between $X(f, m)$, a preprocessed and mean normalized version of dB power spectrum of the utterance x at frame ID m with assumed VTL l, and each item $L_{vN}^{(n)}$ in the reference template $\boldsymbol{V}_N$.

$$d_v^{(n)}(x, m; l) = \left(\frac{1}{\#(\boldsymbol{f}_d)} \sum_{f \in \boldsymbol{f}_d} \Big| X(lf, m) - L_{vN}^{(n)}(f) \Big|^2 \right)^{\frac{1}{2}}. \quad (18)$$

For text independent VTL estimation, a measure $\xi_I(x, \boldsymbol{V}_N; l)$ based on *Diverse Density* [14] is introduced.

$$\xi_I(x, \boldsymbol{V}_N; l) = \sum_{\{n,m,v\} \in T \times M \times V} \exp\Big(-\alpha \max(0, d_v^{(n)}(x, m; l) - \beta)^2\Big) \times \prod_{v \neq u \in V} \Big(1 - \exp\Big(-\alpha \max(0, d_u^{(n)}(x, m; l) - \beta)^2\Big)\Big) \quad (19)$$

where a symbol T represents the set of speaker IDs in the database. Parameters α and β are experimentally determined parameters.The maximum function $\max(0, x - \beta)$ and the bias parameter β are introduced to reduce effects of within-category spectral variations ($\alpha = 2.5, \beta = 0.2$ are used). The first term of Eq.(19) represents the contribution of the same category templates and the product term in the second line represents the mask for suppressing effects of outside category templates.

6.1. Calibration by MRI based VTL

The estimate of the text independent relative VTL $\hat{l}_{TI}(x)$ for a given utterance x is numerically determined using a nonlinear maximization of $\xi_I(x, \boldsymbol{V}_N; l)$. Regression analysis of this estimate $\hat{l}_{TI}(x)$ and the MRI based VTL $\hat{l}_{MRI}(x)$ was conducted using logarithmic conversion of them. Both regression coefficient and intercept were highly significant ($p = 9 \times 10^{-7}$, $p < 10^{-15}$ respectively) and yielded the equation:

$$\hat{l}_{MRI} = \exp\Big(0.6916 \times \log(\hat{l}_{TI}) + 2.701\Big) \ . \quad (20)$$

Right plot of Fig. 3 shows the scatter plot of the text independent relative VTL and the MRI based ones. The R-square is 0.79 and the standard error around this regression line is about 6 mm at $\hat{l}_{TI} = 0$. Note that without introducing time consuming labeling in the runtime process, comparable performance is obtained by introducing *Diverse Density* as the measure.

6.2. Implementation

A high-speed F0 extractor based on higher order waveform symmetry [27] combined with a simple Kalman smoother [28] is designed to calculate F0 [29]. The text independent estimation procedure including preprocessing of the given input utterance is implemented using Matlab [30]. Elapsed time for relative VTL estimation from signal input to estimate $\hat{l}_{TI}(x)$ output is comparable to the signal duration (for example, the elapsed time is 5.2 s for processing a signal of 5.3 s) on a PC (MacBook pro, 2.6 GHz, Intel Core i7, 16 GB memory, OS X 10.9.2) when 44.1 kHz sampled signals are analyzed using a 100 ms Blackman window with 20 ms frame shift. Practically, 40 ms frame shift is acceptable and yields faster then realtime throughput. All the procedures in the proposed method were designed scalable to sampling frequencies higher than 7000 Hz. Introducing noise robustness is one of the further research, since the proposed method is potentially useful for age and gender identification in dialogue systems. Currently, SWIPE' [26] is one useful alternative F0 extractor for such adverse conditions.

7. Conclusion

A highly reproducible VTL estimation method based on vowel database consisting of 385 speakers and a fast, simple and text independent method for VTL estimation are proposed. The estimated VTLs are calibrated based on VTL measured from MRI data and vowel recording of each speaker. Elapsed time of VTL estimation by the text independent method is generally shorter than the duration of the given utterance. The primary application of the propose method is an automatic procedure for assisting a newly proposed "Temporally variable multi-aspect N-way morphing" algorithm [31]. Many other prospective applications (such as [32]) are underdevelopment.

8. Acknowledgements

This work is partly supported by Kakenhi (Aids for Scientific Research) of JSPS 24300073, 24500233, 24650085 and 25280066. The MRI data and recorded sound data used in this study are parts of "ATR vocal tract MRI data for Japanese vowels" that were acquired at Human Information Science Laboratories in Advanced Telecommunications Research Institute International (ATR) and released from ATR-Promotions Co. Ltd. The use of the data is under the license agreement with ATR-Promotions Co. Ltd.

9. References

[1] W. T. Fitch and J. Giedd, "Morphology and development of the human vocal tract: A study using magnetic resonance imaging," *The Journal of the Acoustical Society of America*, vol. 106, no. 3, pp. 1511–1522, 1999.

[2] M. Benzeghiba, R. D. Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, and C. Wellekens, "Automatic speech recognition and speech variability: A review," *Speech Communication*, vol. 49, no. 10-11, pp. 763–786, 2007.

[3] H. Wakita, "Normalization of vowels by vocal-tract length and its application to vowel identification," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 25, no. 2, pp. 183–192, 1977.

[4] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," in *ICASSP-96*, vol. 1, 1996, pp. 346–348.

[5] T. Emori and K. Shinoda, "Rapid vocal tract length normalization using maximum likelihood," in *Eurospeech 2001*, 2001, pp. 1649–1652.

[6] E. Okamoto, T. Irino, R. Nisimura, and H. Kawahara, "Auditory filterbank improves voice morphing," in *Interspeech 2011*, 2011, pp. 2517–2520.

[7] D. R. R. Smith, R. D. Patterson, R. Turner, H. Kawahara, and T. Irino, "The processing and perception of size information in speech sounds," *The Journal of the Acoustical Society of America*, vol. 117, no. 1, pp. 305–318, 2005.

[8] S. R. Schweinberger, C. Casper, N. Hauthal, J. M. Kaufmann, H. Kawahara, N. Kloth, and D. M. Robertson, "Auditory adaptation in voice perception," *Current Biology*, vol. 18, pp. 684–688, 2008.

[9] V. Sorokin and I. Geras'kin, "Vocal-tract length estimation," *Journal of Communications Technology and Electronics*, vol. 58, no. 12, pp. 1292–1301, 2013.

[10] T. Kaburagi, "Estimating the vocal-tract shape from speech spectrum using a sensitivity function," in *Proc. Spring Conf. Acoust. Soc. Japan*, 2014, pp. 293–296, [in Japanese].

[11] G. Ohyama, T. Deguchi, and H. Kasuya, "Construction of Japanese vowel database uttered by native speakers over a wide range of age," in *Proc. Spring Meeting of ASJ*, Mar. 2011, pp. 2–P–15(a), [in Japanese].

[12] B. A. I. Center, *Description of ATR vocal tract MRI data for Japanese vowels*, 1st ed., ATR-Promotions, Kyoto, Japan, Dec. 2012.

[13] H. Takemoto, K. Honda, S. Masaki, Y. Shimada, and I. Fujimoto, "Measurement of temporal changes in vocal tract area function from 3D cine-MRI data," *The Journal of the Acoustical Society of America*, vol. 119, no. 2, pp. 1037—1049, 2006.

[14] O. Maron and T. Lozano-Perez., "A framework for multiple-instance learning," in *Advances in Neural Information Processing Systems (NIPS),*, vol. 10. MIT Press, 1998, pp. 570–576.

[15] Y. Atake, T. Irino, H. Kawahara, L. Jinlin, S. Nakamura, and K. Shikano, "Robust fundamental frequency estimation using instantaneous frequencies of harmonic components," in *Proc. ICSLP-2000*, vol. 2, 2000, pp. 907–910.

[16] H. Kawahara, M. Morise, T. Toda, R. Nisimura, and T. Irino, "Beyond bandlimited sampling of speech spectral envelope imposed by the harmonic structure of voiced sounds," in *Interspeech 2013*, 2013, pp. 34–38.

[17] G. Fant, *Acoustic Theory of Speech Production*. The Hague, The Netherlands: Mouton, 1960.

[18] B. Bozkurt, B. Doval, C. d'Alessandro, and T. Dutoit, "Zeros of z-transform representation with application to source-filter separation in speech," *Signal Processing Letters, IEEE*, vol. 12, no. 4, pp. 344–347, 2005.

[19] I. R. Titze, "Nonlinear source–filter coupling in phonation: Theory," *J. Acoust. Soc. Am.*, vol. 123, no. 5, pp. 2733–2749, May 2008.

[20] J. Dang and K. Honda, "Acoustic characteristics of the piriform fossa in models and humans," *J. Acoust. Soc. Am.*, vol. 101, no. 1, pp. 456–465, 1997.

[21] S. O. Ternström, "Hi-Fi voice: observations on the distribution of energy in the singing voice spectrum above 5 kHz," in *Proc. Acoustics'08 Paris*, 2008, pp. 3171–3176.

[22] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno, "A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0 and aperiodicity estimation," in *ICASSP2008*, 2008, pp. 3933–3936.

[23] B. H. Story, "Technique for "tuning" vocal tract area functions based on acoustic sensitivity functions," *The Journal of the Acoustical Society of America*, vol. 119, no. 2, pp. 715–718, 2006.

[24] T. Kaburagi, T. Takano, and Y. Sakamoto, "Estimating area function of the vocal tract from formants using a sensitivity function and least-squares," *Acoustical Science and Technology*, vol. 34, pp. 301–310, 2013.

[25] M. Kobayashi, R. Nisimura, T. Irino, and H. Kawahara, "Estimated relative vocal tract lengths from vowel spectra based on fundamental frequency adaptive analyses and their relations to relevant physical data of speakers," *Proc. ICA2013, Proceedings of Meetings on Acoustics (ASA POMA)*, vol. 19, no. 1, p. 5aCb44, 2013.

[26] A. Camacho and J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 124, no. 3, pp. 1638–1652, 2008.

[27] H. Kawahara, M. Morise, R. Nisimura, and T. Irino, "Higher order waveform symmetry measure and its application to periodicity detectors for speech and singing with fine temporal resolution," in *ICASSP2013*, 2013, pp. 6797–6801.

[28] P. Garner, M. Cernak, and P. Motlicek, "A simple continuous pitch estimation algorithm," *Signal Processing Letters, IEEE*, vol. 20, no. 1, pp. 102–105, 2013.

[29] H. Kawahara, M. Morise, and K.-I. Sakakibara, "Temporally fine F0 extractor applied for frequency modulation power spectral analysis of singing voices," *Proc. MAVEBA*, pp. 125–128, 12 2013.

[30] Matlab, *version 8.2.0.701 (R2013b)*. Natick, Massachusetts, USA: The MathWorks Inc., 2013.

[31] H. Kawahara, H. Morise, Banno, and V. G. Skuk, "Temporally variable multi-aspect N-way morphing based on interference-free speech representations," in *ASPIPA ASC 2013*, 2013, p. 0S28.02.

[32] M. Sakaguchi, M. Kobayashi, R. Nisimura, T. Irino, and H. Kawahara, "Spectrally estimated vocal tract lengths of singing voices and their contributing factors," *Proc. MAVEBA*, pp. 121–124, 12 2013.