Journal of Computer Science 6 (4): 381-385, 2010
ISSN 1549-3639

Arabic Speaker Recognition: Babylon Levantine Subset Case Study

Mansour Alsulaiman, Youssef Alotaibi, Muhammad Ghulam,
Mohamed A. Bencherif and Awais Mahmoud
Department of Computer Engineering, College of Computer and Information Sciences,
King Saud University, Riyadh, Kingdom of Saudi Arabia

Abstract: Problem statement: Researchers on Arabic speaker recognition have used local data bases unavailable to the public. In this study we would like to investigate Arabic speaker recognition using a publically available database, namely Babylon Levantine available from the Linguistic Data Consortium (LDC). **Approach:** Among the different methods for speaker recognition we focus on Hidden Markov Models (HMM). We studied the effect of both the parameters of the HMM models and the size of the speech features on the recognition rate. **Results:** To accomplish this study, we divided the database into small and medium size datasets. For each subset, we found the effect of the system parameters on the recognition rate. The parameters we varied the number of HMM states, the number of Gaussian mixtures per state, and the number of speech features coefficients. From the results, we found that in general, the recognition rate increases with the increase in the number of mixtures, till it reaches a saturation level which depends on the data size and the number of HMM states. **Conclusion/Recommendations:** The effect of the number of state depends on the data size. For small data, low number of states has higher recognition rate. For larger data, the number of states has very small effect at low number of mixtures and negligible effect at high number of mixtures.

Key words: HMM, GMM, MFCC, Arabic speaker, Babylon, Levantine

INTRODUCTION

The literature on Arabic speaker recognition systems has a good number of researches, though very low compared to English language. Among those researches, unfortunately very few worked on some well known datasets. All others used local recordings, (unavailable for extended or further research by other groups), containing some digits, or some primitive words, just enough to say it is a local dataset. This makes it hard to compare the different systems and their results.

In this study we focus on the Babylon dataset (BBL), which is available from the Linguistic Data Consortium (LDC). We present the results of speaker recognition on two subsets of the dataset using Hidden Markov Models/Gaussian Mixtures Models (HMM/GMM).

The subsets consist of both males and females. We will present the results for the subsets independently of the gender of the speaker then each gender is presented alone.

The system, that we will investigate, is text-independent. The text-dependent systems are easier to deal with than text-independent systems and have higher recognition rates. Nonetheless, we will use text independent because, it fits our forensic application goal and more suitable for the database.

The current trend in the literature is to use GMM for text independent speaker recognition (Bimbot *et al*., 2004). In this study, we are going to use HMM/GMM because, though we are reporting on two subsets of the dataset, but at the end we are going to use the whole dataset, hence HMM/GMM might be able to use the information embedded in the long samples of speech. Current work using HMM shows excellent results (Deshpande and Hoalmbe, 2008; Ilyas *et al*., 2007).

In our investigation, we looked at the effect of varying the system parameters on the speaker recognition rates. The different variations were as follows: We used 1-4 states per HMM model, using 1, 2, 4, 8, 12, 16 and 24 Gaussian mixtures per state. We used 12 and 36 MFCC (12MFCC+ 12ΔMFCC and 12ΔΔMFCC).

MATERIALS AND METHODS

Babylon Levantine LDC dataset: Among the public datasets available from LDC, we have chosen the BBL,

Corresponding Author: Mansour Alsulaiman, Department of Computer Engineering, College of Computer and Information Sciences, King Saud University, Riyadh, Kingdom of Saudi Arabia

because it is formed of spontaneous speech. The use of spontaneous speech is more suitable in many forensic applications. Spontaneous speech is harder to work with than read speech and has its own characteristics and problems (Shriberg, 2005).

The BBL dataset consists of 164 speakers, 101 males and 63 females. It is a set of spontaneous speech sentences, recorded from 164 subjects speaking in Levantine colloquial Arabic. Levantine Arabic is the dialect of Arabic spoken by ordinary people in Lebanon, Jordan, Syria and Palestine. It is significantly different from Modern Standard Arabic (MSA), in that it is a spoken rather than a written language. It includes different word pronunciations and even different words, from MSA, the written and "Official" form of Arabic.

The subjects in the corpus were responding to refugee/medical questions (such as "Where is your pain?", How old are you? , and were playing the part of refugees. Each subject was given a part to play, that prescribed what information, they were to give in response to the questions, but were told to express themselves naturally, in their own way, in Arabic. To avoid priming subjects to give their answer with a particular Arabic wording, the parts were given in English rather than Arabic. All subjects were thus bilingual. The following is part of an example scenario:

> "You are Maraam Samiir Shamali.., you have no children"

Speech data has been recorded using a close-talking, noise-cancelling, headset microphone (the Andrea Electronics NC-65). A Java-based data-collection tool, developed by BBL, was used to do the collection of speech. The audio was recorded in MS WAV, signed PCM. Sampling rate was 16 KHz, with 16-bits resolution.

System specification:

Feature extraction: The set of observations related to each state are well modeled by the Cepstral coefficients, mainly the Mel-Frequency Cepstral Coefficients (MFCC). The Mel-Cepstrum makes use of the auditory system principle, namely, it has a high discriminating power at lower frequencies compared to higher frequencies. Cepstral coefficients are the mostly used features in speaker recognition due to many reasons, the most important one is that they represent well the vocal tract changes and have the ability to contend with convolution channel distortion.

Usually the MFCC are augmented by some parameters that represent the dynamic features of speech, namely their first and second order derivatives, in order to give better recognition rates.

Many proposals are available in the literature to improve the system performance by proposing other features (Grimaldi and Cummins, 2008) or combining MFCC with other features, such as high level features (Shriberg, 2007; Ezzaidi *et al*., 2001). Those proposals did not show much improvement over MFCC, hence we restrict our study to MFCC and their dynamic counter parts, for system simplicity and ease of comparison with other works. In our recognition system, we used 12 and 36 MFCC (12MFCC+ 12ΔMFCC and 12ΔΔMFCC).

There were some concerns about the recorded files in the dataset. For example, (i) the dataset was recorded through a microphone and using a kind of switch (or mouse click) which induced bursts or spikes in all the wave files. We removed all these spikes from our selected files. (ii) The recorded speech amount was not of the same length for all the speakers, as some speakers spoke for a long time compared to others, a ratio of 1-5 was sometimes observed. However, we did not time normalize the speech durations across the speakers. (iii) Some speakers had many 'euh' at each answer, making a non spontaneous manner of talking. We left it as is, to make our data as realistic as possible. (iv) There was too much silence at the beginning and at the end of the files, in the whole dataset. Silence represented a ratio of approximately 1/3 from the full dataset. We removed this silence from the selected sentences using loudness and zero-crossing criteria. (v) Some speakers were not well understood as others, due perhaps to the distance from the microphone.

The obtained wave files were segmented into frames of 20 ms, with an overlap of 10 ms; Each frame is then converted to a set of MFCCs.

HMM models: Hidden Markov Models (HMMs) (Rabiner, 1989) are a well known and widely used statistical method for characterizing the spectral features of speech frames. The assumption underlying HMM is that the speech signal can be well characterized as a parametric random process and the parameters of the stochastic process can be predicted in a precise, well-defined manner, for example, by a mixture of Gaussians.

For speaker recognition, each speaker is modeled by an HMM of a fixed number of states, to which are associated a mixture of Gaussians representing the space of observations (segments of voice) of each speaker.

In (Matsui and Furui 1992) , the authors concluded that "with continuous HMM's, the speaker identification rates are strongly correlated with the total number of mixtures, irrespective of the number of states". A recent study (Deshpande and Hoalmbe, 2008) concluded that "the speaker identification rates using a Continuous Density HMM (CDHMM) are strongly correlated with the number of mixtures per state and the amount of data used for training".

In this study, we used an HMM model per speaker and varied the number of states per model from 1-4, with different number of mixtures per state (# mixtures: 1, 2, 4, 8, 12, 16 and 24), for the two subsets with different amounts of speech and different number of speakers.

Training and testing:

First subset: This subset consists of the first 20 speakers in the database (ID 001-020). The subset consists of 10 males and 10 females. For each speaker, we selected the first twenty sentences. We divided the wave files into two parts, 2/3 for training and 1/3 for testing (average of 32 and 16 sec respectively). We used 12 MFCC and varied the number of states from 1-3 and the number of mixtures per state from 1-24.

Second subset: This subset consists of 60 speakers, 30 males and 30 females. Each speaker had 20 sentences. The selected sentences were the longest wave files existing in the directory of each speaker respectively. The first 14 largest sentences were used for training and the remaining 6 were used for testing. No initial analysis had been done on the quality of the speakers, in order to select the best candidates, since this selection will result in an unrealistic situation in speaker recognition. In real situations, impostors might be anyone, sometimes even from the registered set of users.

Table 1: Subset of 60 speakers (30 males and 30 females) and their corresponding speech durations (Dur.) in seconds

Males				Females			
ID	Dur.	ID	Dur.	ID	Dur.	ID	Dur.
003	137	033	103	001	240	029	244
005	99	035	271	002	182	030	346
007	171	038	150	004	92	034	105
013	113	039	139	006	120	036	269
014	73	040	144	008	50	037	100
015	101	041	146	009	98	210	104
016	105	042	231	010	79	211	76
017	118	043	83	011	114	216	127
018	396	044	123	012	127	219	78
020	180	200	220	019	124	220	59
021	320	201	73	023	121	222	106
022	170	202	61	024	134	227	123
027	178	203	197	025	301	231	96
031	236	204	54	026	184	232	295
032	261	205	66	028	123	238	91

The list of speakers and their corresponding total speech durations in seconds is presented in Table 1, where the ID numbers represent the speaker as mentioned in the dataset document.

RESULTS

The initial idea, of incrementing the number of states against the variation of the number of mixtures, is to apply the investigation of (Deshpande and Hoalmbe, 2008) to the Arabic speaker recognition. The authors mentioned that the best result of speaker identification was obtained with a 2 states single mixture HMM. Our results for the first dataset shows that for a low number of states, we got higher recognition rates, as shown in Fig. 1, which agrees in general with (Deshpande and Hoalmbe, 2008).

From Fig. 1, we can notice that increasing the number of mixtures increases the recognition rate. We also can notice that the one state model has the best result, except at the case of one Gaussian per state.

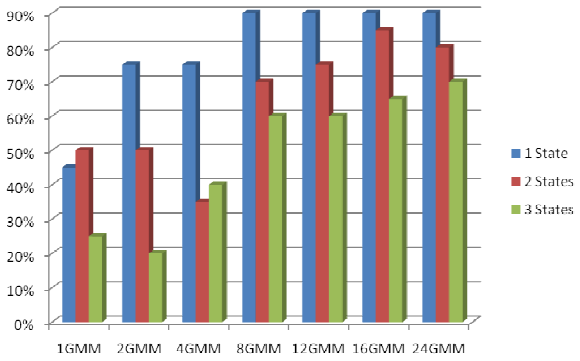

Fig. 1: Recognition rate for 20 speakers (12 MFCCs), against HMM states' increment

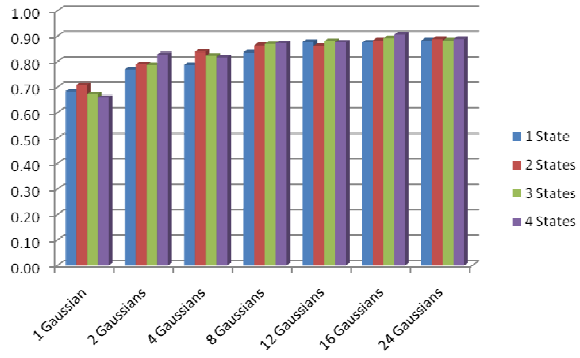

Fig. 2: Recognition rate for 60 speakers (12 MFCCs), against HMM states' increment

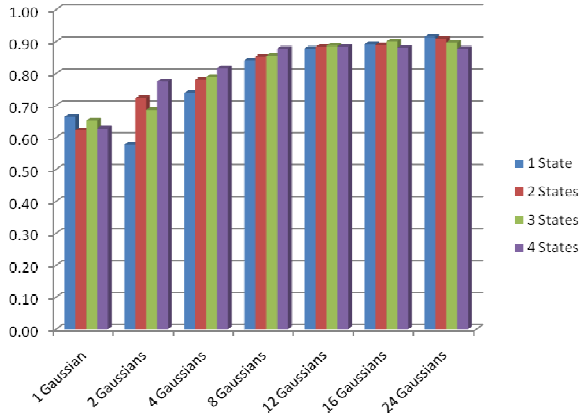

Fig. 3: Recognition rate for 60 male speakers (36 MFCCs), against HMM states' increment

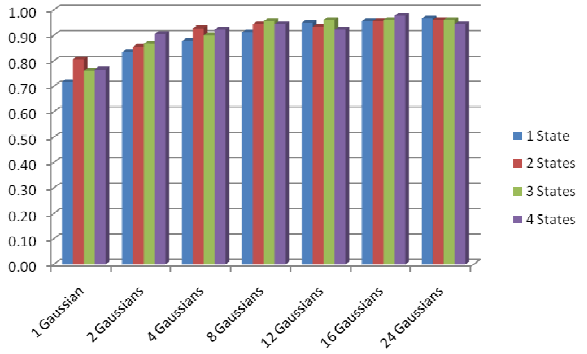

Fig. 4: Recognition rate for 30 male speakers (12 MFCCs), against HMM states' increment

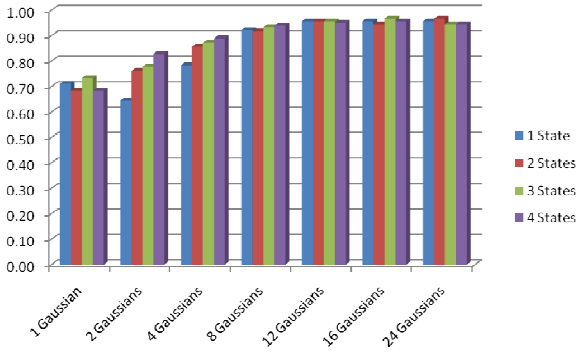

Fig. 5: Recognition rate for 30 male speakers (36 MFCCs), against HMM states' increment

Figure 2 and 3 present the results for the 60 speakers of the second subset, using 12-36 MFCC respectively. Figure 4 and 5 give the result for males of the same subset using also 12-36 MFCCs respectively, while Fig. 6 and 7 give the result for the females of the same subset using 12-36 MFCCs respectively.

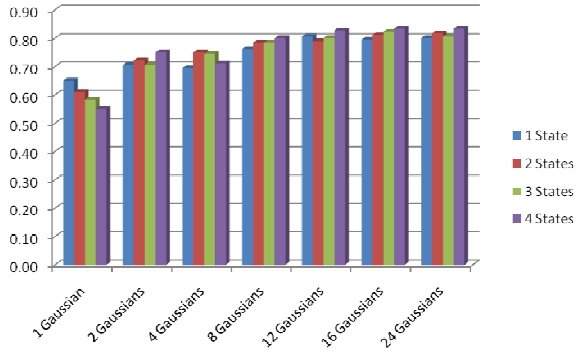

Fig. 6: Recognition rate for 30 female speakers (12 MFCCs), against HMM states' increment

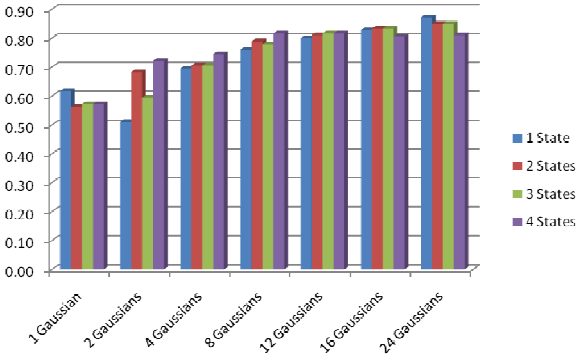

Fig. 7: Recognition rate for 30 female speakers (36 MFCCs), against HMM states' increment

For the second subset, our results do not agree with (Deshpande and Hoalmbe, 2008) conclusion because the recognition rate was higher when we increased the number of states, but they agree with the conclusion of (Matsui and Furui, 1992) where it is clear that the recognition rate depended mainly on the number of mixtures.

DISCUSSION

We can observe, from the Fig. 1-7, that increasing the number of mixtures increased the recognition rate noticeably at the cases of low number of mixtures (1, 2, 4 and 8). For the case of 12 mixtures and more, the number of mixtures did not have any noticeable difference on the results, nor did the number of states. By "noticeable", we mean a variation more than 3%. We also can remark that males had better results than the females, for almost all the variations in the number of states as well as the number of mixtures. Including the ΔMFCC and ΔΔMFCC, had no noticeable effect, on the overall recognition rates.

In Matsui and Furui (1992) concluded that the speaker identification rates are strongly correlated with the total number of mixtures, irrespective of the number of states. Our results lead to the same conclusion, when varying the number of mixtures from 1-12. From 12 mixtures and up it seems that neither the number of mixtures nor the number of states have any noticeable effect. These conclusions are not the same as (Deshpande and Hoalmbe, 2008) conclusion.

There was no noticeable difference in changing the number of mixtures between the result of the 60 speakers and the males or females subsets, the trend is gender independent. Similar conclusion can be reached for effect of changing the number of states.

CONCLUSION

We developed an HMM/GMM based system for Arabic speaker recognition; we used two subsets, one with 20 speakers and the other with 60 speakers, from the Babylon dataset. We investigated the effect of varying the number of states and number of mixtures per state of the HMM model. Our results have shown that, for the larger dataset, there is a strong correlation between the recognition performance and the number of mixtures per state, which favors the findings of (Matsui and Furui, 1992).

For the small dataset, increasing the number of states did not affect the recognition rates as increasing the number of mixtures per state. Moreover low number of state has better results than high number of states. We also found that the recognition rate for male speakers is a bit higher than for the female speakers.

For the larger dataset the number of states has small effect at low number of Gaussians and almost has no effect at high number of Gaussians.

We are currently working on modeling the whole dataset and investigating the effect of the number of states and the number of mixtures per state as well as the type of HMM used.

REFERENCES

Bimbot, F., J.F. Bonastre, C. Fredouille, G. Gravier and I. Magrin-Chagnolleau *et al*., 2004. A tutorial on text-independent speaker verification. EURASIP. J. Applied Sign. Process, 2004: 430-451. DOI: 10.1155/S1110865704310024

Deshpande, M.S. and R.S. Hoalmbe, 2008. Text independent speaker identification using hidden Markov models. Proceeding of the 2008 1st International Conference on Emerging Trends in Engineering and Technology, July 16-18, IEEE Computer Society, USA., pp: 641-644. DOI: 10.1109/ICETET.2008.46

Ezzaidi, H., J. Rouat and D.O. Shaughnessy, 2001.Towards combining pitch and MFCC for speaker identification systems. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.6.3889&rep=rep1&type=pdf

Grimaldi, M. and F. Cummins, 2008. Speaker identification using instantaneous frequencies. IEEE Trans. Audio Speech Language Process., 16: 1097-1111. DOI: 10.1109/TASL.2008.200110

Ilyas, M.Z., S.A. Samad, A. Hussain and K.A. Ishak, 2007. Speaker verification using vector quantization and hidden Markov model. Res. Dev. SCORED. 12: 1-5.

Matsui, T. and S. Furui, 1992. Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMMs. Proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing, Mar. 23-26, IEEE Xplore Press, San Francisco, CA., pp: 157-160. DOI: 10.1109/ICASSP.1992.226096

Rabiner, L.R., 1989. A tutorial on hidden Markov models and selected applications in speech recognition. IEEE Proc., p. 257-286.

Shriberg, E.E., 2005. Spontaneous speech: how people really talk and why engineers should care. Proceeding of the European Conference on Speech Communication and Technology, (CT'05), IEEE, USA., pp: 1781-1784. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.73.3765&rep=rep1&type=pdf

Shriberg, E.E., 2007. Higher-level features in speaker recognition. Lecture Notes Comput. Sci., 4343: 241-259.