GMM based Nonlinear Transformation Methods for Voice Conversion

Hoang Gia Vu*, Jae-Hyun Bae*, Yung-Hwan Oh*
* Voice Interface Lab, Div. of Computer Science, EECS., KAIST.

{ vuhg, jhbae, yhoh }@speech.kaist.ac.kr

Abstract

Voice conversion (VC) is a technique for modifying the speech signal of a source speaker so that it sounds as if it is spoken by a target speaker. Most previous VC approaches used a linear transformation function based on GMM to convert the source spectral envelope to the target spectral envelope. In this paper, we propose several nonlinear GMM-based transformation functions in an attempt to deal with the over-smoothing effect of linear transformation. In order to obtain high-quality modifications of speech signals our VC system is implemented using the Harmonic plus Noise Model (HNM) analysis/synthesis framework. Experimental results are reported on the English corpus, MOCHA-TIMIT.

I. Introduction

Voice Conversion is a technique that modifies a source speaker's utterance to be perceived as if it is produced by another target speaker. There are numerous applications of voice conversion such as personalizing text-to-speech systems, improving the intelligibility of abnormal speech of speakers, and morphing the speech in multimedia applications. Voice conversion consists in spectral conversion and prosodic modification in which spectral conversion has been studied more extensively and obtained many achievements in the voice conversion research community. In this paper, we also deal with the problem of spectral conversion only.

Many approaches have been proposed for spectral conversion including codebook mapping [1], back-propagation neural networks [6], and GMM-based linear transformation [2], [4]. Among them, the GMM-based linear transformation approaches have been shown to outperform other approaches [2], [4], [6].

Our paper is organized as follows. In section 2, we briefly describe the conventional GMM-based linear transformation methods. Then, in section 3, we identify the over-smoothing effect of linear transformation and present several nonlinear transformation methods using Radial Basis Function (RBF) networks. Our experiments on an English database are reported in section 4.

II. GMM-based Voice Conversion

In this section, we briefly describe the widely used GMM-based linear transformation methods proposed in [2] and [4].

Let $x = [x_1\ x_2\ \cdots\ x_N]$ and $y = [y_1\ y_2\ \cdots\ y_N]$ be the time-aligned sequences of spectral

vectors of the source speaker and the target speaker respectively in which each spectral vector is a p-dimensional vector. The goal of spectral conversion is to find a conversion function $F(x)$ that transforms each source vector x_i into its corresponding target vector y_i.

In GMM-based spectral conversion, a GMM is assumed to fit to the spectral vectors

$$p(x) = \sum_{i=1}^{m} \alpha_i N(x; \mu_i, \Sigma_i) \quad (1)$$

where α_i denotes the prior probability of class i and $N(x; \mu_i, \Sigma_i)$ denotes the p-dimensional normal distribution with mean μ and covariance matrix Σ defined by

$$N(x;\mu,\Sigma) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)} \quad (2)$$

The parameters of the model can be estimated by the expectation-maximization (EM) algorithm [8].

In the least squares estimation (LSE) method [2], the following form is assumed for the conversion function

$$F(x) = \sum_{i=1}^{M} P(C_i \mid x)\left[\nu_i + \Gamma_i \Sigma_i^{-1}(x - \mu_i)\right] \quad (3)$$

where $P(C_i/x)$ is the probability that x belongs to the class C_i. The parameters ν_i and Γ_i are estimated from training data by the linear least squares estimation method. However, in (3) the terms μ_i and Σ_i play no special roles in the linear transformation of x. So (3) can be simplified as

$$F(x) = \sum_{i=1}^{M} P(C_i \mid x)[b_i + A_i x] \quad (4)$$

and we also refer to (4) as the LSE method.

An alternative for the LSE method is the joint density estimation (JDE) method proposed in [4] with the conversion function

$$F(x) = E[y \mid x] = \sum_{i=1}^{M} P(C_i \mid x)\left[\mu_i^Y + \Sigma_i^{YX}\left(\Sigma_i^{XX}\right)^{-1}\left(x - \mu_i^X\right)\right] \quad (5)$$

LSE and JDE methods are theoretically and empirically equivalent. Therefore, in this paper we just use the LSE method as the spectral conversion algorithm for our baseline system.

III. Nonlinear GMM-based Voice Conversion Algorithms

Although GMM-based linear transformations have been shown to outperform other methods, our experiments shows that in some cases it is inadequate to model the conversion function by a linear transformation since the correlation between source and target vectors are small. Therefore, we attempt to model the conversion function by a nonlinear transformation function using GMM.

In this research, we present two methods using RBF networks since RBF networks have been shown to possess the property of best approximation [9] and it is easy to incorporate GMM into an RBF network.

The first method is a refinement of the approach proposed in [6]. As illustrated in Figure 1, an RBF network normally consists of 3 layers with p inputs, m hidden nodes, and n outputs. A p-dimensional input vector x is applied to all the basis (response) functions in the hidden layer. The outputs of the hidden layer (i.e., $h_i(x)$) then are linearly combined to form the output of the network

$$y_k(x) = \sum_{i=1}^{m} h_i(x) w_{ki} + w_{k0}, \quad k = 1,2,...,n \quad (6)$$

or simply in matrix form

$$y = W \times h(x) \quad (7)$$

where $h(x)$ is the $(m+1)\times 1$ vector $[1\ h_1(x)\ \cdots\ h_m(x)]^T$, W is a $(m+1)\times n$ weight matrix. The weight matrix W is estimated by the linear least squares estimation method.

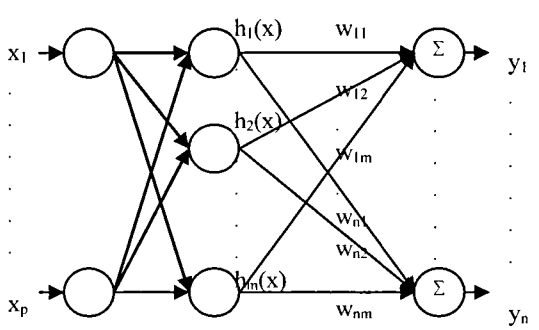

Figure 1: Structure of an RBF network.

In RBF networks, the choice of basis functions plays an important role in the success of approximation problems. The widely used basis functions include Gaussian functions and spline functions whose parameters are determined empirically as in [6]. In this paper, unlike [6], we use a more principled RBF approach presented in [8] in which the basis functions are the normal probability density functions of the GMM of the source speaker. Specifically, we fix the number of basis functions as the number of mixtures of a GMM and then estimate the parameters of the GMM using the EM algorithm. The basis functions then have the form

$$h_i(x) = \exp\left[-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)\right] \quad (8)$$

Note that (8) differs from (2) in the constant term $\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}$ since the basis functions need to be normalized.

The second proposed method is a more generalized version of (4) in which each linear transformation $A_i x + b_i$ is replaced by a nonlinear transformation using the RBF network proposed above. Hence the piecewise linear transformation function in LSE is replaced by a piecewise nonlinear transformation function

$$F(x) = \sum_{i=1}^{M} P(C_i \mid x)[W_i h(x)] \quad (9)$$

where the term $W_i h(x)$ is a nonlinear transformation as in (7), $h(x) = [1\ h_1(x) \cdots h_m(x)]^T$ and $h_i(x)$'s are given in (8) ($i = 1, \cdots, m$).

The parameters of the transformation function are also estimated by using the linear least squares estimation method as in the case of (4).

IV. Evaluation Experiments

1. Experimental Corpus and Features

To evaluate our system, we perform male-to-female and female-to-male conversions using the MOCHA-TIMIT corpus [9]. For each speaker we select 30 sentences as our training set which contains about 6000 vectors. Our evaluation set consists of 10 sentences.

We use Bark-scaled, 16th order line spectral frequencies (LSFs) as our spectral features due to its better interpolation properties compared with other features in voice conversion [4]. The HNM analysis method [3] is employed to compute the LSFs feature.

2. Objective Evaluation

To objectively measure the performance of our system, we used the error measure proposed in [6]

$$E_{LSF}(A,B) = \frac{1}{N}\sum_{k=1}^{N}\sqrt{\frac{1}{p}\sum_{i=1}^{p}\left(L_A^{k,i} - L_B^{k,i}\right)^2} \quad (10)$$

where

A, B: two time-aligned vector sequences
N: number of vectors of each sequence
p: LSFs order
$L^{k,i}$: LSF vector component i in vector k

Table 1 shows the errors for the 3 methods LSE, RBF, and piecewise RBF for various numbers of mixtures. The error between the source and the target spectra is 0.12. Therefore, all the conversion methods succeed in decreasing the errors.

Table 1: Errors between converted and target spectral vectors ($x10^{-2}$). M-F is male-to-female, F-M is female-to-male conversion. m is the number of mixtures.

m	LSE		RBF		Piecewise RBF	
	M-F	F-M	M-F	F-M	M-F	F-M
1	7.82	7.29	10.6	9.8	10.6	9.8
2	7.73	7.15	10.3	9.71	10.0	9.02
4	7.64	7.05	9.98	9.65	9.05	8.1
8	7.54	6.97	9.73	9.32	8.10	7.61
16	7.51	6.97	9.64	9.1	7.83	7.26
32	7.54	6.95	9.51	9.0	7.54	6.94

3. Interpretation

Our experiments show that

- The results of the LSE method are comparable to those reported in literature, e.g.,

[2], [6]. Our errors are slightly higher because the experimental database is not highly time-aligned.

- For small number of mixtures the LSE method outperforms the nonlinear transformation functions. This can be explained that a linear function fits better to the data than does a simple nonlinear function. As the number of mixtures increases, both the errors for linear and nonlinear functions decrease. For large number of mixtures (e.g., 32) the piecewise RBF yields results comparable to or slightly higher than the LSE method.
- The RBF method always results in the highest errors. This can be explained that a simple global nonlinear function (here is a RBF with a small number of basis functions) does not approximate well as a piecewise linear function. To increase the accuracy of the approximation we need a larger number of basis functions which means a large amount of training data. Instead, we can approximate by a piecewise nonlinear function with a smaller number of parameters as in the case of piecewise RBF method.

V. Conclusion

In this paper, we propose two GMM-based nonlinear transformation methods for voice conversion. Experiments show that the piecewise RBF method is comparable to the linear transformation methods and in some cases results in a slightly higher accuracy. Using an RBF network only to model the transformation function is shown to be the worst method.

References

[1] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," *Proc. ICASSP*, 1988, pp. 655-658.

[2] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Speech and Audio Processing*, Vol. 6, No. 2, March 1998

[3] Y. Stylianou, "Harmonic plus Noise Models for speech combined with statistical methods for speech and speaker modification", Ph.D Dissertation, ENST Paris, Jan. 1996

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

[5] A. Kain, "High resolution voice transformation," Ph.D dissertation, OGI, 2001.

[6] G. Baudoin and Y. Stylianou, "On the transformation of speech spectrum for voice conversion," *Proc. ICSLP*, 1996, pp. 1405-1408.

[7] T. Watanabe, T. Murakami, M. Namba, T. Hoya, and Y. Ishida, "Transformation of spectral envelope for voice conversion based on Radial Basis Function networks," *Proc. ICSLP* 2002, pp. 285-288.

[8] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Stat. Soc. B*, vol. 39, pp. 1-22 and 22-38, 1977.

[9] C. Bishop, "*Neural networks for pattern recognition*," Clarendon Press, Oxford, 1995.

[10] A. Wrench, The MOCHA-TIMIT articulatory database, *http://www.cstr.ed.ac.uk/artic/mocha.html*, Queen Margaret University College, 1999.