International Journal of Engineering Research & Management Technology
ISSN: 2348-4039
July-2018 Volume 5, Issue-4
Email: editor@ijermt.org www.ijermt.org

SYNTHESIS AND RECOGNITION OF AUTOMATIC SPEECH USING MFCC AND LPC ANALYSIS

Dinesh Chandra Misra
Research Scholar
NIILM University Kaithal

Anil Kumar
Assistant Professor
NIILM University Kaithal

ABSTRACT
The standard obverse end analysis in speech recognition is a spectral analysis which parameterizes the speech signal into feature vectors; the general set of them is the Mel Frequency Cepstral Coefficients (MFCC). They are grounded on a standard power spectrum estimate which is first imperilled to a log-based transform of the frequency axis (mel- frequency scale), and then decorrelated by using a modified discrete cosine transform. Succeeding an absorbed introduction on speech production, perception and analysis, this work stretches a study of the application of a speech generative model; whereby the speech is synthesized and recovered back from its MFCC representations. The work has been developed into two steps: first, the computation of the MFCC vectors from the source speech files by using HTK Software; and second, the implementation of the generative model in itself, which, actually, represents the conversion chain from HTK-generated MFCC vectors to speech reconstruction.

KEYWORDS: Generative model, Linear Prediction Coding, synthesized speech

INTRODUCTION
The standard obverse end analysis in speech recognition is a spectral analysis which parameterizes the speech signal into feature vectors. The general set of feature vectors used in recognition systems is the Mel Frequency Cepstral Coefficients (MFCC)[1]. These are grounded on a standard power spectrum approximation which is initial subjected to a log-grounded alter of the frequency axis; it results in a spectral picture on a perceptually frequency scale, based on the response of the human perception system[2-5]. After, these are connected by using a modified discrete cosine transform, which allows an energy compaction in its lower coefficients[6-8].

An exciting issue is how much pertinent information connected to speech recognition is misplaced in this analysis[9,10]. Thus, this Paper is worried with synthesizing speech from different parametric representations (MFCCs and Linear Prediction coefficients), and to demeanour an investigation on the perspicuity of the synthesized speech as compared to natural speech[11-15].

According to this purpose, the five principal objects of the Paper

1. Study speech analysis processing and theories based on speech production and speech perception.
2. Study on the implementation of MFCC computation in the Hidden Markov Toolkit (HTK), a typical study and development tool for HMM-based speech recognition.
3. Develop a speech generative model based on the implementation of the conversion chain from HTK-generated MFCC representations to speech reconstruction.
4. Employ objective measures for an intermediate evaluation of the generative model.
5. Present a subjective interpretation of the intelligibility of the synthesized speech.

LPC ANALYSIS OF THE WAVEFORM SPEECH SIGNAL
The algorithm for LPC analysis was implemented in a Matlab function called *wave form analysis.*. This function follows the process shown in Figure 13, implementing the filtered of the speech signal through the pre-emphasis filter, the frame blocking and Hamming windowing and the LPC feature extraction.

The algorithm *waveform_analysis.m* was executed for a LPC analysis of 12th order. The LPC parameters (filter gain, *g*; and LPC filter coefficients, $\{a_i\}$) were computed by using the Matlab function *proclpc.m*, which belongs to Matlab Auditory Toolbox.

In order to show the performance of the different steps involved in LPC extraction process, the following figures were executed for *sa1.wav* file. In Figure 14, the original speech waveform and how is affected after the pre-emphasis filter is illustrated. Figure 15 presents the effect of using a Hamming window, and Figure 15 shows the Linear Predictor spectrum of one frame as compared with its magnitude spectrum.

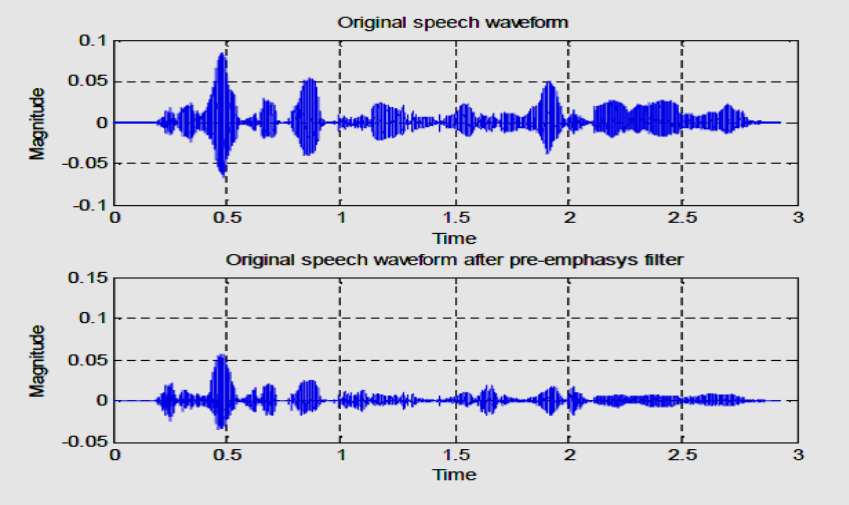

Figure.1 Original speech wavefront and original wavefront after the pre-emphasis filter with coefficient equal to 0.97.

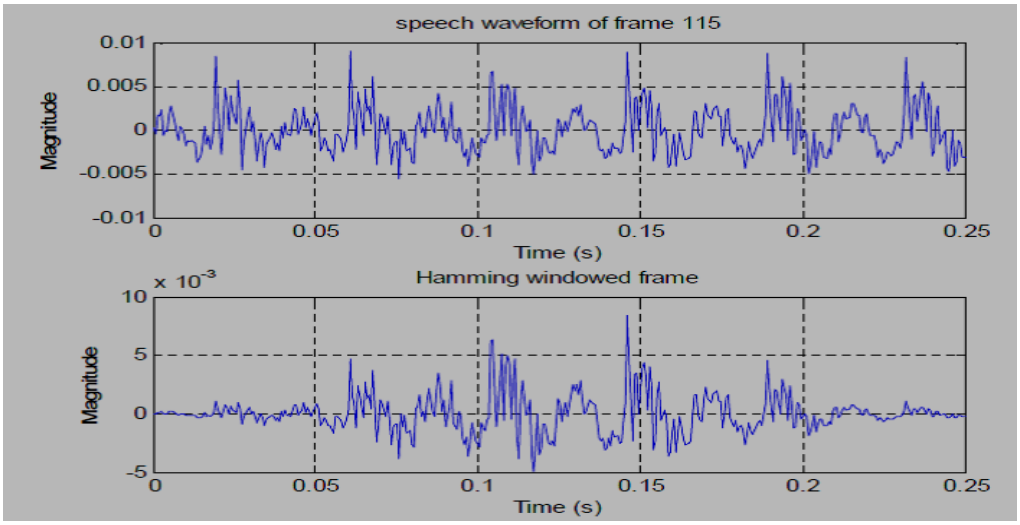

Figure.2 Effect of multiplying one speech frame by Hamming windoe.

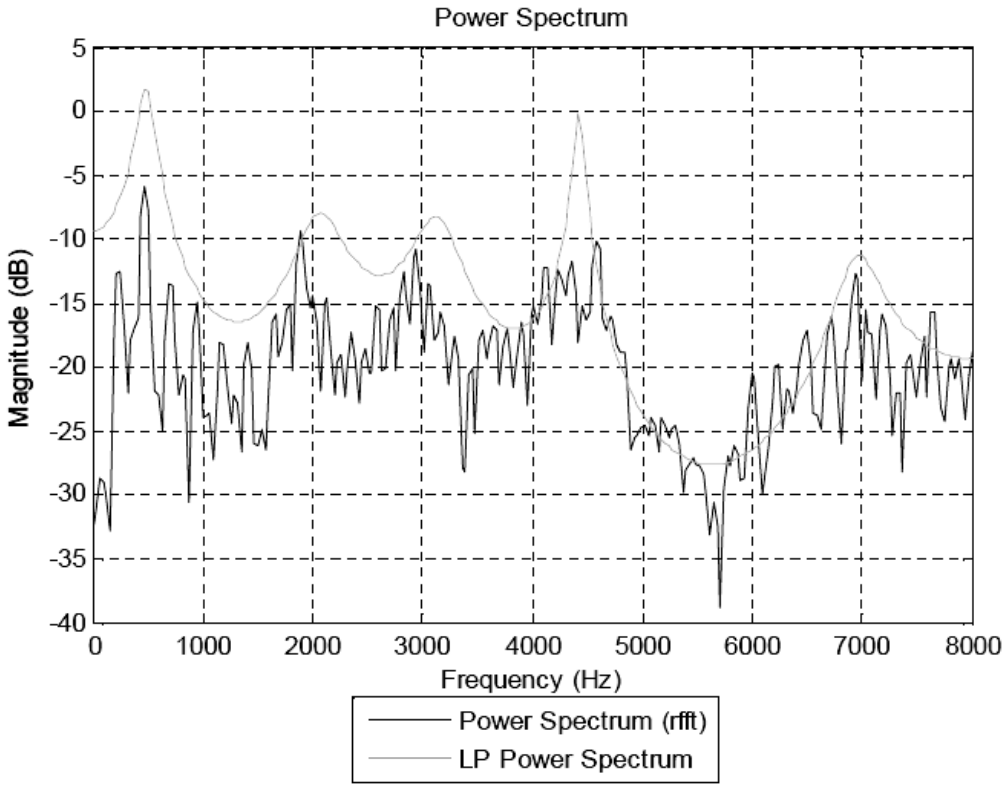

Figure 3: Comparison of the power spectrum computed from LPC coefficients with the original magnitude spectrum (frame 115 of *sa1.wav*)

As one can see in Figure 2, the use of a Hamming window makes that magnitude of the speech frame tapers from the centre of the window to the edges. This fact reduces the discontinuities of the signal at the edges of each frame.

Figure 3 shows the Linear Prediction (LP) power spectrum compared with the magnitude spectrum of a speech frame. One can see that the power spectrum computed from LPC coefficients is actually representing the *spectral envelope* of the magnitude spectrum of this frame. This spectral envelope marks the peaks of the formants of the speech frame.

More examples that illustrate this fact can be added. Figure 4 corresponds to the frames 84 and 176 of the same waveform file (*sa1.wav*).

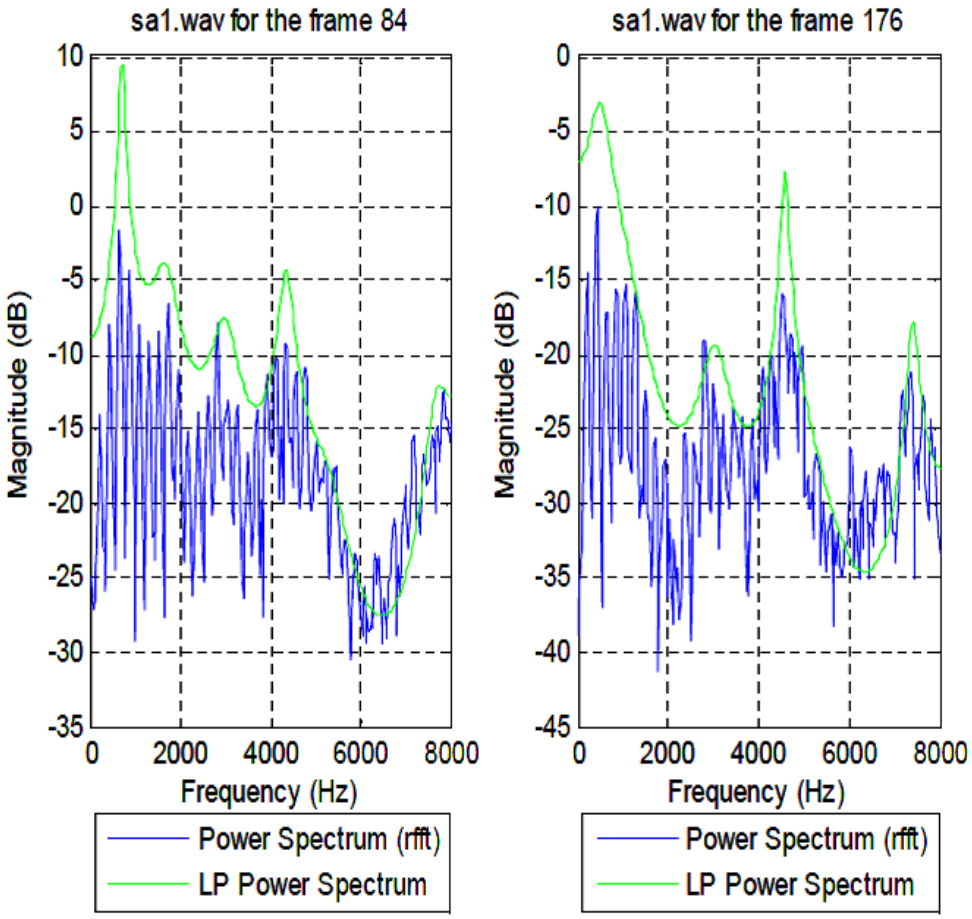

Figure 4: Comparison of the power spectrum computed from LPC coefficients with the original magnitude spectrum (frames 84 and 176 of *sa1.wav*)

ANALYSIS OF TWO APPROACHES FOR THE GENERATIVE MODEL

It deals the computation of the LPC coefficients from the MFCC vectors. It was seen that the LPC coefficients come from the solution of the Yule Walker equations. They can be solved by the autocorrelation method, for which the autocorrelation coefficients must be calculated. In this point, two approaches were proposed to estimate the autocorrelation coefficients based on the IFT of the mel power spectrum. These approaches were implemented in the algorithms *mfcc2spectrum.m* and *mfcc2spectrum2.m*

In this section, the results of both algorithms will be exposed and discussed. So, the power spectrum computed from the LPC parameters as compared with the mel power spectrum will be plotted by executing both algorithms.

Figure 5 is obtained by executing the *mfcc2spectrum.m* function. This algorithm makes a linear interpolation of the mel power spectrum to get samples uniformly spaced in a linear frequency scale in order to use the inverse Fourier Transform.

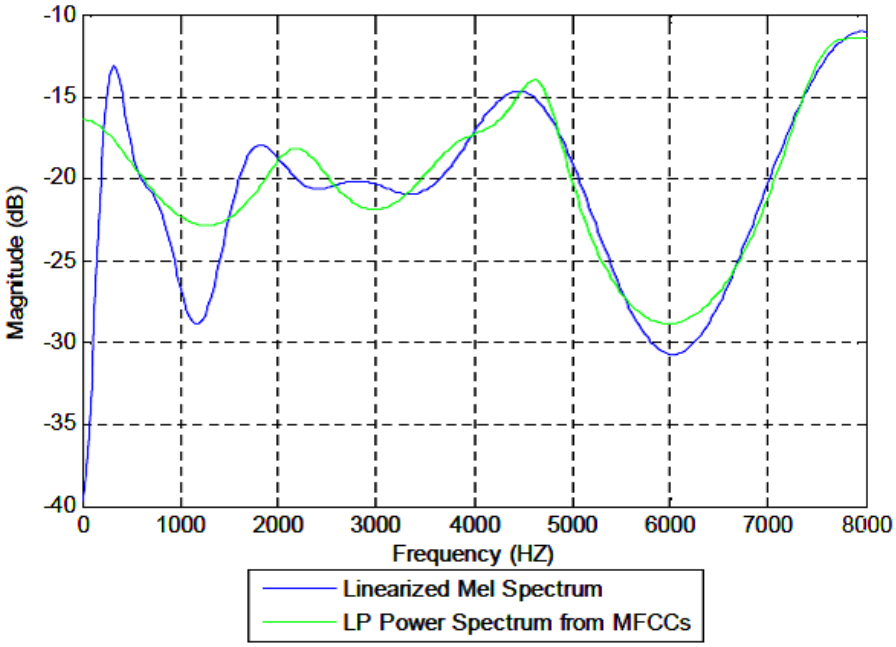

Figure 5: LP power spectrum computed from MFCCs by generative model 1: *mfcc2spectrum.m* (frame from *sa1.wav*)

One can see that the LP power spectrum computed from MFCC coefficients is approximated to the mel power spectrum. Both of them represent the spectral envelope of the magnitude spectrum of the speech frame.

Following Figure 6 is obtained by executing the *mfcc2spectrum2.m* function. This algorithm applies the inverse Fourier Transform directly to the mel power spectrum at frequencies on a mel scale considering their bandwidth.

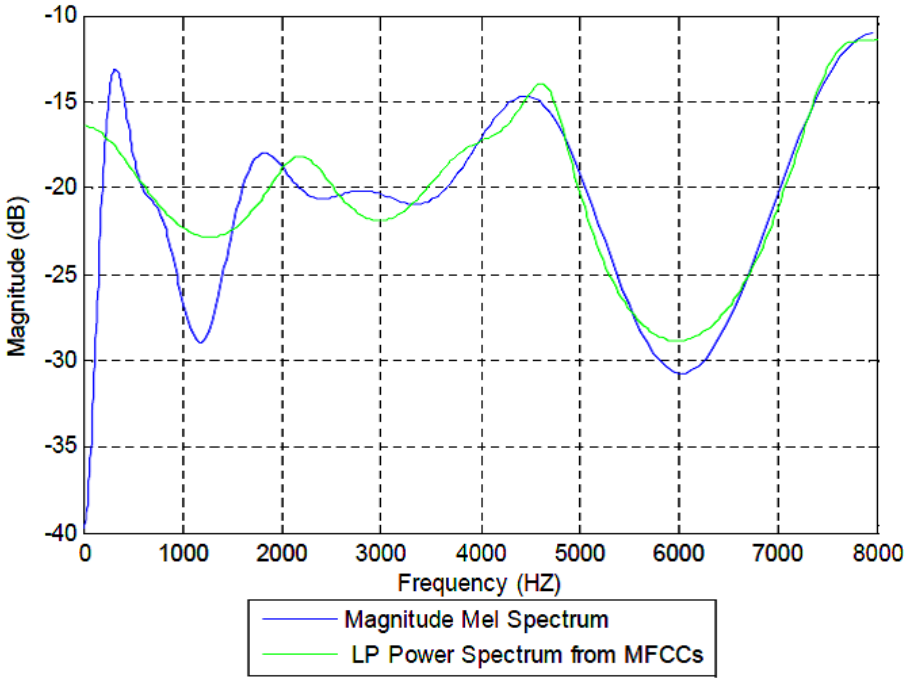

Figure 6: LP power spectrum computed from MFCCs by generative model 2: *mfcc2spectrum2.m* (frame 115 from *sa1.wav)*

The results are equal to the ones obtained with the first algorithm, since both of them can be considered as a linear interpolation in which, finally, the mel power spectrum samples have to correspond to a determined frequency separation.

The algorithm of *mfcc2spectrum2.m* is faster than the *mfcc2spectrum.m*. That is because of, the first one computes the autocorrelation coefficients of one frame in one matrix multiplication; whereas, the second one has to make one linear interpolation for each equally-spaced frequency sample. That is why; the results of the generative model will be performed by using the *mfcc2spectrum2.m* algorithm.

SPECTRAL DISTANCE MEASURE

As was introduced before, the goal of the generative model is to implement a system or method to be able to synthesize speech from its MFCC parametric representation. The goodness of the synthesized speech can be measured by computing the spectral distance between the original signal and the one produced from the MFCC coefficients. For that, the two spectral models used were the one obtained from the LPC coefficients computed from the original signal and the one obtained from the LPC coefficients computed from the MFCCs.

The spectral distance measure and its L_2 spectral norm (*rms log spectraldistance*) were explained in Section 2.3. An algorithm to measure the *spectral distance* between two spectral models was implemented in a Matlab function called *spectral_distance.m*. Several examples will be given to show a graphical comparison between the two spectral models.

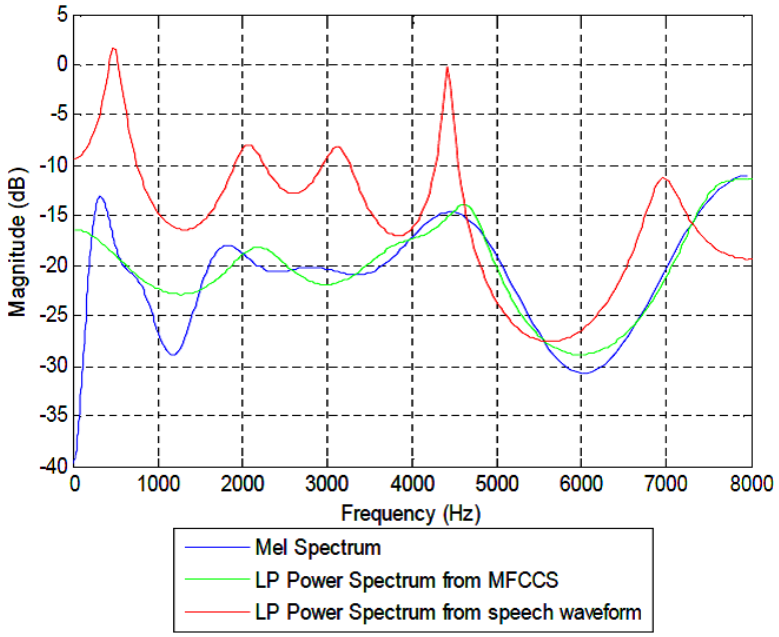

Figure 7: Comparison of spectral models from the original speech waveform and from the MFCC vectors (fame 115 from *sa1.wav*)

It was said before that the LP power spectrum computed from speech waveform as well as from MFCCs coefficients, represented the spectral envelope of the magnitude spectrum of the speech frame. However, in Figure 20, one can see that the harmonics or formants peaks are marked in the LP power spectrum from speech waveform whereas, they are more flattened when is computed from the MFCCs coefficients. This gives a spectral distortion between them of 0.87dB.

Another example can be shown by using the *si648.m* file. Figure 8 illustrates the comparison of the LP spectrums whose spectral distortion computed is of 0.35 dB.

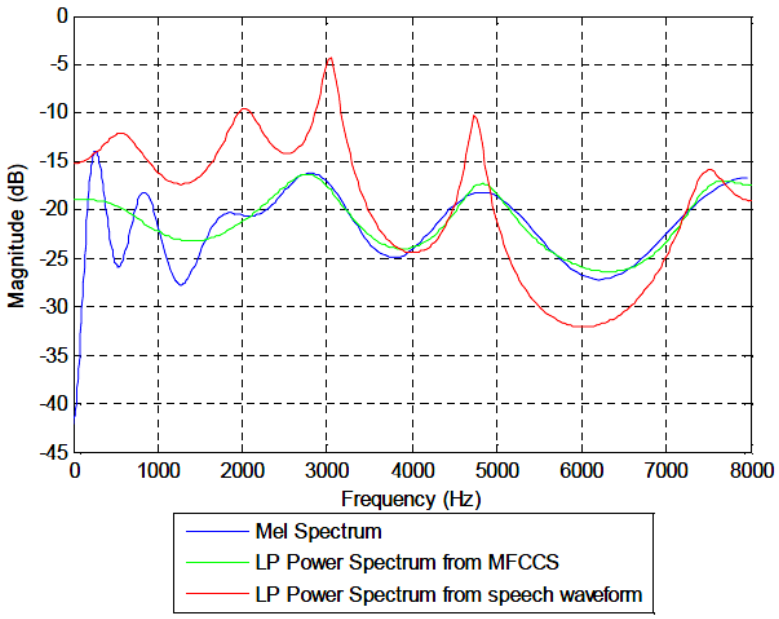

Figure 8: Comparison of spectral models from the original speech waveform and from the MFCC vectors (fame 133 from *si648.wav*)

The generative model can be evaluated more in detail by computing the spectral distance for every frames of each speech waveform file. Hence, it is possible to give an overview of the minimum and maximum spectral distances that were computed by the model. Also, the mean spectral distortion of every speech file is calculated. Table 1 shows the results of these measures.

Table 1: Study of spectral distortion computed between LP power spectrum from original waveform speech signal and the one computed from MFCCs

Source waveform files	Minimum spectral distortion (dB)	Maximum spectral distortion (dB)	Mean spectral distortion (dB)
sa1.wav	0.11	2.09	0.76
sa2.wav	0.14	2.11	0.67
si648.wav	0.09	1.67	0.57
si1027.wav	0.11	2.12	0.62
si1657.wav	0.09	1.71	0.67
sx37.wav	0.10	2.18	0.58
sx127.wav	0.14	2.56	0.68
sx217.wav	0.11	1.83	0.77
sx307.wav	0.16	2.01	0.61
sx397.wav	0.10	1.93	0.72

From the above table, one can extract that he minimum spectral distortion computed is 0.09 dB and the maximum is 2.56 dB. So, the results of the generative model depend of the utterances which have to be synthesized. If one computes the mean of the mean spectral distortion of every speech file can give a mean estimate of the generative model. Doing that, it is possible to say that the generative model has a spectral distortion mean of 0.66 dB. This mean depends strongly on the speech data that were used for the experimental results.

CONCLUSION

The work developed in this Paper consisted of the implementation of a speech generative model; whereby the speech is synthesized and recovered from its MFCC representation. Synthesizing speech from parametric representations allows performing an investigation on the intelligibility of the synthesized speech as compared to natural speech.

The first part of the implementation work consisted of extracting the MFCCs feature vectors from a set of speech waveform files. In the HTK Software, the feature parameterization of speech was performed according to the parameter settings in the configuration file. After, the generative model implemented the conversion chain from HTK-generated MFCC vectors to speech reconstruction.

During the MFCC extraction process, much relevant information was lost due to reduction of the spectral resolution in the filterbank analysis and the next truncation into the MFCC components. However, that allowed recovering a smoothed spectral representation in which phonetically irrelevant detail had been removed. For that, the log mel power spectrum could be computed from its MFCCs by an inverse DCT.

REFERENCES

1. M.A.Anusuya, S.K.Katti (2009), "Speech Recognition by Machine: A Review,"(IJCSIS) International Journal of Computer Science and Information Security, Vol. 6, No. 3, 2009, pp.181-205.
2. Paul A.K., Das D., Kamal M.M.,(2009).“Bangla Speech Recognition System Using LPC and ANN”, Seventh International Conference on Advances in Pattern Recognition, IEEE Xplore, (Kolkata, Feb. 4-6 2009), 171 –174.
3. Sonia Sunny, David Peter S and K Poulose Jacob (2012), “A Comparative Study of Parametric Coding and Wavelet Coding Based Feature Extraction Techniques in Recognizing Spoken Words”, CUBE 2012,Published in September 3–5, 2012, Pune, Maharashtra, India.
4. Thiang ,SuryoWijoyo (2011),. “Speech Recognition Using Linear Predictive Coding and Artificial Neural Network for Controlling Movement of Mobile Robot”, Proc. of. Int. Conf. on Information and Electronics Engineering, IPCSIT vol.6, (IACSIT Press, Singapore), 2011.
5. C. Demiroglu and T. Barnwell (2005), “A Missing-Data Approach to Noise- Robust LPC Extraction for Voiced Speech Using Auxiliary Sensors”, Published in ICASSP 2005.
6. L. S. Chee, Ooi Chia Ai, and S. Yaacob (2009), "Overview of Automatic Stuttering Recognition System," in International Conference on Man- Machine Systems (ICoMMS 2009) Penang, Malaysia, 2009.
7. PreetiSaini ,Parneet Kaur, “Automatic Speech Recognition: A Review” (2013), Published in International Journal of Engineering Trends and Technology- Volume4Issue2- 2013.
8. Nidhi Desai , Prof.KinnalDhameliya , Prof.Vijayendra Desai (2013), “Feature Extraction and Classification Techniques for Speech Recognition: A Review”, Published in International Journal of Emerging Technology and Advanced Engineering (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 3, Issue 12, December 2013.
9. Anusuya, M. A., &Katti, S. K. (2011), “Front end Analysis of Speech Recognition: A review”, International Journal of Speech Technology, Springer, vol.14, pp. 99–145, 2011.
10. S.J.Arora and R.Singh (2012), “Automatic Speech Recognition: A Review, “International Journal of Computer Applications, vol60-No.9, December 2012.
11. UtpalBhattacharjee (2013), “A Comparative Study of LPCC and MFCC Features for the Recognition of Assamese Phonemes”, International Journal of Engineering Research & Technology (IJERT) Vol. 2 Issue 1, January- 2013 ISSN: 2278-0181.
12. Picone, J. W. (1993), Signal Modeling Techniques in Speech Recognition, Proc. IEEE, Japan, 81(9): 1215-1247.

13. Vergin, R. (1998), An Algorithm for Robust Signal Modelling in Speech Recognition, IEEE Transactions on Speech and Audio Processing: 969-972.
14. Vergin, R., O'Shaughnessy, D. &Farhat, A. (1999), Generalized Mel Frequency Cepstral Coefficients for Large-Vocabulary Speaker-Independent Continuous-Speech Recognition, IEEE Transactions on Speech and Audio Processing, 7(5): 525-532.
15. Young, S. (2008), HMMs and Related Speech Recognition Technologies, Springer Handbook of Speech Processing, 2008: 539-557.
16. Pratik K. Kurzekar (2014), A Comparative Study of Feature Extraction Techniques for Speech Recognition System, International Journal of Innovative Research in Science, Engineering and Technology, Vol. 3, Issue 12, December 2014
17. P. Prithvi (2015), Comparative Analysis of MFCC, LFCC, RASTA –PLP, International Journal of Scientific Engineering and Research, Volume 4 Issue 5, May 2016
18. K. M. Shiva Prasad (2017), Speech Features Extraction Techniques for Robust Emotional Speech Analysis/Recognition, Indian Journal of Science and Technology, Vol 10(3), January 2017
19. Jose B. TrangolCuripe (2013), Feature Extraction Using LPC-Residual and Mel Frequency Cepstral Coefficients in Forensic Speaker Recognition, International Journal of Computer and Electrical Engineering, Vol. 5, No. 1, February 2013
20. SonaliGoyal (2017), Issues and Challenges of Voice Recognition in Pervasive Environment, Indian Journal of Science and Technology, Vol 10(30), August 2017
21. Sathish Kumar Selvaperumal (2017), Speech to text Synthesis from Video Automated Subtitling Using Levinson Durbin Method of MEL Frequency Cepstrum Coefficient, International Journal of Video&Image Processing and Network Security, Vol:17 No:01