Enhancing the Performance of a GMM-based Speaker Identification System in a Multi-Microphone Setup

Andreas Stergiou, Aristodemos Pnevmatikakis, Lazaros C. Polymenakos

Autonomic & Grid Computing Group
Athens Information Technology, Athens, Greece
{aste, apne, lcp}@ait.edu.gr

Abstract

In this paper the speaker identification system developed at Athens Information Technology is presented. It is based on the Gaussian Mixture modeling of the Mel-Frequency Cepstral Coefficients of speech. Starting from this basic algorithm, we describe and discuss two significant modifications that have resulted in performance enhancements, in terms of both processing speed and identification accuracy. We present the performance of our system in the recent CLEAR 2006 evaluation workshop and also discuss approaches to further improve our system by fusing decisions derived from a multitude of sensors in a multi-microphone setup.

Index Terms: far-field speaker identification, gaussian mixture models, principal component analysis, microphone arrays.

1. Introduction

Person identification is of paramount importance in security, surveillance, human-computer interfaces and smart spaces. Hence, the evaluation of different recognition algorithms under common evaluation methodologies is very important. Even though the applications of person recognition vary, the evaluations have mostly focused on the security scenario, where training data are few but recorded under close-field conditions. An example of this for faces is the Face Recognition Grand Challenge [1], where facial images are of high resolution (about 250 pixels distance between the centers of the eyes).

The CLEAR person identification evaluations, following the Run-1 evaluations [2] of the CHIL project [3], focus on the surveillance and smart spaces applications, where training can be abundant, but on the other hand the recording conditions are far-field: wall-mounted microphone arrays record speech and cameras mounted on room corners record faces. These two modalities are used, either stand-alone or combined, to recognize people in audiovisual streams. The person identification system implemented in Athens Information Technology operates on short sequences of the two modalities of the far-field data, producing unimodal identities and confidences. The identities produced by the unimodal subsystems are then fused into a bimodal one by the audiovisual subsystem.

This paper discusses our audio-based person identification subsystem and is organized as follows: section 2 details the standard MFCC-GMM approach that forms the basis of our algorithm, while section 3 presents two modifications to this basic scheme that have enhanced its performance, in terms of both speed and recognition accuracy. The CLEAR evaluation protocol and our results are presented in section 4, followed by a discussion of several multi-sensor fusion strategies to further improve performance in section 5. Finally, in section 6 the conclusions are drawn.

2. Speaker identification using GMMs

In the training phase of our system the goal is to create a model for each one of the supported speakers and ensure that these models accentuate the specific speech characteristics of each person. To this end, we first break up the training segments into frames of appropriate size (i.e. duration), with successive frames having a predefined overlap percentage. The samples belonging to each frame are used to calculate a vector of parameters that represents the given frame during the model estimation process. Specifically, a set of Mel Frequency Cepstral Coefficients (MFCC) are extracted from each frame and they are used to model the characteristics and structure of each individual's vocal tract. All MFCC vectors for a given person are collected and used to train a Gaussian Mixture Model (GMM), based on the Baum-Welch algorithm [4]. A GMM is in essence a linear combination of multi-variant Gaussians that approximates the probability density function (PDF) of the MFCC for the given speaker

$$\lambda_k = \sum_{m=1}^{M} w_m N\left(O, \mu_m, \Sigma_m\right) \tag{1}$$

where λ_k is the GMM for the k^{th} speaker and O is the set of training vectors used to estimate it. This model is characterized by the number of Gaussians M that make up the mixture, each having its own weight w_m, mean vector μ_m and covariance matrix Σ_m.

For the identification part, testing samples are again segmented into frames with the same characteristics as the ones created during the training process, and we subsequently extract MFCC's from each frame. To perform identification, each of the GMM's is fed with an array of the coefficients (one row per sample), based on which we calculate the log-likelihood that this set of observations was produced by the given model. The model that produces the highest log-likelihood is the most probable speaker according to the system

$$s = \arg\max_k \left\{L\left(O|\lambda_k\right)\right\} \tag{2}$$

where O is the matrix of observations (MFCC's) for this testing segment and $L\left(O|\lambda_k\right)$ is the log-likelihood that each model λ_k produces this set of observations.

All samples are broken up in frames of length 1024 with 75% overlap. The size of the GMM is fixed at 16 Gaussians and the number of static MFCC's per frame has been set to 12. To this we

concatenate the log-energy of the frame to create 13D vectors, and we also append the delta (first-order derivative) coefficients.

3. Modifications to the basic algorithm

In this section we describe two significant modifications we have applied to the standard MFCC-GMM framework and detail their effect on the system's performance.

3.1. PCA pre-processing of feature vectors

Automatic identification systems are evaluated based on their response time and error rate. It is obviously important to minimize both these numbers, however in many cases it is not easy or even possible to do that and we must settle for a trade-off between speed and identification accuracy. We have addressed this issue by employing the standard Principal Components Analysis (PCA) [5] as a pre-processing step. Specifically, we compute a transformation (projection matrix) for each speaker based on their training data and use that matrix to perform a mapping to the PCA coordinate system prior to GMM calculation. In the testing phase, we compute the log-likelihood of each speaker by first projecting the MFCC vectors to the respective PCA space. Such an approach has been previously reported in [6].

The use of PCA introduces one further degree of freedom in the system, namely the dimensionality of the projection space. It is obvious that by keeping an increasingly smaller number of eigenvalues from the PCA scatter matrix we can reduce this dimensionality accordingly, therefore achieving a significant execution speed increase. The choice of the number of discarded eigenvalues will be ultimately dictated by the truncation error introduced due to the reduction of the projection space dimension. Specifically, if the initial space dimension is d and we discard the q smallest eigenvalues, the truncation error will be equal to

$$e = 1 - \frac{\sum_{i=d-q+1}^{d} \lambda_i}{\sum_{j=1}^{d} \lambda_j} \quad (3)$$

where λ_i is the i^{th} largest eigenvalue.

We have implemented an automatic decision process that determines the number of retained eigenvalues in a way that ensures that the average truncation error across all speakers is no more than 0.2%. The maximum value of q that satisfies this condition is chosen, so that we achieve the greatest speed increase possible while retaining (mostly) optimal identification accuracies. Our experiments indicate that this selection strategy gives a value for q that is at most one above or below the number of eigenvalues that minimizes the error rates. Even if our choice of q leads to slightly sub-optimal solutions, the achieved error rates are still superior to using the standard GMM algorithm approach without PCA pre-processing. We have therefore achieved faster response times as well as enhanced identification performance. Section 4 reports the performance of our system on the CLEAR 2006 evaluation dataset and contrasts it with that of the standard MFCC-GMM classifier.

The efficiency of this selection strategy is illustrated in Table 1 below, based on experiments that have been performed on the CLEAR 2006 evaluation data set. For the 15 sec training case, we can see that although the selected value of 18 eigenvectors is sub-optimal, it is nonetheless very close to the optimal choice in terms of error rate (which is to retain 16 eigenvectors). For the 30 sec training case, our selection strategy clearly yields the best result.

Table 1: *Choosing the optimal value of retained eigenvectors based on the truncation error.*

15 sec training					
# Retained EV	26	22	20	18	16
% Info Retained	100	99.963	99.902	99.804	99.661
1 sec testing	26.43	27.41	27.57	26.92	27.08
5 sec testing	9.49	9.00	9.25	9.73	8.27
10 sec testing	8.65	9.34	7.96	7.96	7.61
20 sec testing	6.74	6.74	5.06	4.49	4.49
30 sec training					
# Retained EV	26	22	20	18	16
% Info Retained	100	99.962	99.9	99.802	99.659
1 sec testing	17.94	18.27	16.64	15.17	17.29
5 sec testing	2.68	3.41	3.41	2.68	2.43
10 sec testing	2.08	2.08	3.11	1.73	2.42
20 sec testing	0.56	1.12	1.12	0.56	0.56

3.2. Deterministic initialization of the EM algorithm

A very crucial step for the creation of a successful GMM is the initialization of its parameters, which will be updated during the iterations of the EM training algorithm. The standard approach is to use the K-Means clustering algorithm to obtain some initial estimates for the Gaussian parameters; this strategy however suffers from the random characteristics of the outcome of K-Means, which in turn lead to a different GMM each time the same data are used for training. Moreover, the identification performance varies considerably across these different models. We have therefore utilized a deterministic initialization strategy for the EM algorithm, based on the statistics of the training data. Specifically, we compute a number of percentiles across all dimensions of the training data set and thus partition the data range in each dimension into as many subsets as the modes of the GMM. The K-Means algorithm is consequently run using the central values of each subset as initial cluster means, and the resulting clustered data are fed into the EM algorithm for parameter fine-tuning.

Our experiments have shown that this strategy gives on average lower error rates than the random K-Means initialization, although there are a few runs using the standard approach that lead to better identification performance. This is illustrated in Table 2, where the reported error rates have been obtained from experiments on the CHIL Run 1 data [2].

Table 2: *Comparison of identification error rates using random K-Means initialization and the proposed deterministic strategy.*

Random initialization statistics (over 30 runs)	
Mean (%)	28.36
Standard deviation (%)	0.75
Best run (%)	27.27
Worst run (%)	30.38
Our strategy (%)	27.75

4. Results in the CLEAR 2006 evaluation

Our speaker identification system has been tested on the CLEAR 2006 data that comprise of speech samples from 26 individuals.

The audio conditions are far-field, in the sense that speech is recorded by wall-mounted microphone arrays. Two training conditions have been defined, with training segments 15 seconds long for the former and 30 seconds long for the latter. Four testing durations are also defined: 1, 5, 10 and 20 seconds long. All these segments contain mostly speech, so a speech activity detection algorithm [7] has not been used. The results of our system in terms of error rate are shown in the rightmost column of Table 3 per training and testing duration. The middle column contrasts the performance of our system with that of a standard GMM classifier of the same complexity. It is clear that the PCA pre-processing step reduces error rates significantly, especially when the training and testing segments have shorter durations. Furthermore, the average identification time is reduced by 39.2% after the PCA preprocessing step is applied.

Table 3: *Performance of our system on the CLEAR 2006 evaluation database.*

15 sec training		
Test duration	GMM	PCA-GMM
1 sec	36.22	26.92
5 sec	11.92	9.73
10 sec	11.08	7.96
20 sec	6.18	4.49
30 sec training		
Test duration	GMM	PCA-GMM
1 sec	22.35	15.17
5 sec	5.11	2.68
10 sec	3.81	1.73
20 sec	0.56	0.56

5. Decision fusion from multiple microphones

In this section we investigate the possible gains from utilizing information from more than one far-field microphones to reach our final decision as to the speaker's identity. The audio sensor setup of all rooms where seminars were recorded in the scope of the CLEAR 2006 evaluations includes at least one NIST MarkIII linear microphone array comprising 64 microphones. The results discussed in the previous section are based on recordings from only the first (leftmost) microphone of those arrays; we have proceeded to repeat the same experiments with different sets of microphones and detail our findings in this section.

5.1. Effect of choosing a different single microphone

We have studied the error rates of our system when both the training and the testing segments were obtained from a different microphone of the array and provide a summary of the results in Table 4. The test microphones have the following mapping to the actual sensors of the array : microphones 1 and 4 are the leftmost and rightmost respectively, microphones 5 and 6 are the two centermost and finally microphones 2 and 3 are located at 1/3 and 2/3 of the array length, respectively.

It is obvious from this table that the error rates vary quite significantly as we move along the length of the microphone array. This variation is especially noticeable as both the training and testing segments increase in duration, as indicated by the computed standard deviation to mean ratios. This large fluctuation arises from the fact that the speaker is not always facing the same microphone throughout the duration of a segment. Furthermore, segments of large duration are usually created as a concatenation of subsegments of contiguous speech without ensuring that the speaker is always at the same position in the room. These results will be reflected in the error rates obtained from all fusion strategies described in the following subsections.

Table 4: *Error rate (%) variation across different single microphones.*

15 sec training				
Test duration (sec)				
Microphone	1	5	10	20
1	26.92	9.73	7.96	4.49
2	25.61	8.76	7.61	3.93
3	26.59	10.95	10.38	6.18
4	24.63	8.27	7.61	5.62
5	26.59	7.06	6.23	3.93
6	26.26	8.52	7.96	6.74
Mean	26.10	8.88	7.96	5.15
Standard deviation	0.85	1.33	1.35	1.20
Standard deviation / Mean	3.24	14.97	16.95	23.34
30 sec training				
Test duration (sec)				
Microphone	1	5	10	20
1	15.17	2.68	1.73	0.56
2	16.97	2.68	2.77	1.68
3	15.33	3.16	2.08	1.12
4	17.46	2.43	2.08	1.12
5	18.27	3.89	2.42	1.12
6	17.94	3.16	3.11	1.68
Mean	16.86	2.82	2.37	1.21
Standard deviation	1.32	0.32	0.51	0.42
Standard deviation / Mean	7.84	11.52	21.46	34.74

5.2. Concatenation of feature vectors from multiple microphones

The simplest fusion strategy is to concatenate the feature vectors obtained from each member of the set of test microphones and evaluate the performance of the system on the whole observation matrix. Table 5 lists the results of applying this procedure on selected sets of 2 and 4 microphones as well as on the full 6 microphone set, where members of each sensor pair lie in symmetric positions with respect to the center of the array.

5.3. Multi-microphone voting on a per segment basis

In this case we process the feature vectors of a segment from each microphone separately and reach as many decisions as the size of the microphone set. The final decision of the system is the most likely speaker according to the majority of the sensors; in case of a tie between two speakers, the one with the highest total log likelihood across all microphones that support him is declared the winner. Table 6 lists the error rates when following this approach for the same sensor sets as in the previous subsection.

Table 5: *Fused error rate (%) after feature vector concatenation.*

15 sec training				
Test duration (sec)				
Microphone Set	1	5	10	20
[23]	23.65	7.79	5.54	3.93
[2356]	25.29	6.57	5.19	2.81
[123456]	24.8	6.33	6.23	2.25
30 sec training				
Test duration (sec)				
Microphone set	1	5	10	20
[23]	15.17	2.43	2.08	0.56
[2356]	15.33	2.43	2.42	0
[123456]	16.64	2.68	2.08	0.56

Table 6: *Fused error rate (%) after per segment voting.*

15 sec training				
Test duration (sec)				
Microphone Set	1	5	10	20
[23]	24.47	7.3	6.23	3.37
[2356]	25.29	5.6	4.5	2.25
[123456]	24.31	6.33	5.54	1.69
30 sec training				
Test duration (sec)				
Microphone set	1	5	10	20
[23]	16.15	2.92	2.08	0
[2356]	15.01	2.19	2.42	0
[123456]	16.15	2.43	2.08	0.56

5.4. Multi-microphone voting on a per frame basis

In this case we compare each feature vector across all microphones and select the microphone that gives the highest log likelihood for each speaker in the given frame. This selection process results in a feature vector set which we process similarly as in the case of a single microphone; the difference is that per-frame log likelihoods for different speakers are not necessarily derived based on the same microphone. Table 7 lists the error rates when following this approach for the same sensor sets as in the previous subsection.

5.5. Comments on the employed fusion strategies

Clearly, the use of multiple microphones greatly reduces the error rates of a speaker identification system. For the case of 15 second long training segments, all three fusion strategies show very promising results, especially the per segment voting scheme. This is not the case for the 30 second training segments, where an improvement is achieved in only a few cases (with respect to the results reported in the CLEAR 2006 evaluations), while there are also instances of performance degradation. This is due to the fact that longer segments are more likely to contain speech coming from different locations inside a room. Hence the probability of some microphones being "bad" choices is minimal; all the microphones are suitable and their combination does not attenuate the effect of some bad choice anymore.

Table 7: *Fused error rate (%) after per frame voting.*

15 sec training				
Test duration (sec)				
Microphone Set	1	5	10	20
[23]	24.47	7.06	4.84	3.93
[2356]	25.45	6.57	4.15	2.81
[123456]	25.29	6.57	5.88	2.81
30 sec training				
Test duration (sec)				
Microphone set	1	5	10	20
[23]	15.33	3.41	2.42	0
[2356]	16.48	2.92	2.77	0.56
[123456]	18.92	3.16	2.42	0

6. Conclusions

We have presented a complete and automatic GMM-based speaker identification system, enhanced with per-speaker PCA pre-processing of feature vectors and a deterministic EM algorithm initialization strategy to reduce error rates and boost execution speeds. After reporting the performance of our system in the recent CLEAR 2006 evaluations using only a single microphone, we have proceeded to describe three simple fusion strategies employing multiple far-field microphones and demonstrate their potential to further increase successful identification rates.

7. Acknowledgements

This work is sponsored by the European Union under the integrated project CHIL, contract number 506909. The authors wish to thank the organizers of the CLEAR evaluations.

8. References

[1] P. Phillips et al., "Overview of the Face Recognition Grand Challenge", CVPR, 2005.

[2] H. Ekenel and A. Pnevmatikakis, "Video-Based Face Recognition Evaluation in the CHIL Project - Run 1", Face and Gesture Recognition 2006, Apr. 2006.

[3] A. Waibel, H. Steusloff, R. Stiefelhagen, et. al, "CHIL: Computers in the Human Interaction Loop", 5th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS), Lisbon, Portugal, April 2004.

[4] D. A. Reynolds and R. C. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models", IEEE Transactions on Speech and Audio Processing, Vol. 3, No. 1, pp. 72-83, January 1995.

[5] M. Turk and A. Pentland, "Eigenfaces for Recognition", J. Cognitive Neuroscience, 3, pp. 71-86, March 1991.

[6] Z. Wanfeng et al., "Experimental Evaluation of a New Speaker Identification Framework using PCA", IEEE International Conference on Systems, Man and Cybernetics, Vol. 5, pp. 4147-4152, October 2003.

[7] J. Sohn, N. S. Kim and W. Sung, "A Statistical Model Based Voice Activity Detection", IEEE Sig. Proc. Letters, Vol. 6, No. 1, Jan. 1999.