INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY

WINGS TO YOUR THOUGHTS…..

Feature Extraction Methods LPC, PLP and MFCC In Speech Recognition

Namrata Dave[1]

[1]G H Patel College of Engineering, Gujarat Technology University, INDIA

Abstract: *The automatic recognition of speech, enabling a natural and easy to use method of communication between human and machine, is an active area of research. Speech processing has vast application in voice dialing, telephone communication, call routing, domestic appliances control, Speech to text conversion, text to speech conversion, lip synchronization, automation systems etc. Nowadays, Speech processing has been evolved as novel approach of security. Feature vectors of authorized users are stored in database. Speech features are extracted from recorded speech of a male or female speaker and compared with templates available in database. Speech can be parameterized by Linear Predictive Codes (LPC), Perceptual Linear Prediction (PLP), Mel Frequency Cepstral Coefficients (MFCC) PLP-RASTA (PLP-Relative Spectra) etc. Some parameters like PLP and MFCC considers the nature of speech while it extracts the features, while LPC predicts the future features based on previous features. Training models like neural network are trained for feature vector to predict the unknown sample. Techniques like Vector Quantization (VQ), Dynamic Time Warping (DTW), Support Vector Machine (SVM), and Hidden Markov Model (HMM) can be used for classification and recognition. We have described neural network in our paper with LPC, PLP and MFCC parameters.*

Keywords: *LPC, MFCC, PLP, Neural Network.*

1. INTRODUCTION

Speech is acoustic signal which contains information of idea that is formed in speaker's mind. Speech is bimodal in nature [1]-[3], Automatic Speech Recognition (ASR) only considers acoustic information contained in speech signal. In noisy environment, it is less accurate [4-5]. Audio Visual Speech Recognition (AVSR) out weights ASR as it uses acoustic and visual information contained in speech.

Speech processing can be performed at different three levels. Signal level processing considers the anatomy of human auditory system and process signal in form of small chunks called frames [6]. In phoneme level processing, speech phonemes are acquired and processed.

Phoneme is the basic unit of speech [7], [8]. Third level processing is known as word level processing. This model concentrates on linguistic entity of speech. The Hidden Markov Model (HMM) can be used to represent the acoustic state transition in the word [6].
The paper is organized as follows: Section 2 describes acoustic feature extraction. In section 3, details of the feature extraction techniques like LPC, PLP and MFCC are discussed. That is followed by description of neural network used for speech recognition in Section 4. Conclusions are given based on survey done on all the three above mentioned methods of speech recognition in last section.

2. BASIC IDEA OF ACOUSTIC FEATURE EXTRACTION

The task of the acoustic front-end is to extract characteristic features out of the spoken utterance. Usually it takes in a frame of the speech signal every 16-32 msec and updated every 8-16 msec [2], [9] and performs certain spectral analysis. The regular front-end includes among others, the following algorithmic blocks: Fast Fourier Transformation (FFT), calculation of logarithm (LOG), the Discrete Cosine Transformation (DCT) and sometimes Linear Discriminate Analysis (LDA). Widely used speech features for auditory modeling are cepstral coefficients obtained through Linear Predictive Coding (LPC). Another well-known speech extraction is based on Mel-frequency Cepstral Coefficients (MFCC). Methods based on Perceptual

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS…..

Prediction which is good under noisy conditions are PLP and RASTA-PLP (Relative Spectra Filtering of log domain coefficients). There are some other methods like RFCC, LSP etc. to extract features from speech. MFCC, PLP and LPC are the most widely used parameters in area of speech processing.

3. FEATURE EXTRACTION METHODS

Features extraction in ASR is the computation of a sequence of feature vectors which provides a compact representation of the given speech signal. It is usually performed in three main stages. The first stage is called the speech analysis or the acoustic front-end, which performs spectra-temporal analysis of the speech signal and generates raw features describing the envelope of the power spectrum of short speech intervals. The second stage compiles an extended feature vector composed of static and dynamic features. Finally, the last stage transforms these extended feature vectors into more compact and robust vectors that are then supplied to the recognizer.

a. Title Mel Frequency Cepstrum Coefficients (MFCC)

The most prevalent and dominant method used to extract spectral features is calculating Mel-Frequency Cepstral Coefficients (MFCC). MFCCs are one of the most popular feature extraction techniques used in speech recognition based on frequency domain using the Mel scale which is based on the human ear scale. MFCCs being considered as frequency domain features are much more accurate than time domain features [9], [10].

Mel-Frequency Cepstral Coefficients (MFCC) is a representation of the real cepstral of a windowed short-time signal derived from the Fast Fourier Transform (FFT) of that signal. The difference from the real cepstral is that a nonlinear frequency scale is used, which approximates the behaviour of the auditory system. Additionally, these coefficients are robust and reliable to variations according to speakers and recording conditions. MFCC is an audio feature extraction technique which extracts parameters from the speech similar to ones that are used by humans for hearing speech, while at the same time, deemphasizes all other information. The speech signal is first divided into time frames consisting of an arbitrary number of samples. In most systems overlapping of the frames is used to smooth transition from frame to frame. Each time frame is then windowed with Hamming window to eliminate discontinuities at the edges [6], [11].

The filter coefficients w (n) of a Hamming window of length n are computed according to the formula:

$$W(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), 0 \le n \le N-1$$

$= 0$, otherwise

Where N is total number of sample and n is current sample. After the windowing, Fast Fourier Transformation (FFT) is calculated for each frame to extract frequency components of a signal in the time-domain. FFT is used to speed up the processing. The logarithmic Mel-Scaled filter bank is applied to the Fourier transformed frame. This scale is approximately linear up to 1 kHz, and logarithmic at greater frequencies [12]. The relation between frequency of speech and Mel scale can be established as:

Frequency (Mel Scaled) = [2595log (1+f (Hz)/700]

MFCCs use Mel-scale filter bank where the higher frequency filters have greater bandwidth than the lower frequency filters, but their temporal resolutions are the same.

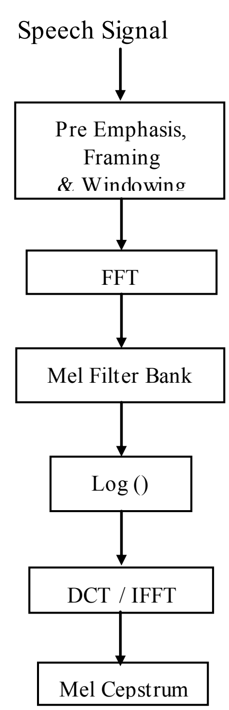

Figure 1: MFCC Derivation

The last step is to calculate Discrete Cosine Transformation (DCT) of the outputs from the filter bank. DCT ranges coefficients according to significance, whereby the 0th coefficient is excluded since it is unreliable. The overall procedure of MFCC extraction is shown on Figure 1.

For each speech frame, a set of MFCC is computed. This set of coefficients is called an acoustic vector which represents the phonetically important characteristics of speech and is very useful for further analysis and processing in Speech Recognition. We can take audio of 2 Second which gives approximate 128 frames each contain 128 samples (window size = 16 ms). We can use first 20 to 40 frames that give good estimation of speech. Total of forty Two MFCC parameters include twelve original, twelve delta (First order derivative), twelve delta-delta (Second order derivative), three log energy and three 0th parameter.

b. Body Linear Predictive Codes (LPC)

It is desirable to compress signal for efficient transmission and storage. Digital signal is compressed before transmission for efficient utilization of channels on wireless media. For medium or low bit rate coder, LPC is most widely used [13]. The LPC calculates a power spectrum of the signal. It is used for formant analysis [14]. LPC is one of the most powerful speech analysis techniques and it has gained popularity as a formant estimation technique [15].

While we pass the speech signal from speech analysis filter to remove the redundancy in signal, residual error is generated as an output. It can be quantized by smaller number of bits compare to original signal. So now, instead of transferring entire signal we can transfer this residual error and speech parameters to generate the original signal. A parametric model is computed based on least mean squared error theory, this technique being known as linear prediction (LP). By this method, the speech signal is approximated as a linear combination of its p previous samples. In this technique, the obtained LPC coefficients describe the formants. The frequencies at which the resonant peaks occur are called the formant frequencies [16]. Thus, with this method, the locations of the formants in a speech signal are estimated by computing the linear predictive coefficients over a sliding window and finding the peaks in the spectrum of the resulting LP filter. We have excluded 0th coefficient and used next ten LPC Coefficients.

In speech generation, during vowel sound vocal cords vibrate harmonically and so quasi periodic signals are produced. While in case of consonant, excitation source can be considered as random noise [17]. Vocal tract works as a filter, which is responsible for speech response. Biological phenomenon of speech generation can be easily converted in to equivalent mechanical model. Periodic impulse train and random noise can be considered as excitation source and digital filter as vocal tract.

c. Perceptual Linear prediction (PLP)

The Perceptual Linear Prediction PLP model developed by Hermansky. PLP models the human speech based on the concept of psychophysics of hearing [2, 9]. PLP discards irrelevant information of the speech and thus improves speech recognition rate. PLP is identical to LPC except that its spectral characteristics have been transformed to match characteristics of human auditory system.

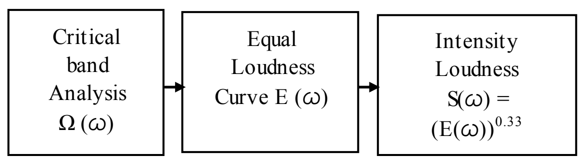

Figure 2: Block Diagram of PLP Processing

Figure 2 shows steps of PLP computation. PLP approximates three main perceptual aspects namely: the critical-band resolution curves, the equal-loudness curve, and the intensity-loudness power-law relation, which are known as the cubic-root.

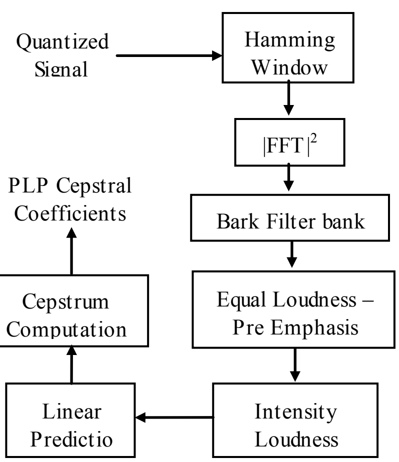

Figure 3: PLP Parameter Computation

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS…..

Detailed steps of PLP computation is shown in figure 3. The power spectrum of windowed signal is calculated as,

$$P(\omega) = \mathrm{Re}(S(\omega))^2 + \mathrm{Im}(S(\omega))^2$$

A frequency warping into the Bark scale is applied. The first step is a conversion from frequency to bark, which is a better representation of the human hearing resolution in frequency. The bark frequency corresponding to an audio frequency is,

$$\Omega(\omega) = 6ln\left[\left[\frac{\omega}{1200\pi} + \left[\left(\frac{\omega}{1200\pi}\right)^2 + 1\right]^{0.5}\right]\right]$$

The auditory warped spectrum is convoluted with the power spectrum of the simulated critical-band masking curve to simulate the critical-band integration of human hearing. The smoothed spectrum is down-sampled at intervals of ≈1 Bark. The three steps frequency warping, smoothing and sampling are integrated into a single filter-bank called Bark filter bank. An equal-loudness pre-emphasis weight the filter-bank outputs to simulate the sensitivity of hearing. The equalized values are transformed according to the power law of Stevens by raising each to the power of 0.33. The resulting auditory warped line spectrum is further processed by linear prediction (LP). Applying LP to the auditory warped line spectrum means that we compute the predictor coefficients of a (hypothetical) signal that has this warped spectrum as a power spectrum. Finally, Cepstral coefficients are obtained from the predictor coefficients by a recursion that is equivalent to the logarithm of the model spectrum followed by an inverse Fourier transform.

The PLP speech analysis method is more adapted to human hearing, in comparison to the classic Linear Prediction Coding (LPC). The main difference between PLP and LPC analysis techniques is that the LP model assumes the all-pole transfer function of the vocal tract with a specified number of resonances within the analysis band. The LP all-pole model approximates power distribution equally well at all frequencies of the analysis band. This assumption is inconsistent with human hearing, because beyond 800 Hz, the spectral resolution of hearing decreases with frequency and hearing is also more sensitive in the middle frequency range of the audible spectrum [9].

4. NEURAL NETWORK

The Generalization is the beauty of artificial neural network. It provides fantastic simulation of information processing analogues to human nervous system. Multilayer feed forward network with back propagation algorithm is the common choice in classification and pattern recognition. Hidden Markov Model, Gaussian Mixture Model, Vector Quantization are the some of the techniques for acoustic features to visual speech movement. Neural network is one of the good choices among all. Genetic Algorithm can be used with neural network for performance improvement by optimizing parameter combination.

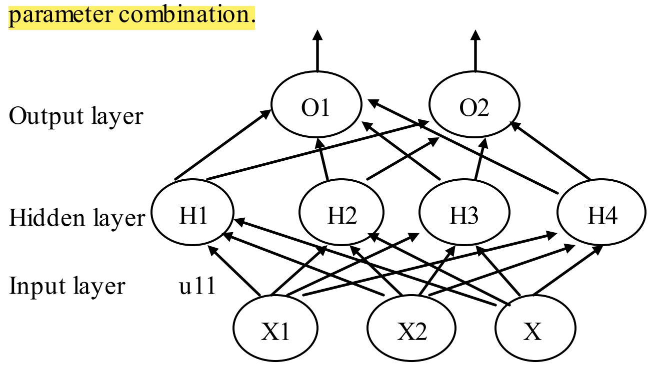

Figure 4: Structure of neural network

We can use multi-layer feed forward back propagation neural network as shown in Figure 4 with total number of features as number of input neurons in input layer for LPC, PLP and MFCC parameters respectively. As shown in Figure 4 Neural Network consists of input layer, hidden layer and output layer. Variable number of hidden layer neurons can be tested for best results. We can train network for different combinations of epochs with goal as minimum error rate.

5. CONCLUSIONS

We have discussed some feature extraction methods and their pros and cons. LPC parameter is not so acceptable because of its linear computation nature. It was seen that LPC, PLP and MFCC are the most frequently used features extraction techniques in the fields of speech recognition and speaker verification applications. HMM and Neural Network are considered as the most dominant pattern recognition techniques used in the field of speech recognition.

As human voice is nonlinear in nature, Linear Predictive Codes are not a good choice for speech estimation. PLP and MFCC are derived on the concept of logarithmically spaced filter bank, clubbed with the concept of human auditory system and hence had the better response compare to LPC parameters.

REFERENCES

[1] Syed Ayaz Ali Shah, Azzam ul Asar, S.F. Shaukat, "Neural Network Solution for Secure Interactive Voice Response," World Applied Sciences Journal 6 (9): 1264-1269, ISSN 1818-4952, 2009.
[2] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," Acoustical Society of America Journal, vol. 87, pp.1738–1752, Apr. 1990.
[3] Corneliu Octavian Dumitru, Inge Gavat, "A Comparative Study of Feature Extraction Methods Applied to Continuous Speech Recognition in Romanian Language," International Symphosium ELMAR, 07-09 June, 2006, Zadar, Croatia.
[4] Tsuhan Chen, Ram Rao, "Audio-Visual Integration in Multimodal Communication," Proc. IEEE, Vol. 86, Issue 5, pp. 837-852, May-1998.
[5] Goranka Zoric, Igor S. Pandzic, "A Real Time Lip Sync System Using A Genetic Algorithm for Automatic Neural Network Configuration," Proc. IEEE, International Conference on Multimedia & Expo ICME 2005.
[6] Goranka Zoric, "Automatic Lip Synchronization by Speech Signal Analysis," Master Thesis, Faculty of Electrical Engineering and Computing, University of Zagreb, Zagreb, Oct-2005.
[7] Andreas Axelsson, Erik Bjorhall, "Real time speech driven face animation," Master Thesis at The Image Coding Group, Dept. of Electrical Engineering, Linkoping University, Linkoping-2003.
[8] Xuewen Luo, Ing Yann Soon and Chai Kiat Yeo, "An Auditory Model for Robust Speech Recognition," ICALIP, International Conference on Audio, Language and Image Processing, pp. 1105-1109, 7-9 July-2008.
[9] Lei Xie, Zhi-Qiang Liu, "A Comparative Study of Audio Features For Audio to Visual Cobversion in MPEG-4 Compliant Facial Animation," Proc. of ICMLC, Dalian, 13-16 Aug-2006.
[10] Alfie Tan Kok Leong, "A Music Identification System Based on Audio Content Similarity," Thesis of Bachelor of Engineering, Division of Electrical Engineering, The School of Information Technology and Electrical Engineering, The University of Queensland, Queensland, Oct-2003.
[11] Lahouti, F., Fazel, A.R., Safavi-Naeini, A.H., Khandani, A.K, "Single and Double Frame Coding of Speech LPC Parameters Using a Lattice-Based Quantization Scheme," IEEE Transaction on Audio, Speech and Language Processing, Vol. 14, Issue 5, pp. 1624-1632, Sept-2006.
[12] R.V Pawar, P.P.Kajave, S.N.Mali "Speaker Identification using Neural Networks," Proceeding of world Academy of Science, Engineering and Technology, Vol. 7, ISSN 1307-6884, August-2005.
[13] Alina Nica, Alexandru Caruntu, Gavril Toderean, Ovidiu Buza, "Analysis and Synthesis of Vowels Using Matlab," IEEE Conference on Automation, Quality and Testing, Robotics, Vol. 2, pp. 371-374, 25-28 May 2006.
[14] B. P. Yuhas, M. H. Goldstein Jr., T. J. Sejnowski, and R. E. Jenkins, "Neural network models of sensory integration for improved vowel recognition," Proc. IEEE, vol. 78, Issue 10, pp. 1658–1668, Oct. 1990.
[15] Ovidiu Buza1, Gavril Toderean1, Alina Nica1, Alexandru Caruntu1, "Voice Signal Processing For Speech Synthesis," IEEE International Conference on Automation, Quality and Testing Robotics, Vol. 2, pp. 360-364, 25-28 May-2006.
[16] Honig, Florian Stemmer, Georg Hacker, Christian Brugnara, Fabio, "Revising Perceptual Linear Prediction ", In INTERSPEECH-2005, pp. 2997-3000. 2005.
[17] Chengliang Li,Richard M Dansereau and Rafik A Goubran , "Acoustic speech to lip feature mapping for multimedia applications", proceedings of the third international symposium on image and signal processing and analysis, vol. 2, pp. 829-832, 18-20 Sept. 2003.