academicJournals

Vol. 8(39), pp. 1911-1918, 18 October, 2013
DOI 10.5897/SRE12.741
ISSN 1992-2248 © 2013 Academic Journals
http://www.academicjournals.org/SRE

Scientific Research and Essays

Full Length Research Paper

Voice onset/offset based local features (VOOLF) for Arabic Speaker recognition

Awais Mahmood* and Mansour Alsulaiman

Speech Processing Laboratory, Department of Computer Engineering, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia.

Received 14 November, 2012; Accepted 16 July, 2013

Local features for any pattern recognition system are based on the information extracted locally. In this paper, a local feature extraction technique is developed, which captures the formant transition and voice onset/off set of a speaker. We named this technique as voice onset/offset local features (VOOLF). These features are extracted in the time spectrum domain by taking the moving average on the diagonal directions. These proposed features are compared with MFCC for speaker recognition system. The results showed that proposed technique perform better than the commonly used MFCC. The proposed method is able to capture the formant transitions and onset/offset of the speaker; hence this resulted in recognition rate higher than the other speech features.

Key words: Voice onset/offset features, local features, Speaker recognition system, Gaussian Mixture Model (GMM).

INTRODUCTION

Speaker Recognition (SR) refers to the process of automatically recognizing a person based on speech information included in the speech signal. The interest in SR has recently increased due to the growing use of speech recognition technologies in various areas. The research efforts in SR are largely focused on developing practical applications, which can be divided into two large classes. The first class of research is focused on controlling the access rights to different systems (information and material systems), and the second focus is in the area of forensics.

Depending on task objective, SR can be divided into two types, which are Speaker Identification (SI) and Speaker Verification (SV) (Sumithra et al., 2011; Alsulaiman et al., 2010). In SI, the unknown voice is assumed to be from a predefined set of known speakers, and it is a one-to-many comparison (Reynolds et al., 2000; Wildermoth et al., 2000). SV is the process of determining whether the speaker identity matches the person he/she claims to be (Reynolds et al., 2000). Speaker verification requires distinguishing a speaker's voice from a potentially large group of voices unknown to the system (Altınçay et al., 2002).

The two major components of any SR are the front-end processing and speaker modeling. Front-end processing converts the input signal into small frames, and then frames are converted into suitable feature space. These features are later fed to the modeling part (Mahmood et al., 2012). Feature Extraction is a very crucial component in SR. The purpose of feature extraction is to extract the speaker specific information in the form of feature vector at reduced data rate. A decent feature set should contain all the components that characterize the speaker (Jayanna et al., 2009). Modeling human voice production and modeling the peripheral auditory hearing are the two categories of the speech features. In the first category, the most popular feature is linear prediction cepstral coefficients (LPCC), while in second category the most popular features are Mel frequency cepstral coefficient (MFCC) and RASTA-PLP. Later many types of features

*Corresponding author. E-mail: awais@ksu.edu.sa; msuliman@ksu.edu.sa.

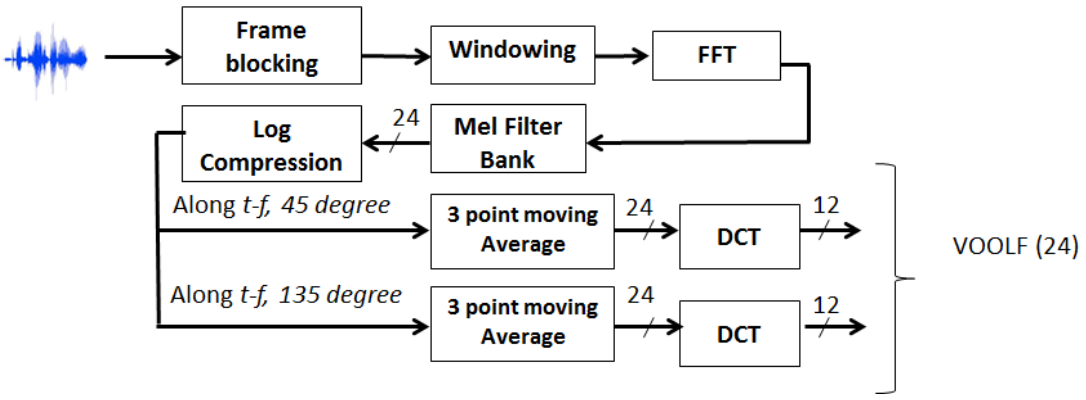

Figure 1. Block diagram of the proposed VOOLF feature extraction method.

were extracted, which mostly falls in these two categories and are described in (Lawson et al., 2011).

Most of the feature extraction techniques were developed for speech recognition system. Nitta (1998) presented a work using local features. The paper describes an attempt to incorporate the functions of the auditory nerve system into the feature extractor for speech recognition. Later, Nitta expanded his earlier work on local features and extracted the features based on orthogonal acoustic-features planes and linear discriminative analysis (LDA) (Nitta, 1999). This method showed significant improvement in comparison to the method described in the earlier paper by the author (Nitta, 1998). The difference from previous research is only the usage of Sobel filter.

Fukuda et al. (2003) described dynamic features for speech recognition, which represent a variation along the time axis and along frequency axis on time spectrum (TS) and/or time cepstrum (TC) pattern (Fukuda, 2004, 2003) and named these features as local features (LF). Later, these variations are converted into pheriphiral features. In Fukuda (2003) the authors went one step ahead of MFCC by taking 8^{th} derivatives along the frequency axis (Fukuda, 2004, 2003) and named it as peripheral features. This approach is different from the work performed earlier by the authors (Nitta, 1998, 1999) in that the later experiments used two acoustic pieces of evidence, which are sharply rising and falling sound and spectral peaks in steady sound. Moreover, the authors did not use these features directly. Instead, they transformed these features into distinctive phonetic features (DPF) using the neural network then they used DPF for speech recognition. Hassan et al. (2011) used the LFs as described in (Fukuda, 2004, 2003) and extracted DPF then used it for phone segmentation.

From above discussion, it can be noted that the researchers, who used LF, only applied these features to speech recognition, and they did not use these LFs directly, rather they transform these features into DPF and peripheral features and then used these features – the speech recognition systems and speech segmentation.

The most commonly used features on speaker recognition were imported from speech recognition systems (e.g., MFCC, LPCC, etc.). Most of the features extraction techniques concentrate on the phoneme characteristics such as spectral energy distribution in different subbands, voiced, unvoiced segmentation, parameterizing the vocal tract (Jayanna et al., 2009). However, most of the papers in SR do not concentrate on the way different speakers pronounce, for example, voice onset/offset etc.

LF has never been used in speaker recognition; hence in this paper we propose local features VOOLF which can capture auditory voice onset/offset local features. VOOLF can also capture the formant transition. These features are extracted in the time spectrum domain by taking the moving average on the diagonal direction.

There are different techniques to model the speaker. The most common technique is Gaussian Mixture Model (GMM) (Reynolds, 1995). An extension of GMM is Universal background model GMM (UBM-GMM) (Anusuya and Katti, 2011). The reason for using UBM-GMM is the amount of training data. GMM-UBM will give better result when the training data is small, if the training data is enough the both GMM and UBM-GMM perform equally well (Anusuya and Katti, 2011). In our case we have enough data for the training so we used GMM. Other popular modeling techniques used by researchers are Support vector machine (SVM) and hidden markov model (HMM) (Anusuya and Katti, 2011).

EXPERIMENTATION

Proposed VOOLF feature

The proposed VOOLF extract two LFs by taking three points moving average (MA) on time-frequency axes, which correspond to capturing the acoustic evidence of formant transitions and onset/offset. Figure 1 shows the block diagram of the proposed

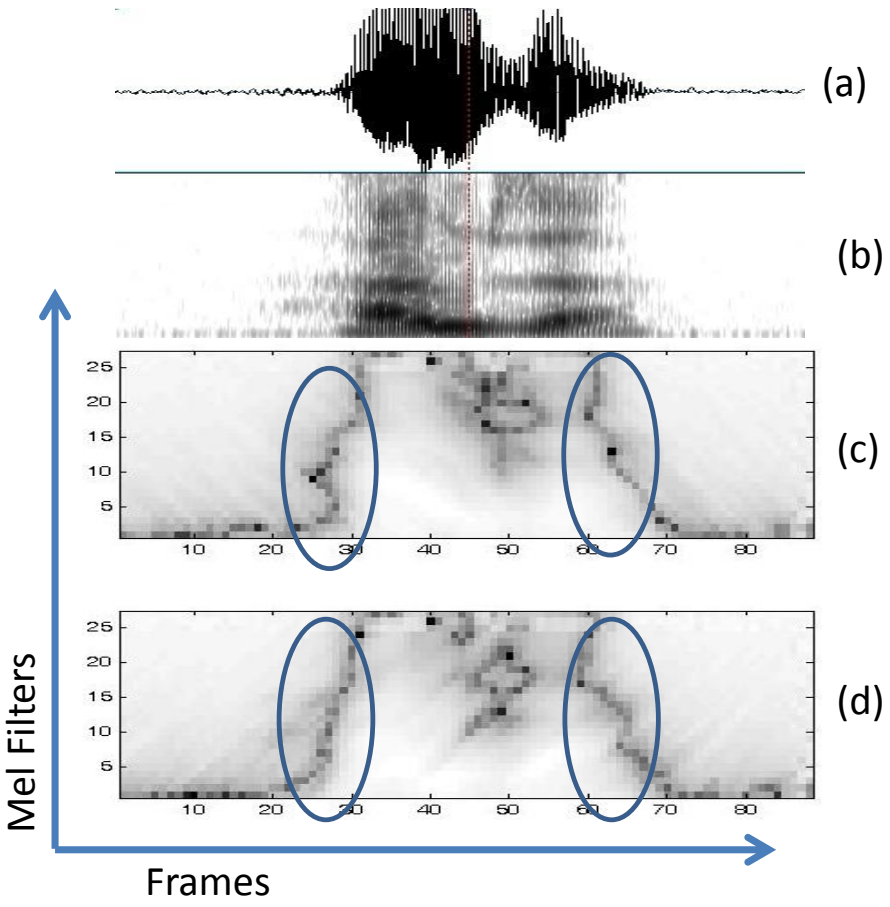

Figure 2. Visualization of the information captured by the proposed VOOLF method. (a) Wave of Arabic alphabet [Hamza], (b) spectrogram, (c) LR on the t-f axis at 45°, (d) LR on the t-f axis at 135.

VOOLF technique. The proposed technique extracts Local Features (LF) on two different time-frequency directions.

First, Fourier Transformation (FT) is calculated for the windowed speech frame. The windowed speech frame is evaluated as

$$s_w(n) = s(n) \times w(n), \qquad \text{for } n = 0,1,2,\ldots, N-1 \tag{1}$$

s(n) and w(n) correspond to the input voice signal and window function, respectively. N corresponds to the number of samples in each frame. Fourier transformation is applied to the windowed signal as in Equation (2).

$$X(k) = \sum_{j=0}^{N-1} s_w(j) \times e^{-j\frac{2\pi jk}{N}}, \quad k = 0,1,2,\ldots, K-1 \tag{2}$$

K is the number of FT points (bins of frequency). After passing the magnitude of FT through the 29-channel mel-filter bank, the log compression is applied. Then two 3-points MA are evaluated as given in Equations 3 and 4.

Along the time-frequency at 45°:

$$c_{t,f}^{45} = \frac{\sum_{i=1}^{3} \left(c_{t-i,f-i} + c_{t+i,f+i} \right)}{2*3} \tag{3}$$

Along the time-frequency at 135°:

$$c_{t,f}^{135} = \frac{\sum_{i=1}^{3} \left(c_{t+i,f-i} + c_{t-i,f+i} \right)}{2*3} \tag{4}$$

After calculating the MA, the dimension of the feature is high. In order to compress and decorrelate these features, discrete cosine transformation (DCT) is applied as shown in Equation 5.

$$c_{t,m} = \sum_{k=1}^{24} d_{t,f}(k) \cos\left[(k-0.5) \frac{m\pi}{24} \right], \quad m = 1, 2,\ldots,12 \tag{5}$$

The capability of VOOLF to capture voice offset/onset of the speaker is presented in Figure 2. Figure 2(a) shows the speech

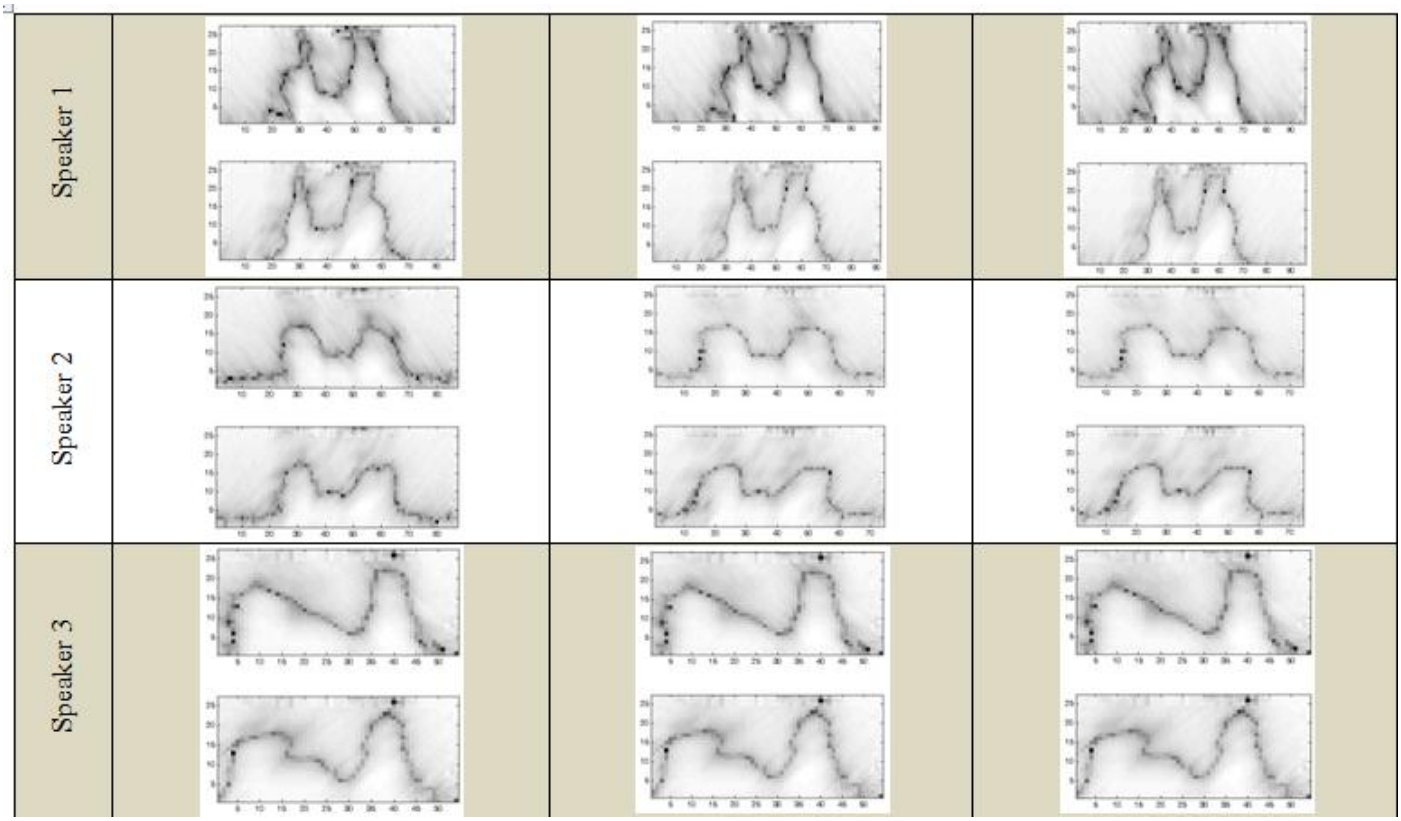

Figure 3. Visualization of the information of three different speaker pronouncing alphabet “hamza” captured by the proposed VOOLF method.

signal, 2(b) presents the spectrogram of the speech signal, 2(c) represent the MA along time-frequency axis (45 degree) and 2(d) represent the MA along time-frequency axis (135 degree). The ability of VOOLF to capture the voice onset and offset of is shown in Figures 2(c), (d). The left circles in Figures 2(c),(d) shows the voice onset whereas the right circles shows the voice offset.

Figure 3 present the VOOLF of the Arabic alphabet “hamza” pronounced by 3 different speakers. Each speaker pronounced the alphabet three times, The VOOLF for each pronunciation is presented in a column. The upper part of each block give the MA taken along 45 degree and lower part give the MA taken along 135 degree on time-frequency axis. Its can be noted that the VOOLF of each speaker are the same in the three columns but is different from one speaker to another, which may explain the good performance of VOOLF.

Database

The problem of having few Arabic speech corpora is described in detail in (Alsulaiman, 2009). In this work we used Arabic alphabet database recorded in King Saud University. The database contains all 29 Arabic alphabets pronounced by 44 speakers. All speakers have recorded 10 utterances of each alphabet. The utterances were used in two different cases. In the first case, seven utterances of each alphabet were used to train the system, and the remaining three were used to evaluate the performance of the system. For second case, five utterances of each alphabet were used in the training and similarly five utterances were used for the testing in order to evaluate the performance of our proposed technique.

RESULTS AND DISCUSSION

Three types of features are evaluated in experiments named as MFCC, LF, and the proposed VOOLF. The results of MFCC will be discussed first. To extract the features, the speech is pre-emphasized then a sliding Hamming window with a length of 25 m and a shift of 10 m was positioned on the signal. The FT was applied on these frames and the magnitude of FT is fed to the Mel-Filters. The number of features for MFCC is 36, which consist of 12 MFCC features, 12 delta and 12 delta features. The speakers were modeled using Gaussian Mixture Models (GMM). The numbers of mixtures per model were 4, 8, 16 and 32. During the experiments, each of the 29 Arabic alphabets was considered separately. For each alphabet, four different numbers of mixtures were used. All the speakers were considered in each experiment.

Table 1 presents the results obtained using MFCC when seven samples were used in the training, and three samples were used for testing. 32 GMMs perform better as compared to 4, 8 and 16 GMMs. For the alphabets ء, م ,ش ,س ,ز ,ثand ه the speaker recognition rate is 100% using all the different numbers of the GMMs. This observation leads us to deduct that it is good to have

Table 1. Recognition rate (%) using MFCC when seven samples are used for training and three samples are used for testing.

Alphabets	4GMM	8GMM	16GMM	32GMM	Alphabets	4GMM	8GMM	16GMM	32GMM
أ	97.73	100	100	99.24	ض	98.48	99.24	100	100
ء	100	100	100	100	ط	99.24	100	100	100
ب	95.45	100	99.24	100	ظ	98.48	100	100	100
ت	96.97	100	100	98.48	ع	98.48	98.48	98.48	98.48
ث	100	100	100	100	غ	98.48	100	99.24	99.24
ج	98.48	98.48	99.24	98.48	ف	99.24	99.24	100	100
ح	99.24	97.73	99.24	99.24	ق	99.24	99.24	100	100
خ	99.24	100	100	100	ك	99.24	100	99.24	99.24
د	98.48	98.48	99.24	99.24	ل	98.48	100	100	100
ذ	98.48	99.24	99.24	99.24	م	100	100	100	100
ر	97.73	96.97	98.48	99.24	ن	97.73	99.24	98.48	98.48
ز	100	100	100	100	ه	100	100	100	100
س	100	100	100	100	و	98.48	98.48	99.24	99.24
ش	100	100	100	100	ى	97.73	99.24	96.97	97.73
ص	99.24	100	100	100					

Table 2. Speaker recognition rate(%) using LF when 7 samples are used for training and 3 samples are used for testing.

Alphabets	4GMM	8GMM	16GMM	32GMM	Alphabets	4GMM	8GMM	16GMM	32GMM
أ	96.21	96.97	97.73	96.97	ض	98.48	97.73	97.73	97.73
ء	97.73	97.73	97.73	97.73	ط	98.48	97.73	97.73	97.73
ب	98.48	97.73	97.73	97.73	ظ	98.48	97.73	97.73	97.73
ت	98.48	96.97	96.21	96.21	ع	99.24	95.45	96.97	96.97
ث	99.24	97.73	97.73	97.73	غ	99.24	100	96.97	97.73
ج	98.48	96.21	96.97	96.21	ف	98.48	100	99.24	97.73
ح	99.24	99.24	99.24	96.97	ق	96.97	97.73	97.73	97.73
خ	100	99.24	99.24	97.73	ك	100	97.73	97.73	96.97
د	100	100	96.97	96.97	ل	100	97.73	97.73	97.73
ذ	97.73	96.21	96.97	96.97	م	100	100	100	97.73
ر	98.48	96.97	96.97	96.21	ن	96.97	99.24	96.97	96.97
ز	100	100	100	97.73	ه	100	100	98.48	97.73
س	99.24	100	100	97.73	و	98.48	96.21	96.97	96.21
ش	100	99.24	97.73	97.73	ى	98.48	99.24	95.45	95.45
ص	99.24	100	100	97.73					

these alphabets in the database to achieve an excellent speaker recognition rate.

Twenty four features were extracted for LF, with 12 for MA along time axis and 12 for MA along the frequency axis. The result when using LF is presented in Table 2. From the Table 2 we can see that as the number of the GMMs increases, the recognition rates decrease. This may be explained that as the GMMs increase, the rising and falling sound is divided into different Gaussians (destroying the pattern preserved by LF) which results in low recognition rate. With a lower number of Gaussians, the rise and fall of sound are maintained, which results in a better recognition rate. LF performed slightly better than MFCC only when using four mixtures for other mixtures it had lower result than MFCC. For the LF, there is no alphabet, which gives a 100% recognition rate when using all numbers of the GMMs. For different values of the GMMs, except 32, few alphabets (ز, م) produced a 100% recognition rate for LF as presented in Table 2. For the proposed VOOLF, 24 features were extracted as described in Section 2. The proposed VOOLF performs better and/or equal to the MFCC and LF. VOOLF achieves 100% accuracy in the case of following alphabets: 'ء ز ظ ل م ه' using all mixtures as shown in

Table 3. Speaker Recognition Rates (%) Using VOOLF.

Alphabets	4GMM	8GMM	16GMM	32GMM	Alphabets	4GMM	8GMM	16GMM	32GMM
أ	98.48	99.24	100.00	99.24	ض	98.48	100.00	100.00	100.00
ء	100.00	100.00	100.00	100.00	ط	98.48	100.00	100.00	100.00
ب	98.48	100.00	100.00	100.00	ظ	100.00	100.00	100.00	100.00
ت	97.73	99.24	98.48	97.73	ع	99.24	97.73	99.24	99.24
ث	99.24	100.00	100.00	100.00	غ	99.24	100.00	99.24	100.00
ج	98.48	98.48	99.24	98.48	ف	98.48	100.00	100.00	100.00
ح	99.24	99.24	99.24	99.24	ق	99.24	100.00	100.00	100.00
خ	100.00	99.24	99.24	100.00	ك	100.00	100.00	100.00	99.24
د	100.00	100.00	99.24	98.48	ل	100.00	100.00	100.00	100.00
ذ	98.48	98.48	99.24	99.24	م	100.00	100.00	100.00	100.00
ر	98.48	96.97	99.24	99.24	ن	99.24	99.24	99.24	99.24
ز	100.00	100.00	100.00	100.00	ه	100.00	100.00	100.00	100.00
س	99.24	100.00	100.00	100.00	و	98.48	98.48	99.24	100.00
ش	100.00	99.24	100.00	100.00	ى	98.48	99.24	97.73	97.73
ص	99.24	100.00	100.00	100.00					

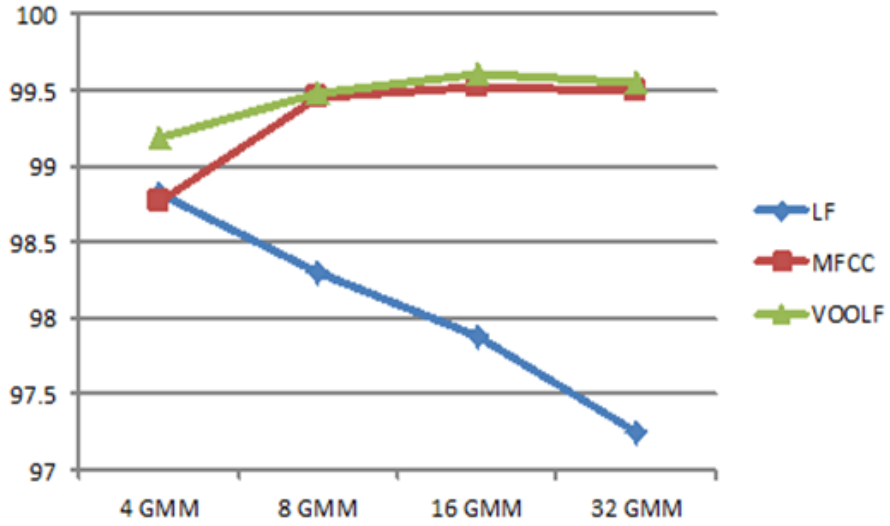

Figure 4. Average accuracy (%) of the systems using three different feature extraction methods with four different Gaussian mixtures, when 7 training and 3 test samples were used.

Table 3. This confirms our observation that there exist some alphabets which clearly describe the speaker identity depending on the speech features used. This finding suggests that if an Arabic speaker recognition database is required, including these alphabets in the script can increase the recognition rate tremendously.

Figure 4 shows the average speaker recognition accuracy (as a percentage) using the three different feature extraction methods with different number of mixtures. From the figure we can see that VOOLF outperformed all the other features. This can be attributed to the ability of VOOLF to capture speaker's speech onset/offset.

All the above experiments were repeated with 5 training samples and 5 test samples. Overall performance decreased yet VOOLF performed better as compared to the MFCC and LF. The best result obtained was when using 8 GMM mixtures, so instead of presenting all the results, we only present the result with 8 GMM mixtures in Table 4.

Figure 5 shows the average accuracies when five samples were used in the training, and five utterances were used for the testing. The best result is obtained when 8 GMM were used. The VOOLF had higher accuracy than the other features for all the number of GMMs. This result emphasizes the result we got with 7

Table 4. Speaker Recognition Rates (%) Using all the feature extraction technique with 8 GMM.

Alphabets	LF	MFCC	VOOLF	Alphabets	LF	MFCC	VOOLF
أ	96.36	95.45	96.36	ض	96.36	95.91	96.36
ء	96.82	97.73	96.82	ط	95.00	94.55	95.00
ب	96.36	95.00	96.36	ظ	96.82	95.91	96.82
ت	94.55	95.45	94.55	ع	98.18	97.27	98.18
ث	99.09	98.18	99.09	غ	97.73	98.18	97.73
ج	95.91	96.82	95.91	ف	97.73	97.27	97.73
ح	97.73	96.82	97.73	ق	95.45	96.82	95.45
خ	99.09	99.09	99.09	ك	96.36	95.45	96.36
د	98.64	98.64	98.64	ل	98.64	99.09	98.64
ذ	98.64	98.64	98.64	م	97.42	97.33	100.00
ر	97.27	97.73	97.27	ن	99.09	98.18	99.09
ز	100.00	100.00	100.00	ه	100.00	99.09	100.00
س	98.64	98.18	98.64	و	95.91	97.73	95.91
ش	98.64	98.64	98.64	ى	96.82	96.36	96.82
ص	98.18	99.09	98.18				

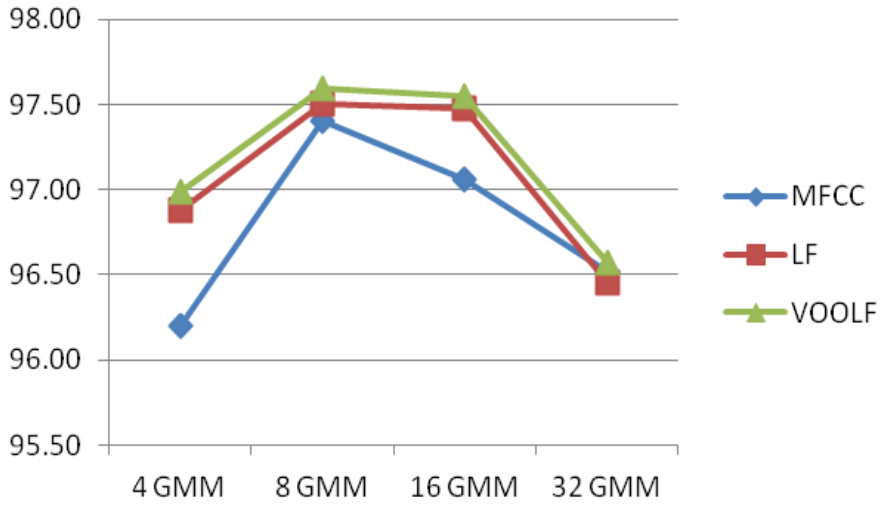

Figure 5. Average accuracy (%) of the systems using three different feature extraction methods with four different Gaussian mixtures when 5 training and 5 testing samples were used.

training samples and 3 testing samples.

Conclusion

A new speech feature VOOLF was proposed for speaker recognition. VOOLF showed higher recognition rate, compared to MFCC and LF, in the experiments with speaker recognition based on Arabic alphabets because of its voice onset/offset capturing capabilities. We showed that VOOLF outperformed MFCC and LF for the different number of GMMs. We also showed that the recognition rate using some specific alphabets can reach up to 100%. So having these alphabets in the text used for recognition will produce higher recognition rate. The high performance of VOOLF can be attributed to its ability to capture the formant transitions and onset/offset of the speaker.

REFERENCES

Alsulaiman M, Alotaibi Y, Mahmood A, Bencherif MA (2009). Survey of Arabic Speaker Recognition. Research report, College of Computer and Information Sciences, King Saud University, Saudi Arabia, 2009.

Alsulaiman M, Alotaibi Y, Ghulam M, Bencherif MA, Mahmoud A (2010). Arabic speaker recognition: Babylon levantine subset case study. J. Comput. Sci. 6:381-385. http://dx.doi.org/10.3844/jcssp.2010.381.385

Altınçay H, Demirekler M (2002). Why Does Output Normalization Create Problems in Multiple Classifier Systems?. Proceedings of CPR2002, 16th International Conference on Pattern Recognition, August 2002, Quebec, Canada.

Anusuya MA, Katti SK (2011). Front end analysis of speech recognition: a review, Int. J. Speech Technol. 14:99-145. http://dx.doi.org/10.1007/s10772-010-9088-7

Fukuda T, Nitta T (2003). A Study on Japanese Distinctive Phonetic Feature Set for Robust Speech Recognition. The 2003 Autumn Meeting of The Acoust. Soc. Japan, September 2003, in Japanese. 1(1-6-5):9-10.

Takashi F, Tsuneo N (2004). Orthogonalized Distinctive Phonetic Feature Extraction for Noise-robust Automatic Speech Recognition. IEICE Trans. Info. Syst. E87-D(5):1110-1118.

Hassan F, Mohammed RAK, Md. Mostafizur R, Mohammad N, Md. Abdul L, Mohammad NH (2011). Local Feature or Mel Frequency Cepstral Coefficients - Which One is Better for MLN-Based Bangla Speech Recognition?, Springer-Verlag Berlin, pp. 154-161.

Jayanna HS, Mahadeva PSR (2009). Analysis, Feature Extraction, Modeling and Testing Techniques for Speaker Recognition. IETE Techn. Rev. 26(3):181-190. http://dx.doi.org/10.4103/0256-4602.50702

Lawson AD, Pavel V, Mark CH, Paul AA, Brandon B, Allen RS (2011). Survey And Evaluation Of Acoustic Features For Speaker Recognition, ICASSP 2011, Prague, Czech Republic, pp. 5444-5447.

Mahmood A, Alsulaiman M, Muhammad G (2012). Multidirectional Local Features for Speaker Recognition. ISMS 2012, February 2012, Kota Kinabalu, Malaysia.

Nitta T (1998). A novel feature-extraction for speech recognition based on multiple acoustic-feature planes. Proc. IEEE ICASSP'98 1:29-32.

Nitta T (1999). Feature extraction for speech recognition based on orthogonal acoustic-feature planes and LDA. Proc. IEEE ICASSP'99 1:421-424.

Reynolds D (1995). Large population speaker identification using clean and telephone speech, Mar. IEEE Sig. Process. Lett. 2:46-48. http://dx.doi.org/10.1109/97.372913

Reynolds DA, Quatieri TF, Dunn R (2000). Speaker verification using adapted Gaussian mixture models, Digital Sig. Process. 10(1-3):19-41. http://dx.doi.org/10.1006/dspr.1999.0361

Sumithra MG, Thanuskodi K, Archana AHJ (2011). A New Speaker Recognition System with Combined Feature Extraction Techniques. J. Comput. Sci. 7(4):459-465. http://dx.doi.org/10.3844/jcssp.2011.459.465

Wildermoth B, Paliwal KK (2000). Use of Voicing and Pitch Information for Speaker Recognition. Proceedings of Australian International Conference on Speech Science and Technology (SST-2000), Canberra, Australia, pp. 324-328.