TEST TOKEN DRIVEN ACOUSTIC BALANCING FOR SPARSE ENROLLMENT DATA IN COHORT GMM SPEAKER RECOGNITION

Jun-Won Suh and John H. L. Hansen

Center for Robust Speech Systems (CRSS)
Erik Jonsson School of Engineering and Computer Science
University of Texas at Dallas, Richardson, Texas, USA
email:jxs064200@utdallas.edu, John.Hansen@utdallas.edu

ABSTRACT

In this study, we address the problem of sparse train/test data for in-set/out-of-set speaker recognition. Sparse enrollment data presents a unique challenge due to a lack of acoustic space coverage. The proposed algorithm focuses on filling acoustic holes and fortifying the acoustic information using the claimed speaker's test token histogram. This scheme is possible by using a GMM model to classify the speaker phone information at the feature level. Parallel GMM training with EM using the most occurring (top) and least occurring (bottom) acoustic feature is called "Top-Down Bottom-Up (TDBU)", and the method employing the acoustic token histogram of test token using the TDBU is called "TDBU using Test Token Histogram (TTH)". Since TTH provides test data histogram information, the most occurred (top) parts in test data fortify the its discriminating ability using same acoustic tokens in enrollment data. The less occurred (bottom) part in test data provide acoustic hole information so that the mismatched acoustic hole between enrollment and test data can be filled in chance. The TDBU-TTH method is evaluated using telephone conversation speech from the FISHER corpus with 5 second train sets. The TDBU-TTH improves on average 2.17% absolute EER over the TDBU, and an average 4.03% absolute EER improvement over GMM-UBM baseline using 2 second test data. The proposed algorithm improvement is a noteworthy stage to compensate for both sparse enrollment data and limited test data.

1. INTRODUCTION

In many scenarios, effective speaker recognition is necessary with short enrollment utterances (5 second) and/or short test utterances (2∼6 second). System performance degrades dramatically with such short enrollment/test data. In this study, we focus on how to fill the acoustic sparseness using a formulated acoustic token histogram information of the test data and an acoustically close speakers' data. The sparse enrollment data results in a unique challenge due to a lack of acoustic phone coverage in the speaker space compared with longer conversational speech data. Therefore, it is highly probable that phoneme mismatch exists between the limited trained acoustic space and input test sequence. We called this phenomenon "acoustic holes" in the acoustic model space. This work has focused on filling acoustic holes using sparse train/test sets.

In-set/out-of-set problem consists of two main parts, closed-set speaker identification and open-set speaker verification. Speaker recognition can be applied in these search applications by classifying the target group of speakers (in-set) and the other non-interest speaker group (out-of-set)[3]. Audio search and speaker diarization is also useful application of in-set/out-of-set for searching target speakers in famous speeches or news audio streams. The extended application can be employed for identifying speakers in a multi-speaker conversation, or for a system that grants security access for a specific group in organizations.

Acoustic tokens can be transcribed along a temporal data stream using a Speaker Independent GMM (S.I.GMM). The GMM represents the most common characteristics of the available speaker data[1, 2]. The goal here is to balance the acoustic distribution of enrollment/test data for each in-set speaker *without* knowledge of the input phoneme sequence. The transcribed speaker's acoustic phoneme-like segments histogram provides the necessary knowledge of what needs to be filled in the speaker model as acoustic holes. The proposed system attempts to achieve a major speaker model impact by employing an acoustic token histogram of the test data(such a histogram matching scheme has not been attempted in the literature for small enrollment data sets; yet there is a parallel idea seen for large train/test sets based on keyword codebooks[4]). If the test data is shorter than the enrollment data, the proposed algorithm focuses on fortifying the expected acoustic token in the test stage. Since the input test stream is labeled the Gaussian mixture index using the S.I. GMM as that used for the in-set train data, it is possible to balance the test data with the available in-set train data and associated cohort speaker data. Therefore, it is not necessary to fill the acoustic holes of the train space if the test data is absent in those locations. For other cases, the longer test data provides further information of acoustic coverage than the enrollment data. A proposed parallel GMM training solution based on the EM algorithm using the data from most occurring and least occurring acoustic features called "Top-Down Bottom-Up (TDBU)" is developed. Also, employing the acoustic phoneme-like token histogram of the test data using the TDBU is called "TDBU using Test Token Histogram (TTH)". This approach incorporates the acoustic information of the test data so that the resulting model fills the acoustic holes and fortifies the expecting acoustic tokens using the parallel training strategy.

This paper is organized as follows. Sec. 2 explains the baseline system for evaluating the proposed algorithm. Sec. 3 presents motivation and a detailed procedure for developing the proposed algorithm. Next, an evaluation and results of the proposed algorithm is presented with a comparison to

This project was funded by AFRL under a subcontract to RADC Inc. under FA8750-09-C-0067.

the baseline system in Sec. 4.2. Finally, conclusions and future work is discussed in Sec. 5.

2. BASELINE SYSTEM

2.1 In-set/Out-of-set Speaker Recognition

Assume we are given a set of in-set (enrolled) speakers, and an organized collected speaker data set $\mathbf{X}_n$, corresponding to each enrollment speaker S_n, $1 \leq n \leq N_{in\text{-}set}$. Let the data $\mathbf{X}_0$ represent all outside non-enrolled speakers in the development set. Each speaker dependent GMM Λ_n, $\{\Lambda_n \in \Lambda, 1 \leq n \leq N_{in\text{-}set}\}$, can be constructed with $\mathbf{X}_n$ using the EM algorithm. In the first stage, called *(closed-set) speaker identification*, we first classify X as one of the most likely in-set speakers Λ^*,

$$\Lambda^* = \underset{1 \leq n \leq N_{in\text{-}set}}{\operatorname{argmax}} \; p(\mathbf{X}|\Lambda_n).$$

In the second stage, called *speaker verification*, we verify whether the observation $\mathbf{X}$ truly belongs to Λ^* or not (i.e., accept/reject).

2.2 GMM-UBM Baseline

The most recognized text-independent system uses the Gaussian Mixture Model (GMM) to represent the out-of-set model for outliers (e.g. UBM), and to adapt a UBM to the speaker in the in-set speaker model with Maximum A Posteriori (MAP) estimation[1, 2]. A speaker model is represented by M component Gaussians trained from the D dimensional observation vector $\mathbf{x}_t$ sequence. A GMM is denoted as $\Lambda_n = (\omega_{nm}, \mu_{nm}, \Sigma_{nm})$, for $m = 1, \ldots, M$ and $n = 1, \ldots, N$ where ω_{nm} is the mixture weight of the mth component unimodal Gaussian density, with each parameterized by a mean vector μ_{nm} and covariance matrix Σ_{nm}, which is assumed diagonal.

2.3 GMM Mixture Tagging(GMT)

The short amount of data requires exploiting information from acoustically similar speakers. Additionally, indexing the short amount of data enables us to supervise the data usage. A GMM is employed to classify the acoustic space represented by each Gaussian. The GMM is built with development and in-set speaker data using the EM algorithm, so we call this the S.I. GMM. The speech observation tokens are tagged with the highest probability mixture of the S.I. GMM. The test observation tokens are also labeled with the mixture index of S.I.GMM.

2.4 GMM-Cohort UBM Baseline

The speaker dependent model is built with MAP using only mean adaptation from the UBM in Sec. 2.2, where the resulting GMM represents a translation of the same Gaussian mixture densities of the UBM. The acoustic holes caused by sparse in-set data are effectively filled with the Cohort UBM[3]. Since the cohort UBM is built with a reduced number of speakers (5 speaker) versus the UBM development speaker set (60 speaker), the resulting Gaussian mixture density represents a more precise acoustic space for the speaker phone information than the UBM. The cohort speakers are selected from non overlapping with UBM development speaker pool, the notation of cohort speaker's observation sequence is defined by $\mathbf{X}_i, 1 \leq i \leq N_{coh\text{-}dev}$. The in-set training speaker's data is defined with $\mathbf{X}_n, 1 \leq n \leq N_{in\text{-}set}$. The speaker model using more cohort speaker data employs more speaker traits, therefore the noble speaker measurement are introduced below to select best speaker groups. Here, the precise speaker similarity measure improves the overall system[5].

A. Speaker Similarity Measure using KL divergence:

Step 1: Using GMT, build Mixture Tag (M_{T_i}) histogram of short duration (5 s) available training data for each enrollment speaker.

Step 2: Select potential cohort data to match enrollment speaker histogram from Step 1.

a) Use Mixture Tagger to tag mixture index for all data for potential cohort speaker. (318 potential development speakers)

b) Select mixture tagged frames from each potential cohort speaker data to match Mixture Tag histogram from Step 1. (This ensures, consistent acoustic representation for input speaker and each potential cohort speaker.)

c) Move to Step 3 for training, Step 4 for distance measurement.

Step 3: Build the GMM for enrollment and potential cohort speakers.

a) Build GMM with EM algorithm for enrollment speaker using 5 s data

b) Using data from each potential cohort speaker, that has been matched to the Mixture Tag histogram, build a GMM to test for cohort distance.

Step 4: Measure the distance between enrollment and potential cohort speakers.

a) Find speaker distance between enrollment speaker and potential cohort speakers.

b) Repeat for all development cohort speakers (318 in our evaluation)

c) Select top number of cohort speakers so that closest speakers are used first, and only Mixture Tag entries that require hole filing data are used.

B. Build the GMM-Cohort UBM:

Step 1: Build the cohort GMM, Λ_n^{cohort}, with the EM algorithm using the observation of top N_{cohort} speakers for each in-set speaker model Λ_n. The system performance using different cohort speaker is studied in [3].

Step 2: Using Λ_n^{cohort} as the initial model, adapt the speaker model via MAP using in-set training data, $\mathbf{X}_n$.

3. PROPOSED ALGORITHM

3.1 Motivation

A speaker recognition system with sparse enrollment data will have a difficult time in decoding a valid speaker's identity given extremely short test data 2 s. The acoustic space of a 5 s in-set speaker's data is too sparse to effectively cover the entire in-set speaker acoustic space. Acoustic holes from sparse enrollment data are filled by exploiting an acoustically similar cohort speakers' phoneme data[3]. A previously proposed system "TDBU"[5] enables us to exploit the specific speaker's acoustic information to fill acoustic holes. The motivation for this study exploits test token histogram information so that in-set speaker model balances for each test data. For exceptionally short test data (2 s), the speaker model should not misrecognize the phones, which have been trained for the enrollment stage. By providing test token histogram information, the speaker model can be robust by assigning more weights on most occurring Gaussian index of test tokens. A longer test utterance (6 s) than the training in-set data can take advantage of deciding which acoustic information is filled, or needs to be filled. The speaker model emphasized by greater acoustic histogram information than in-set train data can provide superior system performance.

The short test observation can be instantaneously categorized and quantized by indexing the most probable Gaussian mixture to represent that part of the acoustic space. Consequently, the emphasis on speaker modeling using the test speaker's acoustic token histogram information results in a better representation of the in-set speaker model for various amount of test data sequence. Note, the only down-side of this approach is that new in-set models need to be generated on-the-fly for new input test set sequences(a small price considering the challenge in 2-6 s train/test sets).

3.2 Top-Down Bottom-Up Speaker Modeling using Test Token Histogram (TTBU-TTH)

The parallel training of TDBU focuses on training the most occurring (top) and least occurring (bottom) enrollment data[5]. Based on enrollment acoustic histogram, the TDBU builds top model for focusing own training data and bottom model for filling acoustic holes . The major difference in using TTH is that it employs the acoustic token histogram information of the test data. Since TTH provides test data histogram information, the most occurring parts in the test data fortify its discriminating ability using the same acoustic phoneme-like tags as in the enrollment data. The least occurring part in the test data provides acoustic hole information so mismatched acoustic holes between enrollment and test data can be filled. The speaker similarity measure was illustrated in Sec. 2.4,and the most acoustically similar set of speakers for each enrollment speaker n, $1 \leq n \leq N$ are selected for each enrollment speaker. Finally, the overall procedure to build the in-set speaker model is summarized as follows:

Step 1: Tagging mixture index on claimant's feature observation (GMT, Sec. 2.3), make a histogram by counting the most frequently occurring acoustic tokens (Top) and the least occurring classes (Bottom).

Step 2: Select cohorts data to match histogram for both classes.

Step 3: Construct cohort GMMs using EM algorithm as $\Lambda_n^{top\text{-}cohort}$ and $\Lambda_n^{bottom\text{-}cohort}$ for each enrolled speaker n. Do a speaker adaptation from initial model, $\Lambda_n^{top\text{-}cohort}$ and $\Lambda_n^{bottom\text{-}cohort}$, to construct the enrolled speaker model Λ_n^{top} and Λ_n^{bottom} with the corresponding top and bottom speaker data using MAP algorithm.

Step 4: Combine Λ_n^{top} and Λ_n^{bottom} to build the final enrollment speaker model.

4. EXPERIMENTAL RESULTS

4.1 FISHER Corpus

An evaluation is performed for in-set/out-of-set speaker recognition with the telephone conversation corpus portion of FISHER. This corpus is selected for minimizing the channel mismatch so that this study focuses on filling acoustic holes using extreme sparse train(5 s)/test(2∼6 s) data. A selected set of 60 speakers comprise the in-set and out-of-set speakers. We make three different groups of in-set/out-of-set speakers to evaluate group size, 15in/45out, 30in/30out, and 45in/15out. As the in-set size becomes larger, the increasing confusability between in-set group causes worse system performance than smaller in-set group. All 60 speakers are devoted to the in-set or out-of-set groups, with 50 randomly chosen combinations for three different groups. A second, independent development set consists of 378 speakers having 30 s of speech data. These speakers are to be used to draw potential cohort data to fill acoustic holes. The analysis window size is set to 20 ms with a 10 ms skip rate. Static 19-dimension 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.

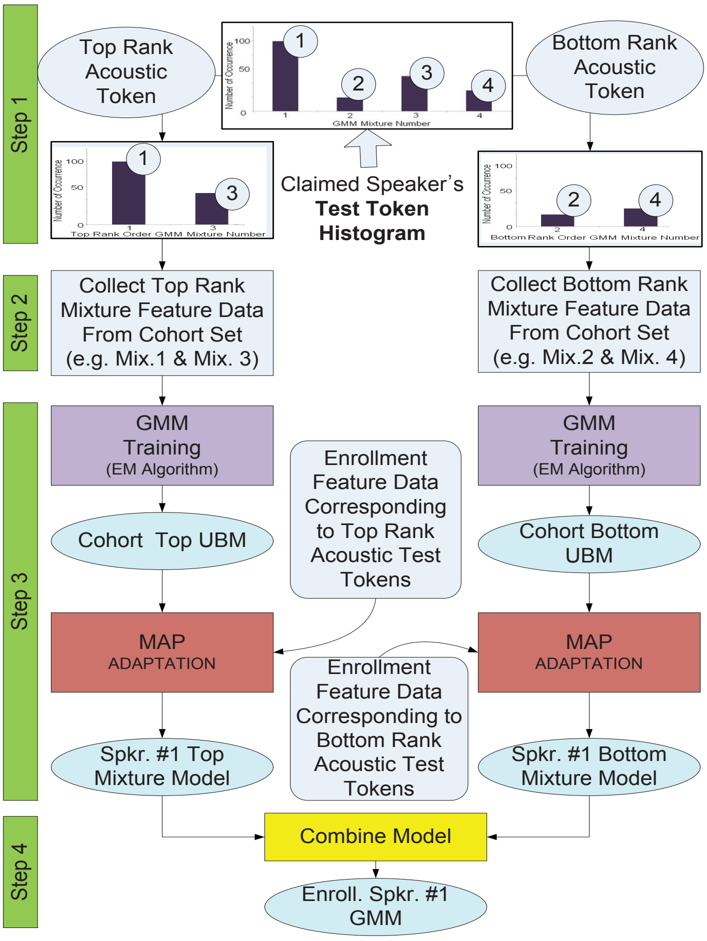
Step 1
Step 2
Step 3
Step 4
Top Rank Acoustic Token
Bottom Rank Acoustic Token
Number of Occurrence
100
50
0
1
2
3
4
GMM Mixture Number
Top Rank Order GMM Mixture Number
Bottom Rank Order GMM Mixture Number
Claimed Speaker's Test Token Histogram
Collect Top Rank Mixture Feature Data From Cohort Set (e.g. Mix.1 & Mix. 3)
Collect Bottom Rank Mixture Feature Data From Cohort Set (e.g. Mix.2 & Mix. 4)
GMM Training (EM Algorithm)
GMM Training (EM Algorithm)
Enrollment Feature Data Corresponding to Top Rank Acoustic Test Tokens
Cohort Top UBM
Cohort Bottom UBM
MAP ADAPTATION
MAP ADAPTATION
Enrollment Feature Data Corresponding to Bottom Rank Acoustic Test Tokens
Spkr. #1 Top Mixture Model
Spkr. #1 Bottom Mixture Model
Combine Model
Enroll. Spkr. #1 GMM

Figure 1: *Block diagram of TDBU-TTH. Each step is described in Section 3.2*

4.2 Evaluations

4.2.1 Baseline System

Each in-set speaker model consists of 32 mixtures representing the short 5 s training data. The parameters for building GMM are set equal to every speaker model and training algorithm. The UBM model will reflect the out-of-set speaker model or outlier, and it built with 60 randomly selected speakers from among the 378 speaker development set. The remaining 318 speakers are used to represent potential cohort speaker pool to fill acoustic holes for the in-set speaker, and we note that this 318 speaker set does not overlap with the 60 speakers used for the UBM. The top 5 cohort speakers are selected from across all potential GMM-Cohort UBM Baseline ,TDBU, and TDBU-TTH system. With these selected cohort speakers, each in-set speaker cohort model is built with 150 s of data (1 development/cohort speaker ≈ 30 s). This cohort model is then adapted with the 5 s in-set training data via the MAP algorithm.

TDBU is first introduced in a previous study[5], and the present TDBU-TTH training method was presented in Sec. 3.2. The primary difference is that the 5 s training data histogram is used to rank the mixture tagged data, as opposed to using the test data histogram. Table 1 shows that the TDBU-TTH improves in-set speaker recognition EER by an average 2.17% absolute over the TDBU, and an average 4.03% absolute EER over the GMM-UBM Baseline system using only 2 s of test data. The impact of test histogram information is superior than TDBU and baseline system.

Table 1: EER(%) performance comparison using 2s test data.

	EER		
	15in/45out	30in/30out	45in/15out
GMM-UBM Baseline	30.62	31.27	31.55
GMM-Cohort UBM Baseline	32.96	32.10	30.43
TDBU	26.71	29.13	32.02
TDBU-TTH	25.27	26.77	29.30

4.2.2 TDBU-TTH

The proposed TDBU-TTH algorithm employs a cohort speaker group of 5 speakers, the same size used for the GMM-Cohort Baseline system. The most occurring (top) and least occurring (bottom) GMM mixture size was fixed at 16 based on heuristic results. The top/bottom GMM is build with supervised data usage depending on test token histogram. The combined weight ratio is set to 7:3 for the top and bottom GMM speaker model for renormalization of overall score weights. The final speaker model combines the top model (fortifying expecting test tokens for training model) and the bottom model (harvesting expecting acoustic hole tokens). By supplying the mixture tagged test data histogram information, the system performance improves EER on average 2.34% over TDBU for 2 and 6 s test data. Fig.2 shows that the equal error rate is reduced by between 2.2%∼6.49% absolute over the GMM-UBM Baseline. Fig.2 also points that a smaller in-set group tends to produce a lower equal error rate. The large in-set group includes more speaker traits into group so unknown speaker or outliers would become a false acceptance speaker. This fact increases the EER in larger in-set group. In summary, the proposed method impacts system performance by focusing the expected acoustic information, and harvesting unseen acoustic knowledge collected at the feature frame level from test data.

5. CONCLUSIONS AND FUTURE WORK

In this study, we have developed a novel strategy to ensure an improved data training balance for an in-set speaker model

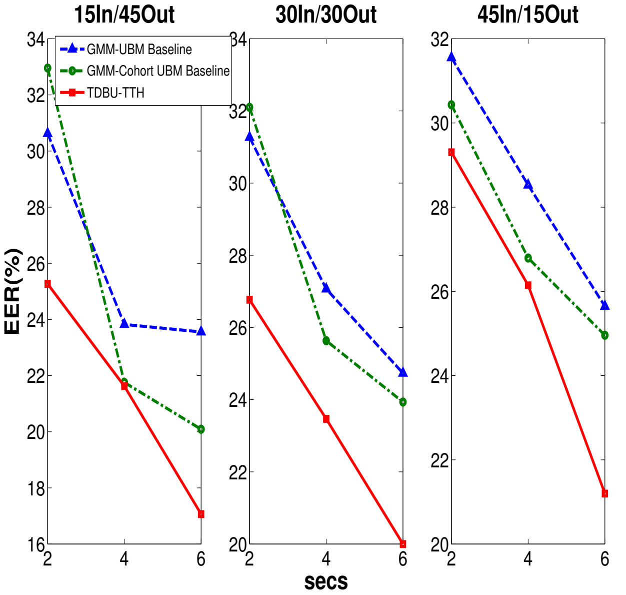

Figure 2: *Performance (in terms of EER(%)) of baseline and proposed algorithm on FISHER, using in-set/out-of-set speaker sizes of 15/45, 30/30 and 45/15.*

using the expected acoustic information from an acoustic token histogram of 2 s test data. The TDBU-TTH strategy improves acoustic hole filling, resulting from the limited in-set speaker data. Evaluations were performed using the "landline telephone channel" from the FISHER corpus to avoid handset variation, and focus on acoustic hole filling. The proposed TDBU-TTH training method improves in-set speaker recognition EER by 2.2∼6.49% absolute with only 2∼6 s of test data. Future work could consider expanding the method to normalize for handset variation effect from the FISHER corpus so that cohort speakers can be selected from any corpus.

REFERENCES

[1] H. Gish, M. Schmidt, "Text-independent speaker identification," *IEEE Signal Process. Mag.*, vol. 11, no. 4, pp. 18-32, Oct. 1994. .

[2] D.A. Reynolds, T.F. Quatieri, R.B. Dunn, "Speaker Verification Using Adapted GAussian Mixutre Models," *Digital Signal Proc.*, vol.10, pp.19-41, 2000.

[3] V. Prakash, J.H.L. Hansen, "In-Set/Out-of-Set Speaker Recognition Under Sparse Enrollment," *IEEE Trans. Audio, Speech, & Language Proc.*, vol. 15, no. 7, pp.2044-2052, Sep. 2007.

[4] C.-C. CHEN, C.-T. CHEN, S.-Y. LUNG, "Efficient Genetic Algorithm of Codebook Design for Text-Independent Speaker Recognition" *IEICE Trans.*, vol.E85-A, No. 11, pp. 2529-2531, Nov. 2002.

[5] J.-W. Suh, P. Angkititrakul, J.H.L. Hansen, "Filling Acoustic Holes Through Leveraged Uncorellated GMMS For In-set/out-of-set Speaker Recognition," *Interspeech-08*, pp. 1905-1908, Sept. 2008

[6] S. Kullback, "Information Theory and Statistics," *New York: Bover*, 1968.