Speaker recognition using Mel Frequency Cepstral Coefficients (MFCC) and Vector Quantization (VQ) Techniques

Jorge MARTINEZ*, Hector PEREZ, Enrique ESCAMILLA
National Polytechnic Institute (IPN)
Jorge.angel.10@gmail.com

Masahisa Mabo SUZUKI
The University of Electro-Communications(UEC)
jusst@fedu.uec.ac.jp

Abstract

This paper presents a fast and accurate automatic voice recognition algorithm. We use Mel frequency Cepstral Coefficient(MFCC) to extract the features from voice and Vector quantization technique to identify the speaker, this technique is usually used in data compression, it allows to model a probability functions by the distribution of different vectors, the results that we achieve were 100% of precision with a database of 10 speakers.

Keywords- Speech processing, Voice, speaker recognition, MFCC, Discrete Fourier Transform, Vector Quantization.

1. Introduction

In this paper we propose a new method using the MFCC and VQ techniques to improve the speaker recognition process.

The voice is the most common method between humans for communication; currently the speaker recognition algorithms are not 100% reliable, since they do not have the truly capability to determine a speaker among others

The use of a better system for speaker recognition can have applications in the daily life, like opening doors and windows or control different devices. If we can trust in the precision of the speaker recognition algorithm, the voice can turn into an essential tool to activate and operate appliances, and even we can think about driving a car with the voice.

According to M. Gray, the first ideas for using speaker recognition systems were proposed in 1966 by S. Saito and F. Itakura of NTT. Then in 1999, Ericsson developed a voice recognition system for their T series mobile phones. In November 2010, Nuance communications acquired PerSay corp. to develop a new method based on biometric features.

The main difficult of a speaker recognition system is the fact that is impossible to say a word exactly in the same way on two different times. It depends on how fast the word is said, and the tone can be different [1].

2. Speech processing

We can divide speech processing in 5 different categories. Speech coding is usually used to encode the voice, for example digitalize a signal voice in MP3 or WAV format. Speech Recognition is the identification of what the speaker is saying, for a example some text processing software has the capability to recognize the speech and translate it into text, this method is used in dictation tasks. Speech Enhancement is used to maximize the voice of a speaker, for example several audio players use contains some filters to enhance the voice in a song. Speech synthesis is the interpretation of text to voice; this system is commonly used when people cannot use their vocal chords or in automatic phone answerers in big companies. Speaker recognition is the process of automatically recognizing who is speaking on the basis of individual information included in speech waves. This technique makes it possible to use the speaker's voice to verify their identity and control access to services such as voice dialing, banking by telephone, telephone shopping, database access services, information services, voice mail, security control for confidential information areas, and remote access to computers. Speaker recognition can be classified into identification and verification.

3. Principles of speaker recognition

Speaker recognition methods are divided in two types, text-independent and text-dependent methods. Text-independent system the system recognize the speaker without having a certain word in a database, the system extracts the singular characteristics of the speaker´s voice, making possible recognition without saying a precise word.

The Text-dependent system recognize the speaker based on some words or phrases that were previously recorded and stored in a database, for example the speaker say a PIN number or his name to activate a device.

The complexity of the process reduces when we use the Text-dependent method, because we have a database to compare the voices.

We decide to use a Text-dependent identification method, for our project.

4. Extracting speech features

*The author is supported by CONACYT Scholarship.

Voice has an infinitive amount of information, we have to determine who is the person speaking based on the features of the person´s voice. An analysis for the voice in time domain will be very impartial. So an analysis in the frequency domain can be a more viable option.

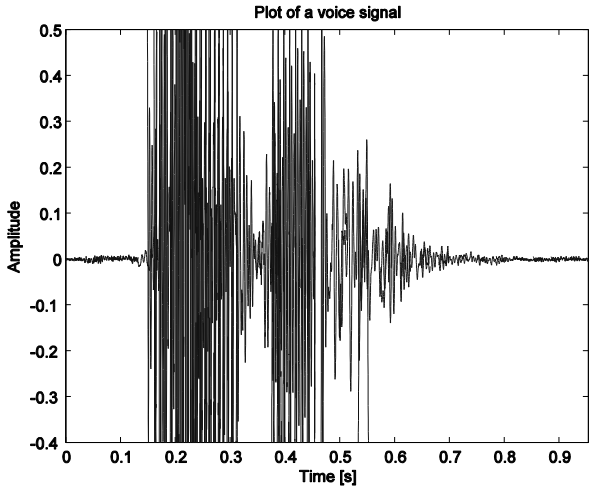


Figure 1. Plot of a voice signal in the domain of time.

We have to convert the speech signal in a digital representation; the speech is a signal that varies with the time, when we examine it in time, between 5ms and 100 its features are fairly fixed. After 0.2 seconds or more, we can see the differences in the speech signals, so a short term spectral analysis is the best way to process the audio signals.

Extract the parametric representation of voice signals is a vital process for the recognition performance.

5. MFCC process

MFCC is a technique based on human hearing behavior that cannot recognize frequencies over 1Khz. MFCC are base on the difference of frequencies that the human ear can distinguish. The signal is expressed in the MEL scale, this scale is based on the perception of the pitches in an equally spaced intervals judged by observers. This scale uses a filter that is spaced linearly at frequencies below 1000 Hz and logarithmic spacing above 1000Hz, in the next paragraphs we will explain the MFCC process.

5.1. Pre-emphasis

In this process we emphasize the higher frequencies; this will increase the energy in the signal at higher frequencies.

5.2. Framing

Is the segmentation of the speech samples in boxes within the range of 20 ms to 40 ms. The voice signal is divide in frames of N samples. Adjacent frames are separated by M(M<N).

Distinctive values used are M = 100 and N= 256.

5.3. Hamming windowing

In signal processing, a window is used when a signal we are interested has a limited length. Indeed, a real signal has to be finite in time; in addition, a calculation is only possible from a finite number of points. To observe a signal in a finite time, we multiply it by a window function.

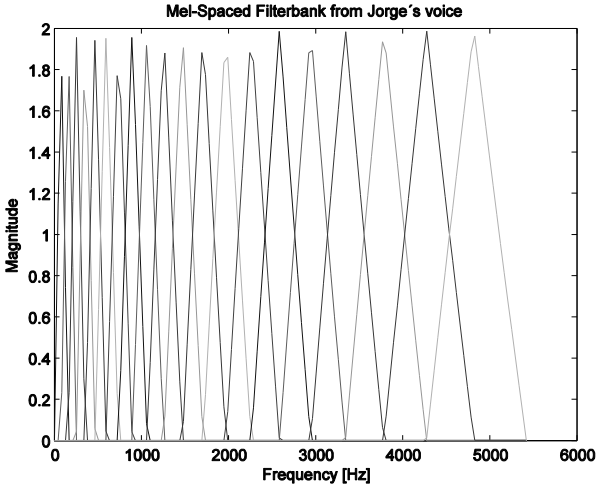
We decide to used Hamming window after doing tests with Triangular window, Rectangular window and Hamming window with values of N = 128, 256 and 512 and M = 50, 100 and 200, the combination of N = 256 and M = 100 gives the best performance. N is equal to the size of the window and M the overlap.

5.4. Fast Fourier Transform

In order to convert every frame of N samples from time domain in frequency domain we have to apply the Fourier Transform to convert the convolution of the glottal pulse and the vocal tract impulse response in the time domain.

5.5. Mel Filter Bank Processing

The range of frequencies in the FFT spectrum is very wide and the voice signal does not follow the linear



scale[2]. To solve this problem we use Mel Filter Bank technique, as shown in figure 2.

Figure 2. Mel Filter Bank from a speaker saying his name.

The filters are used to compute a weighted sum of spectral components to filter the output so the process approaches to the Mel scale. The response of each filter is the given by the frequency magnitude in a triangular shape and is equal to unity at the center frequency and decreases linearly to zero at the center frequency of two adjacent filters. Then, each filter output is the sum of its spectral components filtered. After that we used the next equation to calculate for given frequency f in HZ:

$$F(Mel)=[2595*\log10[1+f]700] \quad (1)$$

5.6. Discrete Cosine Transform

We use this process to change the Mel spectrum to the domain of time using the Cosine Transform (DCT). Doing this we got the Mel Frequency Cepstrum Coefficients, we called the set of coefficients acoustic vectors, consequently each input expression is transformed into a sequence of acoustic vectors.

5.7. Delta Energy and Delta Spectrum

The voice signal and changes frames, as the slope of a formant in their transitions. it is necessary to add features related to the change in the characteristics of cepstral over the time. Adding information from the slopes features is called delta features and adding acceleration features are called double delta acceleration features. We use 12 MFCC, 1 energy feature, 12 delta MFCC features, 12 double-delta MFCC features, 1 delta energy feature and 1 double-delta energy feature, in total 39-dimensional features.

6. Matching process

With Vector quantization technique(VQ) we create a training set of feature vectors, and then we cluster them into a small number of classes that represent each class by a discrete symbol for each class vk, we can compute the probability that it is generated by a given hidden Markov model (HMM) state using Baum-Welch algorithm[3].

In figure 4 we can see that every pair of numbers falling in a particular region, there are 16 regions and 16 red stars, each red star belongs to one region each region is called Voronoi region. The set of all codevectors is called the codebook.

In the training process from our algorithm we generate a VQ codebook for each speaker clustering the acoustic vectors of each speaker in the database [4]. We can see the codeword or centroids in figure 5, they are represented by + red for speaker 1 and with + blue for speaker 2.

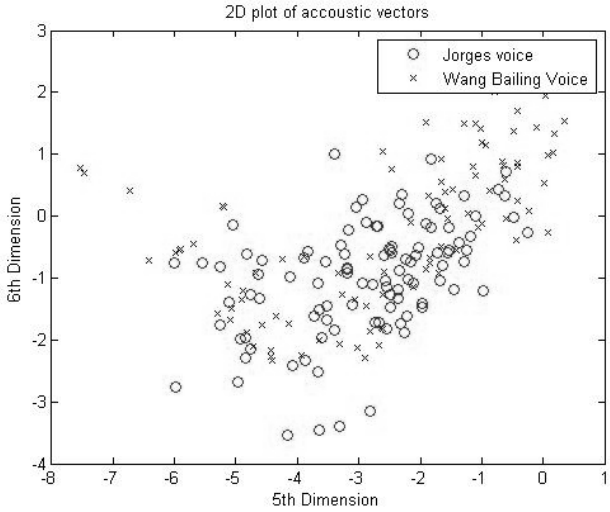


Figure 3. Plot of the data points of the trained VQ codeword of 2 spakers.

The distance of a vector closest to the codeword of a codebook is the named VQ-distortion. In the recognition phase, an input expression of an unknown voice is a “vector-quantized" using each trained codebook and the total VQ distortion is calculated [5]. The speaker in the data base with the smallest distortion is the one that will match with the incoming voice.

7. Tests and Results

The test and train phases were simulated in Matlab. First of all we recorded a database of 20 speakers, we recorded their voices saying their names twice, and we recorded twice because we simulated the train and test phase, we used 22050 as the sampling frequency and 8 bits per sample.

Also we used the English Language Speech Database for Speaker Recognition (ELSDSR) to test our project with 100 speakers.

We decide to used to 20 mel filter banks, we made test using the 10, 20 and 30 filters, the results with 10 filters were 36 mistaken speakers, with 20 filters 18 mistaken speakers and with 30 filters also 18 mistaken speakers.

We made test using codebook sizes of, 1, 2, 4, 8, 16, 32, and 64.

The complete parameters that we used for our algorithm are, sampling frequency of 22050 and 8 bits per sample, the recorded time was 2 seconds, Hamming window with M = 100 and N = 256 for frame blocking process, 20 mel filter banks and a codebook size of 16.

Table 1. Test with 10 speakers.

Amount of speaker	Errors using MFCC and VQ	Errors using LPCC and VQ
10	10%	20%
20	10%	20%
30	13.3%	20%
40	15%	20%
50	16%	22%
60	16.6%	21.66%
70	17.42%	21.42%
80	17.5%	21.25%
90	17.7%	22.22%
100	18%	23%

In this table we compare 10 speakers saying their names.

We achieve 82% percentage of precision using a database of 100 speakers; we compare our system with the traditional technique of LPCC and we got 77% of precision.

8. Conclusion

After analyzing all the information we got from our project, we found that our algorithm improves the speaker recognition task, the main part of our project is the MFCC, and we used these technique instead of LPCC. Using the MFCC we try to represent in a better way the human voice, usually LPCC is used in digital communication, so the main purpose of this technique is not represent the voice, is to compress and transmit the information that contain the voice.

Because the MFCC uses the mel scale, the approximation to the human voice behavior is good, it represent in a better way the voice. We improved the precision in the recognition task; also we increase the number of speaker's database, the original algorithm was only test with 11 speakers.

We found that as long as the number of speakers increase, the number of errors increase, this is because the of distance to each centroid of each speaker can be similar to another, and if we increase the time, we increase the precision, because we have more information to compare, after recording 6 seconds we reach 100% of precision.

After using a codebook size of 64, the best performance is reached, so no matter how much we increase this value, we will not get better results, we will get only more computational cost.

9. References

[1] Lawrence Rabiner and Biing-Hwang Juang, "Fundamental of Speech Recognition", Prentice-HalEnglewood Cliffs, N.J., 1993.

[2] Shannon B.J., Paliwal K.K., "A comparative study of filter bank spacing for speech recognition",Proc. of Microelectronic engi- neering research conference, Brisbane, Australia, Nov. 2003.

[3] Y. Linde, A. Buzo R. Gray, "An algorithm for vector quantizer design" IEEE Transactions on Communications Vol. 28, pp.84-95, 1980.

[4] F. Soong, E. Rosenberg, B. Juang, and L. Rabine, "A Vector Quantization Approach to Speaker Recognition", ATT Technical Journal, vol. 66, March/April 1987, pp. 14-26.

[5] Zhong-Xuan, Yuan Bo-Ling, Xu Chong-Zh Yu, "Binary Quantization of Feature Vector for Robust Text-Independent Speaker Identification in IEEE Transactions on Speech and Audio Processing", Vol. 7, No. 1, January 1999. IEEE, New York, NY, U.S.A.