FLEXIBLE VOICE MORPHING BASED ON LINEAR COMBINATION OF MULTI-SPEAKERS' VOCAL TRACT AREA FUNCTIONS

Yoshiki Nambu, Masahiko Mikawa, and Kazuyo Tanaka

Graduate School of Library, Information and Media Studies, University of Tsukuba, Japan
1-2, Kasuga, Tsukuba-shi, Ibaraki-ken 305-8550, Japan
Email: {ynambu, mikawa, ktanaka}@slis.tsukuba.ac.jp

ABSTRACT

This paper presents a flexible voice morphing method based on conversion using a linear combination of multi-speakers' vocal tract area functions, in which phonological identity is maintained in terms of the overall interpolated area. In this system, the characteristic of vocal tract resonances is separated from that of glottal source waves using AR-HMM analysis of speech. The vocal tract resonances and glottal source wave characteristics are independently morphed. For the morphing of vocal tract resonances, log area vocal tract functions, which are derived from AR coefficients, are normalized and then processed by statistical mapping technique. For glottal source waves, statistical mapping is conducted in the cepstrum domain. Morphed speech is re-synthesized by an AR filter of converted glottal source waves which is re-synthesized using a cepstrum domain conversion. With the proposed morphing system, the continuity of formants and perceptual differences between a conventional method and the proposed method are confirmed.

1. INTRODUCTION

Voice morphing or voice conversion usually means transformation from a source speaker's speech to a target speaker's. Therefore, these techniques are considered to be a kind of point-to-point mapping in a feature space. Our research on voice morphing aims to extend this restriction to area-to-area mapping by introducing multi-speakers of adequate features derived from a speech production model, since that will be useful for such applications as the creation of peculiar voices in animation films.

Since the 1990s, many techniques for voice conversion have been proposed [1-7]. One successful technique is to use a statistical method for mapping a source speaker's voice to a target speaker's in the cepstrum domain [2,3]. However, a weakness of these methods is the discontinuity of formants, due to the fact that the relationship between formant transitions and the time pattern of the power spectral envelope sequence is nonlinear, that is, continuous interpolation of log power spectra does not result in continuous formant transitions. This characteristic behavior will result in a deterioration of the phonological quality. Some improvements of these methods have been proposed to counter this deterioration [5,6].

The proposed method employs an estimated vocal tract area function to avoid such weakness. As is well known [8,9], partial autocorrelation (PARCOR) coefficients can be considered as reflection coefficients of a vocal tract area function, and the local peaks of power spectrum envelopes of vocal tract area functions have a flat level in a certain frequency band for vowels [10]. Moreover, the number of coefficients refers to the number of poles contained in the power spectrum, i.e., formants. Based on these restrictions, interpolation in the vocal tract area domain is considered to provide reasonably continuous transition of formants.

Estimation of the vocal tract area function implies simultaneous estimation of the voice source characteristics. For this purpose we introduce Auto-Regressive Hidden Markov Model (AR-HMM) analysis of speech, which has been proposed for improved AR-modeling of speech [11]. AR-HMM represents the vocal tract resonance characteristics by an AR model and the glottal source wave by an HMM.

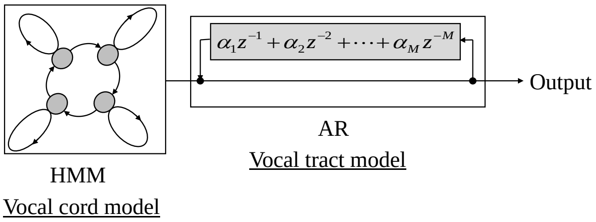
The voice morphing system uses the log vocal tract area functions and a cepstrum sequence of glottal source waves as feature parameters for the linear combination of multi speakers' characteristics. The re-synthesis procedure to obtain morphed speech is as follows; the glottal source wave is synthesized by the synthesis-by-analysis software package STRAIGHT [12], in the cepstrum domain mapping. Output speech is synthesized by AR filtering of the glottal source wave, where AR coefficients are calculated as reflection coefficients of the vocal tract area function obtained by linear combination of the log vocal tract area functions. Before calculating the linear combination, the vocal tract area functions of different speakers are stretched to adjust the length of different vocal tract areas. We show that the interpolated spectral envelopes are reasonable with regard to continuous transition of formants, which are substantially different from those obtained from the cepstrum domain and also improved from using non-stretched vocal tract areas. We confirm by perceptual experiment that the differences can be perceptually recognized.

2. PROPOSED METHOD

2.1 AR-HMM analysis

The AR-HMM model represents the vocal tract characteristics by an AR model and the glottal source wave by an HMM. The AR-HMM model structure is depicted in Fig. 1.

The AR-HMM analysis estimates the vocal tract resonance characteristics and vocal source waves in the sense of maximum likelihood estimation. In this way, components of the vocal tract resonance characteristics and those of the source waves can be naturally separated.



$\alpha_1 z^{-1} + \alpha_2 z^{-2} + \cdots + \alpha_M z^{-M}$
Output
AR
Vocal tract model
HMM
Vocal cord model

Fig. 1: Schematic diagram of AR-HMM for speech analysis

Conventional AR model estimation assumes that the glottal source wave has a Gaussian distribution. This assumption however can become invalid, especially when analyzing speech with a high fundamental frequency, such as that of some female speakers. On the contrary, in AR-HMM estimation, the vocal cord HMM and the vocal tract AR model are alternately estimated using the maximum likelihood method. AR-HMM can estimate the vocal tract features without being biased by pitch harmonics. In addition, since the HMM used here adopts an assumption of ring-states for the glottal source wave, the estimated glottal source can be regarded as an approximation of the glottal source wave. An example of AR-HMM analysis results is shown in Fig. 2.

2.2 Estimation of the vocal tract area function

The power spectrum was calculated from the AR coefficients with AR-HMM analysis, and reflection coefficients (PARCOR) $k_i, i = 1,2,\cdots,n$ of the vocal tract area function were derived from autocorrelation coefficients obtained by IDFT applied to the power spectrum. In the approach described in this paper, before analysis with AR-HMM, a first order adaptive inverse filtering was used for equalization of formants [10].

The vocal tract area function $A_i (i = 1,2,\cdots,n+1)$ was calculated by:

$$\begin{aligned} &\boldsymbol{A_{n+1} = 1} \\ &\boldsymbol{A_i = \frac{1-k_i}{1+k_i} A_{i+1}} \end{aligned} \tag{1}$$

Then, we normalized the vocal tract area functions by dividing by their sum. Finally, we used log normalized vocal tract area functions, in order to prevent vocal tract area functions from becoming negative and AR coefficients from being unstable.

For linear interpolation in the vocal tract area function domain, a formant is expected to be a continuous transition. This is confirmed in Fig. 3, where two spectra transitions are shown; one is a linear interpolation in the cepstrum domain and the other is one in the vocal tract area function domain. It is obvious that the formant transitions are continuous for the graph on the left, i.e., in the vocal tract area function domain.

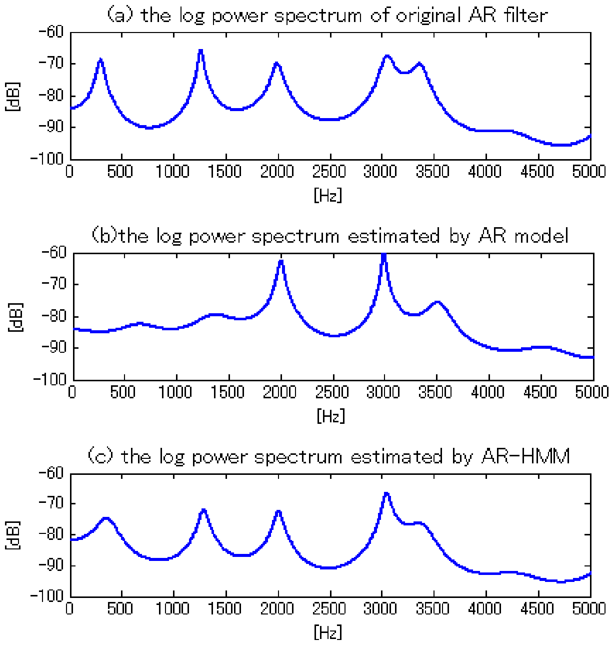


Fig. 2: An example of the results of AR-HMM analysis using a 500-Hz pitch of synthesized speech: The order of AR coefficients is 19, and there are 13 HMM states.

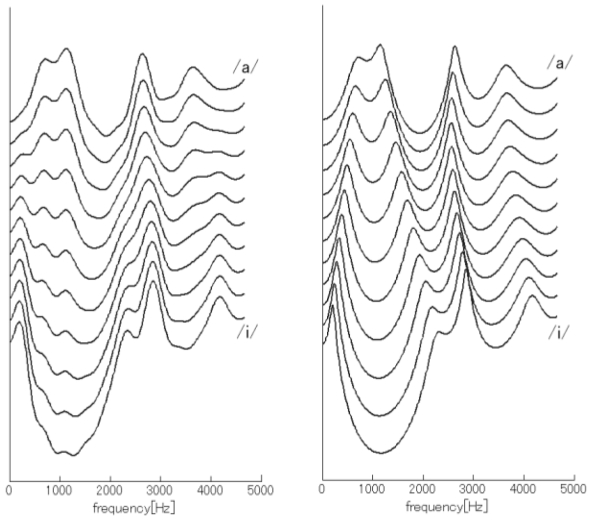


Fig.3: An example of linear interpolation: power spectra sequences obtained using 40 cepstrum coefficients (left), and using log vocal tract area functions (right).

2.3 Conversion function

The voice conversion technique used in the system is basically statistical mapping from a source speaker's voice to a target speaker's. The conversion function is represented as a Gaussian Mixture Model (GMM).

Let us denote the vector analyzed from a source speaker's speech by x, and the corresponding vector analyzed from a target speaker's speech by y. The conversion function $F(\mathbf{x})$ is given as follows.

$$F(\mathbf{x}) = E[\mathbf{y} \mid \mathbf{x}] = \sum_{i=1}^{m} \mathbf{p}_i(\mathbf{x})\left[\boldsymbol{\mu}_i^y + \boldsymbol{\Sigma}_i^{yx}(\boldsymbol{\Sigma}_i^{xx})^{-1}(\mathbf{x} - \boldsymbol{\mu}_i^x)\right] \tag{2}$$

$$\mathbf{p}_i(\mathbf{x}) = \frac{\alpha_i N(\mathbf{x}; \boldsymbol{\mu}_i^x, \boldsymbol{\Sigma}_i^{xx})}{\sum_{j=1}^{m} \alpha_j N(\mathbf{x}; \boldsymbol{\mu}_j^x, \boldsymbol{\Sigma}_j^{xx})} \tag{3}$$

$$\sum_{i=1}^{m} \alpha_i = 1, \quad \alpha_i \geq 0$$

where $N(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The value m is the number of Gaussian components. The vectors $\boldsymbol{\mu}_i^x$ and $\boldsymbol{\mu}_i^y$ denote the mean vectors of the ith Gaussian model estimated from $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively. The matrix $\boldsymbol{\Sigma}_i^{xx}$ denotes the covariance matrix of the i^{th} Gaussian model estimated from $\boldsymbol{x}$. $\boldsymbol{\Sigma}_i^{yx}$ is the cross-covariance matrix, and α_i is the mixture weight of each class.

As described in [2], the GMM-based estimation of the conversion function uses a set of time-aligned $\boldsymbol{x}$ and $\boldsymbol{y}$, $\mathbf{z} = [x^T y^T]^T$ to estimate the parameters of a joint model of Gaussian mixtures. Once the model has been trained, the density of $\boldsymbol{x}$ and $\boldsymbol{y}$ is given by the following.

$$\boldsymbol{\Sigma}_i^z = \begin{bmatrix} \boldsymbol{\Sigma}_i^{xx} & \boldsymbol{\Sigma}_i^{xy} \\ \boldsymbol{\Sigma}_i^{yx} & \boldsymbol{\Sigma}_i^{yy} \end{bmatrix} \quad \boldsymbol{\mu}_i^z = \begin{bmatrix} \boldsymbol{\mu}_i^x \\ \boldsymbol{\mu}_i^y \end{bmatrix} \tag{4}$$

In the method described here, the vocal tract characteristics are converted in the log vocal tract area function domain, and the glottal source wave characteristics are converted in the cepstrum domain.

2.4 Re-synthesis of the converted voice

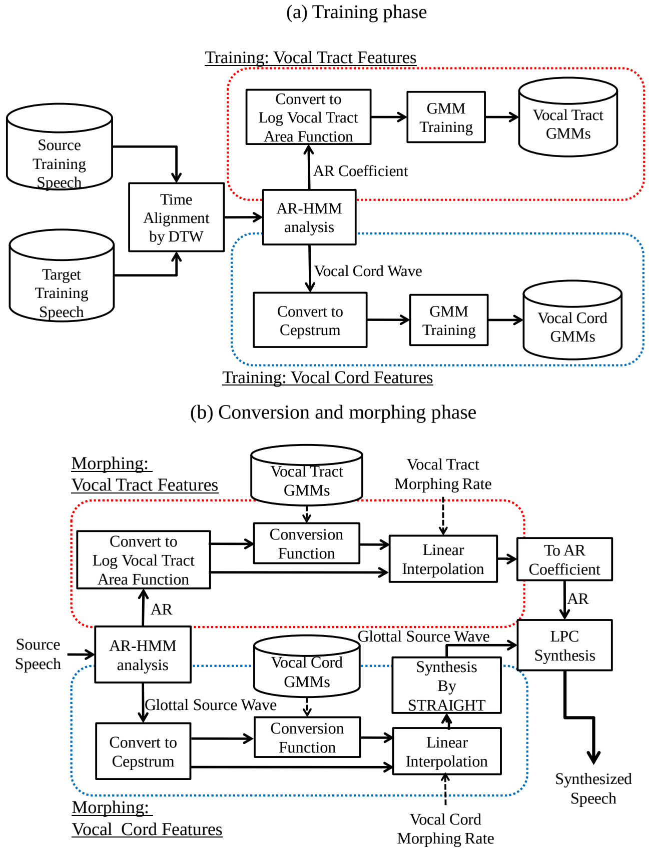
The system overview of the voice conversion process is shown in Fig. 4, where the system consists of a training phase and a conversion phase. The procedure of each phase is as follows.

Training phase:

1) *AR-HMM analysis*: Speech samples with the same phonetic content from both source and target speaker are analyzed, to estimate the AR coefficients for the vocal tract features and the glottal source wave for the vocal cord features. The AR coefficients are transformed to log vocal tract area functions. The glottal source wave is transformed to cepstra.
2) *Feature alignment*: The feature vectors obtained above are time-aligned using dynamic time warping (DTW) in order to compensate for any differences in duration between source and target utterances.
3) *Estimation of the conversion function*: The aligned vectors are used to train a joint GMM whose parameters are then used to construct a stochastic conversion function. The conversion function for vocal tract features and the conversion function for vocal cord features are estimated independently.

Conversion and morphing phase:

1) *AR-HMM Analysis*: As in the training phase, the vocal tract and vocal cord features are estimated using an AR-HMM, but in this case only the source speaker's utterances are used.
2) *Features Transformation*: The GMM-based transformation function constructed during training is now used for converting every source log vocal tract area function and vocal cord cepstrum into its most likely target equivalent.
3) *Linear Interpolation*: The features of morphed speech are obtained using $(1 - \sum_{k=1}^{n} \lambda_k)\mathbf{x} + \sum_{k=1}^{n} \lambda_k F_k(\mathbf{x})$, where $\mathbf{x}$ denotes the original source feature vector, and $F_k(\mathbf{x})$ is the converted feature vectors of speaker k obtained using the conversion functions. The parameter λ_k denotes the morphing rate. Vocal tract features and vocal cord features are interpolated independently.
4) *Synthesis of the source wave*: The source wave for LPC synthesis is synthesized with the STRAIGHT software, using the converted vocal cord cepstrum.
5) *LPC synthesis*: The AR coefficients for LPC synthesis are obtained by the PARCOR coefficients derived from the converted log vocal tract area functions. Finally, we filtered the synthesized source wave with the AR coefficients.


(a) Training phase

(b) Conversion and morphing phase

Fig. 4 Block diagram of the voice conversion system

2.5 Stretch in the direction of vocal tract length

In the case of vocal tract area function space interpolation, it is necessary to take into account the differences in vocal tract length of the speaker. To accommodate this, we stretched vocal tract area functions in the direction of the long axis of the vocal tract before linear interpolation. The procedure is as follows.

1) The vocal tract area functions are approximated by p order polynomials (in our experiments, p =6), and the local maxima and local minima are estimated.
2) The correspondence of the poles between the speakers are selected with minimum distance among those poles.
3) Finally, the vocal tract area function of morphed speech are piecewise stretched and interpolated using the correspondences.

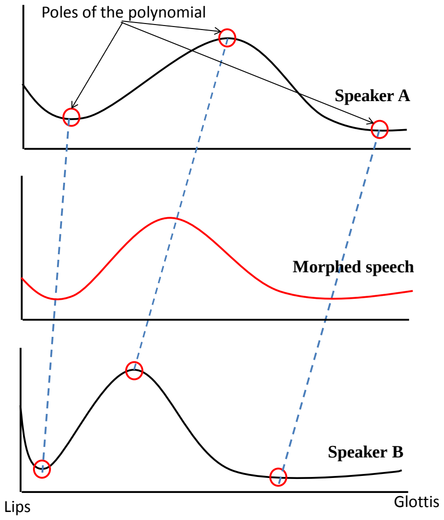


Fig. 5 Stretch in the direction of vocal tract length

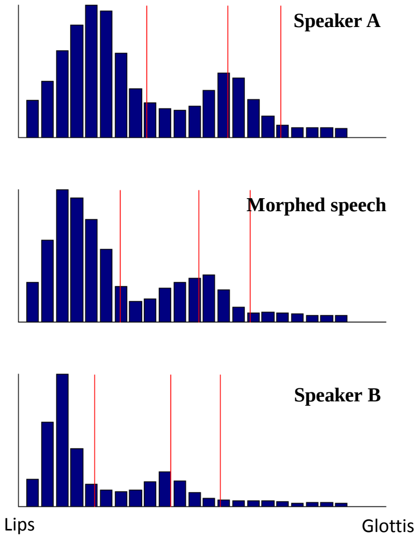


Fig. 6 An example of the piecewise stretch and interpolation of vocal tract area functions

2.6 The method of pitch modification

A basic prosodic transformation was also applied. The fundamental frequencies are modeled by a log-normal distribution. During the training phase, the mean value and variance of the log scale fundamental frequency was calculated for speakers. We estimated the converted fundamental frequency using the following formula.

$$f_0' = \mu_y + \frac{\sigma_y}{\sigma_x} \times (f_0 - \mu_x)$$

where f_0 , f_0' denote log scale fundamental frequencies of before-conversion and after-conversion. The values μ_x , μ_y are the mean log pitch of source and target speakers, respectively. σ_x , σ_y are the variances of log pitch.

3. EXPERIMENTS

3.1 Experimental conditions

The speech sample set used for voice morphing contained 50 sentences in Japanese, each uttered by three male and three female speakers. The sampling frequency was 16 [kHz] and the mean duration of the sentence samples was 4.7 [s]. Forty-five sentences were used for the training of the conversion functions; five sentences were used for the synthesis of the morphed speech.

The number of AR coefficients and HMM states for the AR-HMM were 21 and 13, respectively. The HMM states of the AR-HMM were connected in a ring topology. There were 128 mixtures of GMM for the conversion function. Twenty cepstrum coefficients were used for re-synthesis of the source wave. The following three types of interpolation methods were compared:

(a) Linear interpolation of 40 cepstrum coefficients calculated from the AR coefficients.

(b) Linear interpolation of log scale vocal tract area functions (order 21).

(c) In addition to (b), using the poles of the approximating polynomials of the vocal tract area functions, we carried out a piecewise linear stretch and interpolation.

The morphed speech was synthesized using the methods (a), (b), and (c), changing the morphing rate. Three combinations of source and target speakers were used; male to male, female to female, and male to female.

3.2 Observation of the formants for the morphed speech

We observed the power spectrum for the same analysis frame of the morphed speech between the original male speaker's features and the target female speaker's converted features, when the morphing rate λ =0.5 (Fig. 7). In the case of method (a), interpolation in the cepstrum domain, it can be seen that the power spectra are smoothed and the formants are indistinct. In the case of method (b), the formants are distinct, but each formant is situated at small equal intervals. This is because the form of vocal tract area functions is treated as a uniform tube, as a result of interpolation that doesn't take account of varying vocal tract length. In the case of (c), interpolations that do take into account vocal tract length, each formant is situated appropriately in the middle of the source speaker's and the target speaker's frequency ranges.

3.3 Perceptual test results

We synthesized the morphed speech that was interpolated with speech from six speakers with a constant morphing rate, and conducted a preliminary listening test for differences between results of each method. Firstly, we derived morphed speech from the original speech samples from the six speakers. The listening test result by three subjects indicates that speech derived using the methods (b), (c) was clearer than speech derived by method (a). Next, we derived morphed speech from one male speaker's speech samples as source and converted speech samples that mapped to the other five speakers as targets from the male speaker. As a result, clear differences in the hearing of speech derived using each method were not found. This was possibly because the lower formants were maintained to some extent, and also because we interpolated between the converted speeches, that were of a degraded quality.

4. CONCLUSION

This paper has presented a voice morphing method based on mappings in the vocal tract area space and glottal source wave spectrum that can each be independently modified. These features have been realized using AR-HMM analysis of speech. The research has focused on the phonetic continuity of morphed speech converted from a linear combination of multi-speakers' voices. We have discussed this issue by comparing acoustic features and interpolation techniques. The feasibility of the method has been reasonably confirmed by observing the positions of the formants of the morphed speech, and conducting a preliminary listening test on the morphed speech of six speakers. In future, we will investigate how to improve the quality of voice conversion with interpolation techniques.

Acknowledgement:
We thank Prof. Hideki Kawahara of the University of Wakayama for kindly licensing the use of STRAIGHT. This research is supported in part by a Grand-in-Aid for Scientific Research, C(1) Project No.22500145, from the Japan Society for the Promotion of Science.

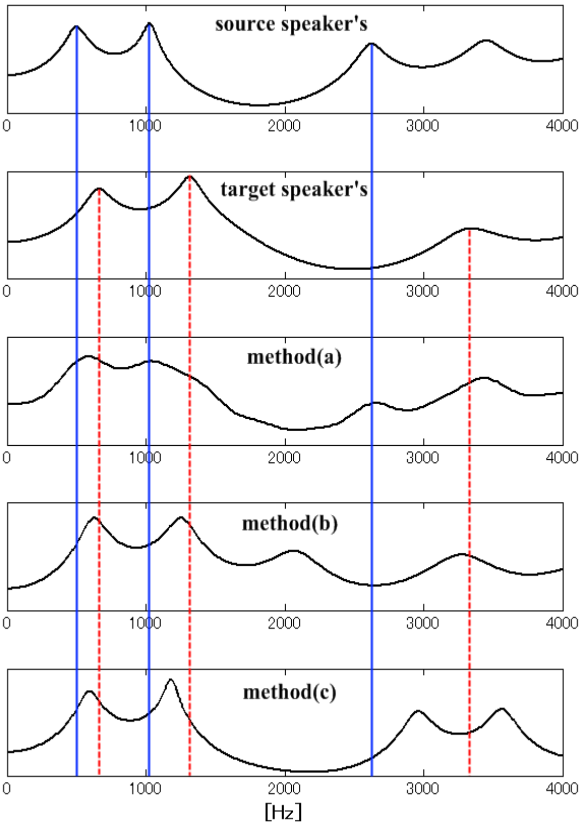


Fig. 7 The power spectrum of synthesized speech, when the morphing rate is 50%: Starting from the top, original speaker's, target speaker's, interpolation in the cepstrum domain (a), interpolation in the log scaled vocal tract area functions (b), and combination with a stretch in the direction of the long axis of the vocal tract (c).

REFERENCES

[1] L.M. Arslan, D.Talkin, "Voice conversion by codebook mapping of line spectral frequencies and excitation spectrum," *Proc. Eurospeech*, pp.1347-1350, 1997.

[2] Y.Stylianou, O.Cappe, "A system voice conversion based on probabilistic classification and a harmonic plus noise model", *Proc.ICASSP*, pp.281-284, 1998.

[3] A.Kain, "Spectral voice conversion for text-to-speech synthesis", *Proc.ICASSP* pp.285-288, 1998.

[4] H. Ye, S. Young, "High Quality Voice Morphing", in *Proc.IEEEICASSP*, pp.9-12, 2004.

[5] W. Percybrooks, E. Moore Ⅱ, "Voice Conversion With Linear Prediction Residual Estimaton", in *Proc.ICASSP*, pp.4673-4676, 2008.

[6] D. Erro, T. Polyakova, A. Moreno, "On Combining Statistical Methods And Frequency Warping for High-Quality Voice Conversion", in *Proc.ICASSP*, pp.4665-4668, 2008.

[7] Z. Shuang, F. Meng, Y. Qin, "Voice Conversion by Combining Frequency Warping with Unit Selection", in *Proc.ICASSP*, pp.4661-4664, 2008.

[8] F. Itakura, S. Saito, "Digital filtering technique for speech analysis and synthesis," *Proc. 7th ICA*, 25C1, 1971.

[9] H. Wakita, "Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms," *IEEE Trans. AU-21*, No.5, pp.417-427, 1973.

[10] T. Nakajima, H. Ohmura, K. Tanaka, S. Ishizaki, "Estimation of vocal tract area functions by adaptive inverse filtering and extraction of articulatory parameters", Proc. of 8th International Congress on Acoustics, London, Vol.1, pp.323, 1974.

[11] A. Saso, K. Tanaka, "Glottal excitation modeling using HMM with application to robust analysis of speech", *Proc. ICSLP*, Vol.4, pp704-707, 2000.

[12] H. Kawahara, I. Masuda, "Spline-based approximation of time-frequency representation in STRAIGHT method", *IEICE Technical Report*, pp19-24, 1997.