Proceedings of the 2007 IEEE Symposium on Computational Intelligence in Image and Signal Processing (CIISP 2007)

Voice conversion using nonlinear principal component analysis

B. Makki, S.A. Seyedsalehi, N. Sadati, *Member, IEEE*, M .Noori Hosseini

***Abstract*— In the last decades, much attention has been paid to the design of multi-speaker voice conversion. In this work, a new method for voice conversion (VC) using nonlinear principal component analysis (NLPCA) is presented. The principal components are extracted and transformed by a feed-forward neural network which is trained by combination of Genetic Algorithm (GA) and Back-Propagation (BP). Common pre- and post-processing approaches are applied to increase the quality of the synthesized speech. The results indicate that the proposed method can be considered as a step towards multi-speaker voice conversion.**

I. Introduction

SPEECH signal possesses two kinds of information; speech message part and speaker individuality part. The objective of voice conversion is to convert the speaker identity; i.e. the characteristics of the speech uttered by a (*source*) speaker are transformed in such a way that it seems as if it was uttered by a different (*target*) speaker. Various applications are available that range from multimedia entertainment to helping the people who have speech organs problems. But, the main application of voice conversion is "personalization" of Text-To-Speech (TTS) systems; because of computational problems, TTS systems are mostly designed to synthesize the speech with the voice of a single speaker. Therefore, it is essential to apply a voice conversion system to convert the synthesized "speaks" to the voice of the other speakers [1].

This technique has been developed in the past and several approaches have been proposed that are mostly based on:

1- vector quantization and codebook mapping [2]
2- statistical methods such as Gaussian Mixture Model (GMM) [3] and Hidden Markov Model (HMM) [4]
3- neural networks [5].

Moreover, some pre- and post-processings such as energy equalization and pitch refinement have been proved to be useful in improving the performance of the voice conversion systems [6].

Reviewing the previous proposed methods, one can notify some challenging problems such as:

1- Loss of speech naturalness. In several VC systems, especially in those which are based on codebook mapping, speech quality of the converted utterances is unsatisfactory and some discontinuities may be generated in the reproduced speech [7].
2- Incomplete conversion. Subjective tests indicate that in some VC systems the speech signal is converted but not to the voice of the desired target speaker [8].
3- Data collection and time alignment. Many VC systems, especially GMM-based ones, are not able to deal with source data without its corresponding aligned target data. However, it is often difficult or impossible to collect a large parallel corpus [9]. Moreover, Dynamic Time Warping (DTW), that is normally applied to obtain the required alignment, causes some distortions especially when the speakers are very different [7].
4- Proper feature selection. The factors conveying information of the speaker individuality are not known precisely and it makes it difficult to select proper features for conversion [10].

Principal Component Analysis (PCA) is one of the best data analysis methods applied in various fields [11-14]. The goal of PCA is to find the principal elements from a set of databases, along which the data exhibits the largest variance. In other words, PCA reduces the dimension of the data by finding a few orthogonal vectors that produce the best representation of the full data. Although PCA has been proved to be the optimal linear transformation for the reconstruction of a dataset, however its efficiency is limited due to its linearity, especially when one should deal with the datasets that exhibit nonlinear characteristics. This problem has been addressed by many researchers and some nonlinear forms of PCA [12], [14] have been recently developed among which, as will be mentioned in the next sections, nonlinear auto-associative neural network is one of the simplest ones [14]. Voice conversion can be carried out if the principal components containing speaker individuality information can be separated from ones that convey information about the message of the speech. This is the idea on which this paper is based and will be explained in detail in the following.

This paper is organized as follows. In section II the schematic of a voice conversion system is presented. A brief

This paper was supported by the Telecommunication Research Center of Iran.

B. Makki is with the Biomedical Engineering Department of Amirkabir University of Technology, P.O. Box: 15875-4413, Tehran, Iran, phone: 009864542350, fax: 009866468186, (e-mail: behrooz_makki@yahoo.com).

S. A. Seyedsalehi is with the Biomedical Engineering Department of Amirkabir University of Technology, Tehran, Iran, e-mail: (ssalehi@cic.aut.ac.ir).

N. Sadati is with the Electrical Engineering Department of Sharif University of Technology, P.O. Box: 11363-9369, Tehran, Iran, (e-mail: sadati@sina.sharif.edu).

M. Noori Hosseini is with the Biomedical Engineering Department of Amirkabir University of Technology, P.O. Box: 15875-4413, Tehran, Iran, (e-mail: monanoori@gmail.com).

review of the PCA and the neural network applied for deriving PCA are given in section III. Then, the training algorithm of the neural network is presented in section IV. Section V describes the conversion procedure and the experimental results are reported in section VI. Finally, section VII concludes the paper.

II. BLOCK DIAGRAM OF A VOICE CONVERSION SYSTEM

As it can be seen in Fig.1, a voice conversion system would require three components:

1- Voice feature extraction. Since LPC (linear predictive coding) and LSF (line spectrum frequency) can be converted to the speech signal, most of the proposed methods employ them as the basic features. However, despite the proper performance of these techniques, the signal processing extracting these parameters causes some distortions that reduce the overall quality of the system [15]. Moreover, they do not correspond to the human's perception of the speech signals [6]. In this perspective and because of the nature of the PCA which reduces the dimension of the signal, there is no need to use them (Although, to some extent, the LPC parameters are useable too). Therefore, we use the fast Fourier transform (FFT) of the speech signal as the input of the neural network.

2- Model estimation. Several alternatives can be applied to derive the mapping function. Here, due to nonlinear and continuous nature of neural networks, a feed-forward neural network is used to capture the voice conversion.

3- Speech synthesis. Finally, the converted features return to the acoustic space and some refinements such as energy equalization and pitch refinement are applied in order to improve the performance of the system.

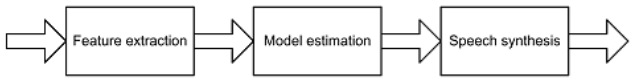

Fig. 1. Block diagram of a voice conversion system.

III. PRINCIPAL COMPONENT ANALYSIS

The significance of PCA has been much discussed [11-14]. It is a powerful tool for feature extraction and data mining which seeks to find the parameters that determine the global characteristics of a multi-dimensional signal like speech. PCA has two distinct versions; linear and nonlinear.

The aim of linear PCA is to find the *m* orthogonal vectors in the L-dimensional data space that best represent the full data; i.e. it is the best projection in the sense of mean-square error between original features and the projected ones. Eq. 1 describes it more precisely:

$$\min \sum_{X} \left[\bar{X} - \sum_{j=1}^{m} (\bar{a}_j^T \bar{X}) \bar{a}_j \right]^2 \quad (1)$$

Where $\bar{X} = [x_1, x_2, ..., x_L]^T$ is the L-dimensional input vector, and $\{\bar{a}_j, j = 1,2,...,m, m \le n\}$ are m orthogonal vectors representing principal components. The limitation of linear PCA is obvious. Since it is based on the covariance matrix of the variables, it can not capture nonlinear relationships that have higher than the second-order statistics. Hence nonlinear component analysis may be needed.

There have been several attempts to define nonlinear PCA [12]. In general, nonlinear PCA is achieved by substituting straight lines with curves. In this case, Eq.1 changes to:

$$\min \sum_{X} \left[\bar{X} - F(G(\bar{X})) \right]^2 \quad (2)$$

Where F and G are nonlinear functions. Due to the lack of unified mathematical structures, there is no exact procedure to extract the principal curves, although some approaches have been proposed [13], [14]. Fig. 2 shows the neural network that is used here to extract and transform the principal components. As it can be seen, it is a feed-forward 8-layer neural network made of sigmoid type neurons in all layers except the last one whose neurons are linear. Eq. 3 shows a sigmoid function in which net_j is the input of the j^{th} neuron in the hidden layer produced by multiplication of the input matrix $\bar{o}$ and the corresponding weighting matrix w_{ij} and θ represents the bias or center of the function.

$$net_j = \sum_i w_{ij} o_i$$

$$y_j = \frac{1}{1 + \exp(net_j - \theta)} \quad (3)$$

There are two groups of neurons in the bottleneck layer; fifty sigmoid-type neurons which must be trained in such a way that they can extract information about the message of the speech. These neurons extract the message of the signal in an unsupervised manner. The other nine neurons represent the code of the speaker identity and are trained supervised. Hence the training consists of three directions which are indicated by dash lines in Fig. 2; weights of the 3rd direction are trained to produce the speaker's code in the bottleneck layer. The weights of the directions 1 and 2 are determined in such a way that if the speak of a speaker and his/her code are given at the input and bottleneck layer, respectively, the input is reconstructed at the output layer. By injecting the code of the speaker, the neural network is forced to separate the principal components of the speaker individuality from the ones containing information about the message. The neural network is trained by combination of GA and BP which will be explained in the next section.

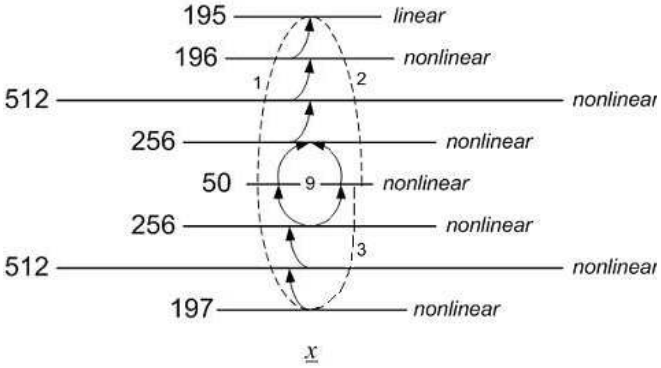

Fig. 2 The feed-forward neural network to extract and transform the principal components.

IV. TRAINING ALGORITHM

GA and BP are the main approaches used to train neural networks. Genetic Algorithm (GA) is an intelligent, stochastic searching approach based on principals of the evolution of biological systems [16], [17]. In contrast to other intelligent approaches, GA does not maintain in the local minima and requires no pre-knowledge or assumptions such as continuity or differentiability but it is time-consuming [16]. On the other hand, using back-propagation, the weights are adapted according to gradient of the error [18], [19]. The shortcoming of back-propagation is that it may fall into local minima that reduce the speed of the training process. As it is expressed in the following, proper combination of these algorithms improves the training process by preventing local minima and, at the same time, accelerates the training process.

In this perspective, the training algorithm is designed as follows. The algorithm starts with GA. After finding the near-optimum weights by GA, training proceeds by BP to reduce the error according to gradient of the mean-square error. In each iteration, the value of the error is compared with the last ones and if its reduction is not significant, it returns to GA to search not the whole weight space, but only around the current weights. Each chromosome of the GA is made of all weights of the neural network and in each iteration, only the best one (the queen) is reproduced. 50% of the next generation is produced by a small random change in the queen and the remaining ones are produced by mutation.

V. CONVERSION

After training the neural network, to convert the speech signal of the source speaker to the target one, the speech signal of the source speaker and the code of the target speaker are given at the corresponding input and bottleneck layers. We have used function $H(z) = 1 - 0.97z^{-1}$ as pre-processing and after conversion, pitch refinement and energy equalization have been applied according to [6]. In the next section, we return to report the results of the experiments illustrating the performance of the proposed method.

VI. EXPERIMENTAL RESULTS

For our experiments, the Farsdat database (the Persian standard speech database) has been used. It consists of vowels of 20 long sentences uttered by 80 adult speakers that have been recorded in almost noise-free condition, sampled at 22.05 KHz, 16 bits. 32 out of 80 speakers are female; the rest are male. A 9-bit binary code is assigned to each speaker. Vowels of 18 sentences are used for training and the rest remain for the test. 512-points Hamming window has been applied for signal segmentation and, since there is no need to above 8 KHz frequency features, the 195 first normalized points of the FFT along with their energy and one bias have been used as the input feature of the neural network. Fig. 3 shows the error function as the training proceeds. To evaluate the efficiency of the proposed training algorithm, the neural network is also trained by GA and BP separately. The results are presented in Fig. 4. Comparing Fig. 3 with Fig. 4, one can notify the efficiency of the proposed method which has accelerated the training process by preventing local minima and has also resulted in less error. In these figures, "speaker code" error refers to the error of the production of the speaker code in the bottleneck layer and the "unsupervised part" error refers to the reproduction of the input feature in the output layer. It should be noted that, although the number of iterations in the proposed method (Fig. 3) and training by GA (Fig. 4-b) are almost equal, each iteration of GA takes time twice as BP. To evaluate the performance of the method, 8 couples of speakers have been selected randomly and have been aligned by DTW. Note that the time alignment is only required to make it possible to perform objective test. The correlation coefficient between the converted speech and the desired one has been computed that was obtained %92.7. To subjectively investigate conversion performance, ABX experiment was carried out. Ten listeners participated in the experiment. They were asked "is X (converted speech) closer to A (the utterance of the source speaker) or B (the target speaker's utterance?". Each listener listened to two sentences of eight couples of speakers. The results indicate that the listeners made the correct responses for %90.6.

VII. CONCLUSION

This paper presented a new multi-speaker voice conversion system. The intelligent extraction and transformation of principal components were carried out by a feed-forward neural network that was trained by combination of GA and BP. The incomplete conversion may be due to two reasons; first, although it has been tried to separate the speaker individuality information from the message of the speech signal, there may be still some information about the speaker identity in the message neurons that causes some confusion in conversion process. The second reason is due to the nature of the NLPCA that looks for global variations lacking of local and tiny information that has been proved to

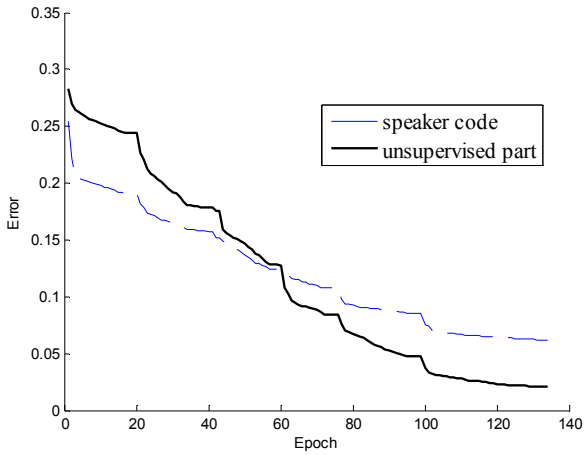

Fig. 3. Training error of the neural network trained by combination of GA and BP.

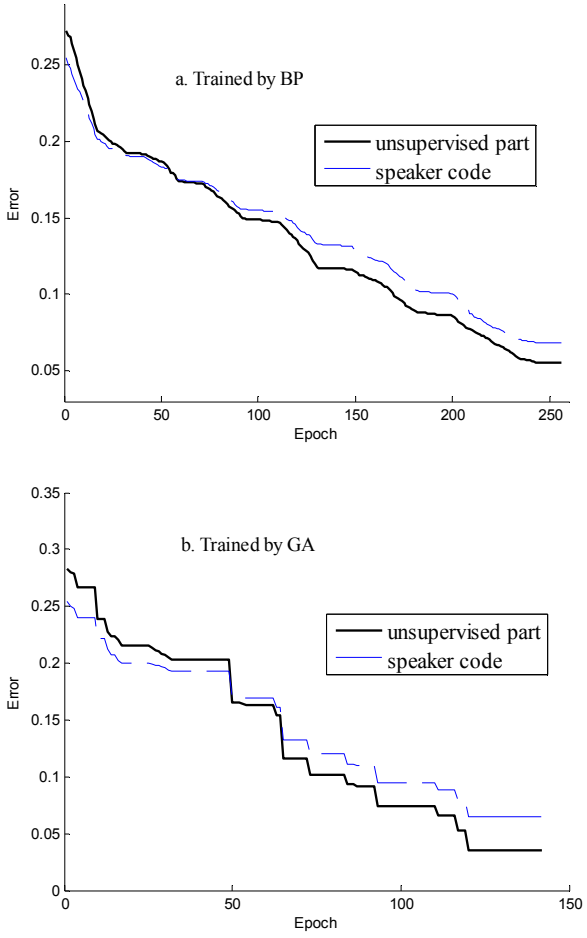

Fig. 4. Training error of the neural network trained by BP (a), GA (b).

be informative in voice conversion. Hence an additive codebook mapping approach to convert the residual signal caused by dimension reduction seems useful to improve the quality of conversion. A comparison between the proposed method and the previously developed ones could give a better insight to the readers but, unfortunately, we could not find any other Persian voice conversion method.

REFERENCES

[1] A. Kain, W. Mocan, "Spectral voice conversion for Text-To-Speech synthesis," Proc. ICASSP, pp. 285-288, May 1998.

[2] G. Zuo, Y. Chen, X. G. Ruan, W. J. Liu, "Learning Mandarin Tone Mapping Codebook for Voice Conversion," Proc. Int. Conf. on Machine Learning and Cybernetics, vol.8 pp. 4824-4828, Aug. 2005.

[3] J. Ma, W. Liu, "Voice Conversion Based on Joint Pitch ans Spectral Transformation with Component Group GMM," Proc. Int. conf. on Natural Language Processing, pp. 199-203, Nov. 2005.

[4] J. Latorre, K. Iwano, S. Furui, "Polyglot synthesis using a mixture of monolingual corpora," Proc. ICASSP, pp. 1-4, Philadelphia, USA 2005.

[5] G. Zuo, W. Liu, "Genetic Algorithm based RBF Neural Network for Voice Conversion," Fifth World Cong. on Intelligent Control and Automation, vol. 5, pp. 4215-4218, Jun. 2004.

[6] L. cheng-Yuan, J. -S. R. Jang, "New refinement schemes for voice conversion," Proc. ICME, vol. 2, pp. 725-728, July 2003.

[7] Y. Hui, S. Young, "Quality-enhanced voice morphing using maximum likelihood transformations," IEEE Trans. on Audio, Speech and language processing, vol. 14, Issue 4, pp. 1301-1312, July 2006.

[8] O. Turk, L. M. Arslan, "Robust Voice Conversion Methods," Proc. IEEE 12th Signal Processing and Communications Applications, pp. 264-267, Apr. 2004.

[9] H. Duxans, A. Bonafonte, "Estimation of GMM in voice Conversion including unaligned data," Proc. EUROSPEECH03, Geneva, Sept. 2003.

[10] M.Narendranath, H. Murthy, S. Rajendran, "Transformation of formants for voice conversion using artificial neural networks Speech communication," vol. 16, pp. 207-216. Nov. 1994.

[11] E. Oja, "Unsupervised Learning in Neural Computation," Theoretical Computer Science, Elsevier, vol. 2, pp. 187-207, 2002.

[12] B. Schoelkopf, A. Smola, "Nonlinear Component Analysis as a Kernel Eigenvalue Problem," Neural Computation, vol. 10, pp. 1299-1319, 1998.

[13] I. K. Fodor, "A Survey of Dimension Reduction Technique," technical report, UCRL-ID-148494, LLNL, 2002.

[14] A. Nicole, C. sandro, "Feedforward Neural Networks for Principal Component Analysis," Computational Statistics and Data Analysis, vol. 33, Issue; 4, pp. 425-437, Jun. 2000.

[15] D. Suendermann, A. Bonafonte, H. Ney, H. Hoege, "A Study on Residual Prediction Technique for Voice Conversion," ICASSP05, vol. 1, pp.13-16, Mar. 2005.

[16] F. Hussein, N. Kharma, R. Ward, "Genetic Algorithm for Feature Selection and Weighting, a Review and Study," 6th Int. Conf. on Document Analysis and Recognition, pp. 1240-1244, Sept. 2001.

[17] R. Myers, E. R. Hancock, "Empirical Modeling of Genetic Algorithms," Evolutionary Computation, vol. 9, No. 4, pp. 461-493, Dec. 2001.

[18] B. G. Vasudevan, B. S. Gohil, V. K. Agarwal, "Backpropagation neural-network-based retrieval of atmospheric water vapor and cloud liquid water from IRS-P4 MSMR," IEEE Trans. on Geoscience and Remote Sensing, Vol. 42, Issue 5, pp. 985-990, May 2004.

[19] Y. Bengio, "Gradient-Based optimization of hyperparameters," Neural Computation, Vol. 12, No. 8, pp. 1889-1900, Aug 2000.