Journal of the Korean Society of Marine Engineering, Vol. 38, No. 2 pp. 182~187, 2014 ISSN 2234-7925 Print
http://dx.doi.org/10.5916/jkosme.2014.38.2.182 ISSN 2234-8352 Online

A nonlinear transformation methods for GMM to improve over-smoothing effect

Yi Geun Chae†

(Received January 21, 2014 ; Revised February 10, 2014 ; Accepted February 17, 2014)

Abstract: We propose nonlinear GMM-based transformation functions in an attempt to deal with the over-smoothing effects of linear transformation for voice processing. The proposed methods adopt RBF networks as a local transformation function to overcome the drawbacks of global nonlinear transformation functions. In order to obtain high-quality modifications of speech signals, our voice conversion is implemented using the Harmonic plus Noise Model analysis/synthesis framework. Experimental results are reported on the English corpus, MOCHA-TIMIT.

Keywords: nonlinear transformation, GMM Method, RBF, Over-smoothing, Piecewise RBF

1. Introduction

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 and others **[1]**. Basically, Voice conversion consists of 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 as shown in **[2]**, back-propagation neural networks, and GMM-based linear transformation. Among them, the GMM-based linear transformation approaches have been shown to outperform other approaches (refer to **[3]-[5]**).

We briefly describe the conventional GMM-based linear transformation methods and also, the over-smoothing effect of linear transformation is presented. And following sections, we describe nonlinear transformation methods using Radial Basis Function (RBF) networks and propose a localized transformation function using RBF networks. We experiments our algorithm with MOCHA-TIMIT and compare with previous method and our method compactly.

2. GMM-based Voice Conversion

Let x 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 vector

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

where α_i denotes the prior probability of class i and

† Corresponding Author: Department of Computer Engineering, College of Engineering, Kongju National University, Seobuk-Gu Cheonan-Daero 1223-24, Cheonan, Chungnam, 331-717, Korea, E-mail: ygchae@kongju.ac.kr, Tel: 041-521-9233

$N(x,\mu_i,\Sigma_i)$ 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. In the least squares estimation (LSE) method, the following form is assumed for the conversion function

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

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 **Equation (3)** the terms μ_i and Σ_i play no special roles in the linear transformation of x. So **Equation (3)** can be simplified as

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

and we also refer to **Equation (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

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

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.

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.

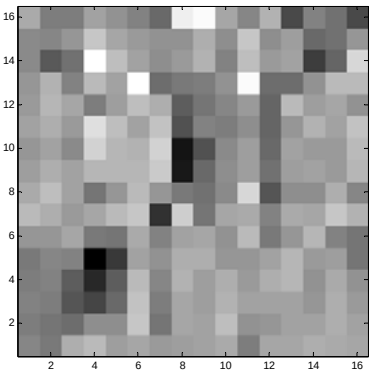

Figure 1: Correlation coefficients of source and target vectors (16^{th} order LSFs) (the darker cell is the lager element)

In **Equation (5)**, the correlation between the source vector x and the target vector y is the term $\Sigma_i^{YX}(\Sigma_i^{XX})^{-1}$. In statistical terms, the correlation coefficients determine the linear association between the two vectors. However, our experiments shows that in many cases the correlation coefficients in this term are very small, meaning that modeling the relationship between x and y by a linear function is inadequate. In our experiments, nearly 90% of the elements have values less than 0.1. Moreover, over 50% of the elements are smaller than 0.01.Due to the small values of this correlation term, the converted vectors, $F(x)$ are usually close to $\sum_{i=1}^{M} P(C_i|x)[\mu_i^y]$. This means that whatever the source vector is, the converted vector is very close to the sum of weighted means of target vectors. As a result, the converted speech seems to be over-smoothed.

3. Nonlinear GMM-based Transformation

3.1. Nonlinear Transformation Function: RBF

A GMM-based nonlinear transformation has been proposed using Radial Basis Function (RBF) networks by **[6]**, which is a refinement of the Baudoin's model **[7]**. RBF network is a nonlinear interpolation technique posing the property of best approximation as shown in **[8]**. **Figure 2** shows the structure of a typical three-layer RBF network with p inputs, one hidden layer containing m nodes corresponding to m basis functions, and q outputs. The input vector $x=[1,x_1,x_2,\cdots,x_p]$ is applied to all the basis functions in the hidden layer. The output of each hidden node $h_i(x),(i=1,\cdots,m)$ is a nonlinear function called a basis or response function. The output of the hidden layer is weighted by a weighted vector $w_k=[w_{k0},w_{k1},\cdots,w_{km}]^T$ (note that w_{k0} is the bias term). Then, the output of the network $y=[y_1,y_2,\cdots,y_q]$ is a weighted sum determined by

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

or simply in matrix form

$$y_i=W\times h(x) \tag{6}$$

where $h(x)$ is the (m+1)-dimensional vector and W is a $(m+1)\bullet q$ weight matrix.

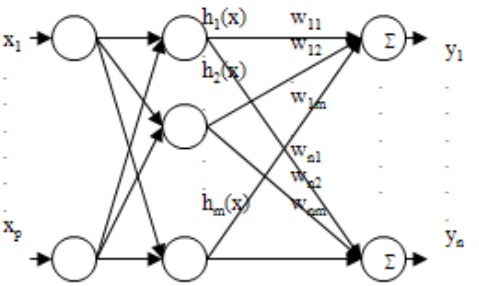

Figure 2: Structure of an RBF network

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 **[7]**. In this paper, unlike Baudoin's model, we use a more principled RBF approach presented by **[9]** 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

$$h_i(x)=\exp[-\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)]. \tag{7}$$

Note that **Equation (7)** differs from **Equation (2)** in the constant term $1/((2\pi)^{p/2}|\Sigma|^{1/2})$ since the basis functions need to be normalized.

The drawback of the RBF method is that it may be difficult to approximate a complex relationship by a global nonlinear function. Therefore, in the next section we propose a refined approach where we approximate the relationship between source and target features by a mixture of locally nonlinear functions, each of which is approximated by an RBF.

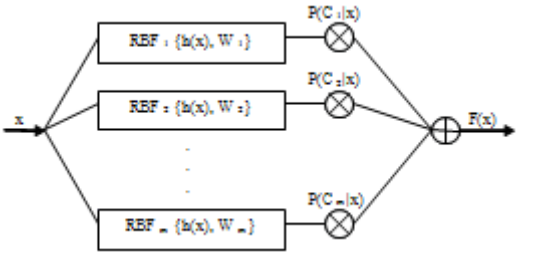

Figure 3: Transformation function using Piecewise RBF. Nonlinear functions are adopted locally. W_i is a linear transformation of $h(x)$ and (ν_i,Γ_i)

3.2. Piecewise RBF

In this section, w e propose a more generalized version of the LSE method where each local linear function is substituted by a local nonlinear function which is modeled by an RBF network. However, since it is hard to determine a different set of basis functions for each local RBF, we use the same set of basis functions $h(x)=[1,h_1(x),h_2(x),\cdots,h_m(x)]$ for every local RBF (see **Figure 2**). Similar to the RBF transformation method above, here the basis function $h_i(x),\ (i=1,\cdots,m)$ is the "normalized" Gaussian distribution density in **Equation (7)**. Consequently, the transformation function has the following form

$$F(x)=\sum_{i=1}^{m}P(C_i|x)[\nu_i+\Gamma_i h(x)] \tag{8}$$

The transformation function $F(\bullet)$ is entirely defined by the p-dimensional vectors ν_i and the $P\times P$ matrices Γ_i for $i=1,\cdots,m$.

These parameters are estimated by linear least squares estimation on the training data so as to minimize the total error

$$\epsilon=\sum_{t=1}^{n}\parallel y_t-F(x_t)\parallel^2 \tag{9}$$

Specially, the least squares optimization of the parameters is the solution of the following set of linear equations

$$y_t=F(x_t)=\sum_{i=1}^{m}P(C_i|x_t)[\nu_i+\Gamma_i h(x_t)] \tag{10}$$

for all $t=1,\cdots,n$. In the matrix form, **(11)** can be written as

$$Y=P\nu+\triangle\Gamma=[P\ \vdots\ \triangle]\begin{bmatrix}\nu\\ \cdots\\ \Gamma\end{bmatrix} \tag{11}$$

Where the two matrix ν and Γ are the unknown parameters of the transformation function. ν is a $m\times p$ matrix, Γ is a $m^2\times p$ matrix and Y is a $n\times p$ matrix containing the target spectral vectors as follows

$$\nu=\begin{bmatrix}\nu_1\\ \nu_2\\ \vdots\\ \nu_m\end{bmatrix},\ \Gamma=\begin{bmatrix}\Gamma_1\\ \Gamma_2\\ \vdots\\ \Gamma_m\end{bmatrix},\ Y=\begin{bmatrix}y_1\\ y_2\\ \vdots\\ y_n\end{bmatrix} \tag{12}$$

P is a $n\times m$ posterior matrix and $\triangle$ is a $n\times m^2$ matrix that depends on the conditional probabilities, the source vectors, and the GMM parameters.

The solution for **Equation (11)** is given by the normal equation

$$\left(\begin{bmatrix}P^TP & \vdots & P^{T\triangle}\\ \cdots & \vdots & \cdots\\ \triangle^TP & \vdots & \triangle^T\triangle\end{bmatrix}\bullet[P\ \vdots\ \triangle]\right)\bullet\begin{bmatrix}\nu\\ \cdots\\ \Gamma\end{bmatrix}=\begin{bmatrix}P^T\\ \cdots\\ \triangle^T\end{bmatrix}\bullet Y \tag{13}$$

The left most matrix in **Equation (13)** is symmetric but not positive definite and thus cannot be inverted using the Cholesky decomposition. Therefore, we exploit SVD to compute its pseudo inverse.

4. Experiments

4.1. Experimental Environments

We use the MOCHA-TIMIT speech database **[8]** to train and evaluate proposed system. For training, 30 sentences were used for each speaker which result in more than 6000 vectors. 10 sentences were used for evaluation. Two male and two female speakers are participated in the experiments. We perform eight conversion tasks, four male-to-female, and four female-to-male conversions. **Table 1** show the speaker combination used in our experiments. There are two important distances in voice conversion: the trans-

formation error $E(t(n), \hat{t}(n))$ and the inter-speaker error $E(s(n),t(n))$ where $s(n), t(n), \hat{t}(n)$ denote the source, target, converted speech respectively. All the errors are conceptual and cannot be measured directly. In this experiment, these errors are approximated by objective measures as

$$E_{LSF}(A,B) = \frac{1}{n}\sum_{i=1}^{n}\sqrt{\frac{1}{p}\sum_{k=1}^{p}(L_A^{i,k}-L_B^{i,k})^2} \quad (14)$$

where A,B is the two sequences of LSF vectors, n is the number of vectors in each vector sequence, p is the order of LPC and $L^{i,k}$is the kth component of ith LSF vector.

Table 1: Speaker Combination.

source/target	M1	M2	F1	F2
M1			√	√
M2			√	√
F1	√	√		
F2	√	√		

To take into account the inter-speaker errors, we define the LSF performance index as **Equation (15)**.

$$P_{LSF} = 1 - \frac{E_{LSF}(t(n),\hat{t}(n))}{E_{LSF}(s(n),t(n))} \quad (15)$$

The Performance index defined as **Equation (15)** uses normalized error to compare the performance of different voice conversion tasks across the different speaker combinations. The performance index P_{LSF} is 0 for a simple copy of source speech to the output without conversion. In the case of producing the exact target speech, P_{LSF} is 1. Although E_{LSF} and P_{LSF} are not the standard measures of error between two speech signals, they can be applied to input and output parameters of the conversion system directly. This is the reason why we use them as objective measures.

4.2. Experimental Results

In the experiments, we investigated the influence of the number of mixture components m on the performance of conversion system. LSE and RBF were used as baseline systems and compared with proposed piecewise RBF method. Experiments were performed while increasing m as 1, 2, 5, 8, 16, 32, 64, 128 with fixed LPC order $p = 16$.

Table 2 shows the experimental results averaged over all speaker combinations. Experimental results show that when the number of mixtures is increased, proposed method gives better results than the baseline systems. This can be interpreted as for the small number of mixtures, linear transformation methods gives better results, however for the large number of mixtures, local non-linear transformation function gives better conversion results by removing the drawbacks of linear and global non-linear transform functions.

Table 2: Performance index P_{LSF} averaged over all speaker combinations

Number of components	LSE	RBF	Piecewise RBF
1	0.35	0.13	0.13
2	0.36	0.14	0.17
4	0.36	0.16	0.23
8	0.37	0.18	0.31
16	0.37	0.19	0.35
32	0.37	0.21	0.38
64	0.38	0.22	0.39
128	0.38	0.23	0.40

5. Conclusions

In this paper, we propose GMM-based piecewise nonlinear transformation methods for voice conversion. Experiments show that the piecewise RBF method is comparable to the linear transformation methods and when the large number of mixtures is

used, the proposed method gives a higher accuracy.

Acknowledgements

Implementation of the algorithm in this paper was a collaboration with Mr. Hoang G. Vu and Dr. Jaehyun Bae. Thank to Mr. Hoang G. Vu and Dr. Jaehyun Bae.

References

[1] S. Moon, “Enhancement of ship’s wheel order recognition system using speaker’s intention predictive parameters”, Journal of Society of Marine Engineering, vol. 32, no. 5, pp. 791-797, 2008.

[2] S. Nakamura, K. Shikano, and H. Kuwabara, “Voice conversion through vector quantization,” Proceedings of Internaltional Conference on Acoustics, Speech and Signal Processing, pp. 655-658, 1988.

[3] Y. Stylianou, O. Cappe, and E. Moulines, “Continuous probabilistic transform for voice conversion,” IEEE Transactions on Speech and Audio Processing, vol. 6, no. 2, 1998.

[4] A. Kain and M. Macon, “Spectral voice conversion for text-to-speech synthesis,” Proceedings of Internaltional Conference on Acoustics, Speech and Signal Processing, pp. 285-288, 1988.

[5] A. Kain, “High resolution voice transformation,” Ph.D dissertation, Oregon Graduate Institute of Science and Technology, 2001.

[6] C. Orphanidou, I. Moroz, and S. Roberts, “Wavelet-based Voice morphing”, Journal of World Scientific and Engineering Academy and Society, vol. 10, no. 3, pp. 3297-3302, 2004.

[7] G. Baudoin and Y. Stylianou, “On the transformation of speech spectrum for voice conversion,” Proceedings of International Conference on Spoken Language Processing, pp. 1405-1408, 1996.

[8] C. Bishop, “Neural networks for pattern recognition,” Clarendon Press, Oxford, 1995.

[9] A. Dempster, N. Laird, and D. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” Journal of Royal Statistical Society, vol. 39, pp. 1-22 and 22-38, 1977.