Spectral Mapping Using Linear Predictive Synthesis Methods for Voice Conversion

Mrs. P.Malathi[1] Pavithra M[2]

[1]Assistant Professor [2]P.G Scholar

[1,2]Department of Electronics & Communication Engineering

[1,2]Prathyusha Institute of Technology and Management, Chennai

Abstract— The objective of voice conversion algorithm is to modify the speech by a particular source speaker so that it sounds as if spoken by a different target speaker. Speech signal is produced by the convolution of excitation and time varying vocal tract system components. These excitation and vocal tract components must be separated from the available speech signal to study these components independently. For deconvolving the given speech into excitation and vocal tract system components, methods based on homomorphic analysis like cepstral analysis are developed. As the cepstral analysis does the deconvolution of speech into source and system components by traversing through frequency domain, the deconvolution task becomes computational intensive process. To reduce such type of computational complexity the Linear Prediction Analysis is developed. The primary objective of prediction analysis is to compute the coefficients which minimize the prediction error.

Key words: Excitation, vocal tract components, cepstral analysis, prediction error

I. Introduction

Voice conversion is the process of adapting the acoustical characteristics of a source speaker according to that of a target speaker. Voice conversion finds its application in areas such as personification of text to speech synthesis, audio based learning tool, audition test, audition customization, audio dubbing and biometric recognition.

Voice transformation should be performed without losing the original speech content including the emotional state of the speaker. Such transformation involves mapping of spectral, excitation, prosodic features and F_0 pattern of a source speaker onto a target speaker's acoustic space.Speech signals are slowly time varying signals (quasi-stationary)[3]. When examined over a short period of time its characteristics are fairly stationary. The information in speech signal is actually represented by short term amplitude spectrum waveform of the speech waveform. This allows us to extract features based on the short time amplitude spectrum from the speech. Voice conversion takes place in two phases. The first one is the training phase and the second is the testing phase. In the training phase the features are extracted. With these extracted features testing phase is carried out.

Different voice conversion systems that employ different methods exists, but they all share the following components:

- A method to represent the speaker specific characteristics of the speech waveform
- A method to map the source and the target acoustical spaces
- A method to modify the characteristics of the source speech using the mapping obtained in the training phase.

Some of the voice conversion algorithms are as follows:

- Statistical technique (Gaussian mixture model (GMM)[1], Hidden markov models (HMM)[2], Principal component analysis(PCA).
- Cognitive technique[8](Artificial neural network),
- Linear algebra technique(SVD)
- Signal processing technique (Vector quantization (VQ)[5],Frequency Warping[4].

In this paper voice conversion is analyzed using linear predictive analysis methods as it is based on parametric estimation [9] and good in recognition in parameterization of voice.

II. Linear Predictive Analysis

Linear predictive(LP) analysis is one of the most widely used speech analysis techniques. In this analysis coding of speech takes place at a low bit rate. Linear prediction is analyzed using auto correlation method. The linear prediction coefficients are used to estimate the vocal tract system function and the error signal function. The basic concept behind linear predictive analysis is that a specific speech sample at the current time can be approximated as a linear combination of past speech samples[6]that is prediction of current sample as a linear combination of past ρ samples form the basis of linear prediction analysis. Where ρ is order of the prediction error.

The prediction concept is represented as,

$$\hat{s}(n) = -\sum_{k=1}^{p} a_k . s(n-k) \qquad (1)$$

Where $a_{k.}$s are the linear prediction coefficient and s (n) is the windowed sequence, given by,

$$s(n) = x(n). \omega(n) \qquad (2)$$

Where ω(n) is the windowing sequence. The prediction error e(n) can be computed by the difference between actual sample s(n) and the predicted sample $\hat{s}(n)$ which is given by,

$$e(n) = s(n) - \hat{s}(n) = s(n) + \sum_{k=1}^{p} a_k . s(n-k) \qquad (3)$$

The primary objective of Linear predictive analysis is to compute the LPC coefficients which minimizes the prediction error e(n).

The total prediction error is given as,

$$E = \sum_{n=\infty}^{\infty} e^2 (n). \qquad (4)$$

III. Computing the Residual Error

The autocorrelation sequence is given as,

$$R(i) = \sum_{n=i}^{N-1} s(n)s(n-i) \qquad (5)$$

For i=1,2,3..ρ and N is the length of the sequence.

This can be represented in matrix form as,

$$R.A = -r \qquad (6)$$

The LP coefficients can be computed as,

$A = -R^{-1}.r$ Where R^{-1} is the inverse matrix

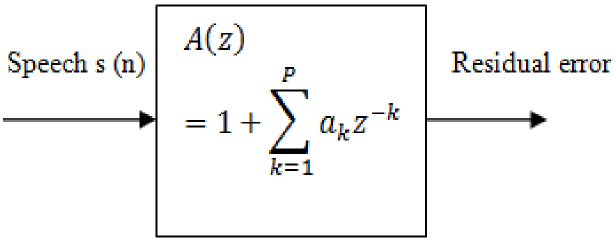

Fig. 1: Computing the residual by inverse filtering

IV. Estimating the Vocal Tract Parameters by Linear Predictive Analysis

LP analysis separates the given short term sequence of speech into its slowly varying vocal tract component represented by LP filter (H(z)) and fast varying excitation component given by the LP residual (e(n)). The LP filter (H(z)) induces the desired spectral shape on the flat spectrum (E(z)) of the noise like excitation sequence. AS the LP spectrum provides the vocal tract characteristics, the vocal tract resonance can be obtained by picking the peaks from the magnitude LP spectrum (|H(z)|).

$$S(z) = E(z).H(z). \qquad (7)$$

Where S(z) is the spectrum of the given short time speech signal.

V. Estimating Pitch from LP Residual

As the LP residual is an error obtained from the LP analysis, it is noisy in nature. The pitch marks are characterized by the sharp peak in the signal. A sharp peak with discontinuity causes an error in the computed residual [7]. The periodicity of the error gives the pitch period of that segment of speech and this can be computed by the autocorrelation method. For unvoiced signal the residual will be like a random noise without any periodicity.

VI. Normalized Error

The normalized error is used to find the prediction order of the speech signal. The normal error can be defined as the total minimum error to the total energy of the signal. The normalized error is given as,

$$V_{(p)} = \frac{E_p}{R(1)} \qquad (8)$$

Where E_p is the total minimum error, R (1) is the total energy of the signal and $V_{(p)}$ is the normalized error.

$$E_p = \sum_{n=-\infty}^{\infty} s^2(n) + \sum_{k=1}^{p} a_k \sum_{n=-\infty}^{\infty} s(n).s(n-k) \qquad (9)$$

In terms of autocorrelation sequence,

$$E_p = R(1) + \sum_{k=1}^{p} a_k.R(k) \qquad (10)$$

VII. Experiment Results and Discussion

A. Database:

Voice conversion techniques need a parallel database in which the source and the target speakers record the same set of utterances. The work presented here is carried out with the help of the NORTH TEXAS VOWEL DATABASE. Each speaker has recorded a set of phonetically balanced utterances of duration less than 1 second.

B. LPC Coefficients:

LPC coefficients are extracted from the corresponding source speaker. The source speaker can be either a male or female speaker. After extracting the LPC coefficients it is convolved with the help of the filter.

C. Excitation Features:

The excitation features are extracted from the target speaker. After extracting the features it is convolved with the LPC coefficients of the source speaker along with the filter. The transformed voice obtained will be the convolution of the source and the target speaker. The transformed voice resembles the characters of both the source and the target speaker.

D. Results:

The results are obtained for different vowels for the male and female voices. The LPC coefficients of the source speaker is given below in the table1.The excitation size of the target speaker is given below in the table 2.

E. Source Speaker:

S.No	Vowels	Size of the LPC coefficients
1.	Heed	92
2.	Had	92
3.	Hood	90
4.	Head	82

Table 1: Source speaker's (male) LPC coefficients

F. Target Speaker:

S.No	Vowels	Excitation size
1	Heed	27828
2	Head	24172
3	Had	23819
4	Hood	23048

Table 2: Targets speaker's (female) excitation size

From the above table it is clear that the vowels with the longest pronunciation duration have the maximum excitation size.

VIII. Source and Target Speaker's Graphs

A. Source Speaker:

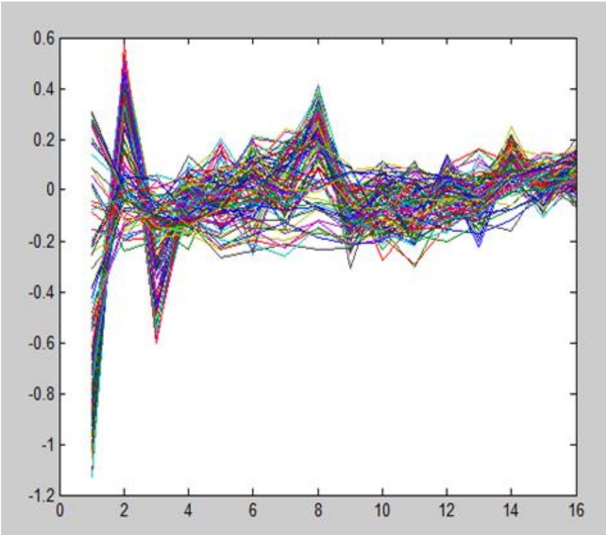

Fig. 2: Male LPC coefficients plot for the word 'had'

The above given graph is with error which is rectified by convolving it with the help of a desired filter.

B. *Target Speaker:*

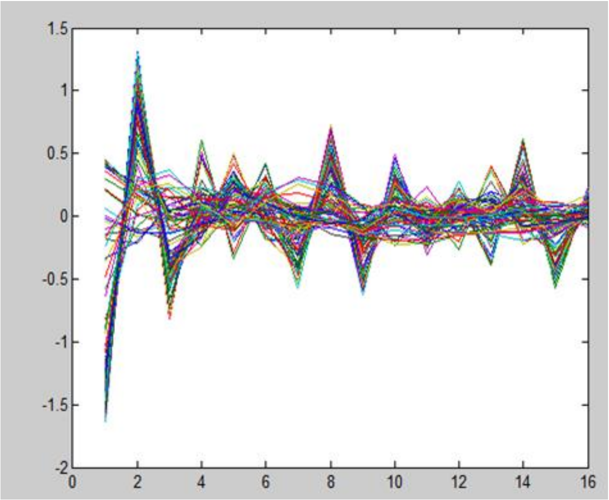

Fig. 3: Female LPC coefficient plot for the word 'had'.

C. *Mean Squared Error (MSE):*

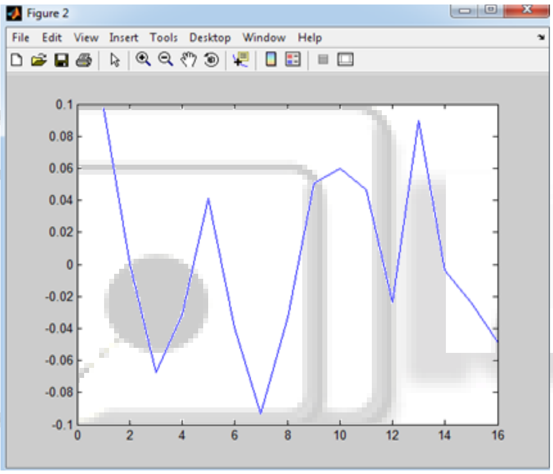

Fig. 4: Mean Squared Error for the word 'had'

D. *Excitation Plot:*

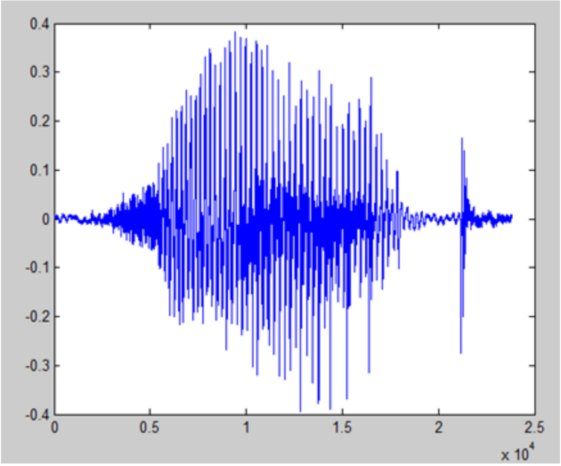

Fig. 5: Excitation plot of the target speaker for the word 'had'

The above given graph is the excitation plot of the female speaker for the word had. By convolving the source speaker's LPC coefficient and target speaker's with the filter we get converted voice, as given in the figure 5.

E. *Converted Voice Plot:*

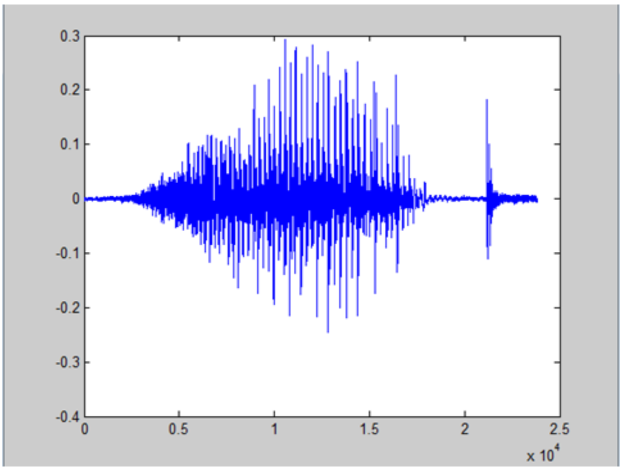

Fig. 6: Plot for Converted Voice.

IX. CONCLUSION

From the above results it is clear that the Linear predictive analysis converts the given voice signal. LPC works good for the vowels. But for fricatives it fails to convert the voice, as it is a parametric estimation. To overcome this problem we go for MFCC which converts voice based on spectral estimation. Our future work is currently focused on converting the voice with MFCC which performs better than LPC.

REFERENCES

[1] Reynolds, D.A., "Speaker identification and verification using Gaussian mixture models. Speech communication 17 (1-2), 91-108. 1995

[2] Kaiyu and Steve young "Continuous Fo modeling for HMM based statistical parametric speech synthesis", IEEE TRAS, ON ASLP, 2011

[3] Urmila shrawankar, "Techniques for feature extraction in speech recognition system: a comparative study".

[4] Daniel erro, Asuncion Moreno. "Weighted frequency warping for voice conversion".

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

[6] David sundermann, HaraldHoge, Antonio, "Text independent voice conversion based on unit selection", 2004.

[7] Alexander Kain & Michael W.Macon, "Design and evaluation of voice conversion based on spectral envelope mapping and residual prediction". International Conference on Acoustics, Speech and Signal Processing, vol 2. Pp 813.816, 2001.

[8] Srinivasa Desai, E. Veera Raghavendra, B. Yegnanarayana, Alan w Black, Kishore parahallad, "Voice conversion using Artificial Neural Networks". Proc. IEEE Int. conf Acoust, speech signal process., Taipei, Taiwan Apr. 2009.

[9] MATLAB 7.5.0(R2007b) of Math Works, Inc, USA,"MATLAB" software's.

[10] North Texas vowel database www.Utdallas.edu/~assmann/KIDVOW1/.