COMBINING FEATURES AT VECTOR LEVELFOR HIGHER SPEED AND ACCURACY OF SPEAKER IDENTIFICATION

M. Afzal, T. Ahmad, M. F. Hayat, H.M. S. Asif and K. H. Asif

Department of Computer Science and Engineering, University of Engineering and Technology Lahore, Pakistan.
Corresponding author e-mail: shmafzal@yahoo.com

ABSTRACT: The study on combination of different feature vectors for the accuracy of Speaker Identification (SI) based on Vector Quantization (VQ) performed well when compared with other paradigms. Feature vectors based on Mel-frequency Cepstral Coefficients (MFCC) and Linear Predictive Codes (LPC) were combined. Texas Instrument and Massachusetts Institute of Technology (TIMIT) database containing 630 speakers were registered and tested for feature level combination studies. LPC feature vector manifested lower accuracy than MFCC when used as such. The combination of feature vectors through sum (MFCC+LPC) and difference (MFCC-LPC) were studied. For 42% of cases of codebook sizes studied (MFCC+LPC) gave higher accuracy than simple MFCC. The accuracy of (MFCC-LPC) combination results was better than simple MFCC for 93% of cases of codebook sizes studied. The accuracy enhancement could be used to reduce the time of speaker identification to half by using half sized codebooks of (MFCC-LPC)feature vectors with same or higher accuracy as compared to codebooks of simple MFCC feature vectors.

Keywords: Speaker Identification, Vector Quantization, Mel-frequency Cepstral Coefficient, Linear Predictive Codes, Speaker Identification Speedup.

(Received 26-11-13 Accepted 17-3-15)

INTRODUCTION

Speech based interaction with computer systems is the most natural because it imposes least constraints on the system user regarding proximity distance, orientation, posture and attention as compared to keyboard, mouse or touchpad. Speech recognition system adaptive to speaker, using a speaker identification front end, has higher accuracy as compared to speaker independent systems. For real-time speech recognition, fast speech adaptationis imperative through accurate identification of speaker's class.(Glaeser and Bimbot,1998) demonstrated a number of real-time applications of automatic speaker identification. Fast growing applications of automatic speaker identification emphasizes speeding up speaker identification systems with high accuracy.

Vector Quantization (VQ) is highly competitive technique for automatic speaker identification compared with Gaussian Mixture Model (GMM) (Kinnunen *et al*., 2006). Accuracy in speaker identification increases with increasing size of speaker model used to represent known speakers registered with the system. However, over fitting effects are frequently observed which decrease identification accuracy after certain size of speaker models (Kinnunen and Li, 2010). The studies of (Afzal and Haq, 2010) show lesser over fitting effect than that listed in (Kinnunen *et al*., 2006) .Also, increase in model size to achieve higher accuracy results in slowing down the SI systems (Kinnunen *et al*., 2006 and Afzal *et al*., 2012). It is due to the fact that pattern matching unit of an SI system consumes most of the time during evaluating codebook of each registered speaker for similarity measure with the test speech sample (Afzal *et al*., 2012). The best similarity measure having codebook's registered speaker is output as the test speaker. The order of codebook or model of registered speaker sorted with respect to similarity measure value from best to worst is termed as rank of registered speaker model.

Combining output of multiple classifiers at measurement level and rank level has been investigated. The sum rule and max rule of combining two classifiers at rank level have also proved to perform the best (Mashao and Skosan, 2006). Neural network classifiers were mixed at measurement level based on LPC and MFCC feature vectors to increase accuracy of speaker identification (Aida *et al*., 2006).

This study was planned to investigate mixing of feature set at feature level before input to a classier has not been explored so far. We used this mixing scheme to increase both speaker identification accuracy and speaker identification speed as compared to two classifiers combined at measurement level or rank level.

MATERIALS AND METHODS

We mixed MFCC vector and LPC vector of each speech signal frame by sum and difference to form MFCC+LPC and MFCC-LPC feature vectors. Simple LPC and MFCC feature vectors were also tested for

comparison purpose. This approach increased accuracy without doubling the speaker identification time.

Background and Related Work: Speaker identification systems have two phases of working namely training phase and testing phase. During training VQ based ASI systems take a sequence of feature vectors, $\tilde{X} = \{\tilde{x}_{\tilde{\imath}} | \, 1 \leq \tilde{\imath} \leq T, \tilde{x}_{\tilde{\imath}} \in \mathbb{R}^d\}$, extracted from speech samples of a speaker and map it to a set or codebook of M code vectors or centroids, $C = \{c_m | \, 1 \leq m \leq M, c_m \in \mathbb{R}^d\}$.

A sequence of feature vectors, defined as $X = \{x_i | \, 1 \leq i \leq T, x_i \in \mathbb{R}^d\}$, was also extracted from test speech samples of a person. Pattern matching unit of our ASI system computed quantization distortion of X with codebook *C* of each registered speaker according to Equation (1) as reported by (Quatieri, 2002). Then the test speaker was identified as the registered speaker whose codebook had minimum distortion.

$$D(X, C) = \sum_{i=1}^{T} \min_{c_m \in C | 1 \leq m \leq M} \left\| x_i - c_m \right\| \tag{1}$$

Where $\left\| x_i - c_m \right\|$ defines Euclidean distance between a test vector, x_i, with a code vector, c_m. Identification of a speaker $\boldsymbol{s}$ of X from N registered codebooks was done according to Equation (2).

$$Speaker \;\; id = s^* \arg\min_{1 \leq s \leq N} (D(X, C_s)) \tag{2}$$

Computation of Equation (2) took most of the time during speaker identification. Therefore, we thought of combining features at feature vector level to increase accuracy which would not increase the identification time as compared to employing multiple classifiers.

Proposed Technique: Feature vectors based on MFCC were mostly used for speaker identification for their higher accuracy(Kinnunen and Li, 2010). We proposed combining of LPC feature vectors with MFCC feature at feature vector level to increase discriminative power of the feature set. This scheme had advantage over combining classifiers at measurement or rank level that it required pattern matching to be performed once for each registered speakerin order to rank them for similarity with a given test sample. This way our scheme increased accuracy without doubling the identification time. We studied LPC and MFCC feature vector combination to enhance accuracy of speaker identification systems. We compared accuracies of speaker identification systems that used sum (MFCC+LPC), difference (MFCC-LPC) feature vectors and, LPC and MFCC feature vectors were used as such.

Data and Experiment: Speech database of TIMIT was used for experimental evaluation of the proposed technique (Garofolo *et al*., 1993). The TIMIT speech data that was originally recorded at 16 kHz sampling frequency was down sampled to 8 kHz using anti-aliasing filters. The TIMIT corpus consisted of speech samples of 630 speakers with 10 files of each speaker. Seven files of 'sa' and 'sx' typein the corpus were concatenated to make system training sample of approximately 23 second duration to build VQ codebooks.

The 'si' files in the corpus contained speech based on different text for each speaker. System testing samples of approximate duration of 8.5 seconds were prepared by assembling together the three 'si' files. This way setup of a text independent speaker identification experiment was carried out.

Working of Feature extraction unit was depicted through block diagram as presented in Figure-1. For extraction of feature vectors, the speech samples both for training and testing were converted into frames of 30 mS overlapping by 33%. Frames with energy less than 12% of average frame energy of each speaker's sample were discarded as silence frames. Hamming window was applied to non-silence frames. These windowed frames were input to algorithms to compute LPC and MFCC feature vectors of size 11 and 12 in respective experimental runs. For LPC Levinson-Dublin filter was used.We useda filterbank of 19 triangular filters for MFCC. Output of triangular filters was compressed by taking log. Then, after applying DCT algorithm, first and last values were discarded to select MFCC vectors of size 11 and 12 for experimentation. All the four types of feature sets namely LPC, MFCC, MFCC+LPC and MFCC-LPC from both the training samples and testing samples were computed once and were stored for repeated experimentations.

We used Linde-Buzo-Gray (LBG) clustering algorithm to construct codebooks of registered speakers for each type of combination of features studied in a study undertaken by (Bei and Gray, 1985). Codebooks of sizes 16, 32, 64, 128, 256, 512 and 1024 were prepared and stored to study the variation in accuracy of the system. Implementation of algorithms for speaker identification system was done in C# programming language. The programs were run on Core™2 Duo CPU E6550@ 2.33GHz based HP Compac DX7400 Microtower. Operating system used was Windows Vista Business.

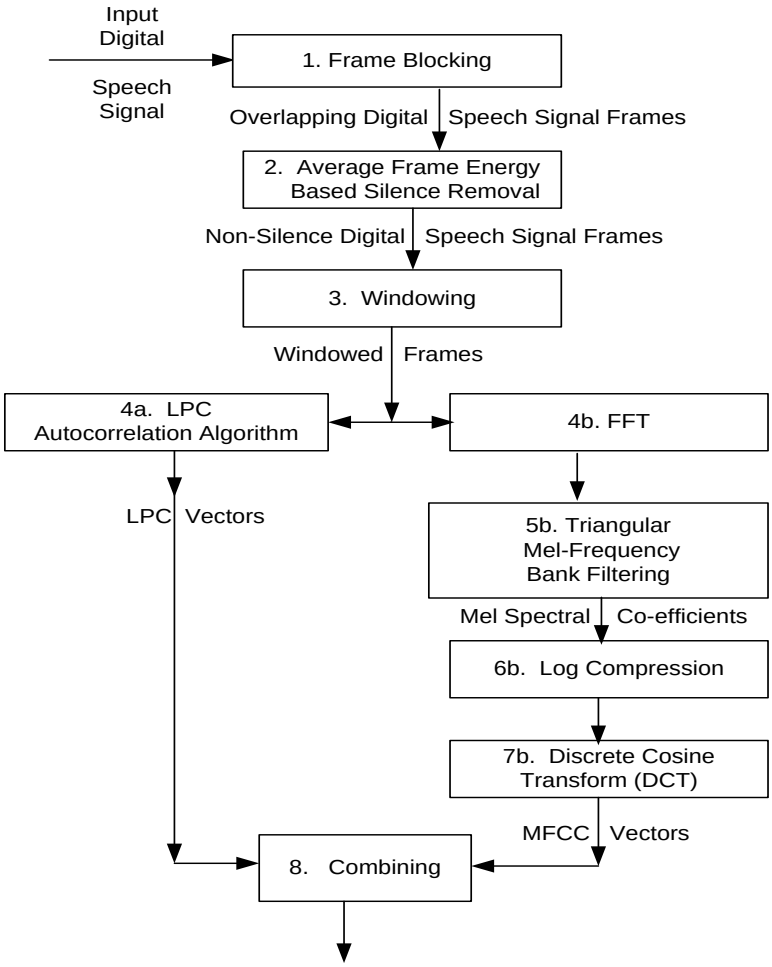

Figure -1: Schematic diagram of steps for extracinglpc and mfcc and combining the feature vectors

RESULTS AND DISCUSSION

We tested all the speakers in the TIMIT database for identification by our developed systems. The training speech data for the systems was small that resulted in over fitting effect for larger codebooks of sizes like 11x1024, 12x512 and 12x1024as was reported by (Kinnunen *et al*., 2006). A set of mapping processes for training codebooks was studied by (Kinnunen *et al*., 2000). However, we used LBG for the purpose of codebook tuning as a final map of $\tilde{T}$ training vectors to M codewords. Our scheme of combining MFCC and LPC features at vector level made system training twice as faster than method of combining two classifiers as performed by (Aida *et al*., 2006; Mashao and Skosan, 2006). We listed results of our test runs for different combinations of feature sets and different sized VQ codebooks for speaker identification accuracy and average identification time per speaker (Table-1). Time consumed by speaker identification process was computed by executing property function of. NET framework called 'System. Date Time. Now' at start and end of the process. We computed total time taken by our ASI system for identification of all 630 TIMIT speakers to calculate the average speaker identification time. Accuracy of ASI of system that used just LPC feature vectors only was lower than 25%. Hence, we did not listed LPC case in details.

However, in general, accuracy of the systems increased with increase in codebook size because larger codebooks tended to capture more detailed knowledge of the speakers' voice into codebooks. In this study, larger sized feature vectors gave higher accuracy as reported by (Afzal and Haq, 2010; and Afzal *et al*., 2012). Mainly, we used MFCC feature vectors because they were considered to be the best in speech processing especially for speaker identification (Kinnunen *et al*., 2006 and Kinnunenand Li, 2010). However, our empirical results showed that discriminative power of MFCC feature vector could be improved by subtracting or adding to it equal sized LPC feature vector extracted from the same speech signal frame (Table-1).

The advantage in accuracy of using the feature combination (MFCC − LPC) instead of MFCC was observed for 93% of times as highlighted with gray shading. The accuracy values of (MFCC+LPC) were also found to be better than MFCC for 42% of cases of the codebook sizes studied.

Actually, computation of minimum distortion, according to Equation (2), was the speed controlling step of ASI systems (Afzal *et al*., 2012; Afzal and Haq, 2010; Kinnunen *et al*., 2006; Kinnunen and Li, 2010) with execution time order of $O(TNMd)$.

Table-1: Showing accuracy and speed of asi systems

Codebook Size *d* x *M*	Correctly identified speakers out of *N*=630with different feature combination			Average. Identification Time (S)
	MFCC	MFCC+ LPC	MFCC-LPC	
11x32	528	530	545	1.15
11x64	603	604	607	2.25
11x128	624	624	624	4.44
11x256	627	625	628	8.81
11x512	628	627	629	17.55
11x1024	624	625	626	36.35
12x32	537	534	554	1.26
12x64	606	608	613	2.48
12x128	622	625	624	4.88
12x256	626	625	628	9.66
12x512	624	623	629	19.25
12x1024	625	622	626	39.95

Techniques of combining multiple classifiers to increase speaker identification accuracy had been studied by many researchers, for example (Mashao and Skosan, 2006 and Aida *et al*., 2006) where they trained the multiple classifiers for different type of feature vectors extracted from training speech samples. Test sample was also converted to respective type of feature vector sequence and then put into the respective classifier. Each classifier then computed similarity measures for each registered speaker model with the given test sample. Consequently the overall identification time was doubled when outputs of two classifiers were combined to achieve some increased overall accuracy as compared to single classifier.

The average speaker identification time would have been doubled as compared to our current study if we had used the methods of combining classifier at rank level as done by (Mashao and Skosan, 2006). Double time would have been resulted from twice computation of total minimum distortion of *X* sequence of *T* feature vectors of *d* elements each for codebooks of *M* code vectors stored for *N* registered speakers. To pick the best match sorting of the total distortion values for *N* codebook would have been extra computation load as well.

The gray shaded and bold faced values in the columns of (MFCC+LPC) and (MFCC-LPC) showed advantage over simple MFCC in Table-1. For the cases of codebooks of sizes 12x256 and 12x512, (MFCC-LPC) feature combination gave better accuracy than larger codebook of 12x512 and 12x1024 that were based on MFCC only. Such cases clearly indicated that feature vector combination at feature vector level could also be used for speeding up ASI systems as well.

Conclusion: Disadvantage of combination of classifiers at measurement level and at rank level for speaker identification speed was highlighted in this study. Instead, combinations of different types of feature vectors extracted from same speech frame and of equal size were proposed to increase accuracy and speed of speaker identification systems. MFCC based feature vectors are considered to be the best performing feature vectors in speech processing, especially in speaker identification. An empirical study of combination of MFCC and LPC at feature vector level was presented with a focus to increase accuracy of SI based on VQ. All 630 speakers of TIMIT speech database were registered and tested for different feature level combination studies. The combination of feature vectors through sum (MFCC+LPC) and difference (MFCC-LPC) were studied in order to achieve increased accuracy. For 42% of the cases of codebook sizes studied in this paper (MFCC+LPC) gave higher accuracy than simple MFCC. The accuracy results of (MFCC-LPC) combination were better than simple MFCC for 93% of cases of codebooks studied. Another strong point of accuracy enhancement is that it can be indirectly used to double the speed of speaker identification system by using half sized codebooks of (MFCC-LPC) with same or higher

accuracy instead of codebooks obtained from simple MFCC vectors.

Acknowledgment: This research work has been completed with the support of the University of Engineering and Technology, Lahore, Pakistan. The TIMIT data was provided by Linguistic Data Consortium, University of Pennsylvania, USA. Their support is gratefully acknowledged.

REFERENCES

Afzal, M. and S. Haq, Accelerating Vector quantization based speaker identification, J. American Sci., 6(11):1046-1050 (2010).

Afzal, M., M.Maud and A. Akbar, Toggling and circular partial distortion elimination algorithms to speedup speaker identification based on vector quantization, Pakistan J. Engin. and Appl. Sci., (11):1-13 (2012).

Aida, K., C. Ardil, and S. Rustamov, Investigation of combined use of MFCC and LPC features in speech recognition systems, World Academy of Science, Engineering and Technology, Azerbaijan, 2(3):74-80 (2006).

Bei, H. andR. Gray, An improvement of the minimum distortion encoding algorithm for vector quantization, IEEE Transactions on Communication, 33(10):1132-1133 (1985).

Garofolo, J., L. Lamel, W. Fisher, J. Fiscus, D. Pallett, D. Dahlgren and V. Zue, DARPA TIMIT Acoustic phonetic continuous speech corpus, (NISTIR Publication No. 4930). Washington DC: US Department of Commerce, (1993).

Glaeser, A. and F. Bimbot, Steps towards the integration of speaker recognition in real world telecom applications, Proceeding of International Conference on Spoken Language Processing, (ICSLP) Sydney, NSW, Australia (1998).

Kinnunen, T., E. Karpove and P. Franti, Real-time speaker identification and verification, IEEE Transactions on Audio and Language Processing, 14 (1):277-288(2006).

Kinnunen, T., T. Kilpelainen, and P. Franti, Comparison of clustering algorithms in speaker identification. Proceeding of International Conference (IASTED) Signal Processing and Communications (SPC 2000) Marbella, Spain,:222–227 (2000).

Kinnunen, T. and H. Li, An overview of text-independent speaker recognition: from features to supervectors, Speech Communication, Elsevier, 52(1):12-40 (2010).

Mashao, D.and M. Skosan, Combining classifier decisions for robust speaker identification. Pattern Recognition, Elsevier, 1 (39):147-155 (2006).

Quatieri, T., Discrete-time speech signal processing principles and practice, Prentice Hall PTR, New Jersey, USA, (2002).