VOICE CONVERSION USING ARTIFICIAL NEURAL NETWORKS

Srinivas Desai†, E. Veera Raghavendra†, B. Yegnanarayana†, Alan W Black‡, Kishore Prahallad†‡

†International Institute of Information Technology - Hyderabad, India.
‡Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA.
srinivasdesai@research.iiit.ac.in, {raghavendra, yegna}@iiit.ac.in, {awb,skishore}@cs.cmu.edu

ABSTRACT

In this paper, we propose to use Artificial Neural Networks (ANN) for voice conversion. We have exploited the mapping abilities of ANN to perform mapping of spectral features of a source speaker to that of a target speaker. A comparative study of voice conversion using ANN and the state-of-the-art Gaussian Mixture Model (GMM) is conducted. The results of voice conversion evaluated using subjective and objective measures confirm that ANNs perform better transformation than GMMs and the quality of the transformed speech is intelligible and has the characteristics of the target speaker.

Index Terms— Voice conversion, Artificial Neural Networks, Gaussian Mixture Model.

1. INTRODUCTION

A voice conversion system morphs the utterance of a source speaker so that it is perceived as if spoken by a specified target speaker. Several approaches have been proposed since the first code book based transformation method developed by Abe et. al. [1]. Researchers have tried to transform only the filter features [2] to get an acceptable quality of voice transformation. But the work presented by [3] proved the need for transformation of excitation features to attain an effective voice morphing system. A variety of techniques have been proposed by researchers for the conversion function, such as mapping code books [1], artificial neural networks [4] [5], dynamic frequency warping [2] or Gaussian mixture model [3] [6] [7] [8]. Reviewing the state-of-the-art references, we can notice that GMM based approaches are most widely used.

GMM based methods, model the joint distribution of source and target speakers speech data and the transformation in GMM follows the equations as shown in section 2.4. As the number of mixture models increases, the performance also increases [9]. The GMM transformation deals with mapping of source to target speaker space, for every feature vector obtained at 5 ms independent of its previous and next frames. Thus it introduces some level of discontinuity. To obtain a smooth trajectory of spectral vectors Maximum Likelihood Parameter Generation (MLPG) [10] is used.

Vocal tract shape between two speakers is non linear and hence Artificial Neural Networks (ANN) based method was proposed as they can perform non-linear mapping [5]. Narendranath et. al. [5] used ANNs to transform the source speaker formants to target speaker formants. Results were provided showing that the formant contour of the target speaker can be obtained using ANN. A formant vocoder was used to synthesize the transformed speech, however, no objective or subjective measures were given as to how good the transformed speech was. The use of radial basis function neural network for voice transformation was proposed in [4] [11]. However, the techniques in [5] and [4] used a carefully prepared training data which involved manual selection of vowels or syllable regions from both the source and the target speaker. This is a tedious task to make sure that the source and the target features are aligned correctly. Our work differs from the earlier approaches using ANN in the following ways:

1. The proposed approach using ANNs make use of parallel set of utterances provided from source and target speakers to automatically extract the relevant training data for mapping of source speaker's spectral features onto the target speaker's acoustic space. Thus our approach avoids any need of manual or careful preparation of data.
2. Subjective and objective measures are conducted to evaluate the usefulness of ANNs for voice conversion.
3. A comparative study is made to show that ANNs perform voice transformation better than that of GMMs.

Our work differs from GMM in the following ways.

- GMMs capture joint distribution of source and target features, whereas ANNs perform mapping of source features onto the target acoustic space.
- GMM based voice conversion systems make use of MLPG, to obtain smooth trajectories, however, the mapping abilities of ANNs provide better transformation results without the need of MLPG.

This paper is organized as follows. Section 2 describes a framework for voice conversion using ANN and GMM. The experiments and results of comparison are briefed in section 3 and conclusions will be finally presented in section 4.

2. FRAMEWORK FOR VOICE CONVERSION

2.1. Database

Most of the current voice conversion techniques need a parallel database [3] [7] [9] where the source and target speakers record the same set of utterances. The work presented here is carried out on CMU ARCTIC databases consisting of 7 speakers. Each speaker has recorded a set of 1132 phonetically balanced utterances [12]. The database includes utterances of SLT (US Female), CLB (US Female), BDL (US Male), RMS (US Male), JMK (Canadian Male), AWB (Scottish Male), KSP (Indian Male). It should be noted that the GMM based voice conversion systems needs about 30-50 parallel utterances to build voice conversion model [7]. Thus, for each speaker we took around 40 utterances as training data and a separate set of 59 utterances as testing data.

To extract features from the speech signal, an excitation-filter model of speech is applied. Mel-cepstral coefficients (MCEPs) are extracted as filter parameters and fundamental frequency estimates are derived as excitation features for every 5 ms [13]. The voice

conversion framework to transform both the excitation and the filter features from source speaker to target speaker's acoustic space is as shown in Figure 1.

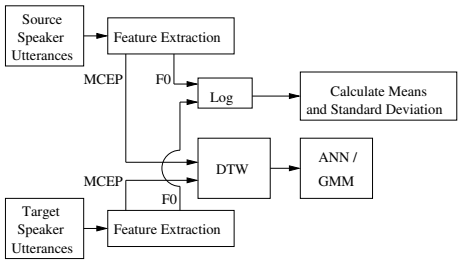

Fig. 1. *Training module in a voice conversion framework.*

2.2. Alignment of Parallel Utterances

25 cepstral coefficients called MCEP's including the zeroth coefficient are extracted for every 5 ms from the recordings of source and the target speakers. Because, the durations of the parallel utterances will typically differ, dynamic time warping (or dynamic programming) is used to align MCEP vectors between the two speakers [6] [7]. Let $X = [\vec{x_1}, \vec{x_2}, .., \vec{x_Q}]$ be the sequence of feature vectors of source speaker. Let $Y = [\vec{y_1}, \vec{y_2}, .., \vec{y_R}]$ be the sequence of feature vectors belong to target speaker. The use of dynamic programming to align X and Y provides us with set of paired feature vectors $(\vec{x_i}, \vec{y_j})$ where $1 \leq i \leq Q$ and $1 \leq j \leq R$, which can be used to capture joint distribution of source and target speakers using GMM. At the same time, this set of paired feature vectors could be used to train ANN to perform mapping from $\vec{x_i}$ to $\vec{y_j}$.

Fundamental frequency estimates are made for both speakers for a frame size of 25 ms with a fixed frame advance of 5 ms. Mean and standard deviation statistics of log(F0) are calculated and recorded.

MCEP's along with F0 can be used as input to Mel Log Spectral Approximation (MLSA) [13] filter to synthesize the transformed utterance.

2.3. Transformation of Excitation features

Our focus in this paper is to get a better transformation of spectral features and compare with GMM based transformation. Hence, we use the traditional approach of F0 transformation as used in GMM based transforms. A logarithm Gaussian normalized transformation [14] is used to transform the source speaker F0 to target speaker F0 as indicated in the equation (1) below.

$$log(f0_{conv}) = \mu_{tgt} + \frac{\sigma_{tgt}}{\sigma_{src}}(log(f0_{src}) - \mu_{src}) \qquad (1)$$

where μ_{src} and σ_{tgt} are the mean and variance of the fundamental frequencies in logarithm for the source speaker, $f0_{src}$ is the source speaker pitch and $f0_{conv}$ is the converted pitch frequency for the target speaker.

2.4. Transformation of spectral features using GMM

In the GMM-based mapping algorithm [9] the learning procedure aims to fit a GMM model to the augmented source and target speaker MCEP's. Formally, a GMM allows the probability distribution of a random variable z to be modeled as the sum of M Gaussian components, also referred to as classes or mixtures. It's probability density function can be written as

$$p(z) = \sum_{i=1}^{M} \alpha_i N(z; \mu_i, \Sigma_i) \qquad \sum_{i=1}^{M} \alpha_i = 1, \ \alpha_i \geq 0 \qquad (2)$$

where $z = [X^T Y^T]$ is an augmented feature vector of input X and output Y. Let $X = [\vec{x_1}, \vec{x_2}, .., \vec{x_Q}]$ be a sequence of Q feature vectors describing the source speaker and $Y = [\vec{y_1}, \vec{y_2}, .., \vec{y_R}]$ be the corresponding sequence as produced by the target speaker. $N(z; \mu_i, \Sigma_i)$ denotes the Gaussian distribution with mean vector μ, covariance matrix Σ and α_i denotes the prior probability that vector z belongs to the i^{th} class. The model parameters (α, μ, Σ) are estimated using the Expectation Maximization (EM) algorithm which is an iterative method for computing maximum likelihood parameter estimates. The computation of the Gaussian distribution parameters is the part of the training procedure.

The testing process involves regression, i.e, given the input vectors, X, we need to predict Y using GMMs, which is calculated as shown in the equation below.

$$\hat{y}_i = E[\vec{y_i}|\vec{x_i}] = \sum_{i=1}^{M} h_i(\vec{x})[\mu_i^{\vec{y}} + \Sigma_i^{\vec{y}\vec{x}}(\Sigma_i^{\vec{x}\vec{x}})^{-1}(\vec{x} - \mu_i^{\vec{x}})] \qquad (3)$$

where

$$h_i(\vec{x}) = \frac{\alpha_i N(\vec{x}; \mu_i^{\vec{x}}, \Sigma_i^{\vec{x}\vec{x}})}{\sum_{j=1}^{M} \alpha_j N(\vec{x}; \mu_j^{\vec{x}}, \Sigma_j^{\vec{x}\vec{x}})} \qquad (4)$$

is the a posterior probability that a given input vector $\vec{x}$ belongs to the i^{th} class. $\mu_i^{\vec{x}}, \mu_i^{\vec{y}}$ denote mean vectors of class i for the source and target speakers respectively. $\Sigma_i^{\vec{x}\vec{x}}$ is the covariance matrix of class i for source speaker and $\Sigma_i^{\vec{y}\vec{x}}$ denotes the cross-covariance matrix of class i for the source and target speakers.

2.5. Proposed method of spectral transformation using ANN

Artificial Neural Network (ANN) models consist of interconnected processing nodes, where each node represents the model of an artificial neuron, and the interconnection between two nodes has a weight associated with it. ANN models with different topologies perform different pattern recognition tasks. For example, a feedforward neural network can be designed to perform the task of pattern mapping, whereas a feedback network could be designed for the task of pattern association. A multi-layer feed forward neural network is used in this work to obtain the mapping function between the input and the output vectors. The ANN is trained to map a sequence of source speaker's MCEP's to the target speaker's MCEP's. A generalized back propagation learning law [5] is used to adjust the weights of the neural network so as to minimize the mean squared error between the desired and the actual output values. Selecting initial weights, architecture of the network, learning rate, momentum and number of iterations play an important role in training an ANN[15] . Various network architectures with different parameters were experimented in this work whose details are provided in Section 3.1.

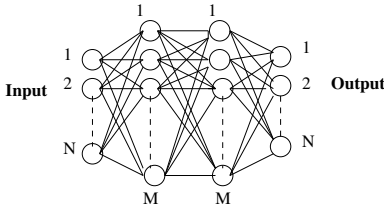

Fig. 2. *Figure showing an architecture of a four layered ANN with N input and output nodes and M nodes in the hidden layers.*

Figure 2, shows the block diagram of an ANN architecture used to capture the transformation function for mapping the source speaker features onto the target speaker's acoustic space. Once the training is complete, we get a weight matrix that represents the mapping function between the source and the target speaker spectral

features which can be used to predict the transformed feature vector for a new source feature vector.

2.6. Evaluation of spectral feature prediction

Mel Cepstral Distortion (MCD) is an objective error measure used which is known to have correlation with the subjective test results [9]. Thus MCD is used to measure the quality of voice transformation [7]. MCD is related to filter characteristics and hence is an important measure to check the performance of mapping obtained by ANN/GMM network. MCD is essentially a weighted Euclidean distance defined as

$$MCD = (10/ln10) * \sqrt{2 * \sum_{i=1}^{24} (mc_i^t - mc_i^e)^2} \quad (5)$$

where mc_i^t and mc_i^e denote the target and the estimated mel-cepstral, respectively.

3. EXPERIMENTS

3.1. ANN architecture

In building ANN based voice conversion system, an important task is to find an optimal architecture for ANN. To experiment with different ANN architectures we considered the source speaker as SLT (US female) and the target speaker as BDL (US male). As described in Section 2.1, for each of these speakers, we considered 40 parallel utterances for training and a separate set of 59 utterances for testing. Given these parallel utterances for training, they are aligned using dynamic programming to obtained paired feature vectors as explained in Section 2.2.

To get an optimal architecture, we have experimented on 3-layer, 4-layer and 5-layer networks. The architectures are provided with number of nodes in each layer and the output function used for that layer in Table 1. For instance, 25L 75N 25L means that it's a 3-layer network with 25 input and output nodes with 75 nodes in the hidden layer. L represents "linear" output function and N represents "tangential" output function.

Table 1. MCD's obtained on the test set for different ANN architectures. (No. of iterations: 200, Learning Rate: 0.01, Momentum: 0.3) Source Speaker: SLT(female), Target Speaker: BDL(male).

S.No	ANN architecture	MCD [dB]
1	25L 75N 25L	6.147
2	**25L 50N 50N 25L**	**6.118**
3	25L 75N 75N 25L	6.147
4	25L 75N 4L 75N 25L	6.238
5	25L 75N 10L 75N 25L	6.154
6	25L 75N 20L 75N 25L	6.151

From the above Table 1, we see that the four layer architecture 25L 50N 50N 25L provides better results when compared with others. Hence, for all the remaining experiments reported in this paper, the four layer architecture (25L 50N 50N 25L) is used.

3.2. Varying the Number of Parallel Utterances for Training

In order to determine the effect of the number of parallel utterances (available for training) on transformation results, we built ANN and GMM systems by varying the training data from 10 to 1073 parallel utterances. Please note that the number of test utterances are always 59. The results of the evaluation are provided for three different systems, namely,

- ANN: ANN based spectral transformation.
- GMM: Traditional GMM based system as explained in section 2.4 [6].
- GMM+MLPG: Traditional GMM based system with MLPG used for smoothing the spectral trajectory [16].

Figure 3 shows the MCD's obtained for both female to male and male to female transformation which indicate that for both the cases, spectral transformation is performed better using ANN than with GMM. It is to be noted that even modest amounts of data, say 40 utterances can also produce an acceptable level of transformation (with MCD: 6.118 for SLT to BDL and MCD: 5.564 from BDL to SLT), the subjective evaluations for which are provided in the Figure 4.

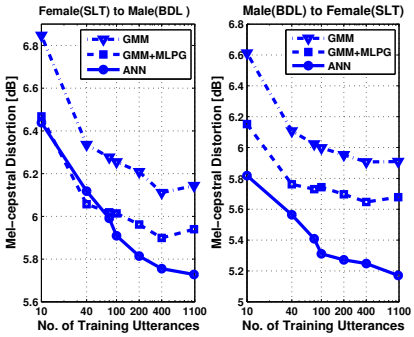

Fig. 3. *Mel-cepstral distortion as a function of the number of files used for training The results quoted for GMM are using 64 mixture components.*

3.3. Subjective Evaluation

In this section we provide subjective evaluation for ANN and GMM voice conversion systems. For these experiments we have made use of voice conversion models built from 40 parallel utterances as it was shown in Section 3.2, that this modest set produce good enough transformation quality in terms of objective measure. We conducted an Mean Opinion Scoring (MOS) test and an ABX test to evaluate the performance of the ANN based transformation against GMM based transformation.

For the ABX test, we presented the listeners with a GMM transformed utterance and an ANN transformed utterance to be compared against X which will always be the original target speaker utterance. To make sure that the listener does not get biased, we have shuffled the position of ANN/GMM transformed utterances i.e, A and B, with X always constant at the end. They were asked to select either A or B, i.e., which was perceived to be closer to the target utterance. A total of 32 subjects were asked to participate in the four experiments below, the results of which are provided in Figure 4(b). Each subject was asked to listen to 10 utterances corresponding to one of the experiments. In the MOS test, listeners evaluated speech quality of the converted voices using a 5-point scale (5: excellent, 4:good, 3:fair, 2:poor, 1:bad), whose results are provided in Figure 4(a).

1. BDL to SLT using ANN + (GMM + MLPG)
2. SLT to BDL using ANN + (GMM + MLPG)
3. BDL to SLT using ANN + GMM
4. SLT to BDL using ANN + GMM

MOS scores and ABX tests indicate that the output from ANN based system outperforms the one from GMM based system, which confirm the results of MCD's obtained in Figure 3. MOS scores also indicate that the transformed output from GMM with MLPG smoothing is perceived better than that transformed using GMM without any smoothing.

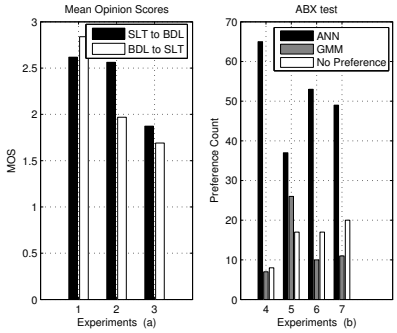

Fig. 4. ***(a)*** *- MOS scores for 1: ANN, 2: GMM+MLPG, 3: GMM.* ***(b)*** *ABX results for 4: ANN, GMM+MLPG(M->F), 5: ANN, GMM+MLPG(F->M), 6: ANN, GMM(M->F), 7: ANN, GMM(F->M)*

3.4. Experiments on Multiple Speakers

In order to show that the method of ANN based transformation can be generalized over different databases, we have provided MOS and MCD scores for voice conversion performed for 10 different pairs of speakers as shown in Figure 5. While MCD values were obtained over the test set of 59 utterances, the MOS scores were obtained from 16 subjects performing listening tests. An analysis drawn from these results show that Inter-gender voice transformation (ex: Male to Female) with MCD: 5.79 and MOS: 3.06 averaged over all experiments is better than Intra-gender (ex: Male to Male) voice transformation with MCD: 5.86 and MOS: 3.0.

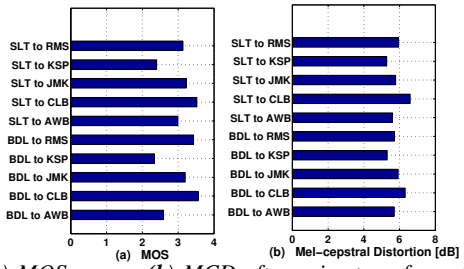

Fig. 5. ***(a)*** *MOS scores,* ***(b)*** *MCD after voice transformation using ANN on 10 different pairs of speakers*

Another result drawn from the above experiments indicate that a voice transformation between two speakers of same accent obtained better voice transformation than that of speakers from different accent. For example, the voice transformation from SLT (US accent) to BDL (US accent) obtained MCD value of 5.59 and a MOS of 3.17, while the voice transformation from BDL (US accent) to AWB (Scottish accent) obtained MCD value of 6.04 and a MOS of 2.8.

4. CONCLUSION

In this paper, we have exploited the mapping abilities of ANN and it is shown that ANN can be used for spectral transformation in the voice conversion framework on a continuous speech signal. The usefulness of ANN has been demonstrated on different pairs of speakers. Comparison between ANN and GMM based transformation has shown that the ANN based spectral transformation yields better results both in objective and subjective evaluation than that of GMM with MLPG. Our future work is currently focused on use of ANNs for voice conversion without the requirement of parallel utterances from source and target speakers.

5. REFERENCES

[1] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," *International Conference on Acoustics, Speech and Signal Processing*, vol. 1, 1988.

[2] H. Valbret, E. Moulines, and J. P. Tubach, "Voice transformation using psola technique," *International Conference on Acoustics, Speech and Signal Processing*, vol. 1, 1992.

[3] A. Kain and M. W. Macon, "Design and evaluation of a voice conversion algorithm based on spectral envelop mapping and residual prediction," *International Conference on Acoustics, Speech and Signal Processing*, vol. 2, pp. 813–816, 2001.

[4] T. Watanabe, T. Murakami, M. Namba, T. Hoya, and Y. Ishida, "Transformation of spectral envelope for voice conversion based on radial basis function networks," *International Conference on Spoken Language Processing*, vol. 1, 2002.

[5] M. Narendranath, H. A. Murthy, S. Rajendran, and B. Yegnanarayana, "Transformation of formants for voice conversion using artificial neural networks," *Speech Communication*, vol. 16, pp. 207–216, Feb. 1995.

[6] Yannis Stylianou, Olivier Cappe, and Eric Moulines, "Statistical methods for voice quality transformation," *Eurospeech*, pp. 447–450, Sept. 1995.

[7] Arthur R. Toth and Alan W Black, "Using articulatory position data in voice transformation," *Workshop on Speech Synthesis*, pp. 182–187, 2007.

[8] Tomoki Toda, Alan W Black, and Keiichi Tokuda, "Acoustic-to-articulatory inversion mapping with gaussian mixture model," *ICSLP*, 2004.

[9] Tomoki Toda, Alan W Black, and Keiichi Tokuda, "Mapping from articulatory movements to vocal tract spectrum with gaussian mixture model for articulatory speech synthesis," in *in 5th ISCA Speech Synthesis Workshop*, 2004, pp. 31–36.

[10] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM based speech synthesis," *International Conference on Acoustics, Speech and Signal Processing*, 2000.

[11] Guoyu Zuo, Wenju Liu, and Xiaogang Ruan, "Genetic algorithm based RBF neural network for voice conversion," *World congress on intelligent control and automation*, 2004.

[12] J. Kominek and A. Black, "The CMU ARCTIC speech databases," *5th ISCA Speech Synthesis Workshop*, pp. 223–224, 2004.

[13] S. Imai, "Cepstral analysis/synthesis on the mel frquency scale," *International Conference on Acoustics, Speech and Signal Processing*, 1983.

[14] K. Liu, J. Zhang, and Y. Yan, "High quality voice conversion through phoneme based linear mapping functions with straight for mandarin," *Fourth International Conference on Fuzzy Systems and Knowledge Discovery*, 2007.

[15] B. Yegnanarayana, "Artificial neural networks," *Prentice Hall of India*, 2004.

[16] T. Toda, A. W. Black, and K. Tokuda, "Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter," *International Conference on Acoustics, Speech and Signal Processing*, vol. 1, 2005.