IDENTIFICATION OF FREQUENCY DOMAIN USING QUANTUM BASED OPTIMIZATION NEURAL NETWORKS

Dhivya bharathi[1], T.Karthikeyan[2], P.Hemalatha[3], P. Joy kinshy[4]

[1]*PG Scholar, Computer Science and Engineering, Knowledge Institute of Technology, Salem*
[2]*Assistant Professor, Computer Science and Engineering, Knowledge Institute of Technology, Salem*
[3]*PG Scholar, Computer Science and Engineering, Knowledge Institute of Technology, Salem*
[4]*PG Scholar, Computer Science and Engineering, Knowledge Institute of Technology, Salem*

Abstract

Voice Conversion (VC) is a zone of speech processing that contract with the conversion of the apparent speaker identity. This paper presents a new conversion method for text to speech and Voice to voice conversions. Text to Speech conversion uses phonematic concatenation for conversion and Voice conversion uses MLP Quantum Neural Network for transformations. The proposed method consists of two stages. In first stage, Features extraction of LSF and pitch residual based data from source and target speaker. In second stage where we use MLP Quantum neural network which is trained to learn the nonlinear mapping function for source to target speech transformation using the features extracted in the first phase. The proposed method can efficiently increase the quality and lack of naturalness of the converted speech. The proposed model is tested by male and female speakers with average duration.

Key Words: *Voice Conversion, Line Spectral Pairs, Quantum Neural Networks, Multi-layer Perceptron's, Phonematic Concatenation, TTS, and GMM.*

----------***----------

1. INTRODUCTION

Text-to-speech (TTS) convention transforms linguistic information stored as data or text into voice. It is far and wide used in acoustic understanding devices for blind people. Potential applications of High Quality TTS Systems are indeed numerous. Here are some examples: Telecommunications services, Multimedia, man-machine communication, Vocal Monitoring, Talking books and toys. A voice conversion system modifies the utterance of a source speaker to be perceived as spoken by a specified target speaker. While the speaker identity is transformed, the speech content (linguistic information) is left unaltered.

There are many applications of voice conversion including customizing voices for TTS systems, transforming voice-overs in adverts and films to sound like that of a well-known celebrity, enhancing the speech of impaired speakers, international dubbing, health-care, multi-media, language education, voice restoration in old documents/movies, to create a speech database in a cost-efficient manner in the field of speech synthesis and to dub character voices in television programs. Generally, VC operation can be divided into two parts, vocal tract parameter conversion and excitation signal conversion. It is believed that speaker characteristics are inherited in the vocal tract parameter. One of the most widely used spectrum conversion methods is statistical parametric approach, Gaussian Mixture Model (GMM) based algorithm. From the other point of view, another transformation paradigm was also conducted, named frequency warping.

This transformation function maps significant positions of the frequency axis (e.g., central frequency of formants) from the source speaker to the target speaker. As this method does not modify the fine spectral details of the source spectrum, it preserves very well the quality of the converted speech. However, the accuracy was less than that of GMM-based VC. Another interest growing in voice conversion field is by using Quantum Neural Networks (QNNs).This interest network will perform nonlinear mapping.

Voice transformation should be performed from source speech to target speech without losing or modifying the original speech content.VC is performed in two phases: the first one is a training phase, in this phase the speech features of both source and target speaker are extracted and appropriate mapping rules are generated for transforming the parameters of the source speaker onto those of the target speaker. In the second testing phase, the mapping rules developed in the training stage are used to transform the features of source voice signal in order to possess the characteristics of the target voice. The vocal tract and prosodic parameters are modified using signal processing algorithms. There are many applications where intra gender and inter gender voice conversion is required.

2. PREVIOUS WORK

In voice conversion, the quality and intelligibility of the Synthetic speech depends on time and spectral expansion, compression, pitch modifications, the glottal excitation, shape and also on the reproduction. Gaussian mixture models (GMMs) is used to partition the acoustic space of the source speaker into overlapping classes called as soft partitions. Using these soft partitions, a continuous probabilistic linear transformation function for vector is obtained. This transformation function contained parametric representation of the spectral envelope and modified the GMM approach in [4]. However, the quality and naturalness of the converted speech signal is found inadequate due to reconstruction of speech signal using the large number of parameters and it results into over smoothing.

Researchers have also provided the solution for over smoothing problem like hybridization of GMM with frequency warping [5] and GMM with codebook mapping. De-sai et.al. [6] have compared the performance of the ANN and GMM and it is reported that the ANN performs better than GMM. it is reported that the structural GMM performs better than GMM. GMM with GA based method [5][6] has been developed using LSP. The conversion function has been proposed using the probabilistic model based on inter-speaker dependencies and cross correlation probabilities between the source and target speaker .Chen et.al have pro-posed improved method [6][7] for VC, in which the input frames are divided into voiced and unvoiced frames. Chen et.al has proposed improved method [4] for VC, in which the input frames are divided into voiced and unvoiced frames.

The voiced frames are mapped from source to target t using GMM and unvoiced frames are stretched or compressed for conversion based on the ratio of vocal tract length (VTL) of source to target. Artificial neural network has been used for VC to exploit the nonlinear relationship between the vocal tract shape of source and target speaker [4].Line spectral pitch has been used for VC where LSP and fundamental frequency (Fo) are used to extract the source and target features of parallel set of data. Using these features the RBF based neural network is trained for an appropriate mapping function that transforms the vocal spectral and glottal excitation cues of the source speaker in to target speaker's acoustic space [3].

3. PROPOSED METHOD

The proposed algorithm consists of two phases. In first phase, we extract the Line Spectral Frequency and pitch residual based features from source and target speaker data. It is followed by the second phase where we use MLP neural network and GMM which is trained to learn the nonlinear mapping function for source to target speech transformation using the features extracted in the first phase.

3.1 Text-To-Speech Conversion by Phonemic Concatenation

Text to speech conversion is as follows:-

- GUI implementation of the Mat lab Module.
- Recording of voice samples through microphone.
- Phoneme extraction by the use of spectrogram.
- Concatenation of phonemes to create any desired word.
- Comparison of concatenated word with the original word. [2]

3.2 LSF, Pitch Residual Based Voice Conversion

When the LSF are in ascending order in the range [0, 1], the resulting filter is guaranteed to be Stable. When two LSF values are close to each other, a spectral peak is likely to occur between them which is useful for tracking formants and spectral peaks. In the training mode the source and target speech is normalized to some predetermined amplitude range and the pitch information is extracted to produce a residual, residual contains vocal tract information which is modeled by LPC. Applying the LPC analysis filter to the residual will result in the vocal tract information being removed leaving lung excitation signal.

LPC produces the results in an unstable synthesis filter. So we convert LPC parameters into LSP. We have mapped the pitch residual and LSP parameters of source speaker on to target speaker using RBF neural network with spread factor of 0.01 and error threshold of 0.00001 respectively. In the transformation phase the LSP and Pitch residual of test speech samples are projected as an input to the trained MLP model, the transformed LSP and pitch residuals are obtained. To resynthesize the speech signal, LSP parameters are reconverted into LPC, transformed speech is reconstructed using LPC synthesis, as a target speech [3].

3.3 MLP QNN Based Transformation

Quantum Neural Networks are more efficient than classical neural network for classification tasks, In that time required for training is much less for QNN. The repetition is not required by QNN; since each component networks learns only one pattern causing the training set to be learned more quickly. A QNN is shown in Figure 2. It has

n inputs in the form of qubits as $\alpha_m|0\rangle + \beta_m|1\rangle$

where $|\alpha_m{}^2| + |\beta_m{}^2| = 1$

The output $|o(k)\rangle$ of the neuron is again a qubit

$$|o(k)\rangle = \ddot{F} \sum_{m-1}^{n} (\hat{w}_m |i_m) \qquad (1)$$

Where $|o(k)\rangle$ is the output of the neuron for the k^{th} input set $\hat{F}$ is an operator that can be implemented by a network of quantum gates. It has to satisfy the unitarity condition that

$$\text{adj}(\hat{F}).\ \hat{F} = I \qquad (2)$$

$\hat{w}_m$ is the weight matrix corresponding to the m^{th} input. It is of the form

$$\hat{w}_m = \begin{bmatrix} w_m^1 & w_m^2 \\ w_m^3 & w_m^4 \end{bmatrix} \qquad (3)$$

i_m is the m^{th} input qubit and is of the form

$$\alpha_m|0\rangle + \beta_m|1\rangle \qquad (4)$$

The basic network considered in this paper is a generalized MLP (GMLP) network, which consists of an input layer, an output layer, a number of hidden nodes, and associated interconnections. Let X and Y be the input and output vectors, respectively. The GMLP network with m inputs, n_h hidden nodes, and n outputs is characterized by the equations

$$x_i = X_i, \quad 1 \le i \le m$$
$$x_i = f\left(\sum_{j=1}^{i-1} w_{ij}\, x_j\right), \quad m < i \le m + n_h + n$$
$$Y_i = x_{i+m+nh}, \quad 1 \le i \le n \qquad (5)$$

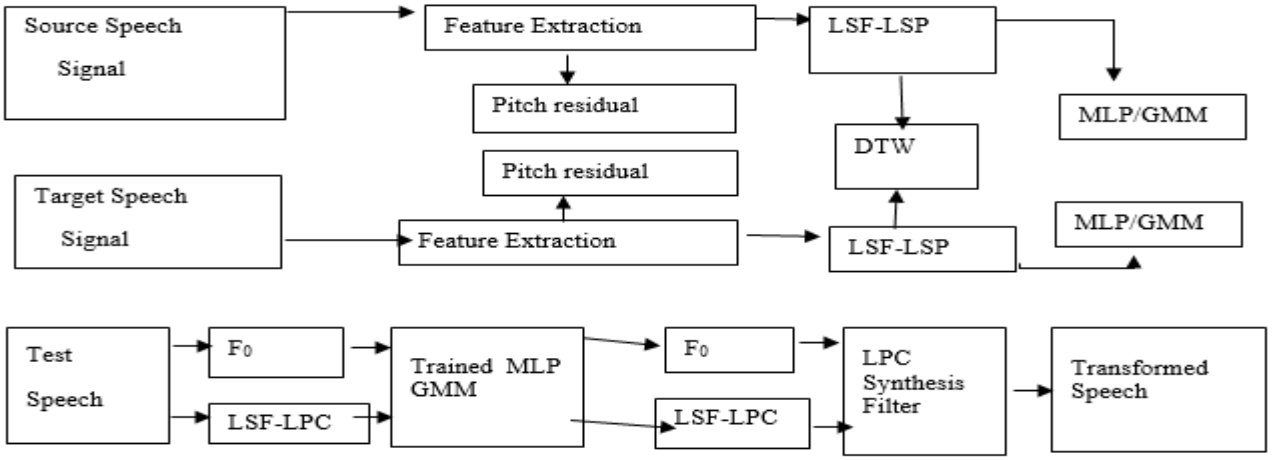

Fig 1: LSF based VC: Training and Testing model

where w_{ij} is the weight of the connection from node j to node i and f is the sigmoid function

$$f(z) = \frac{1}{1+e^{-z}} \qquad (6)$$

The output unit performs weighted sum of hidden units. The MLP Quantum neural network is trained to map LSP as a vocal tract parameters and pitch residual (F_0) as glottal parameters between source and target speaker. Winner-takes-all classification rule is used to adjust the weights of this neural network so as to minimize the mean squared error (MSE) between the desired and the actual output values. For training, the LSP of the source speaker voice are used as input and LSP of target speaker voice are used as a output of the network. The weights of the hidden and output layers of the networks are adjusted in such a way that the source speaker's patterns are converted into the target speaker's patterns. The training step is a supervised learning procedure [1].

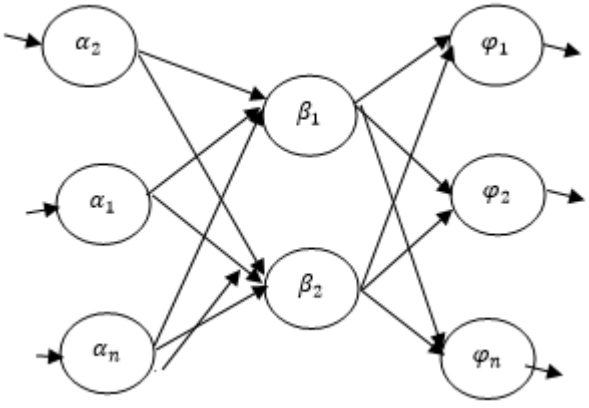

Fig 2: QNN

4. SIMULATION AND RESULT

Our algorithm is able to perform the mapping of the LSP, pitch residual, of source to that of target speaker using MLP with high accuracy. The sample results of our MLP based speech conversion algorithm for inter gender speaker are shown in the figures3. In figure 3, the left column represents the speech signal waveform of source, target and transformed in order. The right column in the figure 3 displays the Spectrogram of source, target speaker and transformed speech in order from top to bottom.

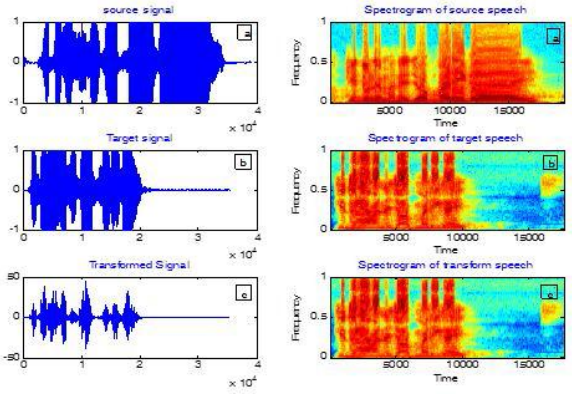

Fig 3: a) Source, b) Target and c) Transformed speech of same sentence waveform of a male-to-female speaker of MLP based speech transformation

The spectrograph is a two dimensional time and frequency graphical plot of the energy present in the signal. Figure3, display the waveforms and spectrograms for male to female voice conversion. From the visual perception of the figure it is clearly seen that the spectrogram of the target speaker is similar to that of the transformed speech. The performance of the QNN based transformation is better than RBF based transformation. The proposed algorithm is able to transform source speech to target speech successfully.

5 OBJECTIVE AND SUBJECTIVE PERFORMANCE

In this section we evaluate our algorithm based on Objective and subjective parameters. We use objective based evaluation parameters (i) mean square error (MSE).

5.1 MSE based Objective Evaluation

In this section we provide the objective evaluation for MLP based VC systems to measure the differences between the target and transformed speech signals. Since many perceived sound differences can be interpreted in terms of differences of spectral features and mean squared error (MSE) are considered to be a reasonable metrics; for both mathematically and subjectively analysis. The MSE between target audio vector *c* and transformed audio vector *b* is calculated as per below equation on a sample by sample basis.

$$E = \frac{1}{N}\sum_{i=0}^{N}[c(i) - b(i)]^2 \qquad (7)$$

The average square difference between two vectors is used to evaluate the objective performance of mapping algorithms.

5.2 Subjective Evaluation

Evaluation of the synthesized speech of desired speaker using MLP and RBF based VC systems by standard subjective test can be attributed due to inefficiency of objective evaluation techniques in defining good objective distance measures which are perceptually meaningful. In this paper subjective evaluation tests have been discussed namely i) Mean Opinion Score.

5.2.1 Mean Opinion Score

To evaluate the overall accuracy of the conversion, a listening test is carried out for evaluating the similarity between the converted voice and the target voice to find the performance of MLP based transformation. 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 results of this MOS similarity tests are provided in figure 6, which indicates that the MLP based voice conversion systems completely characterized the characteristics of the source speaker. The MLP based voice conversion has slightly more resemblance to target speech as compared to the GMM based VC system.

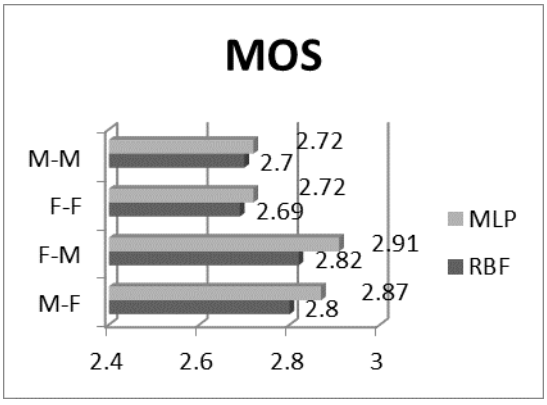

Fig 4: MOS Test

6. CONCLUSIONS

In this paper, we have proposed a novel technique using LSP as spectral features and pitch residual as glottal features. The MLP based and mapping functions are developed to map the acoustical cues of the source speaker according to that of the target speaker. The inter gender and intra gender VC is performed using proposed algorithm. We have experimentally

verified that, the fundamental frequency range of female speech is higher than male speech and vice versa. The comparative performance of RBF and MLP based models are studied using objective and subjective measures .The evaluation results indicate that MLP can be used as an alternative to the RBF based transformation model. The subjective evaluation convinced that the quality and naturalness of the transformed speech can be achieved with the proposed algorithm.

REFERENCES

[1] Tzyy-Chyang Lu, Gwo-Ruey Yu,and Jyh-Ching Juang, "Quantum-Based Algorithm for Optimizing Artificial Neural Networks"IEEE Transactions on Neural Networks and Learning Systems, Vol. 24, No. 8, August 2013.

[2] Tapas Kumar Patra , Biplab Patra , PuspanjaliMohapatra,"Text to Speech Conversion with Phonematic Concatenation", International Journal of Electronics Communication and Computer Technology (IJECCT) Volume 2 (September 2012).

[3] J.H.Nirmal, Suparva Patnaik , and Mukesh A.Zaveri, "Line Spectral Pairs Based Voice Conversion using Radial Basis Function", ACEEE Int. J. on Signal & Image Processing, Vol. 4, No. 2, May 2013.

[4] Chen Zhi ,Zhang Ling-hua"Voice Conversion Based on Genetic Algorithms",Telecommunication 2011.

[5] Daniel Erro, Asunción Moreno ,"Weighted Frequency Warping for Voice Conversion,Speech", Communication2008.

[6] Danwen Peng,Xiongwei Zhang, ,Jian Sun," Voice Conversion Based on GMM and ArtificialNeural Network" 2010 IEEE.

[7] Narpendyah W. Ariwardhani1, Yurie Iribe1, Kouichi Katsurada1 and Tsuneo Nitta1, 2 ,"Voice Conversion For Arbitrary Speakers Using Articulatory-Movement To Vocal-Tract Parameter Mapping" ,2013 IEEE International Workshop On Machine Learning For Signal Processing, Sept. 22–25, 2013.

[8] R. Xuemei and L. Xiaohua, "Identification of extended Hammersteinsystems using dynamic self-optimizing neural networks," IEEETrans.Neural Netw., vol. 22, no. 8, pp. 1169–1179, Aug. 2011.

[9] Srinivas Desai, E.Veera Raghavendra, B.Yegnanarayana, Alan W Black, Kishore Prahallad,"Voice Conversion using Artificial NeuralNetworks"in IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Taipei, Taiwan, April 2009.

[10] Srinivas Desai, Alan W. Black, B. Yegnanarayana, and Kishore Prahallad, "Spectral Mapping Using Artificial Neura Networks For Voice Conversion" IEEE TRANSACTIONS on Audio, Speech, And Language Processing, VOL. 18, NO. 5, JULY 2010.

BIOGRAPHIES

A .DHIVYA BHARATHI is a student pursuing PG Degree in Computer Science and Engineering at knowledge institute of technology, Salem. Research Interest includes Artificial intelligence and Neural Networks contact:dhivi.anand@gmail.com

T. KARTHIKEYAN obtained his Master degree from Central Queensland University, Melbourne, Australia in 2008. He is authored over 10 publications including Artificial Intelligence and Wireless Sensor Networks and he has guided 3 international level projects. He is currently an Assistant Professor in Knowledge Institute of Technology. His research mainly focuses on Neural Networks and Cloud Computing.
Contact: tkcse@kiot.ac.in

PHEMALATHA is a student pursuing PG Degree in Computer Science and Engineering at knowledge institute of technology, Salem. Research Interest includes Data Mining and Artificial intelligence.
Contact:ghemalathaoct8@gmail.com

JOY KINSHY.P is a student pursuing PG Degree in Computer Science and Engineering at knowledge institute of technology, Salem. Research Interest includes Mobile Computing and Artificial intelligence.
Contact:cjoykinshy@gmail.com