ENHANCEMENT OF MISMATCHED CONDITIONS IN SPEAKER RECOGNITION FOR MULTIMEDIA APPLICATIONS

Waleed Fakhr[1], Ahmed AbdelSalam[2], and Nadder Hamdy[3]

Arab Academy for Science and Technology

(1) waleedf@aast.edu , (2) ah_salama@aast.edu , (3) senior member IEEE, nhamdy@ieee.org

ABSTRACT

This paper investigates the performance of an HMM-based text-independent speaker recognition system under different model and feature combinations for matched and mismatched speech coding conditions. The effects of changing the HMM topology and acoustic features is first investigated. Training and testing the models using only the voiced segments of the samples is then considered. The best model structure in each topology is then used to test the effects of speech codecs like G729 at 8 kb/s and G723.1 at 5.3 and 6.3 kb/s, used in multimedia applications, on the performance of both matched and mismatched conditions. To improve the performance in mismatched conditions, a MAP-based adaptation with different amounts of coded training data and a diagonal Affine transform for adapting the coded cepstral features to original PCM cepstral features are investigated. Results have shown that the proposed techniques improve speaker recognition performance and produced comparable results to the matched condition test.

1. INTRODUCTION

Hidden Markov models (HMM's) has proven to be a useful tool for speaker recognition [1], and there is a considerable speaker-recognition activity in industry, national laboratories and universities that make use of HMM technology [2]. Great deal of current applications is on distributed multimedia environments, mobile phones networks, Internet and VoIP [3,4]. There is no fixed rule for choosing the most appropriate acoustic features, and the best HMM topology e.g. the number of states and Gaussian mixtures, but rather it is a matter of trial and error with many heuristics [5].

In the next section, baseline experiments are done using different topologies (Left-to-right and ergodic), with cepstral and delta cepstral features, and different model structures for each topology in order to find the most promising ones. In Section 3 a voiced/unvoiced front end preprocessor is added on both training and testing phases to select only the voiced segments of the data. In section 4, the models are tested under both matched and mismatched coding conditions using the G729 8 kb/s [6], the G723.1 5.3 and the 6.3 kb/s [7] codecs. Section 5 proposes a maximum *a posteriori* "MAP" adaptation technique with different amounts of coded data to adapt the baseline PCM-trained models to enhance the performance of the mismatched test results. Section 6 proposes a method based on a diagonal Affine transform which adapts the coded cepstral features to the PCM cepstral features to enhance the performance in mismatched conditions. All the training and testing experiments are applied using the Hidden Markov Model Toolkit (HTK)[8].

2. BASELINE EXPERIMENTS

The database used in this speaker identification set of experiments includes 30 speakers, each speaker recorded 14 isolated Arabic words repeated 6 times resulting in 84 samples for each speaker and 2520 samples for the whole set, all are coded with PCM 64 kb/s.

12 words (72 samples) from each speaker are used for training and 2 words (12 samples) are used for the testing. Four main topologies are investigated, Left-to-right HMMs, with and without delta-ceps, and ergodic HMMs with and without delta-ceps. Each topology was tried with 520 different model structures (except the ergodic which has 390 structures as it requires more than one state to be applied). The model structures have states ranging from 1 to 4 states, and each state has number of Gaussian mixtures ranging from 1 to 10 mixtures with cepstral coefficients ranging from 8 to 20 (without C_0). These add up to 1820 experiments in total. Average performance against number of Gaussian mixtures curves are plotted in figure(1). It is concluded that the performance increases with the number of Gaussian mixtures until it reaches saturation. It can also be seen that the performance of the ergodic configurations are better, and that the ergodic topology with delta cepstral features has the best results. More detailed results and analysis are found in [9, 10].

3. VOICED SEGMENTS –BASED EXPERIMENTS

To test the effect of considering the voiced parts of the speech on the recognition, a set up is used that comprises a voiced/unvoiced preprocessor added to detect and extract the voiced segments of the samples and removes the unvoiced and silence parts on both the training and testing modules. The used preprocessor is a modification of the one found in [11]. Average performance

against number of Gaussian mixtures curves for the voiced speech experiments compared to the baseline experiments

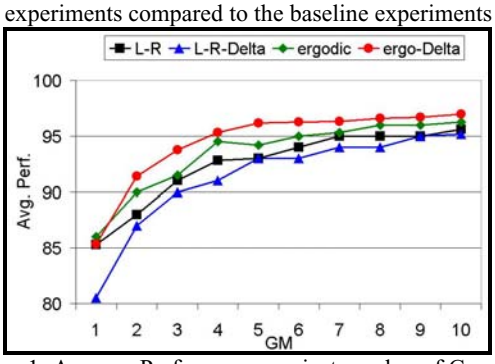

Figure 1: Average Performance against number of Gaussian mixtures curves for baseline experiments

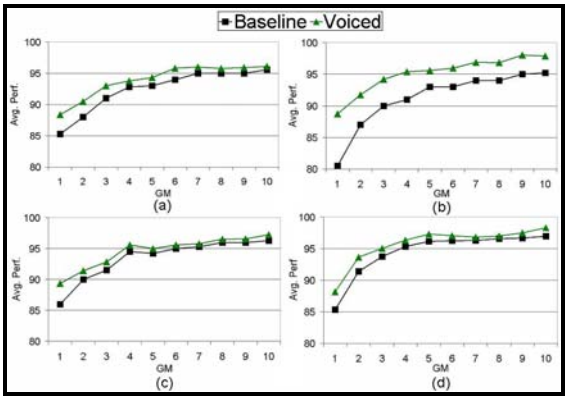

Figure 2: Average performance against number of Gaussian mixtures curves for the voiced speech experiments compared to the baseline experiments: (a) L-to-R topology, (b) L-to-R with delta, (c) Ergodic, and (d) Ergodic with delta

are plotted in Figure (2) where a clear enhancement in performance is evident especially with the delta features.

Based on the above results, the best model structure from each topology will be used in the rest of the next experiments. Best model structure was chosen depending on performance and simplicity [10]. As Gaussian mixture model (GMM) is considered as a special case of (HMM), best model structure from both GMM and HMM on each topology will be used.

Table(1) shows the best GMM and HMM model structure on each topology. The "Model Structure" is represented as (State, Gaussian Mixture , Cepstrums), and the "Model No." represents the number of the model that will be used as reference to it in the following figures.

It is clear that the overall performance of the system has increased by 1-2 % compared to the baseline experiments when the voiced segments are only considered, while the number of model structures that showed performances greater than 95% has increased by 25-35%. This voiced/unvoiced preprocessor is used in all the following experiments.

4. MATCHED/MISMATCHED CODING CONDITIONS

In these experiments we investigate speaker recognition from CELP-coded speech for the G.729 (8 kb/s), and G.723.1 (5.3 and 6.3 kb/s) codecs. For each codec, there are 2 conditions that we tested:

Condition A: This is a "fully matched" case where the training and test data are CELP-coded speech.

Condition B: This is a "fully mismatched" case where the models are derived from PCM-coded speech and the test data are from CELP-coded speech

Table (1): Best GMM model structures

Topology	Model Structure	Performance	Model No.
Cepstrums	(1state, 8 gm, 15 ceps)	96.6 %	M1
Delta Cepstrums	(1state, 9 gm, 12 ceps)	97 %	M2
Left-right	(2 state, 6 gm, 10 ceps)	95.4 %	M3
L-R Delta	(3 state, 4 gm, 13 ceps)	96.6 %	M4
Ergodic	(2 state, 3 gm, 12 ceps)	96.3 %	M5
Ergodic-Delta	(3state, 10 gm, 14 ceps)	98.6 %	M6

Results for experiments based on conditions A and B are shown in Figures (3) and (4) respectively. In the matched conditions, speaker recognition performance corresponded to coder quality, where G729 has best performance and best quality and G723.1 has lowest performance and lowest quality. G723.1 has nearly the same performance for its 2 rates (5.3 and 6.3 kb/s). In the mismatched conditions where the models are trained with PCM data, the 2 codecs changed roles; G729 has the lowest performance while G723.1 has the highest, and in this case G723.1 with rate 6.3 kb/s outperformed the 5.3 kb/s. The degradation in G729 performance in the mismatched condition may be due to the presence of the post-filter, better investigation for the performance of G729 with/without the post-filter can be found in [4]. The performance of both codecs in the mismatched condition is much lower than their performance in the matched condition due to the mismatch between the cepstral features of PCM original signals used in training and the CELP-coded data used for testing. Best performance in the matched and mismatched conditions is achieved using the 'ergodic-delta' model structure.

5. MAP ADAPTATION

A way to enhance the performance is to adapt the existing PCM-based models to the decoded data from the used codecs. We investigate the effect of MAP adaptation on the speaker recognition performance on 2 different conditions:

Condition A: Matched adaptation where the models are adapted only using decoded data from a certain codec and the test data belongs also to that codec

Condition B : Mixed adaptation, where the adaptation data are samples decoded from different codecs

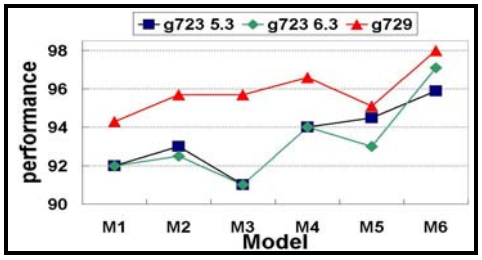

Figure 3: condition "A" experiment results (matched condition)

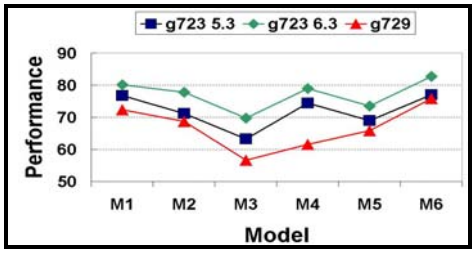

Figure 4: Condition "B" results (mismatched condition)

In both conditions, 2 experiments are made with different amounts of adaptation data. For each speaker 48 and 72 samples from each codec are used for adaptation in the experiments respectively. Figures (5) and (6) show the results of conditions A and B experiments respectively.

MAP Adaptation produced comparable results on both conditions with the matched experiments. Best performance in both conditions was again achieved using the ergodic–delta model structure.

6. AFFINE TRANSFORM

To enhance the performance due to mismatch effects, a diagonal Affine transform is used to map the cepstrum coefficients obtained from the CELP data to the ones obtained from the PCM data. This procedure uses 4 utterances per speaker from the training data to estimate the scaling and shift parameters for the Affine transform. The parameters are estimated for each cepstrum coefficient in the feature vector independently using a least-squares polynomial fitting criterion to obtain a mapping from the CELP data to the PCM data in the cepstrum domain, where the transform obeys the following equation:

$$Y_i(n) = a_i * X_i(n) + b_i$$

Where:
X(n) : frame n of the Celp coded speech cepstum
Y(n) : frame n of the PCM speech cepstum
i : cepstrum coefficient number
a_i : slope parameter for cepstrum coefficient number i
b_i : shift parameter for cepstrum coefficient number i

The average performance for mismatched experiments using G.723.1 5.3 kbps , G.723.1 6.3 kbps and G.729 8 kbps codecs is enhanced as shown in figure(7), figure(8) and figure(9) respectively.

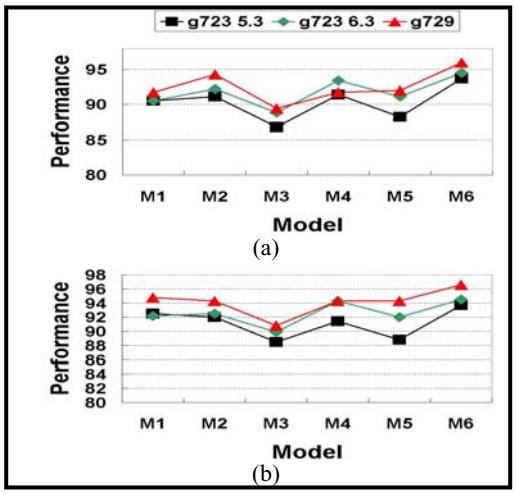

Figure 5: Matched Map Adaptation (a) 48 samples (b) 72 samples per codec for each speaker

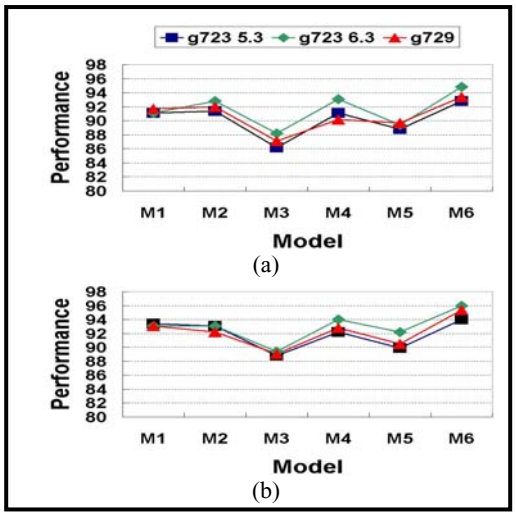

Figure 6: Mixed Map Adaptation (a) 48 samples (b) 72 samples per codec for each speaker

7. CONCLUSIONS

In this paper a comprehensive study on text-independent speaker recognition using different HMM topologies and features is carried out. The addition of a voiced / unvoiced classifier in the front end in order to extract the voiced segments only gave better results on the average compared to considering the whole word. Best model topologies are then selected for the task. These topologies are then used for investigating the effects of matched versus mismatched coding conditions which may occur in distributed speaker recognition applications, e.g., over internet or mobile networks. The experiments have shown significant degradation of performance in mismatched conditions between PCM–trained models and CELP-coded data. Finally, two techniques for improving the performance in these situations were tried, namely; the MAP adaptation strategy, and the Affine

transform strategy. Significant improvements in performance over mismatched conditions for both cases were recorded.

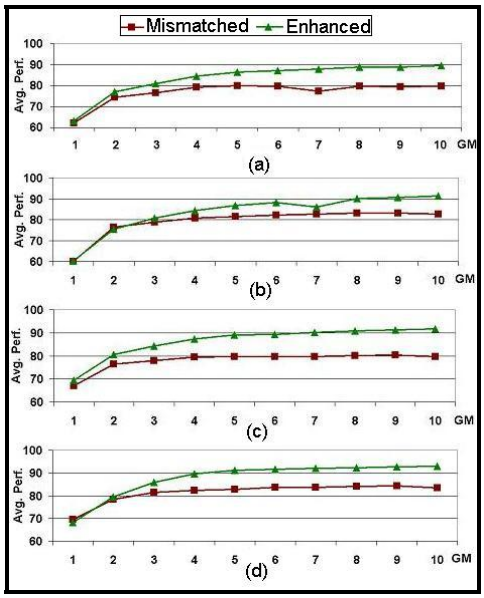

Figure 7. Average performance against number of Gaussian mixtures for mismatched and enhanced speaker recognition experiment using G.723.1 5.3 kbps speech data
(a) Left – Right, (b) Left-Right with delta features, (c) Ergodic , and (d) Ergodic with delta features Model

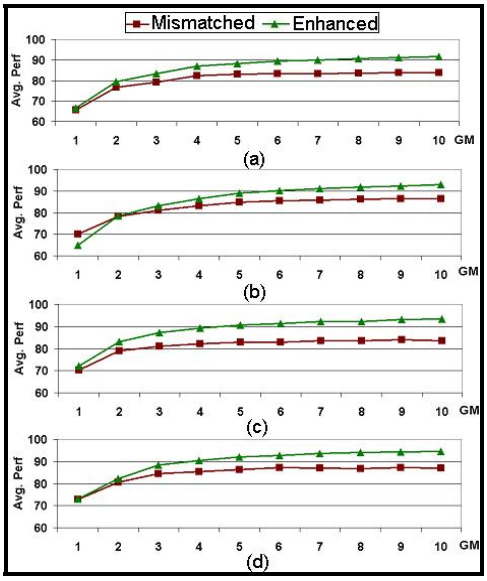

Figure 8.Average performance against number of Gaussian mixtures for mismatched and enhanced speaker recognition experiment using G.723.1 6.3 kbps speech coded data
(a) Left – Right (b) Left-Right with delta features (c) Ergodic , and (d) Ergodic with delta features Model

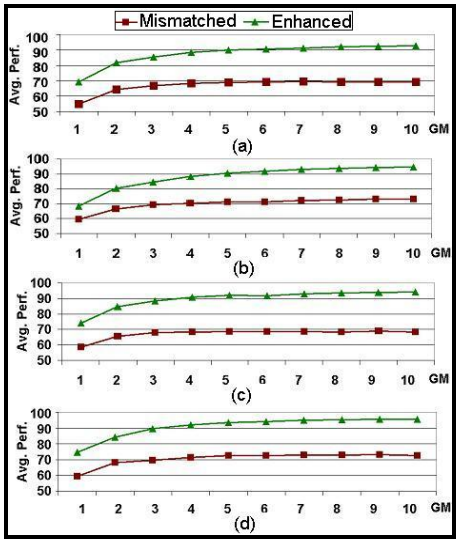

Figure 9.Average performance against number of Gaussian mixtures for mismatched and enhanced speaker recognition experiment using G.729 8 kbps speech data
(a) Left – Right, (b) Left-Right with delta features, (c) Ergodic, and (d) Ergodic with delta features Model

REFERENCES

[1] Bourlard, H., Morgan, N., "Speaker verification, A quick overview", IDIAP research report 98-12, 1998.

[2] Campbell, Joseph P., JR., "Speaker recognition: A tutorial", Proceedings of the IEEE, Vol 85, No.9, September, pp 1437-1462, 1997.

[3] Kuitert, M., Boves, L., "Speaker verification with GSM coded telephone speech", Proc. Eurospeech'97, Vol.2, pp.975-978, 1997.

[4] Quatieri, T.F., Singer, E., Dunn, R.B., Rynolds, D.A., Campbell, J.P., "Speaker and language recognition using speech codec parameters", Proc. Eurospeech'99, Vol.2, pp. 787-790, 1999.

[5] Jurafsky, D., Martin, James H., "Speech and Language processing, An Introduction to natural language processing, computational linguistics, and speech recognition", Prentice Hall, 2000.

[6] ITU-T Recommendation G.729, "Coding of speech at 8 kb/s using conjugate- structure algebraic-code-excited linear prediction," June 1995.

[7] ITU-T Recommendation G.723.1, " Dual rate speech coder for multimedia communication transmitting at 5.3 and 6.3 kb/s," March 1996.

[8] Hidden Markov Model Toolkit (HTK).
http://htk.eng.cam.ac.uk

[9] A.AbdelSalam, W.Fakhr, N.Hamdy, "Text-independent Speaker Recognition using Voiced Segments", Fourth conference on language engineering, Cairo, Oct.2003.

[10] A.AbdelSalam, W.Fakhr, N.Hamdy, "Arabic Text-independent Speaker Recognition", Ain Shams Journal, Sept 2003.

[11] Spectral Analysis and LPC Vocoder, ELCE 532 Course Project, May 2000, Rice University.
http://www.owlnet.rice.edu/~elec532/PROJECTS00/vocode/