International Journal of Physical Sciences Vol. 7(13), pp. 2076 - 2084, 23 March, 2012
Available online at http://www.academicjournals.org/IJPS
DOI: 10.5897/IJPS11.763
ISSN 1992 - 1950 ©2012 Academic Journals

Full Length Research Paper

A technique to overcome the problem of small size database for automatic speaker recognition

Mansour Alsulaiman

Speech Processing Group, Computer Engineering Department, College of Computer and Information Sciences, King Saud University, P.O. Box 51178, Riyadh 11543, Saudi Arabia. E-mail: msuliman@ksu.edu.sa. Tel: 4677102. Fax: 4676990.

Accepted 03 February, 2012

Modeling a system by statistical methods needs large amount of data to train the system. In real life, such data are sometimes not available or hard to collect. Modeling the system with small size database will produce a system with poor performance. In this paper, a method for increasing the size of a speech database is proposed. The method works by generating new samples from the original samples, using combinations of the following methods: speech lengthening, noise addition, and word reversal. To make a proof of concept, a severe test condition is used, in which the original database consists of one sample per speaker, for a speaker recognition system. The system is tested using original samples and the highest obtained recognition rate for mixed genders is 91.41% and that of 93.24% for male only speakers.

Key words: Lengthening of samples, noise addition, word reversal, speaker recognition.

INTRODUCTION

Statistical methods such as Hidden Markov Models (HMM), Gaussian Mixtures Models (GMM), or Neural Networks (NN) are very popular tools for classification. Their main problem is that they need large amount of data to model the system. In real life, such data is sometimes not available or hard to collect. Modeling the system with small size data set will produce a system with poor performance.

There exist several methods to overcome the problem of small size database. One of the well-known solutions is Bagging which was introduced by Breiman (1994, 1996). Bagging (which is the acronym for bootstrap aggregating) works by creating multiple random learning sets from the original learning set. The created sets consist of single and duplicate samples and each of them is the same size as the original set. These sets are fed to weak classifiers and the results of the classifiers are averaged. Bagging goal as introduced by Breiman was to reduce the classification error, but it is also used to deal with small size data in many fields (Lean et al., 2008; Drapper and Beak, 1998). Different enhancements to bagging were introduced by Breiman (2001a, b). Some authors proposed other ways to enlarge the data set that were specific to the data type or application field (Mori et al., 2000; Fei-Fei et al., 2007). Alsulaiman et al. (2010) proposed a method of speech lengthening and word reversal to a database of 25 speakers using HMM as the recognition engine. The results obtained from this method were encouraging. Solving the problem of small data is still a research topic (Cambell 2009). In this paper, we propose a method for generating samples to increase the size of the data set. The method we propose is designed for speech/speaker recognition but it can be tailored for other tasks such as object recognition. The method works by generating new samples from the available samples, using the following methods or combinations of them: speech lengthening, noise adding, and word reversal. These methods or techniques are considered as emulations of different pronunciations by the same speaker (speech lengthening and speech reversal) or different environment of recording (adding different types of noise). All of the changes are done in the time domain, without changing the original characteristics of the speaker.

Part of our method is to perform speech lengthening. Several techniques of speech lengthening are proposed in literature, for example, synchronized overlap-and-add (SOLA) procedure (Roucos and Wilgus, 1985), waveform similarity overlap - and-add (WSOLA) (Erogul and Karagoz, 1998), time domain pitch- synchronized OLA

(TD-PSOLA) (Moulines and Champentier, 1990), segmental lengthening at prosodic boundaries and in accented syllables (Jianten, 2004), etc. However these methods are complex in the sense that they require either estimation of pitch or prosodic boundaries, etc. To make a proof of concept of the proposed method, a severe test condition is used, in which the original database consisted of one sample per speaker for the speaker recognition system. In this paper, we develop a very simple lengthening technique that requires locating only the middle frames of consonants. Once these frames are located, they are copied immediately after the original middle frames.

DATABASE

This research is conducted with a local database recorded at King Saud University, College of Computer and Information Sciences (CCIS), during the year 2007 (Al-Dahri et al., 2008). The database consists of 91 native Arabic speakers, pronouncing the Arabic word "نعم" (/n/,/a/,/ʕ/,/a/,/m/) , which stands for the word "yes" in English, in 5 different occurrences (samples). The main characteristics, of this word, are of two aspects. The first aspect is that approximately all the Arab speakers frequently say "yes" (in Arabic) in any discussion. The second aspect is the richness of this word in the phonetic structure. It contains at the beginning the nasal phoneme [ن] (/n/), at the middle a pertinent phoneme [ع](/ʕ/), and the last phoneme is the bilabial phoneme [م]/m/. It also contains two occurrences of the vowel (فتحة /a/). This richness, plus the fact that it is a commonly pronounced word makes it a good choice for our investigation.

The HMM based system uses the phonemes of the word "نعم", for recognizing the speaker, while the GMM based system models the speaker regardless of the phonemes in the text. In the database, the five original samples are labeled as O_1, O_2, O_3, O_4, and $O_{5.}$

First original sample O_1 is used to generate the new training samples by using the proposed techniques, which will be explained later.

The remaining four original samples O_2, O_3, O4, and O_5 are used for testing the system.

In this work, two sets of speakers are used in the experiments: (i) samples of 50 different speakers (37 adult Male, 5 children and 8 Female) and (ii) samples of 37 male speakers. The speakers recorded their speech samples in one or two sessions. The samples were recorded with 16 KHz sampling rate and 16 bits per sample resolution

MODELING TECHNIQUES

In text dependent applications, where there is a strong prior knowledge of the spoken text, additional temporal knowledge can be incorporated by using HMM, which is a stochastic modeling approach used for speech/speaker recognition. The HMM system is built using the Hidden Markov Toolkit (HTK), which was developed by Steve Young at Cambridge University in 1989. In this research, each phoneme of the word is modeled by one HMM model with every speaker having his own phoneme model. Each phoneme model has three left to right active states; each state has one Gaussian. For a given speaker, each phoneme is modeled differently. These models can be used to find the speaker identity. The silence model is also included in the model set.

The second modeling technique used is GMM. GMM is a state of the art modeling technique that copes more with the space of the features, rather than the time sequence of their appearance. Each speaker is modeled by a GMM that represents, in a weighted manner, the occurrence of the feature vectors. The well-known method to model the speaker GMM is the Expectation-Maximization algorithm, where the updates of the model parameters (Mean, variance and mixture coefficients) are adapted and tuned to converge to a model giving a maximum log-likelihood value. The GMM model is given by the weighted sum of individual Gaussians as

$$p(X|\lambda) = \sum_{i=1}^{M} w_i g(X|\mu_i, \Sigma_i) \qquad (1)$$

where X is a D-dimensional continuous-valued data vector (that is measurement or features), $w_i, i = 1..M,$ are the mixture weights, and $g(X|\mu_i, \Sigma_i)$, $i = 1..M$, are the component Gaussian densities. Each component density is a D-dimensional Gaussian function of the form,

$$g(X|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(X - \mu_i)'\Sigma_i^{-1}(X - \mu_i)\right\}, \qquad (2)$$

with mean vector μ_i and covariance matrix Σ_i. The mixture weights satisfy the constraint that $\sum_{i=1}^{M} w_i = 1$. The model of the GMM is described as:

$$\lambda = (\omega_i, \mu_i, \Sigma_i), \; i = 1..M \qquad (3)$$

FRONT END PROCESSING

This step deals with the extraction of features, where speech is reduced into a smaller amount of important characteristics, represented by a set of vectors, such as the Mel Frequency Cepstral Coefficients (MFCC). The cepstral features are the mostly used feature in speaker recognition, due to many reasons: their robustness to noise distortion, their capability to filter the sound as does the human cochlear system, and their degree of de-correlation. The system uses 25 milliseconds hamming window duration with a step size of 10 milliseconds. In the HMM experiments, 12 MFCC are used, while in the

GMM experiments, 12 and 36 coefficients are used. The 36 coefficients consist of 12 MFCC and their first and second order derivatives.

PROPOSED METHOD AND SAMPLE GENERATION

Combinations of different techniques are used in the proposed method to generate new speech samples from one original sample. The techniques are developed to simulate different instances or circumstances that happen in real life. The techniques produce new samples without changing speaker's features such as pitch, hence keep the speaker's identity. All the samples are generated by modifying the original speech sample O_1 of each speaker in the time domain. The boundaries of the phonemes are detected using the PRAAT software (http://www.fon.hum.uva.nl/praat/, accessed on 05-03-2011). The new samples are generated by any/or combination of the following techniques: speech lengthening, addition of noise at different signal to noise ratio (SNR), and word reversing. These techniques and the sample generation by using them are elaborated one by one in the following subsections.

Speech lengthening

The new samples are generated by copying a small part from speech sample O_1 and then inserting it just after the place it was copied from. This is done on the first, middle, and last consonants of the sample, resulting in three different new samples. The copied part is 20 to 30 milliseconds or 40 to 60 milliseconds. In this technique, we are emulating different pronunciation of the word that is longer in three phonemes.

Samples S_5, S_6 and S_7 are generated by lengthening. They are generated by copying the central part, approximately 20 to 30 mlliseconds, of each phoneme "/n/", "/ʕ/" and "/m/" of the original sample (O_1), then inserting it, just after the place it was copied from, respectively. This group of samples is named conc1.

Samples S_8, S_9 and S_{10} are generated the same way as S_5, S_6 and S_7, but with a longer copied part, 40 - 60 milliseconds. This group is named conc2.

Word reversing

In this technique, four different samples are generated. The first sample is generated by reversing the original sample. The second, third and fourth samples are generated by copying a small part (approximately 20 - 30 milliseconds) from the consonants of the reversed word, then inserting it just after the place it was copied from, in the reversed word. In this technique, we are emulating speech lengthening as in the first technique, but in a new word that is the reverse of the original word.

Different samples are generated by this technique. The first sample in this group is S_{11} and is generated by reversing the sample O_1. The second, third, and fourth generated samples in this group are S_{12}, S_{13} and S_{14}. These samples are generated by copying a part of approximately 20 to 30 milliseconds of each consonant "/m/", "/ʕ/" and "/n/", of the sample S_{11}, then inserting it just after the place it was copied from, respectively. This group is named rev4. S_{11} alone is named rev1. The reversed order of the phonemes is leading to a new Arabic word.

Adding noise at different SNRs

The new samples are generated by adding different types of noise at different noise levels. In this technique, different environments around the speaker and/or different recording equipment are emulated.

A total of six samples are generated in this category. The samples S_{15}, S_{16} and S_{17} are generated by adding the babble noise at 5, 10 and 20 db SNR respectively. This group is named nois1. The three other samples S_{18}, S_{19} and S_{20} are generated by adding the train noise at 5, 10 and 20 db SNR respectively. The selected name for this group is nois2.

A summary of the new generated samples with their method of generation is presented in Table 1. Sixteen new samples are added to the samples of every speaker of the database; hence enlarging the size of the training part of the database by 16 times than its original size. The original sample and different combination of the generated samples are used to train the system in different experiments.

EXPERIMENTAL PROCEDURE

To confirm that the new generated samples contain supplementary information about the speakers, two experiments are performed. In the first experiment, named as E_a, the system is trained with an original sample and four copies of it (S_1 – S_4), and tested with another original sample. The recognition rate was 10%, as expected, which is very low. This is due to the fact, that there was not enough information in one sample. In the second experiment, named as E_b, the system is trained with four generated samples and tested with the original sample of these samples, and 100% recognition rate is obtained. This high rate is due to supplementary or additional information obtained during the training by using the new generated samples. However, this is not a real test, because the system should be tested with other original samples.

Many experiments are performed to evaluate the proposed techniques and, in each experiment, different combinations of generated samples are used for the training of the system. They are divided into three categories according to the techniques used for sample generation. The first category contains the samples generated by lengthening of samples. The second category includes the samples which are generated by the combination of two techniques, lengthening and noise addition. In the third category, samples are generated by lengthening and word reversing. Each category has different number of combinations of the generated samples to train the system. There are three different combinations for the first category, nine combinations for the second category, and that of six for the third. A list of training samples for every experiment of each category is provided in Table 2. Two different speaker recognition systems are used to conduct the experiments. Both systems use MFCC as the feature extraction technique to capture the speaker dependent properties but they have different modeling techniques. The first system use HMM while other use GMM to model the acoustic templates of each registered speaker in both phases of the recognition. The original sample O_1 and different combination of the generated samples are used to train the system every time.

RESULTS

The results of HMM and GMM based recognition are provided in the following subsections, and discussion on the results is presented in the next section.

HMM

The recognition rates with 50 speakers for all the

Table 1. Techniques for generating samples.

Samples label	Group	Method of generation
S5,S6,S7	conc1	A small part of the first, second, and third phonemes which are "ن", "ع" and "م", (approx. 20-30 milliseconds) is copied and inserted just after the place it was copied from.
S8,S9,S10	conc2	A small part of the first, second, and third phonemes which are "ن", "ع" and "م", (approx. 40-60 milliseconds) is copied and inserted just after the place it was copied from.
S11	rev1	Reverse of O1
S11,S12,S13,S14	rev4	S11is as above. The others are generated by copying a small part of the first, second, and third phonemes of S11, which are "م", "ع" and "ن", (approx. 20-30 milliseconds) and inserting it just after the place it was copied from.
S15,S16,S17	nois1	Babble noise at 5, 10 and 20 db is added to the original speech signal.
S18,S19,S20	nois2	Train noise at 5, 10 and 20 db is added to the original speech signal.

Table 2. Training samples for the experiments.

Category	Experiment	Groups for training
Lengthening (First)	L_1	conc1
	L_2	conc2
	L_3	conc1,conc2
Lengthening and Noise (Second)	N_1	conc1,nois1
	N_2	conc1,nois2
	N_3	conc1,nois1,nois2
	N_4	conc2,nois1
	N_5	conc2,nois2
	N_6	conc2,nois1,nois2
	N_7	conc1,conc2,nois1
	N_8	conc1,conc2,nois2
	N_9	conc1,conc2,nois1,nois2
Lengthening and Reversing (Third)	R_1	conc1,rev1
	R_2	conc1,rev4
	R_3	conc2,rev1
	R_4	conc2,rev4
	R_5	conc1,conc2,rev1
	R_6	conc1,conc2,rev4

experiments of each category by using HMM based system are given in Table 3. A comparison of the results is shown in Figure 1. The groups are labeled along x-axis and recognition rates are along y-axis.

Effect of lengthening

Three experiments are conducted in this category, named as L_1, L_2, and L_3. These experiments represent the training of the system by using the samples of conc1 and conc2. The group of training samples conc1 is used in L1, conc2 in L2, and their combination is used in L3. The recognition rates of the experiments L_1, L_2, and L_3 are 70, 63 and 83% respectively. The recognition rate of L3 illustrates that combination of conc1 and conc2 provides better recognition rates than the situation when conc1 and conc2 is used individually.

Effect of adding noise

Nine experiments are conducted in this category, namely N_1, N_2, N_3, N_4, N_5, N_6, N_7, N_8 and N_9. The recognition

Table 3. Recognition Rates (%) for HMM.

Category	Experiment	Recognition rate (%)
Lengthening (First)	L1	70
	L2	63
	L3	83
Lengthening and Noise (Second)	N1	75
	N2	72.73
	N3	72
	N4	84.34
	N5	80.30
	N6	85.35
	N7	75.76
	N8	75.25
	N9	80.30
Lengthening and Reversing (Third)	R1	86.36
	R2	88
	R3	71.21
	R4	89.90
	R5	87
	R6	90

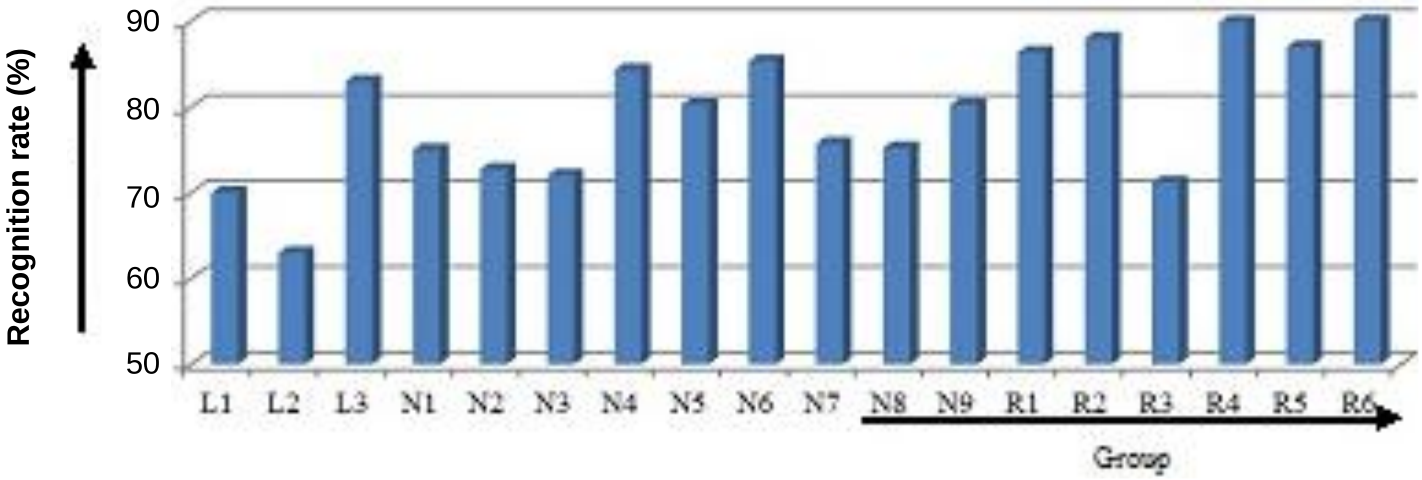

Figure 1. Recognition rates for HMM based recognition system.

rates of the experiments N_1, N_2, N_3, N_4, N_5, N_6, N_7, N_8 and N_9 are 75, 73, 72, 86, 81, 86, 76, 76 and 81% respectively. The group nois1, nois2 and their combination are combined with conc1 in experiments N_1, N_2 and N_3, respectively, while they are combined with con2 in N_4, N_5 and N_6 respectively. In experiments N_7, N_8 and N_9, combination of conc1 and conc2 is used with nois1, nois2 and with their combination respectively. The highest recognition rate achieved in this category is 86% for the experiment N_6 where the system is trained with the samples generated by con2, nois1 and nois2. In this category, conc2 clearly dominates the conc1. There are four results which are more than 80% and conc2 is present in each combination of the groups used for the training of the recognition system.

Adding small amount of noise to the samples simulate different environments without affecting the speakers' characteristics and thereby increase the recognition rate.

Table 4. Recognition rate (%) for GMM with 12 MFCC.

Category	Experiment	12MFCC			
		4 GMM	8 GMM	16 GMM	32 GMM
Lengthening (First)	L1	79.19	81.22	79.70	72.08
	L2	87.50	89.00	83.50	73.00
	L3	90.40	89.39	78.78	68.68
Lengthening and Noise (Second)	N1	88.89	84.85	82.32	74.75
	N2	88.89	89.39	87.88	82.83
	N3	88.38	88.89	86.87	73.74
	N4	88.89	85.86	87.88	83.33
	N5	89.39	88.38	87.37	83.33
	N6	89.90	90.91	87.37	83.33
	N7	89.90	89.39	87.88	82.83
	N8	89.39	90.91	83.84	80.30
	N9	90.40	89.90	88.89	86.36
Lengthening and Reversing (Third)	R1	88.38	87.37	83.33	70.20
	R2	88.89	83.33	84.34	74.24
	R3	90.40	88.38	88.38	78.79
	R4	88.89	83.33	84.34	74.24
	R5	88.00	88.00	83.00	65.00
	R6	87.88	85.35	87.37	68.69

Effect of speech reversal

In this category, six experiments R_1, R_2, R_3, R_4, R_5 and R_6 are performed. The recognition rates of R_1, R_2, R_3, R_4, R_5 and R_6 are 87, 88, 72, 90, 87 and 90% respectively. The groups rev1 and rev4 are combined with conc1 in the experiments R_1 and R_2, respectively, while they are combined with conc2 in R_3 and R_4, respectively, to train the system. In experiments R_5 and R_6, rev1 and rev4 are used with the combination of conc1 and conc2 respectively. The recognition rate of 90% is achieved for this category in the experiment R4 and R6. The training groups for the experiment R4 are conc2 and rev4, and that for R6 are conc1, conc2 and rev4, respectively. Again samples generated by the technique conc2 are part of the set of training samples used for the experiments which provide the maximum result in this category.

GMM

With 50 speakers

Complementary experiments are conducted using GMM. The recognition rates of all experiments of each category are presented in Tables 4 and 5 by using GMM based speaker recognition system. The results of 12 MFCC with 4, 8, 16 and 32 are depicted in Table 4, and the results of 36 MFCC with 4, 8, 16 and 32 GMM are provided in Table 5. It can be observed from the tables that generally the result using 4 GMM outperforms 8, 16 and 32 GMM for both number of MFCC. These results also support the facts which was found during the experiments of the HMM, that combination of more than one technique provides better recognition rates than single technique.

The maximum recognition rates achieved for the first, second and third categories are 90.40% for 12 MFCC with 4 GMM in experiment L3, 90.91% for 12 MFCC with 8 GMM in N6 and N8, and 91.41% for 36 MFCC with 4 GMM in R6. The combination of the groups of training samples for L3 is conc1 and conc2, for N6 is conc2, nois1 and nois2, for N8 is conc1, conc2 and nois2, and for R6 is conc1, conc2 and rev4. It can be seen that the group conc2 is a part of every combination in the experiments which provide maximum result in each category. A comparison between con1, conc2 and combination of con1 and conc2 is depicted in Figure 2. There are different set of experiments along x-axis. In each set, the first, second and third experiments containing group conc1, conc2 and conc1 and conc2, respectively. For instance, conc1, conc2 and conc1 and conc2 are combined with nois1 in the experiments N1, N4 and N7, respectively. The recognition rates are taken along y-axis. The recognition rate of more than 90% is achieved in six different experiments, that is L3, N5, N8, N6, N7, R6. In all these experiment, conc2 alone or with conc1 is a part of the groups used for the training of the system. So it can be concluded that conc2 provides better result than con1 when combined with other techniques

Table 5. Recognition rate in (%) for GMM with 36 MFCC.

Category	Experiment	36 MFCC			
		4 GMM	8 GMM	16 GMM	32 GMM
Lengthening (First)	L1	83.76	79.19	75.63	59.90
	L2	87.50	73.50	66.00	44.00
	L3	82.83	84.85	66.67	49.49
Lengthening and Noise (Second)	N1	85.35	82.83	75.25	56.06
	N2	89.90	83.33	78.79	58.08
	N3	88.89	87.37	78.28	62.12
	N4	88.89	87.88	77.78	68.69
	N5	90.40	86.36	80.30	62.12
	N6	89.39	88.89	79.29	68.69
	N7	89.90	87.88	80.30	59.60
	N8	90.91	85.35	75.25	52.02
	N9	88.89	86.87	73.23	68.18
Lengthening and Reversing (Third)	R1	89.39	85.86	85.86	67.17
	R2	88.89	87.37	87.37	60.1
	R3	86.87	82.83	72.22	48.99
	R4	88.89	87.37	87.37	60.10
	R5	89.00	88.50	78.50	48.50
	R6	91.41	88.89	78.79	67.68

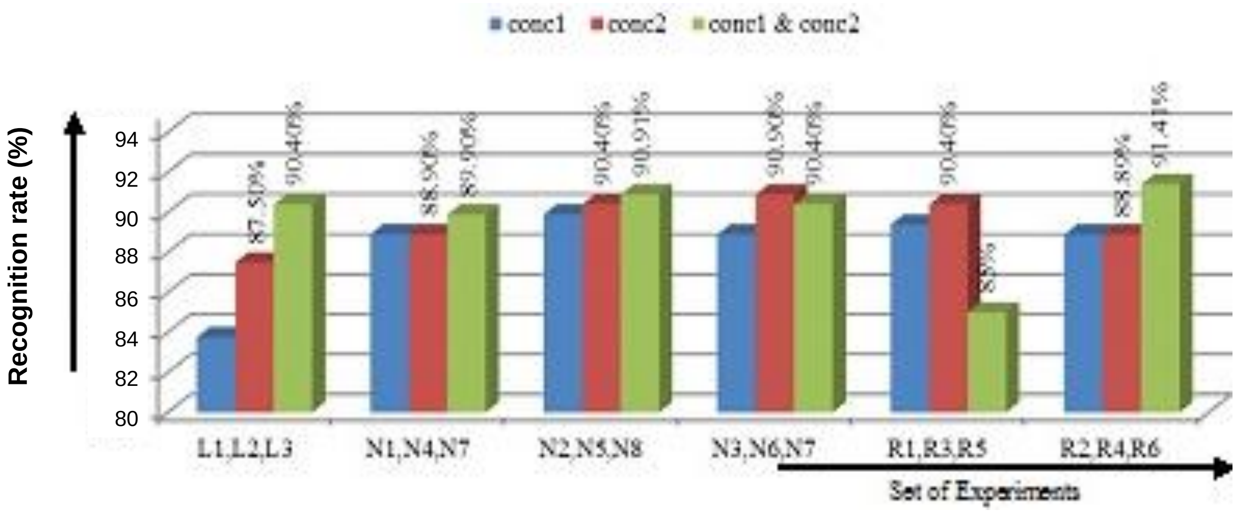

Figure 2. A comparison between conc1, conc2, and combination of conc1 and conc2.

With 37 male speakers

The database did not contain enough females or children to make the composition of male, female, and children acceptable. Hence we opted to test our method on the male only part. All the experiments are performed by using database containing 37 male speakers. The training samples used to conduct each experiment are listed in Table 2. The recognition rate (%) with 12 and 36 MFCC are presented in Tables 6 and 7, respectively. If we compare between Tables 4 and 6, it is observed that recognition rate is significantly improved while using male speakers only. The highest rate of 93.24% is obtained with R_5 experiment for GMM with 12 MFCC. The same phenomenon is true with 36 MFCC. The highest rate of 93.24% is achieved with R_1 experiment.

The higher recognition rate of male only speakers is interesting. It proves that the proposed method works fine while discriminating speakers from the same gender, which, in theory, is more difficult than discriminating

Table 6. Recognition rate in (%) for GMM with 12 MFCC.

Category	Experiment	12MFCC			
		4 GMM	8 GMM	16 GMM	32 GMM
Lengthening (First)	L1	91.22	89.19	83.11	81.76
	L2	90.54	89.19	89.86	83.11
	L3	89.19	90.54	85.14	77.03
Lengthening and Noise (Second)	N1	91.22	89.86	89.86	89.86
	N2	91.22	89.86	90.54	87.16
	N3	91.22	91.89	90.54	88.51
	N4	90.54	87.16	88.51	86.49
	N5	90.54	87.84	87.84	85.81
	N6	91.89	91.22	89.86	85.14
	N7	91.89	91.22	89.19	86.49
	N8	91.89	91.22	86.49	82.43
	N9	91.89	91.22	90.54	89.19
Lengthening and Reversing (Third)	R1	92.57	90.54	91.22	87.84
	R2	91.22	91.89	87.84	87.84
	R3	91.22	89.19	89.19	82.43
	R4	91.89	87.16	86.49	75.00
	R5	93.24	88.51	85.14	79.73
	R6	92.57	89.86	85.81	73.65

Table 7. Recognition rate in (%) for GMM with 36 MFCC.

Category	Experiment	36MFCC			
		4 GMM	8 GMM	16 GMM	32 GMM
Lengthening (First)	L1	90.54	83.11	72.97	43.92
	L2	90.54	79.73	62.16	45.95
	L3	90.54	87.84	70.27	50.68
Lengthening and Noise (Second)	N1	91.22	89.19	83.11	73.65
	N2	90.54	84.46	81.76	59.46
	N3	92.57	89.19	86.49	73.65
	N4	91.22	88.51	80.41	69.59
	N5	92.57	89.86	81.76	68.92
	N6	90.54	90.54	81.08	67.57
	N7	92.57	90.54	87.84	66.22
	N8	91.22	90.54	79.73	55.41
	N9	89.86	89.19	84.46	74.32
Lengthening and Reversing (Third)	R1	93.24	91.22	81.76	67.57
	R2	90.54	87.16	83.11	78.38
	R3	89.86	83.78	74.32	47.30
	R4	90.54	89.19	89.86	62.16
	R5	90.54	89.19	79.73	56.76
	R6	89.86	88.51	86.49	73.65

speakers across genders.

DISCUSSION

Experiment E_a sets the base for this work, since it shows that without enough information in the different samples, the HMM will not be able to build a model and recognize the speaker. Repeating the same sample does not give any new information. Then, by conducting experiments L_1 and L_2, it is proved that by careful modification of a sample, new samples can be generated that would give HMM more information, and allows building an improved model with better recognition rates. From experiment L_3, it can be seen that by complementing one technique of generation with another technique, the recognition rate increased from 63 - 70% to 83% when using HMM.

In the GMM part, the phenomenon of better results when complementing one technique of generation with another technique was not observed and the relation between the results of L_3 compared to L_1 and L_2 depended on the number of MFCCs and GMMs. From Tables 3 to 5, it can be observed that generally GMM gives better results than HMM. When the result of HMM is not good, that is, below 80%, GMM will always give better results. Generally, the results of 4 and 8 GMM are the best and are approximately similar but results with 4 GMM are better in most of the experiments. For higher number of GMMs, the recognition rates decreased as compared to 4 and 8 GMM for all number of MFCCs. In the experiments using a combination of more than one technique, GMM results are always over 80% for any number of MFCC or any number of mixtures. From the different techniques, R_6 mostly gives the best result. This can be due to the fact that R_6 has three different types of information: small lengthening, large lengthening, and reversal. Though R_5 has the same three techniques, but the reversal part is just one sample.

There are six experiments having the recognition rate in the range 83 to 90% for the HMM based recognition system. In the five experiments out of six, samples generated by conc2 are the part of the training samples. While for the GMM based system, six top recognition rates are in the range 90.40 to 91.41% and conc2 is the part of training in each of these experiments. The group conc2 clearly dominates the conc1 in all the cases of GMM, when it is used alone or with other techniques. The small increase in conc1 did not generate enough variation or new information more than the original sample.

CONCLUSION AND FUTURE WORK

A method to increase the size of the database is proposed in this paper. The proposed method is tested by considering the extreme case when only one sample per speaker is available to train the system. The results are very encouraging. The best recognition rate is 91.41% using mixed genders and 93.24% using male only speakers with 4 GMM. The promising results shows that the proposed technique although useful to enlarge the size of database, hence it can be helpful in forensic investigation where only limited information about a person is available. We are working on developing a software program to implement the proposed techniques automatically. The initial results are encouraging (best recognition rate 89%), but overall is below the manual techniques presented in this paper. We are working on improving the automatic techniques. We are also investigating other improvements to the current techniques, such as lengthening not only the consonant parts but also the vowel parts or generating samples with lengthening more than one phoneme.

The word that is used in this paper had some special characteristics that made it an excellent choice plus the fact it is a very common word in daily Arabic conversations. We will investigate applying our methods to other words or other languages.

ACKNOWLEDGMENT

This work is supported by the National Plan for Science and Technology in King Saud University under grant number 08-INF167-02. The author is grateful for this support.

REFERENCES

Al-Dahri SS, Al-Jassar YH, Alotaibi YA, Alsulaiman MM, Abdullah-Al-Mamun KAB (2008). A word-dependent automatic Arabic speaker identification system. IEEE Int. Symp. Sig. Process. Inf. Technol. (ISSPIT), pp. 198-202.

Alsulaiman MM, Mahmood A, Muhammad G, Bencherif M. A. and Alotaibi YA (2010). A technique to overcome the problem of small size database for automatic speaker recognition. Proceedings of the 5th International Conference on Digital Information Management (ICDIM), Lakehead University, Thunder Bay, Canada, pp. 303-308.

Breiman L (1994). Bagging Predictors, Technical Report No. 421, Department of Statistics, University of California (Berkley).

Breiman L (1996). Bagging Predictors. Machine Learn., Springer Netherlands, 24: 123-140.

Breiman L (2001a). Random Forests. Machine Learn., 45(1): 5-32.

Breiman L (2001b). Using iterated Bagging to Debias regressions. Machine Learn., 45(3): 261-277.

Cambell JP (2009). Forensic speaker recognition: A need for caution. IEEE Sig. Process. Mag., pp. 95-103.

Drapper B, Baek K (1998). Bagging in computer vision. Proceedings of the IEEE Computer Society Conference on Computer Vision Pattern Recognition (CVPR), pp. 144-149.

Karagoz I (1998). Time-scale modification of speech signals for language-learning impaired children. Proceedings the 2nd International Biomedical Engineering Days, pp. 33-35.

Fei-Fei L, Fergus R, Perona P (2007). Learning generative visual methods from few training Examples: An incremental Bayesian approach tested on 101 Object categories. Comput. Vision Image Unders., 106(1): 59-70.

Jianten C (2004). Restudy of segmental lengthening in Mandarin Chinese. Proceedings of Speech Prosody, Nara, Japan, pp. 231-234.

Lean Y, Wang S, Lai KK (2008). Credit risk assessment with a multistage neural network ensemble learning approach. Expert Syst. Appl., 34(2): 1434-1444.

Mori M, Susuki A, Shio A, Othsuka S (2000). Generating new samples from handwritten numerals based on point correspondence. Proceedings of the 7th IWFHR, Amsterdam, Netherlands, pp. 281-290.

Moulines E, Chanpentier F (1990). Pitch-Synchronous waveform processing techniques for text-to-speech synthesis using diphones. Speech Commun., 9: 453-467.

Roucos S, Wilgus AM (1985). High quality time-scale modification for speech. IEEE Int. Conf. Acoust. Speech Sig. Process., ICASSP., 85: 493-496.