Analysis of Speech Features Extraction using MFCCs and PLP Extraction Schemes

Research · April 2017

DOI: 10.13140/RG.2.2.29803.28969

4 authors, including:

Si Thu Aung
Mahidol University
4 PUBLICATIONS **9** CITATIONS
SEE PROFILE

Hla Myo Tun
Yangon Technological University
108 PUBLICATIONS **63** CITATIONS
SEE PROFILE

International Journal of Scientific Engineering and Technology Research

ISSN 2319-8885
Vol.06,Issue.15
April-2017,
Pages:2969-2973

www.ijsetr.com

Analysis of Speech Features Extraction using MFCCs and PLP Extraction Schemes

SI THU AUNG[1], HLA MYO TUN[2], ZAW MIN NAING[3], WIN KHINE MOE[4]

Abstract: In this paper, we proposed the features extraction and recognition methods of the speech signal using a combination of alternative methods such as perceptual linear prediction (PLP), Mel frequency cepstral coefficients (MFCCs) and hidden Markov models (HMM). We used the significant features of PLP all of which are combined with based line method like MFCCs for features extraction and HMM is used for speech recognition. The experiments is carried out in the meeting room sound environments using recorded human speech that is used to recognize and shown the signal-to-noise ratio of the input and output sound signals. After finishing that, comparison of the SNR ratio and coefficients of the continuous speech signals using different extraction methods. And a combination of the PLP and MFCCs with HMM is better the conventional methods.

Keywords: Speech Recognition, MFCCs, PLP, HMM And Meeting Room Environments.

I. INTRODUCTION

Most of today's automatic speech recognition (ASR) systems are based on some types of Mel-frequency cepstral coefficients (MFCCs), which have proven to be effective and robust under various conditions.[1]MFCCs is commonly used in the acoustic features extraction, while PLP features are reported to be more robust when there is an acoustic mismatch between training and test data.[2] MFCCs is only superior to PLP in clean and no mismatched conditions whereas the acoustic conditions do not remain constant over the whole data set in many applications, for example, broadcast news transcription, segments with clean speech are intermixed with segments that contain background music or noisy telephone speech. In order to achieve good performance it is desirable to have a feature extraction that is well-suited both for clean and unpleasant acoustic conditions. Thus, the favorable properties of PLP and MFCC have to be combined. In spite of the fact that PLP has been derived independently of the MFCC technique, there are many similarities between the two methods, these in a revised feature extraction algorithm that integrates elements taken from the MFCC procedure into PLP. Even through PLP has been developed based on a psycho-physical finding, and interpret the steps of PLP purely in signal processing terms.

Speech carries many information from different sources. [3] Not all information sources are relevant for a given task. Conventional short-term, spectrum-based speech analysis techniques blindly and faithfully represent most information carrying components that are included in the signal. Then, data-intensive stochastic techniques are commonly applied for reducing the effects of the irrelevant information. However, the sources of nonspeech components are often deterministic, their effect on the speech signal is predictable, and the application of the stochastic techniques appears wasteful; the reduction of irrelevant information in the speech analysis module of the recognizer can increase the efficacy of finite amounts of training data. There are two related speech tasks; speech understanding and speech recognition. Speech understanding is getting the meaning of an utterance such that one can respond properly whether or not one has correctly recognized all of the words. Speech recognition is simply transcribing the speech without necessarily knowing the meaning of utterance.

II. SPEECH SIGNAL PROCESSING

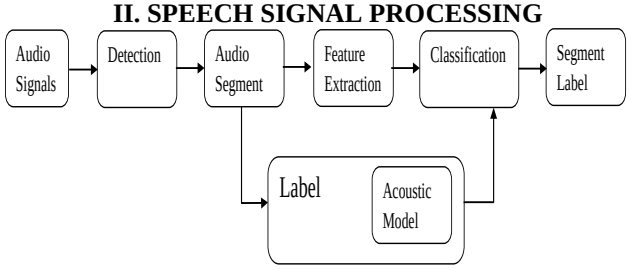

Fig. 1. General block diagram of audio signals processing.

Audio signal processing has basically three main steps as shown in Fig.1. They are preprocessing, feature extraction and recognition or classification. In preprocessing step, audio file is needed to convert the signal using Fourier transform and detect the audio segments. In this paper, we used conversion of audio signal using short-time Fourier transform (STFT). After transformation, the audio signals are extracted with features extraction methods according to its features as shown in Fig.2. Before, recognition or classification this features, they are divided into two sections. First section is used for training process and another is used for testing process. Then, the testing features

are used to matched with the trained data and examine the real acoustic features or error as shown in Fig.3. Finally, these features are used to label the audio segments and some of which are label as broadcast news, meeting room sound, recorded interview sound clips and so on.

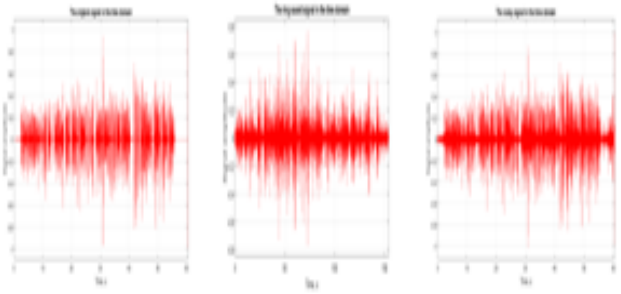

Fig. 2. The sound signals in time domain.

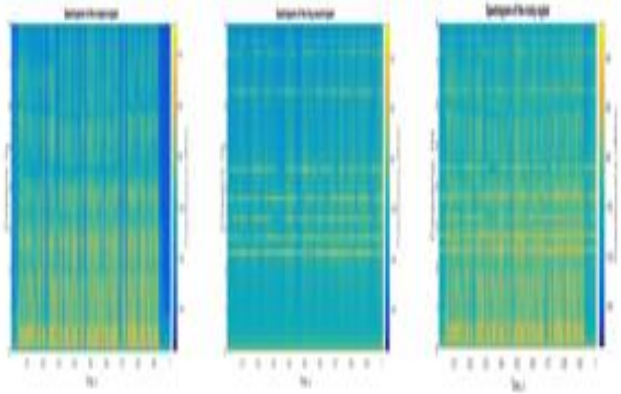

Fig. 3. Spectrogram representation of the sound signals.

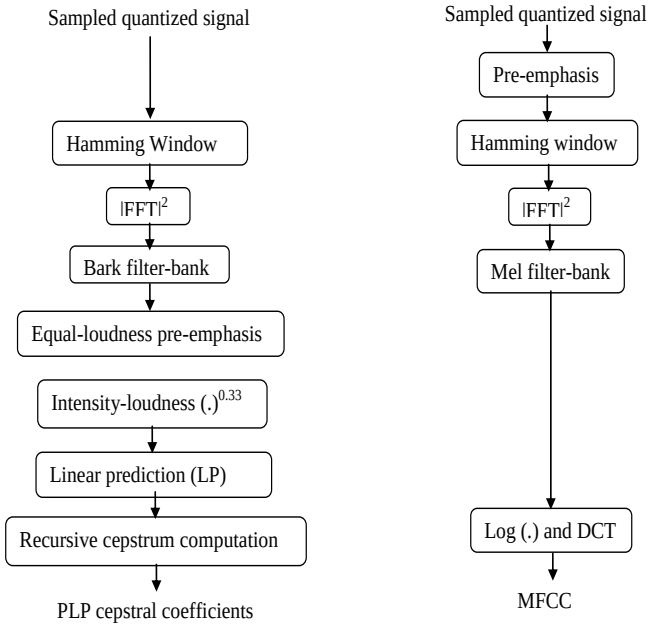

Fig. 4. Step by step procedure of PLP and MFCC.

Fig.4 shows the signal analysis front end of a typical ASR system. The speech waveform which is sampled at 8 or 16 kHz is first differentiated (preemphasis) and cut into a number of overlapping sound segments (windowing), each segments are 25ms long and is shifted by 10ms. A Hamming window is multiplied and the Fourier transform (FFT) is computed for each frame. The power spectrum is changed according to the Mel-scale in order to adjust the frequency resolution to the properties of the human ear. Then the spectrum is segmented into a number of critical bands for a filter bank. The filter bank typically consists of overlapping triangular filters and a discrete cosine transformation (DCT) that is applied to the logarithm of the filter bank outputs and results in the raw MFCC vector. The highest cepstral coefficients are omitted for smoothing and minimize the influence of the pitch which is irrelevant for process of sound signal recognition. The mean of each cepstral component is subtracted, and the variance of each component may also be normalized. Finally, the MFCC vector is increased with time derivates. Additional transformations like linear discriminal analysis (LDA) may further increase the temporal context and together with the discriminance of the acoustic vector. As a result signal analysis provides every 10ms an acoustic vector, which is typically of dimension 25 to 50.

In order to justify the proposed modifications, there is reviewed that PLP and compare it with the MFCC computation. As shown in fig.4, PLP consists of following steps: (i) The power spectrum is computed from the windows speech signal. (ii) A frequency warping into the Bark scale is applied. (iii) The auditorily warped spectrum is convoluted with the power spectrum of the simulated critical-band masking curve to simulate the critical-band integration of human hearing. (iv) The smoothed spectrum is down-sampled at intervals of ≈ 1 Bark. The three steps frequency warping, smoothing and sampling (ii-iv) are integrated into a single filter bank called Bark filter-bank. (v) An equal-loudness pre-emphasis weights the filter-bank outputs to simulate the sensitivity of hearing. (vi) The equalized values are transformed according to power of 0.33. The resulting auditorily warped line spectrum is further processed by (vii) linear prediction (LP). Precisely speaking, applying LP to the auditorily warped line spectrum means that the predictor coefficients of a (hypothetical) signal was computed and that has this changed spectrum as a power spectrum. Finally, (viii), 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.

III. EXPERIMENTS

In this paper, we used two different sound signals for creating overlapping sound event. The first one is 60s recorded human speech signal and another one is 154s ring noise sound. The frequencies of these two sound signals are 16000Hz and 44100Hz respectively. Firstly, these two sound signals are mixed and multiplied with 0.7 noise gain. After that the overlapped sound signal is cut into 25ms long segment and 10ms shifting. Hamming window function is multiplied and fast Fourier transform is computed. Following this, power spectrum is warped according to Mel-frequency scale. Then, the spectrum is segmented into critical bands. The filter bank consists of overlapping

triangular filter for smoothing. Critical band is the band of audio frequency with a second. Finally, discrete cosine transform applied to the algorithm of the filter bank output and results in the raw MFCC vector. This is the extraction of MFCC coefficients.

Fourier Transform of Signal

$$X(e^{j\omega}) = \sum_{k=-\infty}^{\infty} x(k)e^{-j\omega k} \quad (1)$$

Mel-Frequency Scaling

$$\mu(\omega) = 2595.\lg(1+\frac{\omega fs}{2\pi .700Hz}) \quad (2)$$

Cosine Transform

$$\tilde{\mu}(\omega) = \frac{\pi}{\mu(\pi)}.\mu(\omega) = d.\lg(1+\frac{\omega k}{2\pi.700Hz}) \quad (3)$$

$$d = \frac{\pi}{\lg(1+\frac{fs}{2.700Hz})} \quad (4)$$

$$\tilde{\mu}(\omega) = \frac{\pi}{\lg(1+\frac{fs}{2.700Hz})}.\lg(1+\frac{\omega k}{2\pi.700Hz}) \quad (5)$$

For computing PLP coefficients, the power spectrum is computed from windowed speech signal and discrete Fourier transform is also computed. Using Bark filter –bank and this filter-bank consists of three steps; they are frequency warping, smoothing and sampling. An equal-loudness pre-emphasis weight the filter bank outputs for simulating the sensitivity of hearing. Evaluating the intensity and the equalized values are transformed to power of 0.33. Linear prediction (LP) is applied to spectrum for computing predictor coefficients of a signal. Cepstral coefficients are obtained from the predictor coefficients by a recursion which means an inverse Fourier transform. There are many differences between MFCC and PLP and they are filter-bank, equal-loudness preemphasis and the intensity-to-loudness conversion. The speech segment is weighted by the Hamming Window

$$W(n) = 0.54 + 0.46\cos[\frac{2\pi n}{(N-1)}] \quad (6)$$

Where N is the length of the window.

The real and imaginary components of the short-term speech spectrum are squared and added to get the short-term power spectrum

$$P(\omega) = Re[S(\omega)]^2 + Im[S(\omega)]^2 \quad (7)$$

Critical-band spectral resolution.

The spectrum P(ω) is warped along its frequency axis ω into the Bark frequency Ω by

$$\Omega(\omega) = 6\ln\{\frac{\omega}{1200\pi} + [(\frac{\omega}{1200\pi})^2 + 1]^{0.5}\} \quad (8)$$

Where w is the angular frequency in rad/s

$$\varphi(\Omega) = \begin{cases} 0 & \text{For } \Omega<-1.3, \\ 10^{2.5(\Omega+0.5)} & \text{For } -1.3\le\Omega\le-0.5, \\ 1 & \text{For } -0.5<\Omega<-0.5, \\ 10^{-1.0(\Omega-0.5)} & \text{For } 0.5\le\Omega\le2.5, \\ 0 & \text{For } \Omega>2.5, \end{cases} \quad (9)$$

The discrete convolution of □(Ω) with (the even symmetric and periodic function) yields P(ω)samples of the critical-band power spectrum

$$\theta(\Omega_i) = \sum_{\Omega=-1.3}^{2.5} P(\Omega-\Omega_i)\varphi(\Omega) \quad (10)$$

Estimate the temporal derivative of the log critical-band spectrum using regression line through five consecutive spectral values. Nonlinear processing (such as applying threshold or median filtering) can be done in this domain. Re-integrate the log critical-band temporal derivative using a first order IIR system. The pole position of this system can be adjusted to set the effective window size. Currently, we set this value to 0.98, providing an exponential integration window with 3-dB point after 34 frames.

Equal-loudness preemphasis

$$E(\omega) = \left[(\omega^2+56.8\times10^6)\omega^4\right] / \left[(\omega^2+6.3\times10^6)^2\times(\omega^2+0.38\times10^9\right]$$

The sampled □[Ω(ω) is preemphasized by the simulated equal-loudness curve

$$\Xi[\Omega(\omega)] = E(\omega)\Theta[\Omega(\omega)] \quad (11)$$

Equation represents a transfer function of a filter with asymptotes of 12 dB/oct between 0 and 400 Hz, 0 dB/oct between 400 and 1200 Hz, 6 dB/oct between 1200 and 3100 Hz, and 0 dB/oct between 3100 Hz and the Nyquist frequency. For moderate sound levels, this approximation is reasonably good up to 5000 Hz. For applications requiring a higher Nyquist frequency, an additional term representing a rather steep (about -18 dB/oct) decrease of the sensitivity of hearing for frequencies higher than 5000 Hz might be found useful.

The equation would be

$$E(\omega) = \left[(\omega^2+56.8\times10^6)\omega^4\right] / \left[(\omega^2+6.3\times10^6)^2\times(\omega^2+0.38\times10^9)(\omega^6+9.58\times10^{26})\right]$$

Intensity-loudness power law

$$\Phi(\Omega) = \Xi(\Omega)^{0.33} \quad (12)$$

Autoregressive modeling.

In the final operation of PLP analysis, Φ(Ω) is approximated by the spectrum of an all-pole model using the autocorrelation method of all-pole spectral modeling. The inverse DFT (IDFT) is applied to Φ(Ω) to yield the autocorrelation function dual to Φ(Ω) . The IDFT is the better choice here than the inverse FFT, since only a few

autocorrelation values are needed. The first M+1 autocorrelation values are used to solve the Yule-Walker equations for the autoagressive coefficients of the M th-order all-pole model as shown in Figs.5 to 9. The autoregressive coefficients could be further transformed into some other set of parameters of interest, such as cepstral coefficients of all-pole model.

Practical consideration

The spectral sample $\Xi[\Omega(\omega_i)]$ is then given as

$$\Xi[\Omega(\omega_i)] = \sum_{\omega=\omega_{i1}}^{\omega_{ih}} \omega_i(\omega) P(\omega) \quad \textbf{(13)}$$

The equation using the inverse of the angular frequency

$$\omega = 1200\pi \sinh(\Omega/6)$$

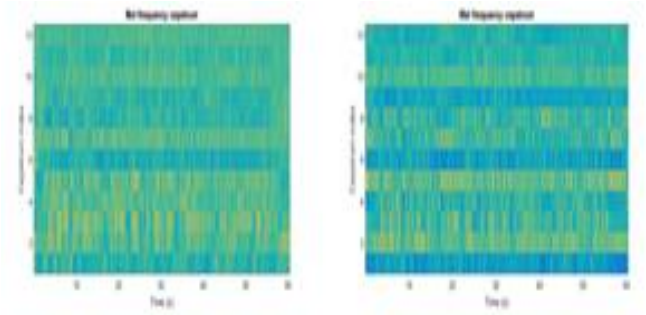

Fig. 5. Mel-frequency cepstral coefficients of the original signals and noisy signals.

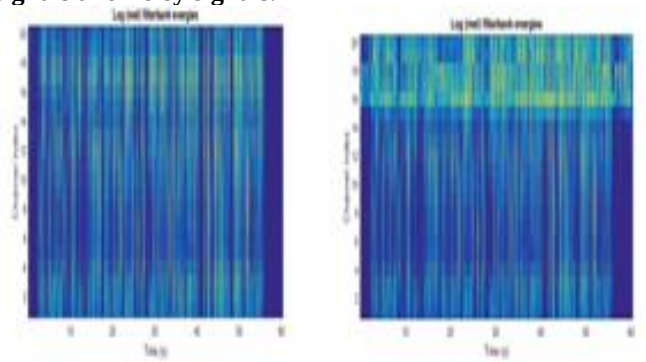

Fig. 6. Log (Mel) frequency coefficients of the original signals and noisy signals.

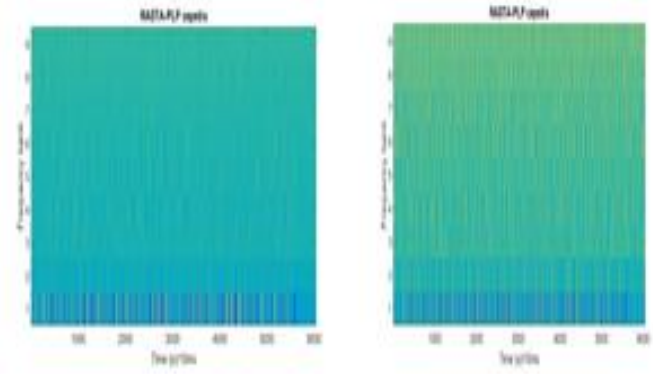

Fig. 7. Perceptual linear predictive (PLP) coefficients of the original signals and noisy signals.

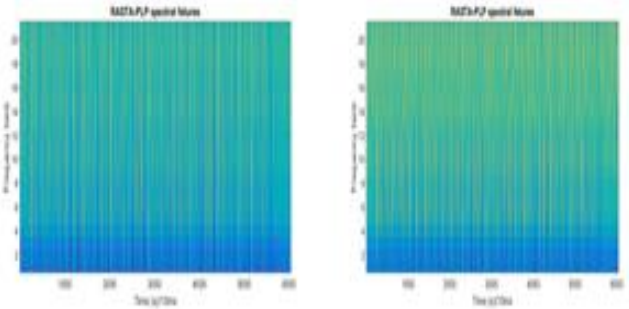

Fig. 8. Perceptual linear predictive (PLP) coefficients of the original signals and noisy signals

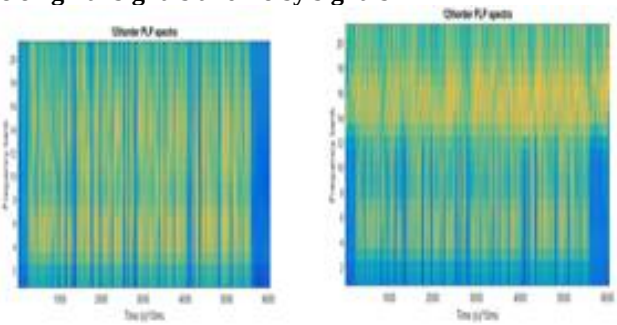

Fig. 9. Perceptual linear predictive (PLP) coefficients of the original signals and noisy signals

TABLE I: Statistics Table For Results

Name	Values
Mel-frequency of original speech	8231.5571Hz
Mel-frequency of noisy speech	10792.3016Hz
Mel-frequency warping of original speech	23.0764Hz
Mel-frequency of warping noisy speech	17.619Hz
SNR input	5.4783dB
SBR output	5.5926dB

IV. CONCLUSION

According to the experiment, spectrogram representation is more suitable for speech signals processing than other methods. Mel-frequency cepstrum coefficients and its log power are shown. Four types of RASTA-PLP features are also mentioned. MFCC is better extraction method than PLP while clean sound signals and no mismatched conditions are needed. So, a combination of these extraction methods yields a good result for features of overlapped sound signals.

V. REFERENCES

[1] Sirko Molau, Michael Pitz, Ralf Schluter, and Hermann Ney," Computing Mel-frequency cepstral coefficients on the power spectrum", Lehrstuhl fur Informatik VI, Computer Science Department, RWTH Aachen-University of Technology, 52056 Aachen, Germany.

[2]Florian Hong, Georg Stemmer, Christian Hacker, Fabio Bruhnara, " Revising Perceptual Linear Prediction (PLP)", Lehrstuhl fur Mustererkennung, Universitat Erlangen – Nurnberg, Germany and ITC-irst-Centro per la Ricera Scientifica e Technologica, Povo di Trento, Italy.

[3]Hynek Hermansky, Nelson Morgan, Aruna Bayya, Phil Kohn," RASTA-PLP speech analysis",US West Technologies, 4001 Discovery Drive, Boulder, CO 80303, International Computer Science Institute, 1947 Center street, Berkeley, CA 94704.

[4]Hynek Hemansky, Nelson Morgan," RASTA Processing of speech", IEEE Transaction on Speech and Audio Processing, vol. 2, no. 4, October 1994

[5]Hynek Hermansky, " Perceptual Linear Predictive (PLP) analysis of speech", Speech Technology Laboratory, Division of Panasonic Technologies, Inc., 3888 state street, Santa Barbara, California 93105, accepted for publication 27 Nov 1989.

[6]D.B.Paul," Speech Recognition Uisng Hidden Markov Models", The Lincoln laboratory Journal, vol. 3, no. 1, 1990.

[7]Daniel Jurafsky and James H.Martin," Hidden Markov Model", Speech and language processing, draft of 1 sep 2014.

[8]L.R.Rabiner, R.W.Schafer," Digital Processing of Speech Signals", 1978 by Bell Laboratories, Incorporated, pp.10-30.

[9]Christophe Levy, Georges Linares and Pascal Nocera, "Comparison of Several Acoustic Modeling Techniques and Decoding Algorithms for Embedded Speech Recognition Systems", 84911 Avignon, France.

[10]Josef Psutka, Ludek Muller and Josef V.Psutka, "Comparison of MFCC and PLP Parameterizations in the Speaker Independent Continuous Speech Recognition Task", University of West Bohemia, Department of Cybernetics, Univerzitni 8, 30614 Pilsen, Czech Republic.