Cepstrum based Voice Transformation using ANN

J.H.Nirmal
Asist.Prof, Dept of ECED
K.J.Somaiya College of Engg
Mumbai

Suparva Patnaik
Asso.Prof Dept of ECED
S.V.National Institute of
Tech, Surat

Mukesh Zaveri
.Asso.Prof, Dept of COED.
S.V.National Institute of
Tech, Surat

ABSTRACT

The basic goal of the voice conversion system to mimics the characteristics of the target speaker voice by keeping the linguistic and paralinguistic information intact. The characteristics of a speaker in speech reflect at different level such as vocal tract, excitation and prosodic parameters. This propose work based on cepstrum which represents the vocal tract and excitation parameters of the speech. This paper proposes the decomposition of the cepstrum by wavelet and mapped the source cepstrum features in to target cepstrum features using Radial basis function neural network. The results are evaluated using subjective and objective measures based on voice quality method and the listening tests prove that the proposed algorithm converts speaker individuality while maintaining high speech quality..

Keywords

Wavelet transforms Voice conversion, Speech cepstrum, and Radial basis artificial neural network.
.

1. INTRODUCTION

The voice conversion is an algorithm which adopt the characteristics of the Target speaker without changing the distinct characteristics of the source speaker [1].It is widely used in many applications such as customization of text to speech, customization of speaking devices for the people with speech impairedment, film dubbing of original actor, personification of text to speech, preparation of virtual clones of popular people for animations and video games, audio customization, dubbing in radio broadcast, adaptive voice conversion for text to speech synthesis, very low bandwidth speech encoding and transmitting, multimedia entertainment, health care industry and also used security and forensic applications etc [2].

Voice conversion system generally carried out using speech analysis and synthesis system. The mapping between the speech signal parameters derived from source and target speaker’s sentences involves four phases: feature Extraction, source to target mapping, source parameters transformations, and resynthesis of speech using the transformed parameters [3]. In feature extraction phase, the speech signal is analyzed for extracting the parameters of the vocal tract and excitation source. The parameters related to the excitation are considered less important than the vocal tract in specifying speaker individuality. The pitch contour is also considered to be an important cue for individuality [3].

The voice conversion problem has attracted a lot of research effort. First, The Vector Quantization (VQ) based Codebook Mapping method is used for converting parameters of the source and target speakers and got an acceptable performance [5]. In [6] The Linear Multivariate Regression and Dynamic Frequency Warping methods are used for mapping the relation between the corresponding non overlapping classes of source and target speaker's acoustic space The limitations of these two methods are the hard partition of the acoustic space that produce the discontinuities affecting to the quality and naturalness of the converted speech. To overcome this limitation Fuzzy Vector Quantization was proposed in [6].Gaussian Mixture Models (GMMs) is used to convert the spectral envelope that considerably improved the quality of converted voice in [7]. Then Kain proposed some improvements to this method [9].Although the GMM based methods successfully solved the hard partition problem but it suffered from the over-smoothing. This method tries to grasp the basic spectral envelope using GMM and retain the spectral detail using a RBF network. Vocal tract length of speakers are different and nonlinear hence artificial neural network is used for the mapping of the source acoustic cues into target acoustic cues [10][11].

The neural network models for developing the mapping functions at different levels (source, system and prosodic levels) are explored because it is good for capturing the nonlinear relations present in the feature patterns of source and target speakers. The objective of this work is (1) Parameterization of vocal tract characteristics and glottal excitation (2) Exploring neural network models for capturing the complex nonlinear relations between source and target speaker features at different levels.

The paper is organized as follows. First, The cepstrum based algorithm is presented next the radial basis function neural network is briefly described finally, results from a formal listening as well as from an objective test are presented to support our conclusions

2. PROPOSED ALGORITHM

In general, voice conversion is performed in two steps. Acoustical cues of speech signals are computed from The training set of two speakers. Estimate the mapping function from the source acoustic space to the target one, which captures the transformation relation between the two speakers' speech. In the conversion stage, the acoustic characteristics of input speech are converted using the trained mapping function, and the desired spectrum of target speaker are synthesized. In this proposed voice conversion method the speech signal is represented as a vocal tract parameters and glottal excitation parameters by cepstrum of the source speaker and the target speaker and decomposed it up to third level by coiflet wavelet transform and mapped it using radial basis function shown in the fig: 1.

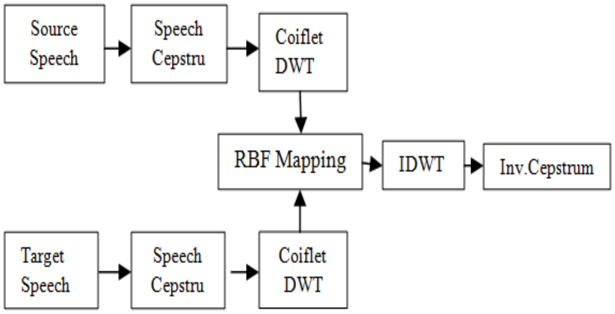

Fig 1: Cepstrum wavelet based ANN voice conversion.

The spectrum of the speech signal is represented by the product of vocal tract parameters and the glottal excitation is given by:

$$S(w)=V(w)G(W) \qquad (1)$$

Here, *S(w)* represents the Fourier transform of speech signal *s(n)*,The *V(w)* and *G(w)* represents the Fourier transform of vocal tract parameters *v(n)* and glottal excitation *g(n)* respectively. The logarithm of the above equation is:

$$log(S(w))=log(V(w))+log((G(w)) \qquad (2)$$

This logarithm spectrum is separated as two parts namely, the log spectral components that vary rapidly with w (high-time components *G(w))* and the log spectral components that vary slowly with w (low-time components, *V(w)).* Hence using an appropriate filter we can separate the two components namely, the excitation spectrum and the vocal tract filter spectrum. This process is called deconvolution. The cepstrum is given by taking the inverse Fourier transform of equation no 2.For the speech synthesis, the excitation function was generated from the knowledge of the pitch period and voiced/unvoiced signal convolved with impulse function [12].

3. RADIAL BASIS FUNCTION NEURAL NETWORK.

Radial Basis function is embedded in a two layer feed forward network where each hidden unit implements a radial activation function. The output unit act as a weighted sum of hidden unit outputs, the input in to RBF network is non linear while output is linear. Due to their non linear approximation properties RBF network are able to model the complex mapping [13]. In order to use the RBF we need to specify the hidden unit activation function, the no of processing units, a criteria for modelling a given task and training algorithm for finding the parameters of the network, finding the RBF weights called as network training set, we optimize the network parameters in order to fit the network output to the given inputs the fit is evaluated by mean of cost function, After training RBF network can be used with data whose underlying statistics is similar to that of the training set, on the training algorithm adopts the network parameters to changing the data statistics.

The activation function of each hidden unit in RBFN computes the Euclidean distance between the input vector a centre of that unit the input data X is an one dimensional input cepstrum frame vector which is transform to the hidden unit and the activation function is symmetrical to the input space and output of each hidden unit depends only on the radial distances between the input vector *s* and centre for the hidden unit, The activation function is nonlinear function and is of many types of Gaussian, The Gaussian activation function can be written as

$$\emptyset j(x) = exp\left(\frac{(x-cj)}{2\sigma^2}\right) \qquad (3)$$

Where x is the training data and $\boldsymbol{\sigma}$ is the width of the Gaussian function, A centre and width are associated with each hidden unit in the network. The weights connecting the hidden and output units are estimating using least mean square method; Finally the response of each hidden unit is scaled by its connecting weights to the output units and summed to produce the overall network output therefore the kth output of the network yk is

$$yk = \sum_{j=1}^{m} \emptyset_j(x) w_{jk} + wo \qquad (4)$$

ϕj(x) is the response of the j_{th} hidden unit w_{jk} is the connecting weight between j_{th} hidden unit and k_{th} output unit

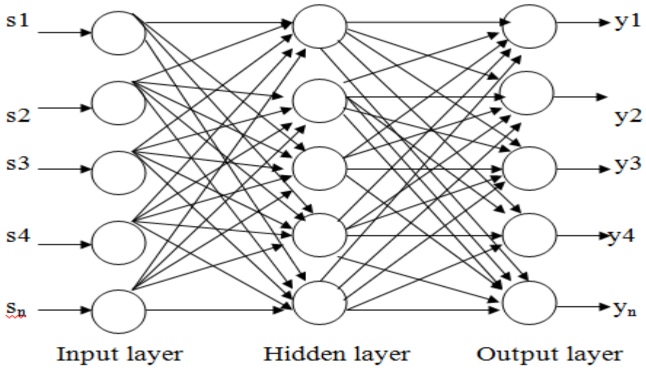

Fig 2: Radial Basis function for voice conversion.

For training, the cepstrum of the source speaker (s)voice and target speaker (y) voice signals are used as input and output of the network respectively. The weights of the hidden and output layers of the networks are adjusted in such a way that the source speaker's pattern is converted into the target speaker's pattern. The training step is a supervised learning procedure.

4. VOICE CONVERSION:

Our voice conversion algorithm is implemented in the following steps

- The source and target speaker uttered the same sentence and split it as training, testing, and validation.
- The source and Target samples are time aligned i.e. resampled so that the length are same.
- The Training and test samples of source and target speech are divided in to frames of size 128 samples and overlapping window size is 80 samples.
- Calculate the cepstrum of every frame of training and test samples and the cepstrum is decomposed up to third level by coiflet wavelet transform.
- The level 1 and 2 details coefficients are set to zero and the remaining wavelet coefficients are normalized.

- For each level of decomposition a RBF is initialised and trained using source and target training samples.
- At each level, the source speaker's test samples' coefficients are projected through the corresponding network and the transformed coefficients are obtained.
- The transformed coefficients are used in order to reconstruct the signal.
- The transformed signal is tested and compared to the target signal to assess the transformation.

5. EXPERIMENTAL WORK AND RESULTS

This proposed work is carried out on our own parallel database consisting of 48 Gujarati, Marathi (Indian languages) short sentences from five male and five female speakers recorded using a high quality microphone (Sony V_120). The speech samples are recorded at a frequency range of 16 KHz the speakers are well trained before capturing the speech corpus. The recorded speech samples are processed labelled and stored.

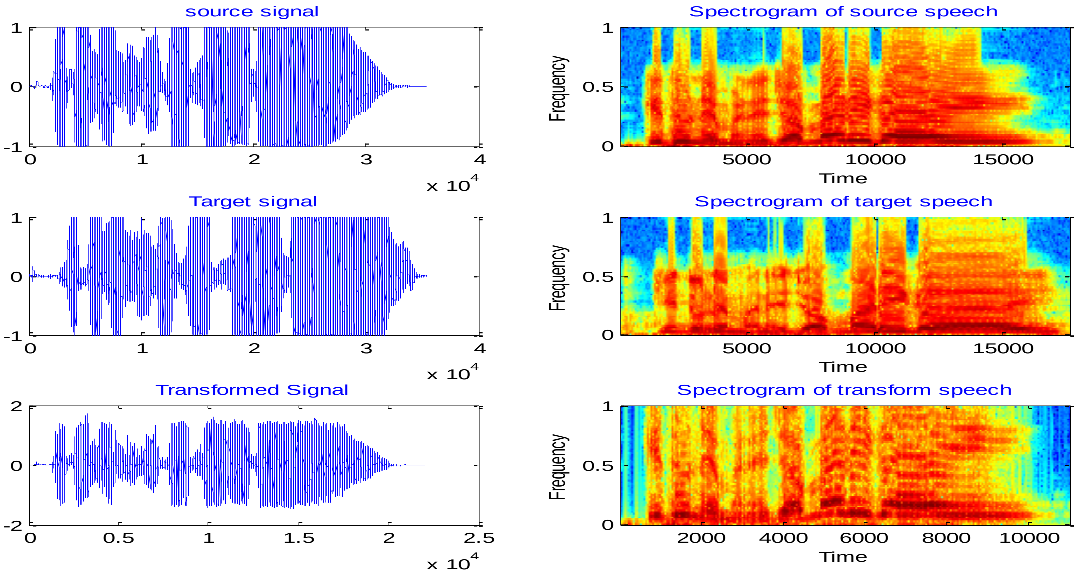

Fig 3 a) source speech b)target speech c)transform speech for voice conversion

We have also tested the performance of our algorithm on CMU ARCTIC database consisting of utterances recorded by 7 speakers; we transformed SLT (U.S. female) to BDL (U.S. male) and BDL (U.S. male) to SLT (U.S. female).The mapping of the cepstrum coefficients of source and target speaker are performed using RBF network .The sample results of our conversion on speech from male-to-male, female-to-female, female-to-male, and male-to-female pairs of speakers are shown in above figure 3. From the visual perception of the above figure 3 it is cleared that the spectrogram of the target speaker is more closer to the transform speech than the source speaker speech.

5. EVALUATION

5.1 Objective Evaluation:

In this section we provide the objective evaluation for cepstral based voice conversion system to measures the differences between the target and transformed speech signals. Since many perceived sound differences can be interpreted in terms of differences of spectral features, mean squared error (MSE) is considered to be a reasonable metric both mathematically and subjectively. The mean squared error between target audio vector *p* and transformed audio vector *s* is calculated as per equation no 1. on a sample by sample basis. The average square difference between two vectors is used to evaluate the objective performance of mapping algorithms shown in table 1.

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

Table 1: MSE between target and transform speech

Mean Square Error			
Male to Male	Male to Female	Female to Male	Female to Female
0.2743	0.0544	0.0815	0.0815

5.2 Subjective Evaluation:

To evaluate the overall accuracy of the conversion, a Mean Opinion Score (MOS) test was carried out for evaluating the similarity between the converted voice and the target voice to find the performance of cepstrum wavelet based transformation followed by RBF Neural network. 5 listeners give scores between 1 and 5 for measuring the similarity between the output of the two voice conversion systems and the target speaker's natural utterances. The results of this average similarity test are provided in Table II, which indicates that the transform speech is more closer to the target speech as compared to source speech

Table II: MOS between transformed and the natural utterances of the target Speakers

Average Similarity Score			
Male to Male	Male to Female	Female to Male	Female to Female
3.89	3.02	3.56	3.81

5. CONCLUSION

In this paper, we propose a novel voice conversion method based on cepstrum wavelet transform followed RBF network for male to male, male to female and female to male voice conversion. Subjective evaluation and objective evaluations are performed on speech quality. The Experimental results show that the proposed algorithm performs better but cannot work for long recording due to the size of the cepstrum and the Artificial Neural Network which is complex and take long time for training.

REFERENCES

[1] Stylianou Y 2009. "Voice Transformation: A survey."Acoustics, Speech and Signal Processing, IEEE International Conference on 2009. ICASSP 2009

[2] A. Kain, " High resolution voice transformation," PhD Thesis, OGI School of Science and Engineering,2001

[3] Lehana P.K, Pande P.C (2011).,"Transformation of short term spectral envelope of speech signal using multivariate polynomial modelling", National conference on communication pp :1-5.

[4] H. Kuwabara and Y. Sagisak,1995 "Acoustic characteristics of speaker individuality: Control and conversion, "Speech Communication, vol.16, pp. 165-173, .

[5] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, 1988 "Voice conversion through vector quantization," in Acoustics, Speech, and Signal Processing 88. ,International Conference on, 1988, pp. 655-658

[6] H. Valbret, E. Moulines and J. P. Tubach,1992 "Voice transformation using PSOLA technique," Speech Communication, vol. II, pp. 175-187,

[7] Shikano, K,Nakamura S,Abe M," Speaker adaptation and voice conversion by codebook mapping" Circuits and Systems, 1991., IEEE International Sympoisum on,vol 1,pp.594-597.

[8] Y. Stylianou, O. Cappe and E. Moulines (1998), Continuous probabilistic transform for voice Conversion,"Speech and Audio Processing, IEEE Transactions on, vol. 6, pp. 131-142.

[9] Y. Kang, Z. Shuang, J. Tao, W. Zhang, and B. Xu I(2005), " A Hybrid GMM and Codebook Mapping Method for Spectral Conversion, " Affective Computing and Intelligent Interaction, pp. 303-310,

[10] Desai, S; Black, A W; Yegnanarayana, B; Prahallad, K.T. 2010 "Spectral mapping using artificial neural networks For voice conversion," IEEE Transactions on Audio, Speech,and Language Processing,vol 18,no.5,pp. 954 -64,

[11] K.S.Rao 2010,,"Voice conversion by a mapping the speaker specific features using pitch synchronous approach" Computer speech and language ,vol 24 issue 3 pp 474-494.

[12] Alan V Opphenheim-1969,"Speech Analysis and Synthesis System based on Homomorphic filtering", The Journal of the Acoustical society of America vol 45 No 2.pp 458-465.