IJITKM Volume 6 • Number 2 • July-December 2013 pp. 133-135

COMPARATIVE STUDY OF SPEECH PARAMETERIZATION TECHNIQUES

Shri Lekha[1], and Ashish Chopra[2]

ABSTRACT: From last four decades, voice communication with computers has been making machines easier for humans to use. It is feasible only if speech signal could be parameterized and evaluated correctly to develop speech understanding system. In this system two components are there, signal processing at front-end and statistical classification at back-end. In this paper, we presents a comparison of techniques for speech signal parameterization such as Mel frequency cepstral coefficient (MFCC), perceptual linear prediction (PLP) in the context of isolated word speech understanding system. Experimental results are computed using standard speech computing device i.e. a headset microphone and sound card along with the help of hidden Markov model toolkit (HTK-3.4.1) in ubuntu 12.4 environment.

1. INTRODUCTION

Feature extraction is the process of extracting the limited amount of useful information from high dimensional data. The goal of feature extraction is to find a set of properties of an utterance that have acoustic correlation to the speech signal, that is, parameters that can somehow be computed or estimated through processing of the signal waveform. Such parameters are terms as features. Speech parameterization techniques divided into two groups:

- Based on Fourier spectrum (MFCC).
- Based on linear prediction spectrum.(PLP)

Mel Frequency Cepstrum Coefficient (MFCC)

We shall explain the step-by-step computation of MFCC [1] in this section as depicted by Figure 1.

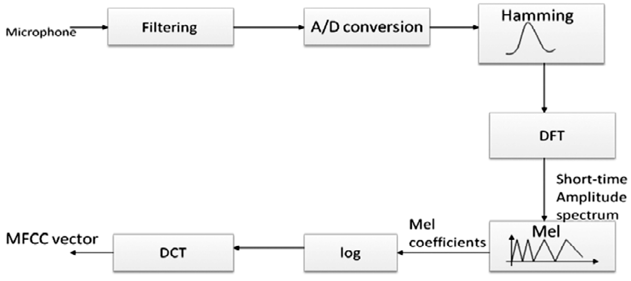

Figure 1: Block Diagram of Speech Analysis Procedure

Steps Involved in Computation of MFCC

1. ***Pre-emphasis:*** The speech signal is sent to a high-pass filter. The z-transform of the filter is: $H(Z) = 1 - a * z^{-1}$ the value of 'a' is usually between 0.9 and 1.0.
2. ***Frame blocking:*** The input speech signal is segmented into frames of 15 to 25 milliseconds with overlap of 30%-70% of the frame size.
3. ***Hamming windowing:*** The next step in the processing is to window each individual frames so as to minimize the signal discontinuities at the beginning and end of each frame . The hamming window is defined as:

$$W(n, \alpha) = (1-\alpha) - \alpha \cos\left(\frac{2\pi n}{N-1}\right), 0 \le n \le N-1$$

4. ***Discrete Fourier Transform (DFT):*** As the filtering operation is applied in the frequency domain, the signal is transformed to frequency domain using Discrete Fourier Transformation (DFT). This transforms the signal from discrete time domain to discrete frequency domain.
5. ***Filter Bank Processing:*** In this block, the energy of the signal in different frequency bands is obtained for further processing. This task is performed by a filter corresponding to a frequency band. The mth filter bank output is given by: $Y_t(m)$, $1 \le m \le M$

The filter can have different shapes triangular as shown in Figure 2, rectangular, Gaussian etc. depending upon the requirement [2].

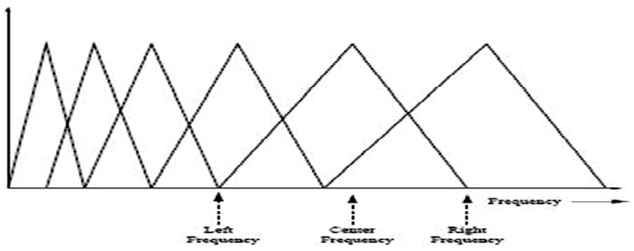

Figure 2: Depiction of Frequency Distribution in Triangular Filter

[1,2] Department of Computer Science and Engineering, Doon Valley Institute of Engineering and Technology
[1]*E-mail: shri.sagwal86@gmail.com,*
[2]*E-mail: ashishchoprakkr@gmail.com*

6. ***Logarithm:*** The next step comprises of computing the logarithm of the square magnitude of the coefficients. Moreover, logarithm performs the dynamic ompression, making feature extraction less sensitive to variations in dynamics.
7. ***Discrete Cosine Transform:*** The final procedure for the Mel Frequency Cepstrum Coefficient (MFCC) consist of performing the Inverse DFT on the logarithm of the filter bank output. The inverse DFT reduces to a Discrete Cosine Transformation (DCT). The DCT has property to produce highly uncorrelated features.

2. PERCEPTUAL LINEAR PREDICTION (PLP)

Steps Involved in Computation of PLP

- Perform frame blocking and windowing on the speech signal.
- Compute the discrete Fourier transform (DFT) and its squared magnitude.
- Integrate the power spectrum hence computed within overlapping critical band filter responses.
- Pre-emphasize the spectrum to simulate the unequal sensitivity of the human ear to different frequencies.
- Compress the spectral amplitudes by taking the cube root after integration.
- Perform an inverse discrete Fourier transform (IDFT).
- Perform spectral smoothing on the critical band spectra using an autoregressive model derived from regression analysis.
- Use an orthogonal transformation like the KLT or the DCT to compute uncorrelated PLPCC.
- Optionally filtering can be performed to equalize the variances of the different cepstral co-efficient.

3. COMPARISON OF MFCC AND PLP

In this section MFCC and PLP are compared on the basis of steps involved in processing the speech. MFCC uses the Mel filter banks to model the hair spacing along the basilar membrane of the ear while PLP uses the Linear Predictive (LP) analysis and Bark scale to model the auditory-like spectrum [3] as depicted by the Figure 3 MFCC analysis computes cepstral coefficients from the log Mel-filter bank using a discrete cosine transform. However in PLP analysis the critical-band spectrum is converted into a small number of LP coefficients through the application of an inverse DFT to provide autocorrelation coefficients. From the LP coefficients, cepstral coefficients are computed, which form the final static feature vector.

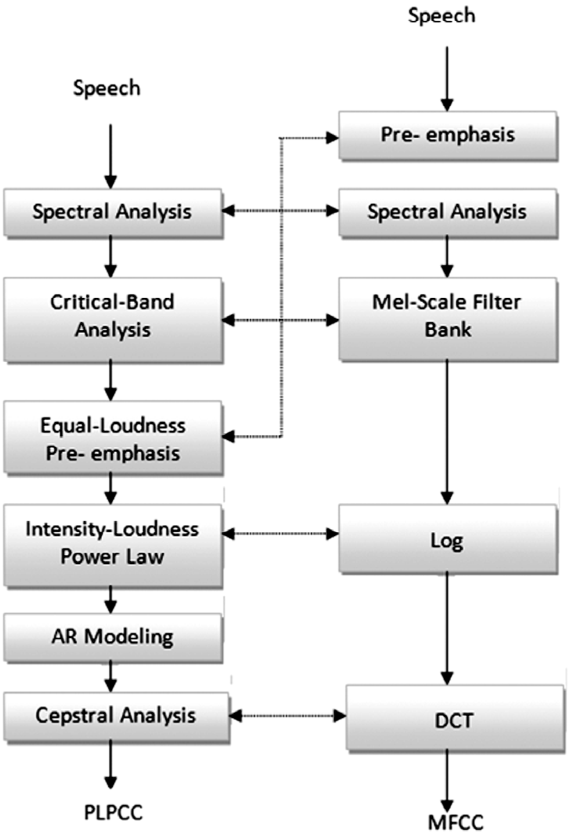

Figure 3: Comparisons of MFCC and PLP

4. EXPERIMENTS AND RESULTS

- ***Experimental Setup***

Many public domain software tools are available for the research work in the field of ASR such as Sphinx from Carnegie Mellon University [4], hidden Markov model toolkit (HTK) from Cambridge University [5]. For our experiment, we have used hidden Markov toolkit HTK-3.4.1 in Ubuntu 12.4. This experiment consist of an evaluation of the system using room condition, and standard speech capturing hardware such as sound card and a headset microphone. Sampling frequency of the signal is 16000 Hz with sample size of 8 bits. To implement it successfully, transcript preparation and dictionary preparation are the most important steps.

- ***Results with Different Dictionary Dize Using MFCC and PLP***

In this experiment, the accuracy of system is observed by varying the size of vocabulary (50 words, 80 words, 120, 150 and 200 words). We have applied speech parameterization techniques both MFCC and PLP on each vocabulary size data. Finally, we achieved better accuracy in case of PLP as shown in Fig. 4:

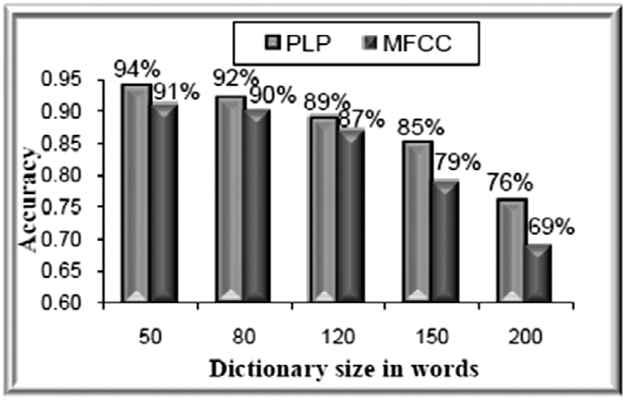

Figure 4: Experiment with Different Vocabulary Dize Using MFCC and PLP

5. CONCLUSION

MFCC and PLP are the two main feature extraction techniques which have been used in most state-of-the-art ASR systems. In this paper, we have presented the comparison of these two feature extraction techniques with various vocabulary sizes. PLP shows 3-4% more accuracy than MFCC in typical Indian office conditions having fan and computer noise.

REFERENCES

[1] Claudio Becchetti and Lucio Prina Ricotti, "Speech Recognition Theory and C++ Implementation", *John Wiley & Sons*.

[2] J.W. Picone, "Signal Modeling Technique in Speech Recognition", *Proceedings of the IEEE*, **81** (9), pp. 1215-1247, 1993.

[3] Ben Milner, "A Comparison of Front-End Configurations for Robust Speech Recognition", *IEEE Transaction of Acoustics, Speech and Signal Processing*, pp.797-800, 2002.

[4] SPHINX: An open source at CMU: http://cmusphinx.sourceforge.net/html/cmusphinx.php.

[5] Hidden Markov Model Toolkit (HTK-3.4.1): http://htk.eng.cam.ac.uk.