INTERSPEECH 2009 BRIGHTON

Voice Morphing based on Interpolation of Vocal Tract Area Functions Using AR-HMM Analysis of Speech

Yoshiki Nambu [1], *Masahiko Mikawa* [1], *Kazuyo Tanaka*[1]

[1] Graduate School of Library, Information and Media Studies, University of Tsukuba, Japan
{ynambu, mikawa, ktanaka}@slis.tsukuba.ac.jp

Abstract

This paper presents a new voice morphing method which focuses on the continuity of phonological identity overall inter- and extra-polated regions. Main features of the method are 1) to separate the characteristic of vocal tract area resonances from that of vocal cord waves by using AR-HMM analysis of speech, 2) interpolation in a log vocal tract area function domain and 3) independent morphing for the vocal tract resonances and vocal cord wave characteristics. By the morphing system constructed on a statistical conversion method, the continuity of formants and perceptual difference between a conventional method and the proposed method are confirmed.

Index Terms: morphing voice, phonological continuity, speech synthesis, inter- and extra-polation

10.21437/Interspeech.2009-494

1. Introduction

Voice morphing is a technique for continuously modifying a source speaker's speech to a target speaker's, whereas voice conversion usually means transformation from a source speaker's speech to a target speaker's. Therefore, in voice morphing, the interpolated voice is required to maintain phonological identity even between the two speakers' utterances. The present research focuses on this aspect of phonological continuity in the overall interpolated section and also that of an extrapolated region, since that will be useful for such applications as creation of peculiar voices in animation films.

Since 1990s, many techniques for voice conversion are proposed[1-7]. One successful technique is to use a statistical method for mapping from a source speaker's voice to a target speaker's in the cepstrum domain[2,3]. However, weakness of this type of methods is discontinuity of formants because 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 (i.e. cepstra) does not result in continuous formant transitions. For example, when one power spectrum has a peak (formant) $f1_a$ and the other has a peak $f1_b$, the interpolated spectrum mediated between the two will have two weak peaks $f1_a$ and $f1_b$. This type of characteristics will result in a deterioration of the phonological quality. Some improved methods have been proposed to counter this deterioration [5,6], for example, to employ line spectrum frequencies (LSF) for the interpolated features.

In this paper, we employ an estimated vocal tract area function to avoid such weakness. As is well known[8,9], 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 the certain frequency band for vowels [10]. Also, the number of the coefficients refers to the number of the 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 means simultaneous estimation of the voice source characteristics. For this purpose we introduce AR-HMM(Auto-Regressive Hidden Markov Model) 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 vocal cord wave by an HMM. It has been confirmed previously that using AR-HMM analysis, vocal tract spectrum envelopes can be precisely estimated, even for a speech wave with a high fundamental frequency.

The proposed voice morphing system introduces the log vocal tract area functions and a cepstrum sequence of vocal cord wave as feature parameters of the inter- and extra-polations, based on a statistical conversion method[2,3]. The feature parameters are then converted to cepstra, and finally output speech wave is synthesized by the synthesis-by-analysis software package STRAIGHT[12]. In this system, therefore, vocal tract characteristics and vocal cord characteristics are processed independently to obtain inter- and extra-polations.

We show that the interpolated spectral envelops are reasonable with regard to continuous transition of formants and the extrapolated spectral envelops are substantially different from those obtained from the cepstrum domain. Also we confirm that the difference can be perceptually recognized.

2. Proposed Method

2.1. AR-HMM analysis

As described above, the AR-HMM analysis estimates the vocal tract resonance characteristics and vocal source waves in a sense of maximal likelihood estimation. Therefore, components of the vocal tract resonance characteristics and those of the source waves can be naturally separated.

The AR-HMM represents the vocal tract characteristics by an AR model and the vocal cord wave by a Hidden Markov Model(HMM). Fig.1 depicts the AR-HMM model structure.

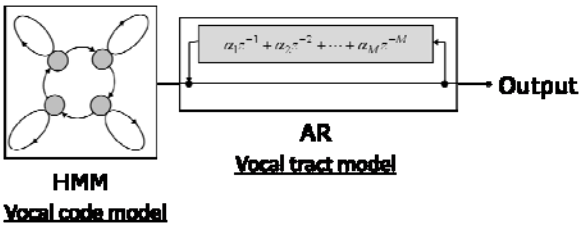

Fig. 1: Schematic diagram of AR-HMM for speech analysis.

Conventional AR 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 the 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 the ring-states for the glottal source wave, the estimated glottal source can be regarded as an approximation of the vocal cord wave. Fig.2 shows an example of AR-HMM analysis results.

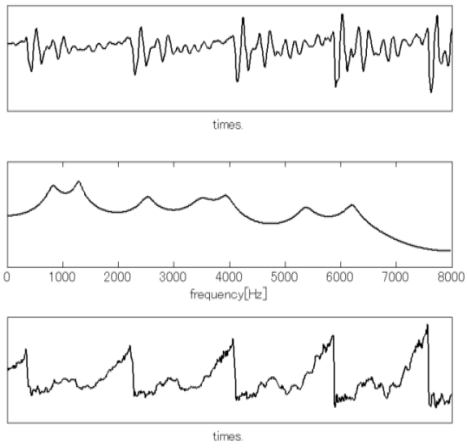

Fig. 2: An example of the results ogAR-HMM analysis: original signal (top), log power spectrum estimated by AR-HMM (middle), vocal cord wave estimated by AR-HMM (bottom). The order of AR coefficients is 19, and the number of HMM states is 15

.

2.2. Estimation of vocal tract area function

The power spectrum was calculated from AR Coefficient with AR-HMM analysis, and reflection coefficients (PARCOR) $k_i, i = 1,2,\cdots n$ of the vocal tract area function was derived from autocorrelation coefficients obtained by IDFT of the power spectrum. 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

$$A_{n+1} = 1$$

$$A_i = \frac{1-k_i}{1+k_i} A_{i+1}$$

Then, we normalized the vocal tract area functions by dividing by sum of the vocal tract area functions. Finally, we used log normalized vocal tract area functions, in order to prevent vocal tract area functions to become negative, and AR coefficients to be unstable.

By linear interpolation in vocal tract area functions domain, formant is expected to be a continuous transition. This is confirmed by Fig.3, where two spectra transitions are shown: one is a linear interpolation in cepstrum domain and the other is that in vocal tract area function domain. It is obvious that the formant transitions are continuous in the left side, i.e., in the vocal tract area function domain.

2.3. Conversion function

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

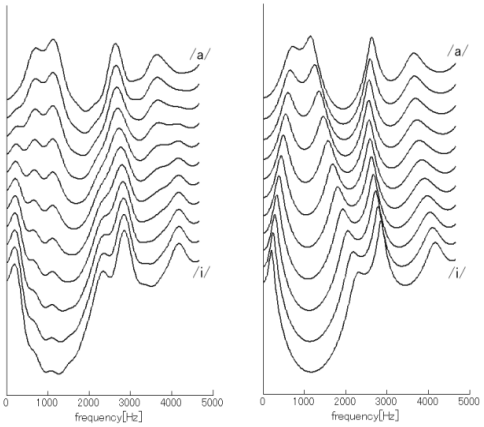

Fig.3: An example of linear interpolation: power spectra sequence obtained using 40 cepstrum coefficients(left), and that using log vocal tract area functions(right).

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

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

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

$\boldsymbol{\mu}_i^x$ and $\boldsymbol{\mu}_i^y$ denotes mean vector of i^{th} Gaussian model estimated from $\boldsymbol{x}$ and $\boldsymbol{y}$. $\boldsymbol{\Sigma}_i^{xx}$ denotes covariance matrix of i^{th} Gaussian model estimated from $\boldsymbol{x}$. $\boldsymbol{\Sigma}_i^{yx}$ is cross-covariance matrix.

As described in [#], GMM-based estimation of 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.

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

In this paper, the vocal tract characteristics is converted in log vocal tract area functions domain, and the vocal cord wave characteristics is converted in cepstrum domain.

2.4. Re-synthesis of converted voice

The system overview of voice conversion process is shown in Fig. 4, where the system consists of a training phase and 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 vocal cord wave for the vocal cord features. The AR coefficients are transformed to log vocal tract area functions. The vocal cord 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 difference 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 then build a stochastic conversion function. The conversion function for the vocal tract features and the conversion function for the vocal cord features are estimated independently.

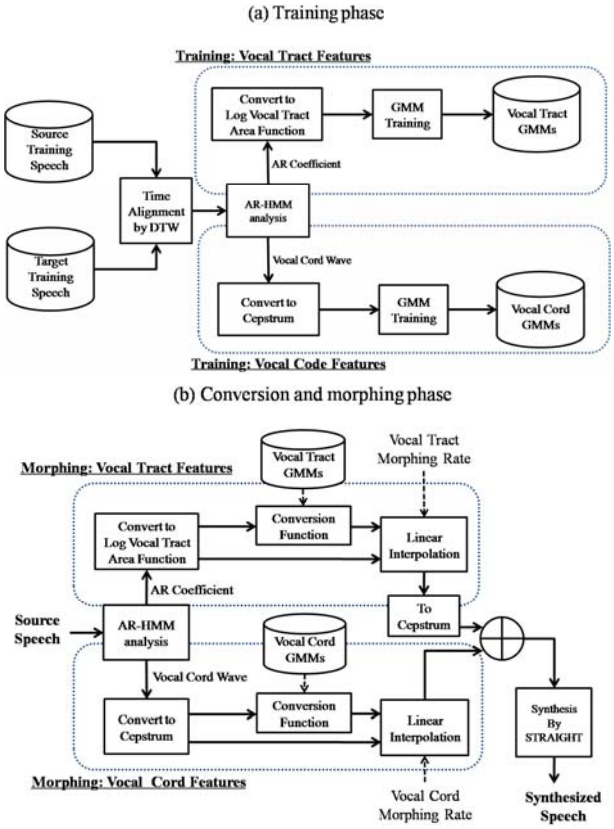

Fig. 4 Block diagram of the voice conversion system.

Conversion and morphing phase:

1) *AR-HMM Analysis*: As in training phase, the vocal tract and the 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 functions built during training is now used for converting every source log vocal tract area functions and vocal cord cepstrum into its most likely target equivalent.
3) *Resynthesis*: The features of morphed speech are obtained from $(1-\lambda)\mathbf{x}+\lambda F(\mathbf{x})$, where $\mathbf{x}$ denotes original source features, and $F(\mathbf{x})$ is the converted features obtained using the conversion functions. λ denotes morphing rate. The vocal tract features and the vocal cord features are interpolated independently. The PARCOR coefficients are obtained from the converted log vocal tract area functions, and the vocal tract cepstrum is obtained from these via power spectrum. We combained this cepstrum and the converted vocal cord cepstrum, and synthesized the morphed speech with STRAIGHT.

The method of pitch modification is conversion of average of log fundamental frequencies.

$$f_0' = \frac{\mu_y}{\mu_x} \times f_0$$

where f_0, f_0' denote log fundamental frequencies of before-conversion and after-conversion, and μ_x, μ_x are the mean log pitch of source and target speakers, respectively.

3. Experiments

3.1. Experimental condition

The speech sample set used for the voice morphing contained 50 sentences in Japanese, each uttered by two male and two female speakers. The sampling frequency was 16[kHz] and the average duration of the sentence samples was 4.7[sec]. 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 mixtures of GMM for the conversion function was 64. The following three types of features were compared:

(a) Conventional LPC analysis coefficients, order of 19.
(b) 40 cepstrum coefficients calculated from the AR coefficients, and 40 cepstrum coefficients calculated from the vocal cord wave, both estimated by AR-HMM analysis.
(c) Log vocal tract area functions (order of 19) and 40 cepstrum coefficients calculated from the vocal cord wave, both estimated by AR-HMM analysis.

The morphed speech was synthesized by using features (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. Evaluation of conversion quality

In order to evaluate conversion quality, we compared target speaker's original speeches and synthesized morphed speeches when morphing rate was 100%, using a log power spectral distortion measure. The results are shown in Table 1.

Table.1 Spectral distortions average of all speakers' combinations: (a)conventional LPC only, (b) cepstra calculated from AR coefficients and vocal cord waves by AR-HMM analysis, and (c)cepstra calculated from log vocal tract area functions and vocal cord waves by AR-HMM analysis. (c)-1and (c)-2 respectively indicate the vocal tract features and the vocal cord features only in (c). (p)morphing rate 0% and (q)morphing rate 100%.

	(p)0%[dB]	(q)100%[dB]	(q)-(p)[dB]
(a)	7.23	6.24	-1.00
(b)	7.15	5.77	-1.38
(c)	7.28	5.90	-1.38
(c)-1	7.28	6.35	-0.93
(c)-2	7.28	6.86	-0.42

It can be observed that (b), (c) resulted in a slightly larger range of reduction of the distances than (a), showing that the speech produced by independent conversions with AR-HMM is closer to the original target speaker's speech than the speech produced by conversions with only LPC. Also, using log vocal tract area functions for conversion results in similar conversion quality as arises from using the cepstrum.

3.3. Observation of the formant transitions

We observed the formant transitions associated with changes in the morphing rate by synthetic morphed speech. Fig. 5 shows the change patterns of the power spectrum for the same analysis frame of the morphed speech, when morphing rates changed from 0% to 100% to 10% each time. In the case of interpolation in the cepstrum domain, it can be seen that the positions of the formants were unchanged. Hence when morphing rate is 50%, the formants appear in mixed positions of 0%-morphing formants and 100%-morphing formants. In contrast, for the case of the interpolation of the log vocal tract area function, it can be seen that the positions of the formants are continuously changed according to the morphing rate.

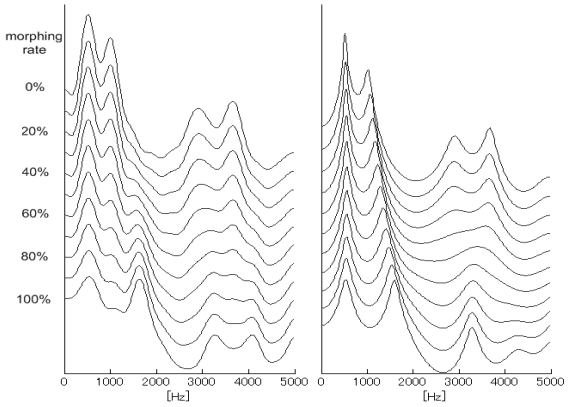

Fig.5 Changes in the power spectrum, when morphing rates changed from 0% to 100%: interpolation in the cepstrum domain(left), and interpolation in the log vocal tract area function(right).

In addition, we investigated the cases for the extrapolation. Fig. 6 shows the changes of the power spectrum when morphing rates change from -100% to 200% in the rate of 10%. From this figure, it can be found that the extrapolation in the cepstrum domain just enhances or dehances local peaks (formants) without changing their positions, whereas that obtained in log vocal tract area moves formant positions. Therefore, their voice quality will be different in each other.

We have conducted a preliminary listening test for this difference, and confirmed that the difference between the morphing speeches in the cepstrum domain and those in the log vocal tract area function domain can be clearly recognized in the case of near 200% morphing rate.

4. Conclusion

We have proposed a novel method for voice morphing, where characteristics of vocal tract resonances and those of vocal cords can be independently modified and formants are continuously changing by inter- and extra-polations in the log vocal tract area function domain. These features have been realized by using the AR-HMM analysis of speech. The feasibility of the method has been reasonably confirmed by observing spectral changing patterns in a continuous rate of morphing, and conducting a preliminary listening test in case of extrapolative morphing.

Acknowledgement

We thank Prof. Hideki Kawahara, Univ. Wakayama, for kindly licensing STRAIGHT. This research was supported in part by Grant-in-Aid for Scientific Research, JSPS.

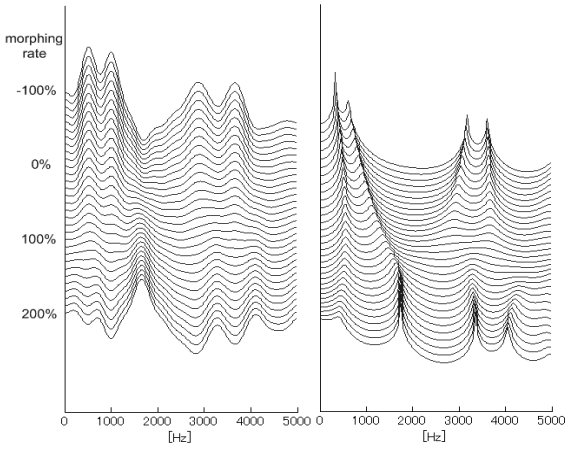

Fig.5: Changes in the power spectrum, when morphing rates varied between -100% to 200%: interpolation in the cepstrum domain(left), and interpolation in the log vocal tract area function(right).

5. 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.