# GMM and i-vector based speaker verification using speaker-specific-text for short utterances

B. Bharathi
Department of Computer Science and Engineering
SSN College of Engineering
Kalavakkam, Chennai 603110
Tamil Nadu, India
Email: bharathib@ssn.edu.in

T. Nagarajan
Department of Information Technology
SSN College of Engineering
Kalavakkam, Chennai 603110
Tamil Nadu, India
Email: nagarajant@ssn.edu.in

***Abstract*—In speaker recognition tasks, one of the reasons for reduced accuracy is due to closely resembling speakers in the acoustic space. In order to increase the discriminative power of the classifier, the system must be able to use only the unique features of a given speaker with respect to his/her acoustically resembling speaker. This paper proposes a technique to reduce the confusion errors, by finding speaker-specific phonemes and formulate a text using the subset of phonemes that are unique, for speaker verification task using GMM-based approach and i-vector based approach. Experiments have been conducted on speaker verification task using speech data of 50 speakers collected in a laboratory environment. The experiments show that the Equal Error Rate (EER) has been decreased by 4% and 4.5% using speaker-specific-text when compared to conventional GMM and base line i-vector based technique respectively.**

## I. Introduction

Gaussian Mixture Modeling (GMM) and Hidden Markov Modeling (HMM) techniques have been successfully used in many classification tasks. Maximum Likelihood Estimation (MLE) and Expectation Maximization (EM) algorithms can be used to estimate the model parameters efficiently. However, a major drawback in this type of modeling techniques is that the modeling is carried out in isolation, i.e., the modeling technique, when modeling a class, does not consider the information from other classes. In other words, out-of-class data is not used to optimize the classifier performance. This may lead to poor models with parameters that are common to other classes, in addition to the unique parameters of a class. This may increase the classification ( or confusion) error. Better classification accuracy can be achieved if the training technique is able to capture the unique features of a class, i.e., the features that discriminate a class from other classes, efficiently.

Many research works have been reported in the literature to increase the classification accuracy of a classifier by increasing the discriminative power of the classifier. Such techniques can be grouped into mainly two classes as follows:

1) Discriminating the classes in the feature level itself by identifying and removing the common features between two classes under consideration.
2) Adjusting the model parameters themselves such that two classes, in the feature space itself, are well separated.

In this paper we propose and emphasize the power of considering only the unique characteristics of each class, in a speaker verification task, at phoneme-level.

To improve the discriminative qualities of Gaussian mixture models, several approaches have been proposed. Universal Background Model-Gaussian Mixture Model (UBM-GMM) is a popular one among them. UBM is a base model from which all speaker models are adapted by a form of Bayesian adaptation [1]. A UBM is built from a large data set containing all probable speakers. During training, speaker specific model is adapted from this UBM by performing Maximum A Posteriori (MAP) adaptation. In [2], GMMs have been built for each speaker discriminatively based on the available positive and negative examples for each speaker. In this approach [2], speaker models are trained by moving the mean values of the mixture components in such a way as to maximize the likelihood of speaker data while also minimizing the likelihood of negative examples for the speaker.

Minimum Classification Error (MCE) approach for speaker verification is proposed in [3]. In this approach [3], all the competing speakers are used to evaluate the score of the anti speaker which is found to be effective. However, it is not practical for verification test over a large population. The i-vector systems have become the state-of-the-art technique in the speaker verification field [7]. They provide an elegant way of reducing the large-dimensional input data to a small-dimensional feature vector while retaining most of the relevant information. The technique was originally inspired by Joint Factor Analysis framework introduced in [8]. In [8], a model has been developed to solve the problem of channel and session variabilities in speaker recognition task, referred as Joint Factor Analysis (JFA). This proves to be an efficient technique for handling channel and session variability.

In our proposed work, the classes are discriminated at the phoneme level, i.e., acoustically dissimilar phonemes of a speaker when compared to his/her closely resembling speakers were derived. During testing, the speaker-specific-text is used thereby the classification accuracy is increased[5]. Even though our proposed work uses same channel for all the speakers during training as well as testing and the session

978-1-4799-2827-9/13/$31.00 ©2013 IEEE

variability is also not considered, we used i-vector approach, is to see the effect of acoustically dissimilar phonemes (unique phonemes) in a classification task. In this paper, GMM-based approach and i-vector based approach are used for speaker verification task using speaker-specific-text.

The organization of this paper is given below. The next section describes the importance of speaker-specific-text in our proposed system. Our proposed system using GMM-based approach is described in Section III. The introduction about i-vector approach is presented in Section IV. Section V describes the details of our speech corpus, and experimental setup of the proposed system. Section VI deals with the performance analysis of the proposed technique on a speaker verification task. Finally, section VII concludes the paper.

## II. Relevance of speaker-specific-text

In [4], a GMM-based technique was proposed to equip a classifier to capture the unique features of a class and to make decisions based on the unique features alone. During testing, feature vectors that are unique to a class have been derived and used thereby the classification accuracy is increased. One of the drawbacks is that, if the test utterance does not contain reasonable number of unique features, then the discrimination power cannot be ensured. Another drawback is that the unique features have to be identified from the test utterances during testing thus increases the computation time. If the speaker is able to utter the word which contains only the unique features then the computation time will be reduced. Even though the unique feature vectors are known, one cannot expect a speaker to utter speech segments, that contain these features alone. On the other hand, if we know unique phoneme list apriori, one can formulate a text to be uttered using such phonemes alone.

In this proposed work we investigate the effect of a subset of phonemes, that are unique to a speaker in the acoustic sense on a speaker verification task. The proposed technique involves three main steps:

1) To find out confusing speakers for each speaker
2) To derive acoustically dissimilar phoneme set for each speaker when compared to his/her confusing speakers
3) To Perform testing using speaker-specific-text

The proposed technique was experimented in [5] on speaker identification task using TIMIT speech corpus. The authors of [5] have demonstrated the improvement in classification accuracy, by considering only one confusing speaker for each of the speakers, during training as well as testing phases. In [5], one of the reasons identified for misclassification is that the acoustically dissimilar phonemes set is derived using average of log-likelihood values. If some of the phonemes have less number of examples, then considering the statistical parameter like the mean of likelihoods, is not appropriate. In [5], since the TIMIT speech corpus is used, many phonemes have very less number of examples (even just two) it is not appropriate to use the mean value and this might have led to false set of phonemes as unique. This error can be avoided by creating our own phonetically balanced speech corpus [1]. Hence, in this proposed work we have created our own speech corpus in which we have made sure that all the phonemes have reasonable number of (minimum 30) examples.

In [5], only one confusing speaker is considered for each of the speaker. During testing, if the confusing speaker is not present in the first position, we will not have the chance to improve the performance. On the other hand, if we consider more than one confusing speaker for each of the speakers, then common set of unique phonemes can be derived from all of the confusing speakers. One may assume that these phoneme set is, to certain extent, unique to the other speakers too. Let us consider a closed-set speaker recognition task, with $N$ speakers. For any speaker, in a given set of $N$ speakers, unique phonemes can be derived in the following two ways:

1) Considering the rest of the $N-1$ speakers as competing speakers.
2) Considering a smaller set of speakers (say $N1$ speakers, where $N1 \ll N$) as competing speakers.

In case(i), when $N$ is very large, deriving unique phonemes is computationally expensive. It is reasonable to assume that most of the speakers in the total set $N$ will not be acoustically closer to the test speaker. Considering this reason, in our work, only a subset of speakers is considered. Since the intention here is to improve the classification accuracy of GMM-based technique, conventional GMM testing can be used to derive this subset by considering $N1$-best results of GMM technique.

For each pair of speakers, where the pair consists of the intended speaker and one of the competing speakers, the unique phonemes can be derived as follows: Given the speech segments for each of the phonemes and the model (GMM) for the speakers in the pair, the unique phonemes (or acoustically dissimilar phonemes) can be derived by comparing the acoustic likelihoods. In our proposed work $N1$ competing speakers have been considered. Therefore, a common set of unique phoneme is derived from all of the competing speakers, then one may assume that these phoneme set is, to certain extent, unique to the other speakers too. The proposed technique is experimented on speaker verification task using GMM-based approach and i-vector based approach.

## III. GMM-based approach using speaker-specific-text

In GMM-based approach, for each speaker, his/her confusing speakers are considered as impostors. Using the confusing speakers training utterances imposter models have been created for each speaker. The test utterance (which has been created using speaker-specific-text) of the claimed speaker is tested with claimed speaker model and their corresponding impostor model. The score have been calculated by finding out the difference between the log-likelihood value of claimed speaker and their impostor model. If the score is greater then

[1]NIST SRE corpora cannot be used for the proposed approach due to the reason that our approach requires speech data to be collected for speaker-specific-text

the desired threshold then the speaker is accepted or otherwise the speaker is rejected.

## IV. I-VECTOR APPROACH

In i-vector based system, a variable length speech pattern is projected onto a low-dimensional linear subspace. The basis vectors of this subspace are estimated from the EM algorithm. This low dimensional representation of a speech utterance is termed as the i-vector (identity vector). The main idea in traditional JFA, is to find two subspaces which represent the speaker and channel-variabilities, respectively. The channel space contains some information that can be used to distinguish between speakers. For this reason, the authors[10] propose a single space that models the two variabilities and named it as the total variability space. The basic assumption is that a given speaker- and channel-dependent GMM super vector M can modeled as follows:

$$M = m + Tw \tag{1}$$

where
$m$ - is a speaker- and channel-independent super vector (UBM super vector is a good estimate of m),
$T$ - is a low rank matrix, which represents a basis of the reduced total variability space
$w$ - is a standard normal distributed vector

$T$ is named the total variability matrix. The components of $w$ are the total factors and they represent the coordinates of the speaker in the reduced total variability space. These feature vectors are referred to as identity vectors or i-vectors. The feature vector associated with a given recording is the MAP estimate of $w$, whose calculation is explained in [11]. The matrix $T$ is estimated using the EM algorithm described in [11].

## V. EXPERIMENTAL SETUP

In our proposed work, we have created our own speech corpus. We have collected 142 English sentences (from TIMIT corpus), that have enough number (minimum 30) of examples for all the 45 phonemes. The number of phonemes taken for this work is 46, including silence. The speech data has been recorded using 16kHz sampling rate. Speech utterances have been collected from 50 speakers which includes 43 female speakers and 7 male speakers. The speakers age group is between 20 to 35. Each utterance is approximately of 3 second duration. The entire speech data have been automatically segmented at phoneme-level using Forced Viterbi alignment algorithm[6] .

For each speaker, among 142 sentences, 130 utterances are used for training and 12 utterances are used for testing. For each speaker, a GMM with 128 mixture components has been trained, considering Mel-frequency cepstral coefficients (13 static + 13 dynamic + 13 acceleration) as the features.

To find out the confusing speakers, the training utterances of each speaker have been tested with all the speaker models. Leave-one-out procedure has been used. For each speaker, $N1$ confusing speakers have been derived based on sorted log-likelihoods (for this work, $N1$ = 3). This process is repeated for all the speakers and a confusing speakers list is derived.

To derive speaker-specific-text of a speaker, the common phonemes (i.e., corresponding speech segments) of the speaker and his/her confusing speaker, available in the training utterances, are tested with his/her model and his/her confusing speaker model. Average log-likelihood of each phoneme is computed for the first speaker and the confusing speaker. Based on the sorted average log-likelihoods, the first twenty phonemes have been considered as acoustically dissimilar phonemes. For each speaker with respect to his/her closely resembling speakers different subset of acoustically dissimilar phonemes are derived. The same process is repeated for the phonemes of all the speakers. For each speaker, common acoustically dissimilar phonemes have been derived by considering three confusing speakers. For each speaker, the speaker-specific-text have been formulated using six common acoustically dissimilar phonemes.

The first two steps involved in our proposed system specified in Section II are common for both GMM-based approach and i-vector approach.

In GMM-based approach, For each speaker, a GMM with 128 mixture components has been trained for claimed speaker model. Similarly, using the confusing speakers training utterances imposter models have been created for each speaker.

For i-vector approach, gender-independent 128 mixture component UBM is built using the training utterances of all the 50 speakers. For each speaker, 142 training utterances are concatenated and the total variability matrix are estimated using the concatenated training utterances. The dimensionality of i-vectors is 400. The i-vectors have been extracted from the test utterances of each speaker.

When the system is tested using speech utterances that correspond to speaker-specific-text, the confusion error is found to be reduced considerably than that of the conventional GMM-based approach and baseline i-vector based classification technique, as discussed below.

## VI. PERFORMANCE ANALYSIS

Speaker verification performance is compared between the utterances with acoustically dissimilar phonemes and without considering the acoustically dissimilar phonemes. To derive speaker characteristics, the constraint that is set in our work is that the test utterances (words) should have at least six phonemes. Each phoneme may have approximately 80ms duration. For conventional GMM-based speaker verification task, each test utterance is divided into 500ms speech signal and given for testing. This 500ms speech signal may contain both acoustically similar and dissimilar phonemes (segments correspond to any silences (more than 100ms) are not considered). The number of speakers taken for this experiment is 50. Similarly, testing is done using speaker-specific-text. The performance of speaker verification task using conventional

GMM-based approach and proposed GMM-based approach using speaker-specific-text have been tabulated in table I.

TABLE I
SPEAKER VERIFICATION PERFORMANCE USING CONVENTIONAL GMM-BASED APPROACH AND PROPOSED GMM-BASED APPROACH USING SPEAKER-SPECIFIC-TEXT

| S.No | Method | EER |
|---|---|---|
| 1 | Conventional GMM | 7% |
| 2 | GMM-based approach using speaker-specific-text | 3 % |

From Table I, it can be noted that there is a 4% reduction in Equal Error Rate using GMM-based approach using speaker-specific-text.

Speaker verification performance is compared between the utterances using speaker-specific-text and without considering speaker-specific-text using i-vector approach. The size of the test utterance is 500ms. The number of speakers taken for this experiment is 50.

TABLE II
SPEAKER VERIFICATION PERFORMANCE USING BASELINE I-VECTOR APPROACH AND PROPOSED I-VECTOR APPROACH USING SPEAKER-SPECIFIC-TEXT

| S.No | Method | EER |
|---|---|---|
| 1 | Baseline i-vector | 37% |
| 2 | i-vector approach using speaker-specific-text | 32.5 % |

From Table II, it can be noted that there is a 4.5% reduction in Equal Error Rate using i-vector based approach with speaker-specific-text.

From Table I and Table II, we can notice that the EER is considerably higher in the case of i-vector based approach. The proposed approach using the conventional GMM-based approach, the EER is only 7% as specified in row 1 of Table I. This shows that i-vector approach may not be ideal when the duration of the test utterance is short. In any sense, in both GMM-based or i-vector based technique, if the test utterance is from speaker-specific-text, the performance is found to be better.

TABLE III
SPEAKER VERIFICATION PERFORMANCE OF DIFFERENT SPEECH UTTERANCE DURATION USING BASELINE I-VECTOR APPROACH

| S.No | Speech utterance duration (in seconds) | EER |
|---|---|---|
| 1 | 3 | 11% |
| 2 | 36 | 6 % |

Speaker verification performance of the system is analysed using different test utterance size using baseline i-vector approach have been tabulated in Table III.

From Table III, it can be noted that, if the utterance length decreases, speaker verification performance degrades at an increasing rate using i-vector based approach [13], [14].

## VII. CONCLUSIONS

In this paper, we have proposed to use speech utterances that correspond to a speaker-specific-text for speaker verification task using GMM-based approach and i-vector based approach. Here, the speaker-specific-text is formed using the unique phonemes of a speaker, in other words, a set of phonemes that are acoustically dissimilar when compared with that of a competing (acoustically closely resembling) speakers. The experimental results show that, in any approach whether GMM-based or i-vector based approach, if the test utterance is from speaker-specific-text, the performance is found to be better.

## REFERENCES


[1] D.A. Reynolds, T.F. Quatieri and R.B. Dunn, "speaker verification using adapted Gaussian mixture models", Digital Signal Processing, vol. 10, pp. 19-41, 2000.

[2] Srikanth M R, Hema A Murthy, "Discriminative training of Gaussian Mixture Speaker models : A New Approach", Proceedings of National conference on Communications, IIT Chennai,2010

[3] Chi-Shi Liu, Chin-Hui Lee, Biing-Hwang Juang , A.E. Rosenberg, "Speaker recognition based on minimum error discriminative training", in proceedings of IEEE International conference on Acoustics, Speech, and Signal Processing, Vol 1, pp 325-328, April 1994.

[4] C. Arun Kumar, B. Bharathi and T. Nagarajan, "A discriminative GMM technique using product of likelihood Gaussians", IEEE TENCON pp.16, 2009.

[5] B. Bharathi, P. Vijayalakshmi, T. Nagarajan, "Speaker identification using utterances correspond to speaker-specific-text", IEEE Students technology symposium(Techsym) pp. 171-174, 2011.

[6] F.Brugnara, D. Falavigna, and M. Omologo, "Automatic segmentation and labeling of speech based on hidden markov models", Speech Communication, vol. 12, no.4, pp.357-370,1993.

[7] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification", IEEE Transactions on Audio,Speech, and Language Processing, pp.788-798,2011.

[8] Kenny, P., Boulianne, G., Ouellet, P. and P. Dumouchel. "Joint factor analysis versus eigenchannels in speaker recognition", IEEE Transactions on Audio, Speech and Language Processing, pp. 1435-1447, May 2007.

[9] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of inter-speaker variability in speaker verification", IEEE Transactions on Audio,Speech and Language Processing, vol. 16, no. 5, pp.980 988, July 2008.

[10] N. Dehak, R. Dehak, P. Kenny, N. Brummer, P. Ouellet, and P. Dumouchel, Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification, in INTERSPEECH, Brighton, UK, Sept 2009.

[11] Glembek, O., Burget, L., Matejka, P., Karafiat, M., and Kenny, P., "Simplification and Optimization of i-vector extraction", in proceedings of IEEE International conference on Acoustics, Speech, and Signal Processing, pp.4516-4519,May 2011

[12] B. Bharathi, Nagarajan Thangavelu, "A two-level approach for speaker recognition using speaker-specific-text", in proceedings of National conference on communications, at IIT Delhi, 15-17, Feb, 2013.

[13] A. Kanagasundaram, R. Vogt, D. Dean, S. Sridharan, and M. Mason, "i-vector Based Speaker Recognition on Short Utterances", in Interspeech, 2011.

[14] Anthony Larcher, Pierre-Michel Bousquet, Kong-Aik Lee, Driss Matrouf, Haizhou Li, Jean-Franois Bonastre "I-vectors in the context of phonetically-constrained short utterances for speaker verification", ICASSP, pp.4773-4776, 2012.