Continuous Voice Morphing Using Separated Vocal Tract Area Functions and Glottal Source Waves

Kazuyo Tanaka\textsuperscript{1}, Yoshiki Nambu\textsuperscript{2}

\textsuperscript{1}Faculty of Library, Information and Media Science, University of Tsukuba
\textsuperscript{2}Graduate School of Library, Information and Media Studies, University of Tsukuba

1-2 Kasuga, Tsukuba, Ibaraki, 305-8550, Japan
\textsuperscript{1}ktanaka@slis.tsukuba.ac.jp

Abstract- This paper presents a flexible voice morphing method, which is based on a conversion using a linear combination of the vocal tract area functions estimated from speech signals. The method focuses on the continuity of the phonological identity of the overall interpolated area. The main features of the method are 1) to separate characteristics of the vocal tract resonances from those of glottal source waves using AR-HMM analysis of speech, 2) independent morphing of the vocal tract resonances and glottal source wave characteristics, and 3) a non-linear interpolation in a log vocal tract area function domain. The method employs a statistical mapping on the log vocal tract area function domain and the cepstrum domain for the glottal source wave. We establish that a morphing system constructed from the proposed method improves the continuity of formants and the speech quality in the intermediate morphing rate.

Keywords- Voice Conversion; AR-HMM Analysis; Statistical Mapping; Continuity Of Phonological Identity

I. INTRODUCTION

Voice morphing or voice conversion usually refers to a technique that transforms a source speaker’s speech to a target speaker’s speech. Since the 1990s, many voice morphing techniques have been proposed \cite{1-12}. These techniques are generally restricted to a point-to-point mapping in a feature space, even when they deal with multi-speakers voices. The purpose of our research on voice morphing is to extend this restriction to area-to-area mapping in the feature space. We believe that this extension will be useful in applications such as creating peculiar voices in animated films.

One successful technique for voice morphing uses a statistical method for mapping in the cepstrum domain \cite{2-3}. However, a weakness of these methods is the discontinuity of formants, due to the fact that the relation-ship between formant transitions and the time pattern of the power spectral envelope sequence is nonlinear. In other words, the continuous interpolation of log power spectra (i.e., cepstra) does not result in continuous formant transitions. For example, when one power spectrum has a peak \( f_{1a} \), and the other has a peak \( f_{1b} \), the interpolated spectrum will have two weak peaks \( f_{1a} \) and \( f_{1b} \). This characteristic behavior will result in a deterioration of the phonological quality. Some improvements of these methods have been proposed to counter this deterioration \cite{6, 9, 12}, for example, employing line spectrum frequencies (LSF).

In this paper, we employ the estimated vocal tract area function and glottal source wave to avoid the weakness described above. Partial autocorrelation (PARCOR) coefficients can be considered to be reflection coefficients of a vocal tract area function in a certain acoustic condition \cite{13-15}, and the number of coefficients refers to the number of poles contained in the power spectrum, i.e., the formants. Based on these restrictions, interpolation in the vocal tract area domain is considered to provide reasonably continuous formant transitions. Conventional voice conversions are basically conducted in the power spectrum domain, even when using formant-like parameters. Therefore, the voice conversion using the vocal tract area function itself is an original experiment.

Estimating the vocal tract area function implies simultaneously estimating the voice source characteristics. For this purpose, we analyze speech using the autoregressive hidden Markov model (AR-HMM) \cite{16}. The AR-HMM uses an AR model to represent the vocal tract resonance characteristics and a HMM to represent the glottal source wave.

The proposed voice morphing system introduces log vocal tract area functions and cepstrum coefficients of the glottal source wave as feature parameters for combining speakers’ characteristics. The feature mapping is based on the conventional statistical method \cite{2, 3}. The resynthesis procedure to obtain the morphed speech is as follows: the voice source wave is synthesized by the analysis and synthesis software package, STRAIGHT \cite{17}, in the cepstrum domain mapping; output speech is synthesized by AR filtering of the voice source wave, in which AR coefficients are calculated from log vocal tract area function that is obtained by a linear combination of plural speakers’ log vocal tract area functions. This synthesis procedure is original and the techniques are well matched. We show that the interpolated spectral envelopes are reasonable with regard to the continuous transition of formants, which are substantially different from those obtained from the cepstrum domain. We evaluate the speech quality of the morphed speech using an opinion score, and confirm that

© American V-King Scientific Publishing

- 67 -
the quality of the interpolated morphed speech is improved by the proposed method.

II. FEATURE EXTRACTION METHOD

A. AR-HMM Analysis

The AR-HMM model is obtained by combining an AR process with a HMM, introduced to represent the glottal source wave. The AR-HMM analysis estimates the vocal tract resonance characteristics and vocal source waves by means of the maximal likelihood estimation (MLE). Therefore, components of the vocal tract resonance characteristics and those of the source waves can be naturally separated.

Conventional AR analysis assumes that the glottal source wave has a Gaussian distribution. However, this assumption can be invalid, especially when analyzing speech with a high fundamental frequency, such as that of some female speakers. In contrast, the AR-HMM analysis assumes ring-states HMM for the glottal wave, and alternately estimates the glottal source HMM and the vocal tract AR model using the MLE method. AR-HMM can estimate the vocal tract features without being biased by pitch harmonics [18].

B. Vocal Tract Area Function Estimation

We calculate the power spectrum from the AR coefficients of the AR-HMM analysis, and autocorrelation coefficients by inverse DFT of the power spectrum. The reflection coefficients (i.e., PARCOR), \( k_i, i = 1, 2, \ldots, I \), of the vocal tract area function can be calculated from those autocorrelation coefficients. In the approach described in this paper, we first use a first-order adaptive inverse filtering before the AR-HMM analysis to equalize the formant level [14].

We calculated the vocal tract area function \( A_i, i = 1, 2, \ldots, I+1 \) using the following equation [7]:

\[
A_{i+1} = 1 - \frac{k_i}{1 + k_i} A_i \tag{1}
\]

Then, we normalize the vocal tract area functions by dividing by the sum of the area values. Finally, we use log-normalized vocal tract area functions to prevent the vocal tract area functions from becoming negative and AR coefficients from being unstable.

For a linear interpolation in the vocal tract area function domain, a formant is expected to be a continuous transition. This is confirmed in Fig. 1, which shows two spectral transitions: the first is a linear interpolation in the cepstrum domain, and the other is a linear interpolation in the vocal tract area function domain. It is obvious that in the left side graph the formant transitions are not clearly changed, whereas they are continuous for the graph on the right, i.e., the transition calculated from the vocal tract area function domain.

III. CONVERSION PROCEDURE

A. Estimation of Conversion Function

The voice conversion technique used in the system is a statistical mapping from a source speaker’s voice to a target speaker’s voice. The conversion function is represented on the basis of a Gaussian Mixture Model (GMM).

\[
F(x) = \sum_{m=1}^{M} \alpha_m \mathcal{N}(x; \mu_m^x, \Sigma_m^{xx}) \tag{2}
\]

Let us denote the vector derived from a source speaker’s speech by \( x \), and the corresponding vector derived from a target speaker’s speech by \( y \).

As described in [2], the GMM-based estimation of the conversion function uses a set of time-aligned \( x \) and \( y \) vector values, \( z = [x^T \; y^T]^T \), to estimate the parameters of a joint model of Gaussian mixtures. Once the model has been trained, the \( x \) and \( y \) density of the \( m \)th component cluster is given by the following:

\[
\Sigma_m = \begin{bmatrix} \Sigma_m^{xx} & \Sigma_m^{xy} \\ \Sigma_m^{yx} & \Sigma_m^{yy} \end{bmatrix}, \quad \mu_m = \begin{bmatrix} \mu_m^x \\ \mu_m^y \end{bmatrix} \tag{3}
\]

The vectors \( \mu_m^x \) and \( \mu_m^y \) denote the mean vectors of the \( m \)th cluster of the Gaussian model estimated from \( x \) and \( y \), respectively. The matrix \( \Sigma_m^{xx} \) denotes the covariance matrix of the \( m \)th cluster of the Gaussian model estimated from \( x \), \( \Sigma_m^{yy} \) is the cross-covariance matrix.

The conversion function \( F(x) \) is given as follows:

\[
p_m(x) = \frac{\alpha_m \mathcal{N}(x; \mu_m^x, \Sigma_m^{xx})}{\sum_{j=1}^{M} \alpha_j \mathcal{N}(x; \mu_j^x, \Sigma_j^{xx})} \tag{4}
\]

where \( \mathcal{N}(x; \mu, \Sigma) \) denotes a normal distribution with mean...
vector, $\mu$, and covariance matrix, $\Sigma$. $M$ is the number of Gaussian components and $\alpha_m$ is the weight of each component cluster.

In the system, the conversion process consists of two stages: the first is the conversion of the glottal source wave characteristics and the second is the conversion of the vocal tract area resonance characteristics. The feature vector for the vocal tract characteristics is the log vocal tract area function and that for the glottal source wave characteristics is the cepstra.

B. Conversion Procedure

The system overview of the voice conversion process is shown in Fig. 2. The system can be divided into a training phase and a conversion phase. The procedure for each phase is as follows.

1) Training Phase:
   
   a) **AR-HMM analysis**: Speech samples with the same phonetic content uttered by a source speaker and a target speaker are analyzed by AR-HMM in order to estimate the AR coefficients for the vocal tract features. The glottal source waves are obtained by inverse filtering with the AR coefficients. The AR coefficients are transformed into log vocal tract area functions [19]. The glottal source wave is transformed into cepstra.

   b) **Feature vector alignment**: The feature vectors obtained in Step 1) are time-aligned using dynamic time warping (DTW) in the log vocal tract area function domain to have the optimal phonetic matching between the source and the target speech utterances. As a result, the order of the feature vector of a source speaker’s speech is the same as that of a target speaker’s. Here the feature vector of source speaker’s speech is $x$, and that of target speaker’s is $y$, as described in Equations (2) and (4).

   c) **Estimation of the conversion function**: The time aligned vectors, $x$ and $y$ are used to estimate a joint GMM as described in Section III A. The parameters of the joint GMM are then used to construct a stochastic conversion function indicated in Eq. (3). The conversion function for vocal tract features and the conversion function for glottal source wave features are estimated independently.

2) Conversion Phase

   a) **AR-HMM analysis**: Similar as in the training phase, the vocal tract and glottal source wave features are estimated using AR-HMM, for the source speaker’s utterances.

   b) **Features transformation**: The GMM-based transformation function constructed in the training Step 3) is now used to convert the source log vocal tract area functions and glottal source wave cepstra into the most likely target equivalents.

   c) **Linear interpolation**: The morphed speech features are obtained using

   $$ y = (1 - \sum_{k=1}^{K} \lambda_k) x + \sum_{k=1}^{K} \lambda_k F_k(x) $$

   where $x$ denotes the original source feature vector, $k$ indicates speaker number, and parameter $\lambda_k$ denotes the morphing rate. Vocal tract features and glottal source features are interpolated independently.

   d) **Synthesis of the glottal source wave**: The source wave for LPC synthesis is synthesized by the STRAIGHT software using the converted glottal source wave cepstra.

   e) **LPC synthesis**: The AR coefficients for the LPC synthesis are obtained by the PARCOR coefficients derived from the converted log vocal tract area functions. Finally, we filter the synthesized source wave with the AR coefficients and obtained the converted speech.

(a) Training phase

![Diagram of training process]
C. Warping of the Vocal Tract Area Function in the Direction of Vocal Tract Length

In the case of vocal tract area function domain interpolation, the differences in the lengths of the vocal tracts of individual speakers need to be taken into consideration, particularly between male and female speakers. Using a simple linear interpolation of vocal tract area functions, the narrow places of articulation between vocal tract area functions are not corresponding adequately as illustrated in Fig. 3. To solve this problem, we calculate warping functions in the direction of vocal tract length using dynamic programming (DP) to make the optimal match with each other, as shown in Fig. 4, where the vocal tract area functions are nonlinearly interpolated. This interpolation is calculated for every frame.

Let us denote the vocal tract area function of L order vector on frame t by \( v_l(t) \), \( l=0,1,2,...,L-1 \), the warping function calculated using DP by \( W_l(t) \), and the vector converted from \( v_l(t) \) using the conversion function by \( F(v_l(t)) \). Then, the interpolated vocal tract area function \( v'_k(t) \) is given as follows:

\[
v'_k(t) = (1 - \lambda)v_l(t) + \lambda F(v_{W_k(t)}(t)) \tag{6}
\]

where \( k \) is given by following:

\[
k = (1 - \lambda)l + \lambda W_l(t) \tag{7}
\]

In the DP calculation, a restricted path type is used to compute the distance efficiently and to prevent unnatural warping trajectories.

---

**Fig. 3** Interpolation in the vocal tract area function domain. Nonlinear interpolation is necessary for adequate correspondence of the place of articulation.

**Fig. 4** Illustration of the warping function for the DP matching between two vocal tract area functions.
D. Modification of the Pitch Pattern

We also applied a basic prosodic transformation to the pitch. 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 \(f_0\) and \(f_0'\) denote the log scale fundamental frequencies before the conversion and after the conversion. The values \(\mu_x\) and \(\mu_y\) are the mean log pitch of the source and target speakers, respectively. The values \(\sigma_x\) and \(\sigma_y\) are the variances of the log pitches.

IV. EXPERIMENTS

A. Experimental Condition

The audio database used for the evaluation contained 50 sentences in Japanese, each uttered by two male and two female speakers. Five sentences were used for the synthesis of the morphed speech and the evaluation. Forty-five sentences were used to train the conversion functions. The sampling frequency was 16[kHz].

The number of AR coefficients and HMM states for the AR-HMM were 22 and 10, respectively. The HMM states of the AR-HMM were connected in a ring topology. Twenty-four cepstrum coefficients were used for the re-synthesis of the source wave.

In this experiment, following three methods were compared:

Method (a): The conversion and linear interpolation of 40 cepstrum coefficients, calculated from the STRAIGHT spectrum. This method is a baseline in the experiments.

Method (b): The conversion and linear interpolation of log scale vocal tract area functions (Order 22).

Method (c): The conversion and non-linear interpolation using the DP warping towards the vocal tract length of the log scale vocal tract area functions (Order 22).

The morphed speech for the evaluation was synthesized according to method (a), (b) and (c), while changing the morphing rate. Three combinations of source and target speakers were used: male-to-male (m2m), male-to-female (m2f), and female-to-female (f2f). The female-to-male pair was not tested.

B. Observation of Formant Transitions

We observed the formant transitions associated with changes in the morphing rate of the synthetically morphed speech. Fig. 5 shows the change patterns of the power spectrum for the same analysis frame of the morphed speech.

C. Evaluation by Spectral Distortion between the Original Speech and Morphed Speech

First, in order to evaluate the quality of the morphed speech after using the proposed method, we calculated the distortion between the original speech and the morphed speech in log power spectrum domain. The spectral distortion \(D\) was defined as follows:

\[
D = \frac{1}{T} \sqrt{\sum_{t=1}^{T} \left| P_x(t) - P_y(t) \right|^2}
\]

where \(P_x(t)\) and \(P_y(t)\) indicated the log power spectral
sequence of the original speaker’s speech. $P_y(t)$ was a time warped spectral sequence of a target speaker’s speech, where the length of the utterance of the target speaker’s speech was adjusted to that of the original speaker’s speech.

The result for the case of 100% morphing rate is shown in Table 1, which indicates that the distortions for the proposed method (c) are bit lower for all cases of conversions, m2m, m2f and f2f than those of the baseline method (a).

<table>
<thead>
<tr>
<th>Original speech</th>
<th>m2m [dB]</th>
<th>m2f [dB]</th>
<th>f2f [dB]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Converted by Method (a)</td>
<td>4.54</td>
<td>4.27</td>
<td>4.25</td>
</tr>
<tr>
<td>Converted by Method (c)</td>
<td>4.46</td>
<td>4.23</td>
<td>4.18</td>
</tr>
</tbody>
</table>

D. Evaluation of Speech Quality

In order to evaluate the quality of the morphed speech, we performed subjective evaluation experiments and compared the results from methods (a) and (c) in terms of the opinion score. Eight listeners, who were all native Japanese speakers aged between twenty and forty, participated in the experiment. The opinion score used for the evaluation was a 5-point scale (1: bad, 2: poor, 3: fair, 4: good, 5: excellent). Three sentences which were not included in the training data were used to evaluate the results. To evaluate the quality of the converted speech in the intermediate rate of morphing, we synthesized both the 100% and the 70% morphing-rate speech samples for the experiment. These synthesized samples were presented randomly to the listeners through a headphone.

The experimental results are shown in Fig. 6. In the case of the 100%-morphing-rate speech, the difference of MOS scores between method (a) and method (c) depended on the combination of source and target speakers. In the cases of m2f and f2f, the scores by method (a) are bit higher than that by method(c). However, according to the statistical test, the average score for all combinations of speakers showed no significant differences between methods (a) and (c), in significance level 5%.

In the case of the 70%-morphing-rate speech, the scores for method (c) were higher than the scores for method (a) across all combinations of speakers. The average score across all combinations of speakers showed significant differences. These results mean that the proposed method maintained the continuous change of the formants in the morphed speech and improved the speech quality, which is the main advantage of the method.

These opinion scores do not always coincide with the power spectral distortions $D$, shown in Table 1. The reason is considered that the values of $D$ were calculated in the equal weight for the overall frequency range, that is, the same weight for formant bands and for the other bands. And also the relation between the pitch harmonics and local peaks of the spectral envelope is not always adequate. These points need to be investigated in the feature by increasing test speech samples and detailed analysis of synthesized speech.

V. CONCLUSIONS

This paper presented a voice morphing method that focuses on the continuity of phonological identity in overall interpolated area. In the proposed method, the characteristics of vocal tract resonances and those of glottal source waves were separated using the AR-HMM analysis, and modified independently. In the preliminary experiment, it was shown that the formant patterns changed continuously using interpolations in the log vocal tract area function domain against those of the cepstrum domain. The feasibility of the method has been confirmed by statistical conversion of several test speech samples, and reasonably established by observing the changing patterns of formants for the morphed speech. We also established that the proposed method improves the quality of the morphed speech in the interpolated area. From these results, we expect that the morphed speech can be synthesized by linear combination of multiple speakers’ voice characteristics.
ACKNOWLEDGMENT

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, Project No. 22500145, from the Japan Society for the Promotion of Science.

REFERENCES


Kazuyo Tanaka has been a professor at the University of Tsukuba, Tsukuba, Japan since 2002. He received the B.E. degree from Yokohama National University, Yokohama, Japan, in 1970, and Dr. Eng. degree from Tohoku University, Sendai, Japan, in 1984.

He was a Research Officer at Electrotechnical Laboratory (ETL), Ministry of International Trade and Industry, Japan, from 1971 to 2001, where he was working on speech analysis, synthesis, and recognition, and also served as Chief of the Speech Processing Section. In 2001, he was a leader of the Speech and Auditory Signal Processing Research Group at the National Institute of Advanced Science and Technology (AIST), Tsukuba, Japan. From 2005 to 2006, he served as a Chairperson of the Speech Committee of the IEICE. His current interests include digital signal processing, spoken document processing, and human information processing. He is a Fellow of the IEICE, and a member of IEEE, ISCA, Acoustical Society of Japan, and Japan Society of Artificial Intelligence.

Yoshiki Nambu received the B.S. degree and M.S. degree both from University of Tsukuba, Tsukuba, Japan, in 2009 and 2011, respectively. He is now with Nifty Corporation.