Spectral Transformation using Artificial Neural Network

Prof. Deepali A. Khandekar[1], Prof. Priyanka Gurao[2]
[1,2]EXTC Department
K.C.C. E M.S.R
Thane, India
[1]khalsode.deepali@gmail.com, [2]priyankagurao@gmail.com

Abstract*—*Voice conversion method aims to transform the source speech signal such that the output (transformed) signal will be perceived as produced by target speaker. This paper presents a new method of voice conversion which is referred to as Spectral Transformation using Artificial Neural Network. The process of voice conversion involves transforming acoustic cues such as spectral parameters which includes the vocal tract and fundamental frequency. Spectral parameters representing the vocal tract shape, is very important parameter for speaker identity. In this paper we propose Artificial Neural Network (ANN) technique to map spectral features of a source speaker to that of a target speaker. The results of VC are evaluated on the basis of subjective and objective techniques, the experimental results conclude that ANNs perform better spectral transformation and the quality of the transformed speech is intelligible and has the characteristics of the target speaker.

Keywords*— *Voice conversion, Artificial Neural Network.

I. INTRODUCTION

VC is a method for converting source speaker utterance in to target speaker utterance. This technique has many applications in abundant aspects, including health care, text to speech, forensic applications, international dubbing, animation and entertainment, multimedia and language education. For the implementation of voice transformation two problem need to be consider what features are extracted from the essential speech signal and how to modify these features in a way so that the transform speech signal imitate target speakers voices. Extracting speaker specific characteristics such as vocal tract transfer function, glottal excitation and prosodic parameter.

Vocal tract parameter transformation has been widely used for voice personality conversion. Spectral envelope is very important parameter in VC system, until now, many approaches have been proposed by speech researchers for spectral envelope conversion such as codebook mapping[1], dynamic frequency warping[2],prosodic transformation[3], GMM[4], adaptive hmm[5],ANN[6],sub band processing [7],class mapping and unit selection techniques[8].

Pitch is another important parameter for including abundant information of speaker [9]. Conventional method of pitch transformation mainly includes mean values [10] and Gaussian model [11].f0 sequences are converted by a simple linear function but the relationship between source pitch and target pitch are actually nonlinear. In this paper Artificial Neural Network (ANN) based method was proposed to convert the spectral envelope and pitch. Vocal tract is nonlinear so ANN has magnificent agility and auto adaptability in nonlinear mapping.

This paper is organized as follows; the next section describes the proposed method in detail; section 3 describes the VC experimental setup. Experiment results and conclusions are reported in section 4 respectively.

II. ANN FOR VOICE CONVERSION

A. Artificial Neural Network(ANN)

An Artificial Neural Network (ANN) is a computer program that can recognize patterns in a given collection of data & produce a model for that data. The behaviour of ANN resembles to the human brain, its capability to learn, recall and generalize from training patterns. The characteristic of each node is simple, a network system consisted of numerous nodes have powerful self-organizing and self-learning ability, besides high-quality tolerance and prediction. 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 with supervised learning for multi-layer forward network shown in Figure.1, also known as the generalized delta rule is used. Error data at the output layer is "back propagated" to earlier ones, allowing incoming weights to these layers to be updated. Back propagation provides computationally efficient method for changing the weights in fed forward network with differentiable activation function units. As gradient descent method is used, it minimizes the total squared error of the output computed by network. What makes this algorithm different than the others is the process by which the weights are calculated during the learning network. The ANN training algorithm is given in Figure.2 where initially the weights w are chosen with random numbers between -0.5 and 0.5 or between -1 to 1 to get best results. The input vector is chosen randomly and the signal is propagated forward through the network. The local gradient for both output & hidden layer is computed and the weights are updated. The weight change is due to the combination of current gradient and the previous gradient. A momentum is added to the updated weight formula to achieve faster convergence

The network architectures with different parameters were experimented whose details are provided in next section.

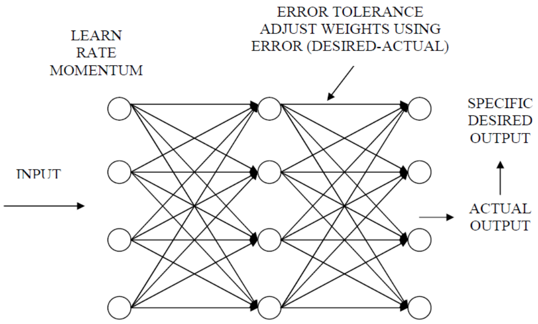

Figure.1 Back propagation network

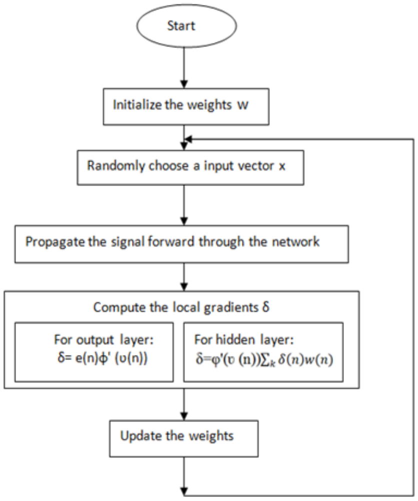
Start
Initialize the weights w
Randomly choose a input vector x
Propagate the signal forward through the network
Compute the local gradients δ
For output layer:
δ= e(n)ϕ' (υ(n))
For hidden layer:
δ=φ'(υ (n))$\sum_k \delta(n)w(n)$
Update the weights

Figure.2 ANN algorithm

B Voice Conversion Experimental Setup

The Figure.2 & 3 provide experimental setup of voice conversion system for training & testing modes. In this paper the work is carried out using parallel database where source and target speaker record same set of utterances. The work presented here is carried out on CMU ARCTIC databases which consists of 7 speakers; SLT(US Female), CLB(US Female), BDL(US Male), RS(US Male), JMK(Canadian Male), AWB(Scottish Male), KSP(Indian Male). For training; spectral and excitation parameters of the source and the target speakers were extracted as shown in Figure.3. Mel-cepstral coefficients (MCEPs) are extracted as spectral parameters and the fundamental frequency (F_0) as excitation parameters for every 5ms [11]. Even though the source & target speakers have spoken the same utterance, they vary in duration; therefore dynamic programming (Dynamic time warping) is used to align MCEP vectors between two speakers [12]. The use of dynamic programming provides with a set of paired feature vectors which is use to train ANN to perform mapping.

The fundamental frequency (f0) uses the cepstrum method to calculate the pitch period for frame size of 25ms. Mean and standard deviation statistics of log (F0) are calculated. For testing the MCEP and f0 are extracted from a test speaker in the similarly way as extracted during the training and appropriate mapping is applied using trained ANN, transformed MCEP and f0 are synthesize into speech by passing them through the Mel Log Spectral Approximation (MLSA) [11] filter as shown in Figure.4.

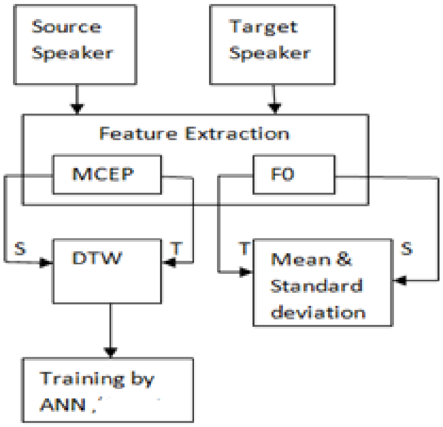

Figure.3 Training model

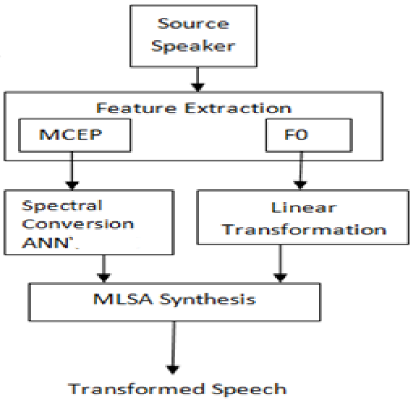

Figure.4 Testing model

III. EXPERIMENTS

A. Objective Evaluation for ANN:

For ANN based voice conversion system, the important task is to find the optimal architecture of an ANN. For a VC system based on ANN technique, we have considered two sets: Set 1: Transformation of SLT (US female) to BDL (US male), Set 2: Transformation of BDL (US male) to SLT (US female). For this experimental work, we have taken 3-layer, 4-layer and 5-layer ANNs. The architectures are provided with the number of nodes in each layer and the activation function used for that layer.

For example, 25L 75N 25L means that it is a 3-layer network with 25 input and output nodes and with 75 nodes in the hidden layer. Here, L represents "linear" activation function and N represents "tangential (tan h)" activation function. For an ANN architecture, the no. of parameters to be computed is calculated as follows: Suppose the ANN architecture is

25L50N 25L, the number of parameters (25*50)+(50*50)+(50*25)+(50+50+25) =5125. From Table.2, we see that the four layered architecture 25L 50N 50N 25L (with 5125 parameters) provides better results when compared with other architectures.

TABLE.1: MCD for different architectures of an ANN model

Sr. No	Architecture	No. of parameter	MCD	
			Set 1	Set 2
1	25L 75N 25L	3850	7.807	7.506
2	**25L 50N 50N 25L**	**5125**	**7.026**	**6.502**
3	25L 75N 75N 25L	9550	7.869	8.406
4	25L 75N 4L 75N 25L	4529	8.531	8.412
5	25L75N 10L 75N 25L	5435	8.630	8.549

B. Subjective Evaluation

To evaluate the overall accuracy of the conversion, a Mean Opinion Score (MOS) and similarity test were carried out for evaluating the similarity between the converted voice and the target voice to find the performance of ANN based transformation. 10 listeners give scores between 1 and 5 for measuring the similarity between the output of the two VC systems and the target speaker's natural utterances. The result of MOS test is provided in Figure.5 & the average similarity test is given in the Table.2 which indicates the ANN based voice conversion system performance.

TABLE 2: Average similarity scores

Transformation Method	Avg. Similarity Score	
	SLT to BDL	BDL to SLT
ANN	2.95	3.05

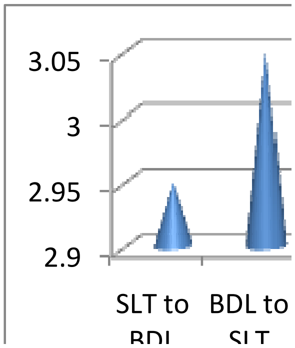

Figure.5 MOS Scores

CONCLUSION

In this paper, ANN based method was used to mapped the MCEP and f0 of the source speaker to that of target speaker. We have analyzed ANN performance based on objective and subjective evaluation. In objective evaluation we have consider different architectures of ANN model for which we have got different MCD parameters and the four layered architecture 25L 50N 50N 25L (with 5125 parameters) provides better results when compared with other architectures.

REFERENCES

[1] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, “Voice conversion through Vector quantization,” in Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing, 1988, vol. 1, pp. 655–658.

[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] 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.

[4] Y. Stylianou, O. Cappe, and E. Moulines, “Statistical methods for voice quality transformation,” in Eurospeech, 1995, pp. 447–450.

[5] E. K. Kim, S. Lee, and Y. H. Oh, “Hidden markov model based voice conversion using dynamic characteristics of speaker,” in European Conference On Speech Communication And Technology, 1997, pp 1311–1314.

[6] 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, 1995.

[7] O. Turk and L. Arslan, “Subband based voice conversion,” in ICSLP, 2002.

[8] D. Sundermann, H. Hoge, A. Bonafonte, H. Ney, and S. Narayanan, “Text independent voice conversion based on unit selection”.

[9] L. M. Arslan, “Speaker transformation algorithm using segmental codebooks”, *STASC Speech Communication*, 28(3), pp. 211-226, 1999.

[10] S.Desai,E. V. Raghavendra, B. Yegnanarayana, A. W. Black, and K. Prahallad,“Voice conversion using artificial neural networks,” in Proceedings of IEEE Int.Conf. Acoust., Speech, and Signal Processing, 2009.

[11] S. Imai, “Cepstral analysis synthesis on the mel frequency scale,” in Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing.

[12] YannisStylianou, Olivier Cappe, and Eric Moulines, Statistical methods for voice quality transformation,.*Eurospeech*, pp. 447.450, Sept. 1995