euronoise

vowels recognition using mellin transform and plp-based feature extraction

Mahdi Jamaati, Hossein Marvi and Milad Lankarany

Technical University of Shahrood, 12345 Shahrood, Iran
mahdi.jamaati@gmail.com

Feature extraction for speech recognition is a subject of major interest today. Different feature have been investigated in speech recognition systems. The Mel – Frequency Cepstral Coefficient (MFCC) and Perceptual Linear Prediction (PLP) have usually reported to have yielded good performance. The Mellin transform, and the restriction version called Scale transform, can represent a signal in terms of scale. In this paper, a new method is presented which combines feature extracted from improved cepstrum and mellin transform with the PLP features. Preliminary experiment show that this approach posses promising result.

1 Introduction

One of the first decisions in any pattern recognition system is the choice of what features can be used and how exactly to represent the basic signal that is to be classified, in order to make the classification task easiest. Speech recognition is a typical example. Through more than 30 years of recognizer research, many different feature extractions of the speech signal have been suggested and tried. The most popular feature representation currently used is the Mel-Frequency Cepstral Coefficients (MFCC). Another popular speech feature representation is known as Perceptual Linear Prediction (PLP).

Vowel recognition has been used in speech recognition system for large vocabulary and isolated vowel recognition. The vowels sound produced is determined primarily by the position of the tongue, but the positions of the jaw, lips, and to a small extent, the velum, also influence the resulting sound . Although the length of a speaker's vocal tract is dependent on the positions of the lips and the larynx, to a first approximation it may be regarded as constant. Much of the variability between the voices of men, women, and children is due to differences in the mass of the vocal folds and the length of the vocal tract. For a given vowel, these differences lead to significant differences in both the Glottal Pulse Rate (GPR), perceived as voice pitch, and the frequencies of the most prominent spectral peaks (formants).[4]

The modulation frequency (pitch) and formant frequencies of a spoken vowel depend on the speaker. From a temporal point of view, the signals of the same real vowels pronounced by the same person, but with different pitch are not equal. The vocal tract of every person can be different only in length between different persons. Since studying a signal only from a temporal point of view cannot reveal all the information that it carries, studying the same signal in other domains can be helpful in revealing other. The mathematical tool allows to pass from representation in time to representation in frequency. So the idea is to take the spectrum of the signal (only positive frequency), extract the envelope and apply the scale transform to the envelope and finally calculate plp parameters.

Since the envelope of the discussed signals stays the same (with a different compression factor for different length vocal tracts), making a scale transform gives us a magnitude distribution identical for all this signals. [1]

This paper proposed a new method which combined the feature extracted from improved cepstrum method and mellin transform with PLP features.

The organism of this paper is as follows. In section 2 the cepstrum method are described. In section 3 definition of Mellin and scale transform are given. A PLP algorithm presented in section 4. The Mellin – PLP based feature is introduced in section 5. Recognition experiments are shown in section 6, followed by a conclusion in section 7.

2 Cepstrum method

There are different algorithms and ideas on how to extract a spectral envelope (Channel Vocoder, LPC, cepstrum). One of the well known methods is cepstrum analysis. The cepstrum (backward spelling of "spec") method allows the estimation of a spectral envelope starting from the Fourier transform of the signal.

At the beginning zero padding and Hanning, Hamming or Gaussian windows can be used depending on the number of points used for the spectral envelope estimation. Then Fourier transform is taken as follow.

$$X(k)=\sum_{n=0}^{N-1}x(n)W_N^{kn}=\left|X(k)\right|e^{j\varphi_x(k)},\quad k=0,1,...,N-1 \tag{1}$$

By taking logarithm and performing an IFFT of $\hat{X}(k)=\log X(k)$, which yields complex cepstrum

$$\hat{x}(n)=\frac{1}{N}\sum_{k=0}^{N-1}X(k)W_N^{-kn} \tag{2}$$

The real cepstrum is given by

$$\hat{X}_R(k)=\log\left|X(k)\right| \tag{3}$$

And performing an IFFT of $\hat{X}_R(k)$, which leads to the real cepstrum

$$c(n)=\frac{1}{N}\sum_{k=0}^{N-1}\hat{X}_R(k)W_N^{-kn} \tag{4}$$

Since $\hat{X}_R(k)$ is an even function, the inverse discrete Fourier transform of $\hat{X}_R(k)$ gives an even function *c(n),* which is related to the complex cepstrum $\hat{x}(n)$ by $c(n)=\frac{\hat{x}(n)+\hat{x}(-n)}{2}$.

Figure 1 illustrates the computational steps for the computation of the spectral envelope from the real cepstrum. The real cepstrum *c(n)* is the IFFT of the logarithm of the magnitude of FFT of the windowed sequence *x(n)*. The lowpass window for weighting the cepstrum *c(n)* is given by

$$\omega_{LP}(n) = \begin{cases} 1 & n = 0, N_1 \\ 2 & 1 \le n < N_2 \\ 0 & N_1 < n \le N-1 \end{cases} \quad with \quad N_1 \le N/2. \tag{5}$$

The FFT of the windowed cepstrum $c_{LP}(n)$ yields the spectral envelope

$$C_{LP}(k) = FFT[c_{LP}(n)] \tag{6}$$

which is a smoothed version of the spectrum *X (k)* in dB. [2]

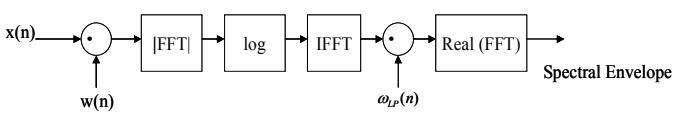

Figure 1 Spectral envelope computation by real cepstrum analysis.

3 The Mellin and Scale Transform

The mellin transform is an integral transform defined as

$$M_f(s) = \int_0^\infty f(t)\, t^{s-1} dt \tag{7}$$

where *s* is a mellin parameter. By setting $s = \sigma + j2\pi\beta$ in Mellin transform we can obtain

$$M_f(\sigma + j2\pi\beta) = \int_{-\infty}^{\infty} f(e^{-t})\, e^{-\sigma t}\, e^{-j2\pi\beta t} dx \tag{8}$$

in the other words, mellin transform of *f(t)* is identical to the Fourier transform of $e^{t\beta} f(e^t)$. The scale transform is a particular restriction of the Mellin transform when $s = 1/2 - jc$, with $c \in \Re$. It can be defined as:

$$D_f(c) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(t) e^{(-jc-1/2)\ln t} dt \tag{9}$$

Therefore

$$D_f(c) = F[e^{t\beta} f(e^t)] \tag{10}$$

The key property of the scale transform is the scale invariance. This transform is said to be scale-invariant, thus meaning that the signals differing just by a scale transformation (compression or expansion with energy preservation) have the same transform magnitude distribution. A scale modification is a compression or expansion of the time axis of the original function that preserves signal energy. Thus, a function *g(t)* can be obtained with a scale modification from a function *f(t),* if $g(t) = \sqrt{a} f(at)$, with $a \in \Re$ and $a > 0$. Given a scale modification with parameter *a*, the scale transform magnitude of the original signal and scaled signals are

$$\left|D_g(c)\right| = \left|D_f(c)\right| \tag{11}$$

4 PLP algorithm

PLP was originally proposed by Hynek Hermansky in 1989 [8] as a way of warping spectra to minimize the differences between speakers while preserving the important speech information. This method essentially provides a more auditory like feature, based on a linear prediction technique. For the feature to be more consistent with human hearing, the spectral analysis must be formed on a warped frequency scale, or in such a way that particular frequency regions are more frequency sensitive than others. One of the advantages of PLP is to include the attributes of the psychological processes of human hearing into analysis.

Perceptual Linear Prediction (PLP) use combination of several engineering approximations of psychology of human hearing processes. Critical band analysis simulated by an auditory-based warping of the frequency axis is derived from the frequency sensitivity of human hearing. In original approach, Bark scale warping function is employed [5]:

$$F_{Bark} = 6\ln\left[\frac{f}{600} + \sqrt{\left(\frac{f}{600}\right)^2 + 1}\right] \tag{12}$$

In PLP filter bank analysis window shape is designed to simulate critical bank masking curves that is showed in Figure2.[9] Both allocate more filters to the lower frequencies, where hearing is more sensitive. In order to compensate the unequal sensitivity of human hearing at different frequencies, the next processing stage in PLP analysis simulates equal loudness curve, such as [5]:

$$E(\omega) = \frac{(\omega^2 + 56.8*10^6)\omega^4}{(\omega^2 + 6.3*10^6)^2(\omega^2 + 0.38*10^9)} \tag{13}$$

Next processing stage called intensity-loudness power law models the non-linear relation between the intensity of sound and its perceived loudness. That a cubic root compensation of critical band energies as follow is applied:

$$l(\omega) = E(\omega)^{\frac{1}{3}} \tag{14}$$

Next step is determined lp parameters of speech that is used in our paper for estimate all pole model of input signal. Last block in general plp algorithm is post processing filter that product plp feature for speech recognition purpose. Plp block is shown in figure 3. [5]

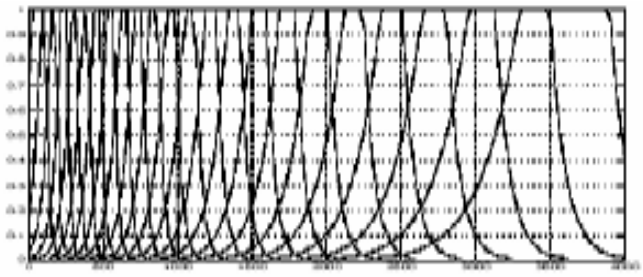

Figure 2. Bank of bark for f_{sample}=8khz

5 Mellin – PLP based feature

In vowels recognition, it is desired that to map the same vowels to the same class and different vowels in distinctive class. The block diagram of mellin – PLP based feature is shown in figure 4.

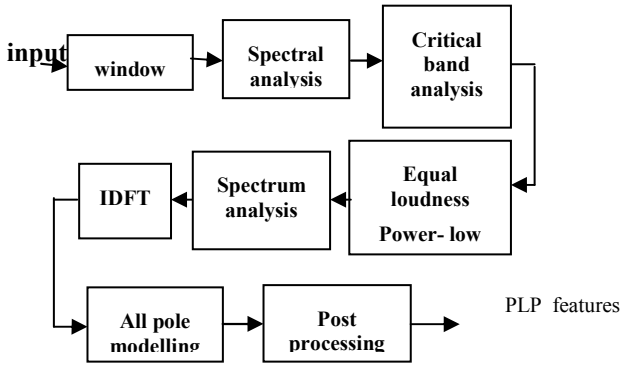

Figure 3. Block diagram of PLP

The system can be viewed as a sequence of steps: In the first step, the spectrum of speech signal is taken. Then the envelope is extracted. But, instead of real cepstrum, we used the improved cepstrum method, which apply the power function (i.e. $10^{|fft(x)|}$), to obtain the spectral envelope of windowed signals. By applying the improved cepstrum method causes to reach better results than original cepstrum method and other methods such as phase vocoder and linear predictive coding (LPC).

In the second stage, was calculated the scale transform magnitude of the spectral envelope. We used the differential of scale transform instead of scale transform. By using this, we have better representation for same vowels in each class.

Moreover, by using the fast Mellin transform [6] algorithm we can achieve the scale magnitude quickly. Energy normalization of scale magnitude has been done in this step.

In the third step, the plp algorithm which is described in section 4 is applied to the output of previous step. Ten plp coefficients are computed for each vowel. Classification of the vowels has been done in the last step.

The purpose of this step is to determine to which class, a given input vowel sample belongs. This is based on a set of features extracted from third step, which make up the feature vector. The classifier uses these features to assign an input vowel to the correct class. A template matching involves a comparison of an average of features, computed on the test pattern, to a collection of stored average for each of the classes in training which is know as pattern. In our experiment a simple template matching, where the whole pattern is compared with a references pattern by measuring the Euclidean distance between features means are used.

6 Recognition Experiments

In order to assess the effective of the proposed mellin – PLP based feature for speech recognition, experiment were conducted on a vowel data base which has been taken from [8]. Five vowels are used in the experiment, which are shown in table 1.

Each vowel is represented by only ten coefficients or in same cases with less than ten coefficients, which are achieved by proposed algorithm.

For testing process, as mentioned before, the Euclidean distance is used. Table 2 shows the accuracy rates of vowel recognition and confusion matrix of vowels which are obtained using the proposed feature.

From this table, we can see that most of tested vowel has an accuracy of %100 but only when the vowel "aa" is tested, the accuracy rate is %90.

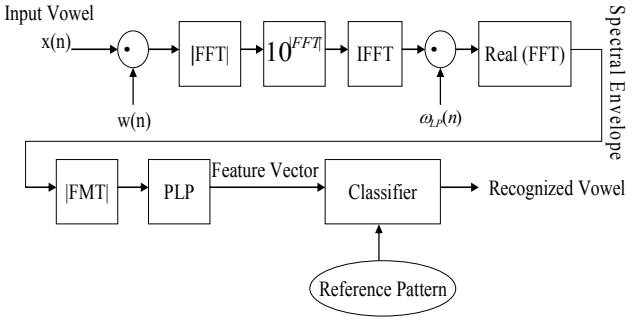

Figure 4 Vowels recognition by using proposed algorithm

Vowel	Example
ii	beet
uu	foot
aa	hot
ee	bet
oo	bought

Table 1 Vowels used in the experiments.

Test \ Result	ii	uu	aa	ee	oo
ii	24	0	0	0	0
uu	0	24	0	0	0
aa	3	0	21	0	0
ee	0	0	0	24	0
oo	0	0	0	0	24
Accuracy Rate	%100	%100	%87	%100	%100

Table2 The accuracy rate and confusion matrix of vowel recognition

7 Conclusion

A novel feature extraction has been suggested in this paper. This method combine feature extracted from improved cepstrum and Mellin transform with the PLP features. It is shown in this paper that each vowel can be represent by only ten coefficients or even less than 10 coefficients in same cases.

Furthermore, experimental results indicate that an accuracy rate of around %100 can be achieved by using these proposed features in application of vowel recognition.

References

[1] Antonio De Sena and Davide Rocchesso , "A Study on using the Mellin transform for vowel recognition" .

[2] D. Arfib, F. Keiler, and U. Z¨olzer. "Source-filter processing". In U. Z¨olzer, editor, *Digital Audio Effects*, pages 299–461. John Wiley and Sons, Ltd., Chichester Sussex, UK, 2002.

[3] Lawrence Rabiner Biing-Hwang Juang , "Fundamentals of speech recognition".

[4] David R. R. Smith and Roy D. Patterson, "The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age".

{5] Petr Motlicek, “Feature Extraction in Speech Coding and Recognition”, March 19, 2003, Report of PhD research internship in ASP Group, OGI-OHSU.

[6] Antonio De Sena and Davide Rocchesso, 2007. "A fast mellin and scale transform". EURASIP Journal on Applied Signal Processing. Volume 2007 , Issue 1 (January 2007) Page: 75

[7] http://www.utdallas.edu/~assmann/KIDVOW

[8] H.Hermansky. "Perceptual Linear Prediction (PLP) Analysis of speech". The Journal of the acoustical Society of America, 87(4): 1738-1752,1990.

[9] Petr Motlicek, “Feature Extraction in Speech Coding and Recognition”, March 19, 2003, Report of PhD research internship in ASP Group, OGI-OHSU.