Effect of MFCC Normalization on Vector Quantization Based Speaker Identification

M.Hassan Shirali-Shahreza
Virtual Education Graduate College
Amirkabir University of Technology
hshirali@aut.ac.ir

Sajad Shirali-Shahreza
Department of Computer Science
University of Toronto
shirali@cs.toronto.edu

Abstract**—***Mel Frequency Cepstral Coefficients (MFCC) are widely used in speech recognition and speaker identification. MFCC features are usually pre-processed before being used for recognition. One of these pre-processing is creating delta and delta-delta coefficients and append them to MFCC to create feature vector. Another pre-processing is coefficients mean normalization.*

In this paper, the effect of these two processes on the accuracy of a Vector Quantization (VQ) speaker identification system is compared. Additionally, it is shown that coefficient variance normalization, which is less common, can improve the accuracy.

Keywords**—** *Normalization, Mel Frequency Cepstral Coefficients (MFCC), Speaker Recognition, Vector Quantization (VQ)***.**

I. Introduction

Speech is one of the primary and natural communication ways between humans. People usually use speech in their interactions. So speech is a good candidate for human users to interact with computers.

Speech is mainly used for two applications: user identification and interaction between humans and computers. In user identification, the goal is to identify or verify a user's identity which are known as speaker identification and speaker verification. To do this, the user's speech should be compared with stored user's samples. For interaction applications, the goal is to recognize what the user said which is known as speech recognition or generate a speech to communicate which is known as speech synthesis.

Although speaker identification and speech recognition seems different at first, they have many similarities. Both of these problems are pattern recognition problems. The raw speech signal is not a good choice to be used as feature vector in these applications. So a series of features should be extracted from speech signal and then used as feature vectors for subsequent processes.

In the speech processing tasks such as speech recognition or speaker identification, the speech signal is examined during a time interval, because the speech information is primarily conveyed in short time periods of speech [1]. Usually speech signal is converted into a number of frames with length of 10-20 milliseconds. Then a feature vector is extracted from each window. The feature extraction phase is common between speech recognition and speaker identification and similar features are used for both purposes. Two common features are Mel Frequency Cepstral Coefficients (MFCC) and Linear Predictive Coding (LPC). These two features are widely used for both speech recognition and speaker identification [2] [3]. It seems that MFCC is better and more popular than LPC [4].

The first step in calculating MFCC is converting the windowed frame of speech into the frequency domain by using the Discrete Fourier Transform (DFT). This is usually done by using Fast Fourier Transform (FFT) algorithm. Then, the log magnitude of complex Fourier coefficients is calculated. After that, a mel scaling is applied on these coefficients to reduce the effects of high frequencies. The mel scaling is created from human audition system [1].

Finally, the Inverse Discrete Fourier Transform (IDFT) is applied on the coefficients to obtain the Cepstral domain coefficients. The first 10-20 low order coefficients are usually used as the feature vector of the frame, because most of the information required for speech recognition and speaker identification is in these coefficients [5]. Some of the speaker identification methods use the first order and second order derivatives of MFCC coefficients as additional features [6]. These derivatives are known as delta and delta-delta coefficients. The previous researchers suggest that appending these derivatives to the feature vector can improve the accuracy of speech recognition and speaker identification [7].

MFCC can be processed further and then be used as feature vectors. A common example is Cepstral mean subtraction [6]. In Cepstral mean subtraction, the means of MFCC features are estimated from speech signal and then this mean value is subtracted from MFCC. We will name this process as mean normalization in this paper. This process can remove linear, static and time invariant noises from the signal [7]. The goal of this paper is to evaluate the effect of using delta features and mean normalization on the accuracy of vector quantization speaker identification.

Moreover, a new type of normalization is proposed and compared with these two processes. The variance normalization, which previously used in speaker verification by using Support Vector Machines (SVM) [7], is tested for use in vector quantization speaker identification. More details about this type of normalization and reasons for using it are described in next section. In the next section, the vector quantization based speaker identification system which is used to obtain results is described. The implementation details of this system and test data information are provided in section 3. The experimental results and their analyses are mentioned in section 4. The last section will conclude the paper.

978-1-4244-9991-5/11/$26.00 ©2011 IEEE

II. VECTOR QUANTIZATION BASED SPEAKER IDENTIFICATION

The main idea of vector quantization speaker identification is to create a codebook from feature vectors of each speaker and then scoring a test speech with respect to these codebooks to find the best matching speaker for the test speech. The main advantage of these systems is their speed. A more formal description of these systems is followed:

In these systems, a codebook is created for each speaker in the training phase and it is used as a model for that speaker. The codebook is created by clustering the feature vectors of speaker's training samples into M clusters. The centers of these clusters are then used as the codebook $C=\{c_1,\ldots,c_M\}$. Although any clustering algorithm can be used, the LBG clustering algorithm [8] is usually used in speaker identification systems [3], because it is fast and its implementation is simple.

In the identification phase, a match score function is used to calculate the score of an unknown speech sample with respect to each speaker's model. Then the speaker that its model has the best score is selected as the speaker of that sample. Usually the average quantization distortion [10] is used as the match score of an unknown speech X which has N frames with a codebook C of size M:

$$D\left(X,C\right)=\frac{1}{N}\sum_{i=1}^{N}\min_{c_j\in C}\left\|x_i-c_j\right\| \qquad (1)$$

III. IMPLEMENTATION DETAILS AND TEST DATA

The system described in previous section is usually used as the base system in vector quantization systems. This base system is used in this paper to compare the effect of different processes and normalization of MFCC on the accuracy of a vector quantization based speaker identification system.

The TIMIT database [9] is used in this paper. TIMIT database is a well known database in speech processing and used for evaluation of different speaker identification and verification systems [6]. There are 630 speakers in TIMIT database where each speaker said 10 sentences. The speakers are from 8 major dialect region of USA. In this paper, a subset of TIMIT is used. The first two male and two female speakers are selected from each region, so, there are 32 speakers in the test set. The sentences said by each speaker are divided into three groups: SA, SI and SX. Each speaker said two SA, three SI and five SX sentences. The SA sentences are the same for all speakers, but the SI and SX sentences are different for different speakers.

The TIMIT database is designed for speech recognition, so the test and train methodology which described in the documentation are applicable for speech recognition tasks. When TIMIT is used for evaluation of a speaker identification or verification system, the sentences which are said by each speaker must be divided into training and testing sets. In this paper, the SA and SX sentences are used as training sentences and the SI sentences are used to evaluate the system and calculate the accuracy of the system. So there are 224 training and 96 testing samples.

A tool available in the CMU SPHINX [11] for extracting the MFCC features is used to extract the MFCC features. Each sample speech is converted into 25 ms frames with 10 ms shifts and the first 15 MFCC are calculated for each frame. These MFCC features are then processed by a processor written in C++ to apply normalizations and delta calculation and create feature vectors. The LBG clustering algorithm [8] is implemented in C++ and used to cluster training data and create speaker's codebooks. Another program, written in C++, is used to calculate the average quantization distortion between a test sample and speaker's codebooks and find the speaker of sample.

The accuracy of system is defined as the number of testing samples identified correctly divided by the total number of testing samples. The accuracy is calculated for two tests. In the first test, a codebook is created for each test sample and this codebook is used to calculate the match score. This is done to represent a system which requires less comparison time such as real-time systems, because the codebook length can be very smaller than the full length sample. In the second test, the full test sample is used for calculating the match score. So the second test is more accurate, but requires more comparison time.

IV. RESULTS AND ANALYSIS

Three feature selection choices are compared in this paper. The first choice is to use or to remove the first MFCC feature. Some methods such as [3] and [12] remove the first MFCC from the feature vector while other methods like [4] keep it in the feature vector. The second choice is to append the delta and delta-delta coefficients to the feature vector, as done in methods like [6]. The third choice is the type of MFCC normalization. The mean normalization is usually used in methods like [7], while there are methods such as [6] that do not use mean normalization. In this paper, a new type of normalization is also proposed. Different MFCC have different variances. If a dimension has more variance, then it has more impact on the cluster centers in algorithms like LBG. To remove this effect, the variances of different coefficients can be normalized to one. This type of normalization was previously proposed for another reason in [7]. In [7], the SVM is used for classification and because the SVM is sensitive to the variances of different features, the variances are normalized. In this paper, the variance normalization is proposed to improve the quality of codebook generation.

The above mentioned choices for feature selections create 12 different types of feature selection. The accuracy of the speaker identification system using each of these 12 feature selection policies are provided in Tables 1, 2 and 3. All of the tables are reporting the same results, but the results are provided in different compositions so that the desired feature selection choice can be easily evaluated. Table 1 shows that removing the first coefficient is improving the accuracy in most cases, especially for small codebooks. These results justify why most methods are removing the first coefficients.

978-1-4244-9991-5/11/$26.00 ©2011 IEEE

TABLE I
COMPARING ACCURACY RESULTS OF USING FIRST MFCC COEFFICIENTS WITH REMOVING FIRST MFCC COEFFICIENT

Test Sample Type	Speaker Model Type	Number of Clusters	Without Delta						With Delta					
			No Normalization		Mean Normalization		Mean & Variance Normalization		No Normalization		Mean Normalization		Mean & Variance Normalization	
			Use First Coefficient	Remove First Coefficient	Use First Coefficient	Remove First Coefficient	Use First Coefficient	Remove First Coefficient	Use First Coefficient	Remove First Coefficient	Use First Coefficient	Remove First Coefficient	Use First Coefficient	Remove First Coefficient
Test Sample Codebook	Speaker Codebook	8	54.2	85.4	35.4	70.8	82.3	83.3	60.4	61.5	36.5	45.8	79.2	83.3
		10	66.7	91.7	36.5	68.8	87.5	90.6	67.7	70.8	43.8	50.0	77.1	87.5
		16	82.3	97.9	56.3	85.4	94.8	96.9	82.3	86.5	59.4	65.6	92.7	90.6
		32	95.8	99.0	70.8	93.8	97.9	96.9	97.9	95.8	71.9	80.2	95.8	95.8
		64	99.0	99.0	83.3	96.9	97.9	97.9	99.0	99.0	89.6	90.6	99.0	96.9
Full Test Sample	Speaker Codebook	8	87.5	97.9	66.7	84.4	95.8	93.8	89.6	86.5	65.6	75.0	94.8	92.7
		10	92.7	95.8	69.8	90.6	99.0	100	91.7	84.4	64.6	76.0	93.8	92.7
		16	97.9	100	77.1	94.8	97.9	99.0	95.8	93.8	79.2	79.2	97.9	95.8

Table 2 shows that using the delta coefficients decreases the accuracy in most cases, expect for the some cases that the first coefficient is used in the feature vector. These results are consistent with results reported in [4] and [13].

Table 3 is comparing the effect of different normalization methods. As mentioned in [12] and [4], the samples of TIMIT are clean and hence the mean normalization does not improve the accuracy and even decrease the accuracy. The results in Table 3 are also supporting this justification. The results of Table 3 also show that the variance normalization which is proposed in this paper can greatly improve the accuracy in comparison with mean normalization. The results of mean and variance normalization are comparable to or better than the results of no normalization. So, it seems that this type of normalization has advantages of mean normalization, while it does not degrade the performance on clean samples. Testing this type of normalization on a noisy database such as those used in [6] is required for more precise judgment about this type of normalization.

V. CONCLUSION

Mel Frequency Cepstral Coefficients (MFCC) are widely used in speech recognition and speaker identification and verification systems. There are a number of choices in creating feature vectors from MFCC such as removing or keeping the first coefficient, using or not using the delta coefficients and finally using or not using the mean normalization.

In this paper, the effects of these choices on the accuracy of a vector quantization based speaker identification system are assessed on a subset of TIMIT database. Additionally, the variance normalization of MFCC is proposed to improve the accuracy. While some previous works have reasons and justifications for the feature selection they used, they usually lack numerical comparison between different choices and only report the best case.

The results of this paper, which are consistent with previous reported results and justifications on feature selection choices, show that removing the first coefficient can improve the accuracy and using the delta coefficients does not improve the accuracy and even decrease the accuracy.

TABLE II
COMPARING ACCURACY RESULTS OF USING DELTA COEFFICIENTS WITH NOT USING DELTA COEFFICIENTS

Test Sample Type	Speaker Model Type	Number of Clusters	Use First Coefficient						Remove First Coefficient					
			No Normalization		Mean Normalization		Mean & Variance Normalization		No Normalization		Mean Normalization		Mean & Variance Normalization	
			Without Delta	With Delta	Without Delta	With Delta	Without Delta	With Delta	Without Delta	With Delta	Without Delta	With Delta	Without Delta	With Delta
Test Sample Codebook	Speaker Codebook	8	54.2	60.4	35.4	36.5	82.3	79.2	85.4	61.5	70.8	45.8	83.3	83.3
		10	66.7	67.7	36.5	43.8	87.5	77.1	91.7	70.8	68.8	50.0	90.6	87.5
		16	82.3	82.3	56.3	59.4	94.8	92.7	97.9	86.5	85.4	65.6	96.9	90.6
		32	95.8	97.9	70.8	71.9	97.9	95.8	99.0	95.8	93.8	80.2	96.9	95.8
		64	99.0	99.0	83.3	89.6	97.9	99.0	99.0	99.0	96.9	90.6	97.9	96.9
Full Test Sample	Speaker Codebook	8	87.5	89.6	66.7	65.6	95.8	94.8	97.9	86.5	84.4	75.0	93.8	92.7
		10	92.7	91.7	69.8	64.6	99.0	93.8	95.8	84.4	90.6	76.0	100	92.7
		16	97.9	95.8	77.1	79.2	97.9	97.9	100	93.8	94.8	79.2	99.0	95.8
		32	100	99.0	83.3	82.3	99.0	97.9	100	100	97.9	86.5	100	99.0
		64	100	100	87.5	90.6	100	99.0	100	100	99.0	88.5	100	97.9

TABLE III

COMPARING ACCURACY RESULTS OF USING DIFFERENT NORMALIZATION METHODS

Test Sample Type	Speaker Model Type	Number of Clusters	Use First Coefficient						Remove First Coefficient					
			Without Delta			With Delta			Without Delta			With Delta		
			No Normalization	Mean Normalization	Mean & Variance Normalization	No Normalization	Mean Normalization	Mean & Variance Normalization	No Normalization	Mean Normalization	Mean & Variance Normalization	No Normalization	Mean Normalization	Mean & Variance Normalization
Test Sample Codebook	Speaker Codebook	8	54.2	35.4	82.3	60.4	36.5	79.2	85.4	70.8	83.3	61.5	45.8	83.3
		10	66.7	36.5	87.5	67.7	43.8	77.1	91.7	68.8	90.6	70.8	50.0	87.5
		16	82.3	56.3	94.8	82.3	59.4	92.7	97.9	85.4	96.9	86.5	65.6	90.6
		32	95.8	70.8	97.9	97.9	71.9	95.8	99.0	93.8	96.9	95.8	80.2	95.8
		64	99.0	83.3	97.9	99.0	89.6	99.0	99.0	96.9	97.9	99.0	90.6	96.9
Full Test Sample	Speaker Codebook	8	87.5	66.7	95.8	89.6	65.6	94.8	97.9	84.4	93.8	86.5	75.0	92.7
		10	92.7	69.8	99.0	91.7	64.6	93.8	95.8	90.6	100	84.4	76.0	92.7
		16	97.9	77.1	97.9	95.8	79.2	97.9	100	94.8	99.0	93.8	79.2	95.8
		32	100	83.3	99.0	99.0	82.3	97.9	100	97.9	100	100	86.5	99.0
		64	100	87.5	100	100	90.6	99.0	100	99.0	100	100	88.5	97.9

Finally, the mean normalization decreases the accuracy on TIMIT database which is a clean database. The proposed variance normalization can retrieve the accuracy decreased by mean normalization and provides results comparable to or better than results of no normalization. Further tests on noisy databases can better asses this type of normalization.

REFERENCES

[1] L. R. Rabiner, and B. H. Juang, *Fundamentals of Speech Recognition,* Prentice-Hall, 1993.

[2] D. Chow and W.H. Abdulla, "Robust speaker identification based on perceptual log area ratio and Gaussian mixture models," *Proceedings of INTERSPEECH 2004,* pp. 1761-1764.

[3] T. Kinnunen, E. Karpov, and P. Franti, "Real-time speaker identification and verification," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 14, no. 1, January 2006, pp. 277-288.

[4] D.J. Mashao, and M. Skosan, "Combining classifier decisions for robust speaker identification", *Pattern Recognition,* vol. 39, no. 1, January 2006, pp. 147-155.

[5] Gish, H. and Schmidt, M., "Text-independent speaker identification," *IEEE Transactions on Signal Processing.,* vol. 11, no. 4,1994, pp. 18-32.

[6] J. Ming, T.J. Hazen, J.R. Glass, and D.A. Reynolds, "Robust Speaker Recognition in Noisy Conditions," *IEEE Transactions on Audio, Speech and Language Processing.,* vol. 15, no. 5, July 2007, pp. 1711-1723.

[7] Wan, V., *Speaker Verification using Support Vector Machines,* PhD Dissertation, Department of Computer Science, University of Sheffield, 2003.

[8] Y. Linde, A. Buzo, and R.M. Gray., "An algorithm for vector quantizer design," *IEEE Transactions on Communications.,* vol. 28, no. 1, 1980, pp. 84–95.

[9] J. S. Garofolo, et al., *The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus,* NIST 1986.

[10] F. Soong, A. Rosenberg, L. Rabiner, and B. Juang, "A vector quantization approach to speaker recognition," *Proceedings of ICASSP 1985,* April 1985, pp. 387-390.

[11] CMU Sphinx Group, *CMU Sphinx,* http://cmusphinx.sourceforge.net/.

[12] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication,* vol. 17 no. 1-2, 1995, pp. 91-108.

[13] D. A. Reynolds, "Experimental evaluation of features for robust speaker identification," IEEE Transactions on Speech, Audio, and Language Processing, vol. 2, no. 4, October 1994, pp. 639-643.