A Cohort - UBM Approach to Mitigate Data Sparseness for In-set/Out-of-set Speaker Recognition

Vinod Prakash, John H.L. Hansen

Center for Robust Speech Systems (CRSS)
Erik Jonsson School of Engineering & Computer Science
University of Texas at Dallas
Richardson, Texas 75083-0688, U.S.A.
vinod.prakash@colorado.edu, john.hansen@utdallas.edu

Abstract

In this study, the problem of identifying in-set versus out-of-set speakers is addressed. Here the emphasis is on low enrollment and test data durations, in a text-independent mode. In order to compensate for the limited enrollment data (5 sec), a method is proposed that utilizes data from speakers that are acoustically close to a particular in-set speaker. A speaker specific model is obtained by adaptation of a base model that is built using data from such speakers. The performance of the proposed algorithm is evaluated using the TIMIT database with an adapted GMM classifier (GMM-UBM) employed as the baseline system. Experimental results show a consistent increase in system performance, with a relative improvement ranging from 10.57-58.33% depending on in-set speaker size and test data duration.

Index Terms: in-set/out-of-set speaker recognition, cohort speakers, data sparseness.

1. Introduction

For in-set/out-of-set speaker recognition problems, the enrollment data for classifier model development is obtained from each individual in a group of speakers, referred to as the in-set group. When given an unknown test utterance, the recognizer is required to produce a binary decision as to whether the test utterance came from a speaker belonging to the in-set group or not. This problem is a simplification of the open-set case where the recognizer is required to identify a specific speaker within the in-set group or declare that the test utterance is from an out-of-set (i.e., unknown) speaker; (In the context of speaker verification an out-of-set speaker is referred to as an imposter). Previous studies in this area have focused on using discriminative training [11], clustering [12] or neighborhood information [13] among the in-set speakers.

When statistical models are used to construct representations for the speakers, it has been observed that the raw likelihood scores are not very reliable during the decision process [3, 4, 8]. If the decision rule in the recognizer depends not only on the individual speaker models, but also on a model for the out-of-set speakers, this is referred to as score normalization [5]. Two main approaches have emerged to model the out-of-set speakers [2], the *world model* and *cohort model* based schemes.

The *world model* (background model, universal background model (UBM)) is constructed by pooling data from a large number of potential out-of-set speakers. Alternatively, in the *cohort model* approach, a set of potential imposters is created for each enrolled speaker and score normalization is performed using a statistic of the likelihood scores of these imposters. Score normalization experiments in GMM based open-set recognition are described in [6, 7]. A cohort model normalization scheme using a pooled cohort has been described in [9].

In this study, we focus on the problem of in-set speaker recognition with low enrollment (5 sec) and test material (2-8 sec), with in-set group sizes ranging from 9 - 45 speakers. When the enrollment data is this low, it is expected that the phoneme coverage for a speaker will be incomplete, resulting in “acoustic holes” in the speaker’s model space. When a test token for an enrolled speaker contains phonetic content that was not seen during training, it results in a low likelihood score, and possibly a wrong decision for that speaker. (i.e., phonemes seen in the test set, but not in the training data for the same speaker,cause the speaker model to be rejected). The adapted GMM approach [1] alleviates this problem to a certain extent. Here the speaker models are derived by adaptation from a world model. Unseen test data has a similar impact on both the speaker and world models resulting in cancellation of the influence of such data. Viewed in a negative light, this setup discards information from unseen acoustic data for a speaker. Also, since there is very little data available for speaker model adaptation, scores of test tokens from imposters who are acoustically very distinct from the enrolled speakers will be comparable to that of the background model, and such imposters are not decisively rejected.

Most studies so far have used an enrolled speaker’s cohorts to model potential imposters. In this paper, we propose to use the Cohorts to fill the acoustic holes in an in-set speaker’s training space. An example of work along these lines for Speaker Identification is given in [10]. There, information from the cohorts is utilized by merging cohort models, while in our work we utilize information from the cohorts at the feature level.

The remainder of this paper is structured as follows, the next section covers the objective formulation of the problem and contains a brief overview of the baseline system. Sec. 3 contains details about the proposed algorithm. Experimental results are provided in Sec. 4. Discussion and Analysis of the results is done in Sec. 5, with overall conclusions in Sec. 6.

This work was funded by grants from the U.S. Air Force Research Laboratory, Rome NY, under contract F30602-03-0110, and by the University of Texas at Dallas under project EMMIT.

2. Objective Formulation

We assume we are given a set of N in-set (enrolled) speakers in a system, and the collected data $\vec{X}_n$, corresponding to each enrolled speaker S_n, $1 \leq n \leq N$. Let the data $\vec{X}_0$ represent all other non-enrolled speakers in the development set. Each speaker dependent statistical model $\{\Lambda_n \in \mathbf{\Lambda}, 1 \leq n \leq N\}$ can be obtained from $X_n = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_{t_n}\}$ where t_n denotes the total number of samples belonging to speaker S_n.

If $\vec{O}$ denotes the sequence of feature vectors extracted from the test utterance, then the problem of identifying an in-set versus out-of-set speaker requires that we perform two statistical stages. In the first stage, called *speaker identification* or *speaker classification*, we first classify $\vec{O}$ into one of the most likely in-set speakers, Λ^*, e.g.,

$$\Lambda^* = arg \underset{1 \leq n \leq N}{max} \; p(\vec{O}|\Lambda_n). \tag{1}$$

In the second stage, called *speaker verification* or *outlier verification*, we verify whether the observation $\vec{O}$ truly belongs to Λ^* or not (accept/reject). In general, this stage is formulated as a problem of statistical hypothesis testing where the *null* hypothesis H_0, represents the hypothesis that $\vec{O}$ really belongs to model Λ^*, against the competitive hypothesis H_1, that represents the hypothesis where $\vec{O}$ is actually "not" from model Λ^*. The likelihood ratio test is given by:

$$\frac{p(\vec{O}|\Lambda^*)}{p(\vec{O}|\Lambda_0)} \begin{cases} \geq \gamma & \text{: accept } H_0, \\ < \gamma & \text{: reject } H_0 \text{ (accept } H_1\text{).} \end{cases} \tag{2}$$

where γ is a threshold, Λ_0 is a competitive model (*out-of-set model*), and $p(\cdot|\cdot)$ is the likelihood generated from each model.

2.1. Baseline GMM-UBM System

In practice, it is impossible to have a true out-of-set model for the competitive speaker class, otherwise we could define such a speaker model as one class in the training phase. The conventional strategy assumes another special class, or speaker independent model, as a universal representative of the out-of-set speakers (e.g., UBM). Gaussian Mixture Model (GMM) with Maximum A Posteriori (MAP) speaker adaptation has become the dominant approach in text-independent speaker recognition [1]. A speaker independent model, or Universal Background Model (UBM), is trained from the development speaker set by the Expectation Maximization (EM) algorithm. The probability density function (*pdf*) of a GMM having M-Gaussian components for D-dimensional observation vectors $\vec{X}$ is defined as:

$$p(\vec{X}|\Lambda_0) = \sum_{m=1}^{M} \omega_{0m} G_{0m}(\vec{X}), \tag{3}$$

where ω_{0m} is the weight of the m-th component, and G_{0m} is the Gaussian probability density function with mean μ_{0m} and covariance matrix Σ_{0m}, which is assumed diagonal. For each target in-set speaker, a speaker dependent GMM ($\Lambda_n : \{\hat{\omega}_n, \hat{\mu}_n, \hat{\Sigma}_n\}$) can be created by MAP adaptation of the UBM parameters $\{\omega_0, \mu_0, \Sigma_0\}$ with training data $\vec{X}_n$ via the following formula [1]:

$$\hat{\mu}_{nm} = \frac{\eta_m}{\eta_m + r} E_m(\vec{X}_n) + \frac{r}{\eta_m + r} \mu_{0m}, \tag{4}$$

where η_m is the weight assigned to the m-th component in the UBM, and r is a relevance factor which depends on the parameter and controls the balance of adaptation. For all experiments in this study, only the component means were adapted and the relevance factor was fixed at 16.

3. Proposed Algorithm

3.1. Motivation/Idea

Since a UBM is constructed by pooling data from several speakers; to accurately model data from this pooled set of speakers a huge GMM usually containing 100's of mixture components needs to be constructed. Alternatively, if individual speaker models are built using data from only that speaker, these models usually contain 32 Gaussian components, with an intuitive understanding that one component would be used to cover approximately each phoneme. We note that some low energy phonemes are typically discarded due to the prescence of a speech-silence detector. The central idea behind the proposed scheme is that, if, for each in-set speaker a model is built by pooling data from "acoustically close" speakers, that is from that in-set speaker's cohort set, then if this model were to be MAP adapted using the limited enrollment data, the resulting GMM (which will have no more than 64 components) should be far more representative of the speaker than MAP adaptation from a general UBM (which would have a much larger number of pdf components). In addition, if more development data is available for speakers in the cohort set, we are more likely to be able to fill in acoustic holes in the training space when only 5 sec of data is available for the in-set speaker.

3.2. Steps

The procedure followed to construct a speaker model for in-set speaker n, $1 \leq n \leq N$, is:

- For each development speaker i, Construct a GMM(Λ_i^{dev}) using the training data for that development speaker. $1 \leq i \leq N_{dev}$.
- Score each of the above models using the training data $\vec{X}_n$ for the inset speaker:

$$S_i = p(\vec{X}_n|\Lambda_i^{dev}), 1 \leq i \leq N_{dev} \tag{5}$$

- Sort the scores S_i and pick the top N_{co} speakers corresponding to the top N_{co} scoring models. N_{co} ($\ll N_{dev}$) is the number of cohorts that are used to fill the acoustic holes for in-set speaker n. These speakers form the cohort set Ω_n^{cohort} for this in-set speaker.
- Pool together the data of the selected cohorts and construct a cohort GMM for Λ_n^{cohort} for in-set speaker n.
- Using Λ_n^{cohort} as an initial model in Eq. (4) obtain the in-set speaker model Λ_n.

If Ω_{dev} is the set of all development speakers, construct the set:

$$\Omega_{out} = \Omega_{dev} - \bigcup_{1 \leq n \leq N} \Omega_n^{cohort} \tag{6}$$

Data from a randomly chosen subset of speakers from Ω_{out} is pooled to construct a model for the out-of-set speakers. (This model is used for score normalization in both the Baseline and Proposed systems.)

3.3. Expected Performance

Borrowing terminology from [14], Table 1 summarizes (heuristically) the expected behavior of the Baseline (a conventional GMM-UBM scheme) and the proposed cohort model schemes for different test scenarios. The in-set speaker is taken to be a "sheep", a default speaker type who dominates the population and for whom systems perform nominally well. The out-of-set speakers are taken to be either "sheep" or "wolves" (speakers who are particularly successful at imitating other speakers).

Speaker	Phn Overlap	System-0	System-1	System-2
In-S	Yes	Accept	Accept	Accept
In-S	No	Accept	X	Accept
Out-S	Yes	Reject	Reject	Reject
Out-S	No	Reject	X	Reject
Out-W	Yes	Reject	Accept	Accept
Out-W	No	Reject	X	Accept

Table 1: *Expected Decisions of Ideal(System-0), Baseline(System-1) and Proposed (System-2) systems for different kinds of out-of-set speakers and differing overlap between train and test phones. S:sheep, W:wolf, X:Unknown*

From Table 1, The baseline and the proposed systems are expected to perform comparably if the out-of-set speakers are sheeps and there is overlap between the train and test phonemes. When there is no overlap, the behavior of the baseline system is unpredicatable(*'X'* in the table), i.e., neither the accept nor the reject hypothesis is consistently favoured. Assuming that the proposed algorithm successfully fills the acoustic holes in the training space, under the no-overlap condition, correct(incorrect) results are expected when the out-of-set speakers are sheeps(wolves).

4. Experiments

4.1. Experimental Setup

4.1.1. TIMIT

A set of 60 male speakers was randomly selected as a speaker sample space. These 60 speakers serve both as in-set speakers and out-of-set speakers (imposters) depending on the experimental set. In particular, three different sizes of in-set speakers are considered (e.g., 15, 30, and 45). For example, 15 speakers were randomly selected from the speaker sample space as the in-set speakers, with the remaining 45 speakers taking the role of imposters ('15in/45out'). Similar to other Round-Robin test procedures, different combinations of in-set and out-of-set speakers were also selected, resulting in four distinct '15in/45out' groups, two distinct '30in/30out' groups, and two (with some overlap) '45in/15out' groups. The training and test speech data of each speaker were randomly selected and concatenated from the original TIMIT database, with no data overlap and initial and trailing silence removed. The training data was limited to approximately 5 seconds worth of speech, while test data was created for 2, 4, 6, and 8 seconds worth of speech. The remaining 378 male speakers, each having about 30 secs of data, were used as development data.

4.2. Front-end Processing

The speech analysis frame rate is set to 30 ms with a 10 ms skip rate. Speech is pre-emphasized with the filter $(1 - 0.95z^{-1})$. Nineteen-dimensional Mel-Frequency Cepstral Coefficients (MFCC) are extracted and used for statistical modeling. Silence and low-energy speech parts are removed using an energy based detection technique (e.g., frames that have higher energy than the pre-defined threshold are selected).

4.3. Evaluations

A UBM containing 32 Gaussian components is constructed by randomly selecting 60 speakers from the development set. The GMM construction starts using VQ codebooks with several updated iterations, and the GMM parameters are consequently adjusted with EM iterations. All speaker models used in our experiment have 32 Gaussian components. (For the given experiment we found that choosing 32 Gaussian components gave the highest peformance for the Baseline system). This UBM is used to model the out-of-set speakers for both the Baseline and Proposed systems.

The speaker models for the Baseline system are obtained by MAP adaptation using eq. (4) of the above UBM. Cohort sets for the in-set speakers are selected from the remaining 318 (378-60) male speakers. After experimenting with a number of cohort set sizes, a cohort set size (N_{co}) of 10 was fixed for all in-set speakers. The steps given in 3.2 were carried out to construct the speaker models for the proposed algorithm. Figure 1 shows the obtained EER for the baseline and proposed systems for three different in-set/out-of-set configurations of 15in/45out, 30in/30out and 45in/15out.

From Figure 1, consistent and in some cases, esp. for the 6/8 sec test condition, dramatic improvement in performance is observed. It is clear that with 5 secs of training data, acoustic holes in the speaker production space will be present. This has been observed in earlier studies [11, 12, 13]. The proposed algorithm provides measurable improvements for 2 secs of test data, we see that the resulting EER's have decreased (relative improvement between 10.96 and 18.18%). For the 4 secs case the improvement is more pronounced (26.73 - 38.23%). The EER improvements for the 6 and 8 sec case are quite impressive.(e.g., absolute improvement between 1.94 and 5 %), the corresponding relative improvement in EER is in the range 30.43 - 58.33%. Since the order of improvement for the 2 sec case is quite different from the 4 - 8 case, this suggests that for a test size of 2 secs the classifier structure should be different than for the 4 - 8 test sizes.

5. Discussion

Some observations and directions for future work are:

- Currently the size of the cohort set is fixed (at 10) for all in-set speakers. Performance should be better if this size were chosen separately for each speaker depending on, for example the cross-verification scores eq. S_n (5). This would ensure a more consistent closeness of each cohort speaker selected for the present in-set speaker.
- This is a limited study in the sense that only corpora recorded under clean conditions have been used. The effectiveness of the proposed method should be assessed on noisy corpora under varying channel mismatch conditions.
- The performance of the method using various score normalization criteria [6, 11] could also be investigated. (This becomes important in case the size of the development set is not large enough, and hence it may not be possible to perform score normalization using a world model.)

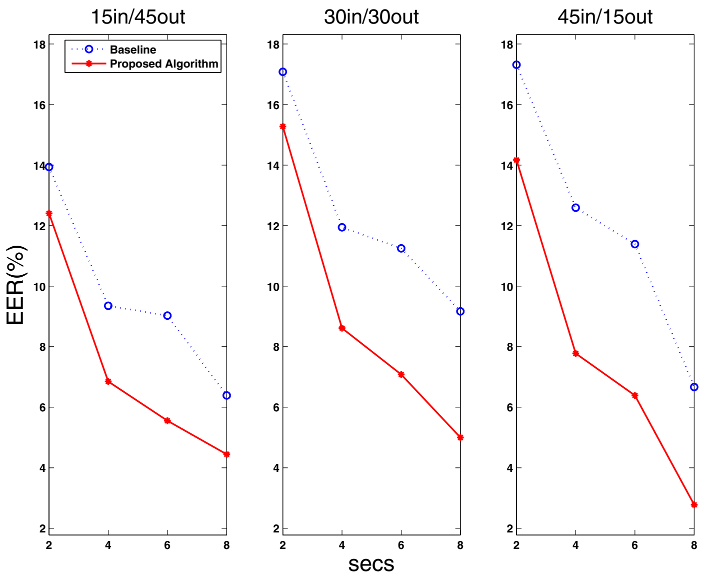

Figure 1: *Performance (in terms of EER(%)) of baseline and proposed algorithm on TIMIT, using in-set/out-of-set speaker sizes of 15/45, 30/30 and 45/15.*

6. Conclusions and Future Work

In this paper, we have studied the problem of identifying in-set versus out-of-set speakers under low train-test conditions. We proposed an algorithm that uses an in-set speaker cohort set to make up for the sparse (e.g., 5 sec per speaker) enrollment data. Investigations on a clean speech database show consistent improvement for the proposed method over a GMM-UBM baseline.

Future work will primarily focus on choice and composition of the cohort-set and the use of score normalization techniques (other than world-model based).

7. Acknowledgements

We would like to thank Dr. Pongtep Angkititrakul of CRSS at the University of Texas at Dallas for his insightful comments during preparation of this manuscript.

8. References

[1] Reynolds, D. A., Quatieri, T. F., and Dunn, R. B., “Speaker verification using adapted Gaussian Mixture Models”, Digital Signal Processing, vol. 10, 2000.

[2] Auckenthaler, R., Carey, M. and Lloyd-Thomas, H., “Score Normalization for Text-Independent Speaker Verification Systems”, Digital Signal Processing, vol. 10, 2000.

[3] Higgins, A., Bahler, L. and Porter, J. O., “Speaker verification using randomized phrase prompting”, Digital Signal Processing, vol. 1, 1991.

[4] Rosenberg, A. E., DeLong, J., Lee, C.-H., Juang, B.-H. and Soong, F. K., “The use of cohort normalized scores for speaker verification”, Proc. ICSLP 92, vol. 2, 1992.

[5] Liu, W., Isobe, T. and Mukawa, N. “On optimum normalization method used for speaker verification”, Proc. ICSLP 98, 1998.

[6] Sivakumaran, P., Fortuna, J., and Ariyaeeinia, A., “Score normalisation applied to open-set, text-independent speaker identification”, Proc. Eurospeech’03, 2003.

[7] Fortuna, J., Sivakumaran, P., Ariyaeeinia, A. and Malegaonkar, A., “Open-set speaker Identification using adapted Gaussian mixture models”, Proc. Interspeech’05, 2005.

[8] Reynolds, D. A., “Comparison of Background Normalization methods for text-independent speaker verification”, Proc. Eurospeech’97, 1997.

[9] Rosenberg, A. E. and Parthasarathy, S., “Speaker background models for connected digit password speaker verification”, Proc. ICASSP 96, vol. 1, 1996.

[10] Charlet, D., “Neighborhood-adapted GMM for speaker recognition”, Proc. Odyssey 2004 Speaker and Language Recognition Workshop, 2004.

[11] Angkititrakul, P. and Hansen, J.H.L., “Discriminative in-set/out-of-set speaker recognition”, IEEE Trans. Speech & Audio Processing, (accepted, to appear 2007).

[12] Angkititrakul, P., Hansen, J.H.L. and Bagahaii, S., “Cluster-dependent modeling and confidence measure processing for in-set/out-of-Set speaker identification”, Proc. ICSLP 04, 2004.

[13] Angkititrakul, P. and Hansen, J.H.L., “Identifying in-set and out-of-set speakers using neighborhood information”, ICASSP-04, Vol. 1, 2004.

[14] Doddington, G., Liggett, W., Martin, A., Przybocki, M. and Reynolds, D. A., “Sheep, Goats, Lambs and Wolves: a statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation”, Proc. ICSLP-1998, paper 0608, 1998.