MODIFIED MEL-FREQUENCY CEPSTRUM COEFFICIENT

Li Tan and Montri Karnjanadecha

Department of Computer Engineering
Faculty of Engineering
Prince of Songkhla University
Hat Yai, Songkhla
Thailand, 90112
E_mail: litan212@hotmail.com, montri@coe.psu.ac.th

ABSTRACT

This paper describes the principle of MFCC feature extraction and the knowledge of human auditory masking effect in order to introduce a modified-MFCC feature extraction that can improve the robustness of speech recognition systems.

KEYWORDS

MFCC (Mel-frequency Cepstrum Coefficient), HMM (Hidden Markov Model), MPEG (Moving Pictures Expert Group), SMR (Signal to Masking Ratio)

1. INTRODUCTION

The structure of a typical speech recognition system mainly consists of preprocessing, feature extraction, training and recognition. Because of the unstability of speech signal, feature extraction of speech signal becomes very difficult. There exist different features between each word. For each word there are differences among different person, such as the differences between adults and children, male and female. Even for the same person and the same word there also exists changes for different time. Nowadays, there are several feature extraction methods used in speech recognition systems. All of them have good performance when used in clean condition. In the adverse condition, we still can’t find a good way in speech recognition system. Compared with them, human auditory system always has good performance under clean and noisy condition. So a way to research our auditory system and use the result in speech recognition system is developed.

There are two major approaches to feature extraction: modeling human voice production and perception system. For the first approach, one of the most popular features is the LPC (Linear Prediction Coefficient) feature. For the second approach, the most popular feature is the MFCC (Mel-Frequency Cepstrum Coefficient) feature. Both features work well in clean but not so in adverse environments.

Here we decide to pursue the auditory system-based approach MFCC and human auditory masking-effect to develop a noise robust front-end since human auditory-system seems to be rather resilient to various types of noise.

In the paper, the basic idea of MFCC is introduced firstly. The analysis of our auditory masking effect is presented after it. Then the modified-MFCC method used to improve robustness is introduced. Finally the conclusions of the paper are presented.

2. PRINCIPLE OF MFCC

Because of the known variation of the ear’s critical band-widths with frequency, filters spaced linearly at low frequencies and logarithmically at high frequencies have been used to capture the phonetically important characteristics of speech. Pols [3] showed that the first six eigenvectors of the covariance matrix for Dutch vowels of the speakers, expressed in terms of 17 such filter energies, accounted for 91.8 percent of the total variance. The direction cosines of this eigenvectors were very similar to a cosine series expansion on the filter energies. Additional eigenvectors showed an increasing number of oscillations of their direction cosines with respect to their original energies. This result suggested that a compact representation would be provided by a set of mel-frequency cepstrum coefficients. These cepstrum coefficients are the result of a cosine transform of the real logarithm of the short-time energy spectrum expressed on a Mel-frequency scale.

In MFCC, the main advantage is that it uses mel-frequency scaling which is very approximate to the human auditory system. The basic steps are: Preprocessing and FFT, Mel-frequency scaling, Cepstrum. The following paragraphs would give the specific explanation for each step.

2.1 Preprocessing and FFT

Preprocessing mainly includes framing, windowing and pre-emphasis. Speech signal is a
kind of unstable signal. But we can assume it as stable signal during 10–30ms. Framing is used to cut the long-time speech to the short-time speech signal in order to get relative stable frequency characteristics. Windowing is mainly to reduce the aliasing effect, when cut the long signal to a short-time signal in frequency domain. The windowing function that we usually use are hamming, hanning and cosine. FFT is used to convert speech signal from time-domain to frequency-domain. For speech our auditory system seems to be very sensitive to the frequency characteristics. The perceptual attributes just like loudness, Pitch, Timbre and so on seem to have a strong correlation with the physical properties of Intensity, Fundamental frequency, Spectral shape. All of them mainly depend on the frequency characteristics of speech signal, although the connection between them is complex. So nowadays speech signal processing is mainly based on the frequency-domain analysis.

2.2 Mel-frequency scaling and Cepstrum

Researchers have undertaken psychoacoustic experimental work to derive frequency scales that attempt to model the natural response of the human perceptual system, since the cochlea of the inner ear acts as a spectrum analyzer. The complex mechanism of the inner ear and auditory nerve implies that the perceptual attributes of sounds at different frequencies may not be entirely simple or linear in nature. AT&T Bell Labs has contributed many influential discoveries in hearing, such as critical bands. The cochlea in our auditory system acts as if it was made up of overlapping filters having bandwidths equal to the critical bandwidth [2]. So the skill of frequency scaling is used to map linear frequency into human perception. Mel-frequency scale is such a kind of perceptually motivated scale, which is linear below 1khz, and logarithmic above. One mel is defined as one thousand of the pitch of a 1khz tone. As with all attempts, it is hoped that the mel scale more closely models the sensitivity of the human ear than a purely linear scale and provides for greater discriminatory capability between speech segments. Mel-scale frequency analysis has been widely used in current speech recognition system. It can be approximated by equation (1):

\[ B(f) = 1125 \ln(1 + f/700) \]  

Where \( B \) is the Mel-frequency scale, \( f \) is the linear frequency.

Given that the DFT of the input signal in equation (2):

\[ X_a[k] = \sum_{n=0}^{N-1} x[n] e^{-j2\pi kn/N}, 0 \leq k \leq N \]  

And we define mel-frequency filterbank with \( p \) filters \( m_j (j=1,2,\ldots,p) \), where filter \( m \) is triangular filter shown in the figure 1.

![Figure 1. Mel-frequency Filterbank](image)

Each FFT magnitude coefficient is multiplied by the corresponding filter gain and the results accumulated. It can be computed as equation (3):

\[ m_j = \sum_{k=0}^{N-1} |X_a[k]|^2 H_j[k], 0 \leq j \leq p \]  

Where \( H_j[k] \) is the transfer function of filter \( j \). The mel-frequency cepstrum is then the discrete cosine transform of the \( p \) filter outputs. It’s described as equation (4):

\[ c_i = \frac{2}{N} \sum_{j=1}^{p} m_j \cos \left( \frac{\pi}{N} (j - 0.5) \right) \]  

For speech recognition, normally only the first 13 Cepstrum coefficients are used. [1]

3. AUDITORY MASKING EFFECT

Frequency masking is a phenomenon under which one sound cannot be perceived if another sound close in frequency has a high enough level. The first sound masks the other one. Frequency-masking have been determined empirically, with complicated models that take into account whether the masker is a tone or noise, the masker’s level, and other considerations.

If without a masker, a signal is inaudible if its SPL(sound pressure level) is below the absolute threshold of hearing. The absolute threshold of
hearing is a function of frequency [2] that can be approximated by equation (5):

\[
T_q(f) = 3.64(f / 1000)^{-0.8} - 6.5 e^{-0.6(f / 1000 - 3.3)^2} + 10^{-3}(f / 1000)^4 \text{ (dB SPL)}
\]  

(5)

Where \( f \) is the linear frequency. It’s plotted in the figure 2.

Tone-masking noise has been determined empirically that noise with energy \( En(dB) \) masks a tone at bark frequency scale \( b \) if the tones energy is below the threshold. It’s shown in equation (6):

\[
T_T(b) = E_n - 6.025 - 0.275 i + S_m(b) \text{ (dB SPL)}
\]  

(6)

Where the \( S_m(b) \) is given by equation (7):

\[
S_m(b) = 15.81 + 7.5(b + 0.474) - 17.5\sqrt{1 + (b + 0.474)^2} \text{ (dB)}
\]  

(7)

The bark frequency scale \( b \) is one class of critical band scales just like Mel-frequency scale. The bark scale ranges from 1 to 24 Barks, corresponding to 24 critical bands of hearing.

Also noise-masking tone has been determined that a tone at critical band number \( i \) with energy \( E_T(dB) \) masks noise at bark frequency \( b \) if the noise is below the threshold given by equation (8):

\[
T_N(b) = E_T - 2.025 - 0.175 i + S_m(b) \text{ (dB SPL)}
\]  

(8)

The following figure shows both the absolute threshold of hearing and the masked threshold of a tone at 1 kHz with a 69 dB SPL. The combined masked threshold is the sum of the two in the linear domain given by equation (9).

\[
T(f) = 10\log_{10}(10^{0.170f} + 10^{0.170f})
\]  

(9)

Which is approximately the larger of the two and is plotted in figure 3.

In addition to frequency masking, there is a phenomenon called temporal masking by which a sound too close in time to another sound cannot be perceived.

The knowledge of auditory masking effect has been used in perceptual coding successfully just like MPEG (Moving Pictures Expert Group) technique. MPEG file compresses the bit rate according to the auditory masking effect. It finds the peak SPL and calculates the global masking threshold in each sub-band. Then the information SMR (Signal to Masking Ratio) is used for bit allocation in each sub-band, where SMR is the subtraction of SPL and global threshold masking. Global masking threshold considers absolute threshold, tone-masking noise threshold, noise-masking tone threshold. [4] Here auditory masking effect plays an important role in compressing the bit rate. Then the method that uses global threshold of each sub-band as the speech feature can be considered. The signals whose SPL is under the threshold can be thrown away. The effect of noise signal that is under the level of the threshold would be ignored and the robustness of the system can be improved. But here the global masking threshold considers each sample of the signal. It means the big calculations and isn’t suitable for speech recognition system. To get the trade-off between the calculation and robustness of processing, we consider the method described in the following.
4. MODIFIED MFCC

Here we consider using auditory masking effect to modify MFCC and improving robustness of speech recognition system. The basic structure includes Preprocessing and FFT, Masking Threshold, Decimation, Mel-frequency scale, Cepstrum. It’s shown in figure 4.

The differences between Modified-MFCC and original MFCC are the part of Masking-Threshold and Decimation. The basic idea of these two parts is:

1). After the step FFT, the peak value of spectrum in each critical band is calculated. According to the empirical formula of noise masking tone (Here we assume that the peak value is tone), the masking threshold of the peak value is determined. The final masking threshold would consider both of the masking threshold and absolute threshold of hearing. The larger one between them is selected for modifying spectrum.

2). After getting the threshold, the spectrum of signal is compared with the value of threshold. The spectrum whose value is less than threshold are deleted. And the new spectrum is obtained after deleting.

Finally the same method used in MFCC is used to get the feature of the signal. It throws away the spectrums that are useless in our auditory system. The effect of noise that is under the threshold is ignored and then the robustness of system is improved. The limitation of calculation is also considered.

5. CONCLUSIONS

The method described here is based on both of MFCC and auditory masking effect. MFCC uses Mel-frequency scale that is approximate to the critical band of auditory system and have good performance in the current speech recognition system. And the masking effect is already used in speech coding system. But the big calculation makes it unsuitable to be used in speech recognition system. Here we only consider the masking effect of peak value to get the trade-off between the calculation and the performance. Also we can consider different way for using the masking effect. Another method considered is to calculate the threshold of the spectrum in each critical band that is most sensitive to our auditory system. So the effect of the noise to the speech is reduced in such frequency.

The further experiment would be done on the idea. It will be based on the reference HMM-based recognizer that is implemented by HTK toolkit. And the speech database in the Aurora Project would be used as both training and testing data. [5] Then it can be compared with other methods and be evaluated.

REFERENCES

[1].Steve Young and Dan Kershaw “the HTK Book(for HTK version 3)” Copyright Microsoft corporation, July,2000