NOISE ROBUST SPEAKER IDENTIFICATION USING BHATTACHARYYA DISTANCE IN ADAPTED GAUSSIAN MODELS SPACE

Kshitiz Kumar, Qi Wu, Yiming Wang, Marios Savvides

Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213 USA
Email: {kshitizk, qwu, yimingw}@ece.cmu.edu, marioss@andrew.cmu.edu

ABSTRACT

This is a study on the issue of noise robustness of text independent Speaker Identification (SID). Over the past years, SID technology has emerged as extremely important tool with applications in security and authentication. The current technology works well in presence of matched acoustic conditions for training and testing but the performance shows immediate loss in mismatched conditions. Our broad approach in this work is to map features to models and then do classification in the space of models. In particular, our algorithm is works in the space of adapted Gaussian Mixture Models, where we use Bhattacharyya Shape to measure closeness of models. We show our approach to be robust to noise in SID evaluations. We tested our approach on speech corrupted by white and music noise and found it to be very advantageous in low SNR conditions.

1. INTRODUCTION

Over the past years, Speaker Identification (SID) technology has emerged as an extremely important tool[7]. In SID, we exploit an identifying property of voiceprint of a person that makes it difficult for someone to obtain the voiceprint of someone else. Thus, voice serves as a unique biometric identity of a person. SID system has thus far been deployed in many commercial applications where it is desired to authenticate a transaction or provide secure access to information. SID tasks can be classified as either being text dependent or text independent. In our study, we focus on the later task, where the SID system does not constrain the user to memorize the passwords as is required in text dependent tasks. Text independent SID system frees the users from traditional password based authentication which required them to memorize their passwords and thus completely eliminates any cost associated with fetching forgotten passwords[2]. SID system has been deployed in banking, telecommunication, forensic and many other biometric based applications [6].

SID is thus a task of identifying a person from the speech modality of the person. It is assumed that the person belongs to one of the valid subjects in the enrollment database. SID system extracts specific features from the speech of a speaker and represents those features in terms of a model. During identification SID system hypothesizes a speaker for a test utterance, that is based on the closest matching of the utterance against different speaker models. The above procedure works reasonably well under matched acoustic conditions in training and testing but close acoustic matching is only guaranteed under lab conditions and there exists an acoustic mismatch in real-life conditions. Thus, SID systems needs to be robust to the mismatch in training and testing. Acoustic mismatch can be due to background noise which is typically white noise or music noise, else interfering speaker or reverberation in testing environments. In this study, we propose to provide a algorithm to provide robustness to SID system against white noise or music noise.

Gaussian mixture models (GMM) [2][7] are currently the dominant approach for modeling in text-independent SID systems. We use GMM based SID as our baseline SID, there we observe that GMM based approach works extremely well in matched acoustic conditions but SID accuracy deteriorates very sharply in presence of noise. We discuss this baseline approach in Section 2. Next, we present our approach in 3. Section 4 discusses our experimental setup for SID. We present our results and discuss their implications in Section 5, we present some analysis behind the success of our approach in Section 6, Section 7 concludes our findings in this work.

2. GAUSSIAN MIXTURE MODEL

Gaussian mixture models (GMM) [2][7] are currently the dominant approach for modeling in text-independent SID systems. In this approach, the broad idea is to individually learn a probability distribution on the training data for each of the speakers in the database and then perform maximum a posterior based identification. The distribution used is a mixture of gaussians with a given number of mixtures, where the number of mixtures is usually obtained from an evaluation task. A GMM is parameterized by

$$\Lambda_n = \{(w_k^n, \mu_k^n, \Sigma_k^n), \forall k = 1 \dots K\} \tag{1}$$

where, for n^{th} speaker, w_k^n is the weight of k^{th} mixture, μ_k^n is the corresponding mean vector, Σ_k^n is covariance matrix and K is the total number of Gaussian mixtures. The GMM parameters in (1) are obtained with Expectation Maximization (EM) algorithm on training features. The below equation evaluates the likelihood with respect to the GMM.

$$P(O = o_t|\Lambda_n) = \sum_{k=1}^{K} \frac{w_k^n}{(2\pi)^{D/2}|\Sigma_k^n|^{1/2}} exp\left\{-\frac{1}{2}(o_t - \mu_k^n)^T \Sigma_k^{n-1}(o_t - \mu_k^n)\right\} \tag{2}$$

where, O constitutes the set of observations in (2). Given a GMM model for all the speakers in our database, we can perform maximum a posterior decoding to build our hypothesis about the test speaker.

3. BHATTACHARYYA DISTANCE IN GAUSSIAN MIXTURE MODEL SPACE

Our work along the approach to be presented later in this section is guided by two primary motivations. The first motivation stems from the results reported in [2]. There, during SID testing a Bhattacharyya distance (BhD) metric was evaluated for a test utterance against training speaker models. BhD metric is evaluated on probability distributions, so for evaluating BhD in [2], a GMM distribution was fit to the test utterance and BhD was evaluated using that test distribution individually against the training models. Hypothesized speaker was the one with minimum BhD. There, it was also shown that Bhattacharyya shape (BhS) measure which is Bhattacharyya distance between distributions when assuming means of the distributions to be identical, provided better SID performance than that using BhD. The above approach requires fitting a probability distribution on test utterances as well but for SID tasks the test utterances are usually short and fitting distributions suffers from limited data. In our approach, we intend to provide a way to extend this problem of limited data for learning distribution from test utterances.

The second motivation for our work is related to the more important problem that we are focussing on in this paper, that is robustness to noise. In the traditional GMM based approach for SID, a posterior probability is the measure to decide the hypothesis speaker. In the case of matched training and testing conditions, the above method is guaranteed to be optimal. But in presence of mismatch in training and testing, we empirically observed that a posterior probability based score was very sensitive to the mismatch. Even a few noisy features changed the score significantly and lead to smaller SID accuracy. GMM based SID does classification in the space of features but in our approach we will be doing classification in the space of GMM models which we hypothesize and analyze to be more robust to noise than traditional GMM based approach.

Based on the motivations above, we next present our approach for SID. In our approach, we intend to develop a measure between training and test models. The measure used will be Bhattacharyya Shape (BhS) measure. We first generate a Universal Background Model(UBM) which is a Gaussian mixture model trained by EM algorithm on a large amount of speech. Next, we generate a GMM model for each of the speakers in our database by Bayesian adaptation of training utterance features from UBM, these individual speaker's models models are indicated as Λ_n in Fig. 1. Generating speaker's models from UBM is advantageous because the UBM then acts as a prior distribution and provides a way to conveniently score the unseen features in testing data. The required theory and complete update equations in Bayesian adaptation for generating the speaker's models from UBM have been detailed in [1][4]. Update equations for adapting means of Gaussian mixtures are as noted in (3)(4)(5). We essentially adapt one Gaussian mean vector at a time. Adapted mean is a weighed combination of mean of UBM model and an expected mean as in (4). The expected mean is obtained by posterior weighing of the features for the k^{th} Gaussian mean as in (3). The α controls the weighting of the UBM mean and expected mean. The weights of Gaussian mixtures as well as covariance terms were adapted. We refer to [4] for complete equations describing updates for GMM mixture weights and covariance matrices.

$$
\begin{aligned}
Pr(k|o_t) &= \frac{w_k P_k(o_t|\Lambda_n)}{\sum_{k=1}^{K} w_k P_k(o_t \Lambda_n)} && (3)\\
E_k(O) &= \frac{\sum_{t=1}^{T} Pr(k|o_t) o_t}{\sum_{t=1}^{T} Pr(k|o_t)} && (4)\\
\hat{\mu}_k &= \alpha E_k(O) + (1-\alpha)\mu_k && (5)
\end{aligned}
$$

Generating UBM and adaptation of the training utterances from the UBM to generate speaker's model constitute the training phase of our SID system. The testing phase of SID follows an approach similar to that in training. During testing, we need to assign a label to a test utterance from the set of speakers in our database. After feature extraction from the test utterance, the features are adapted from the UBM and a test model is generated. That model is indicated as Λ in Fig. 1. We have thus far generated a test model in the testing phase and we already had training models from training phase. The next step in identification is to find the training model which is closest to the test model. We define closeness in terms of Bhattacharyya shape (BhS) as in (6) between Gaussian Mixture Models for defining how well two Gaussian distributions are close in their space.

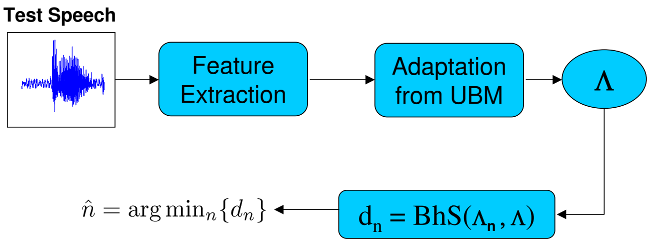
Test Speech
Feature Extraction
Adaptation from UBM
Λ
$\mathrm{d_n} = \mathrm{BhS}(\Lambda_n, \Lambda)$
$\hat{n} = \arg\min_n\{d_n\}$

Figure 1: Speaker Identification Flowchart during Testing

$$
BHD(p,q) = -ln\left(\int_{\mathscr{R}^D} \sqrt{p(x)q(x)}\, dx\right) \qquad (6)
$$

The distribution p and q in BhS metric in (6) have identical mean $\{\mu_k^n\}$ as is required in definition of BhS in [2]. In general the means will not be same for two distributions, but then one of the distribution can be changed to center it at the mean of the other distribution. After having defined a measure of closeness, we seek the training model which minimizes $BhS(\Lambda_n, \Lambda)$ as in Fig. 1. Thus $\hat{n}$ in (7) indicates the hypothesis speaker corresponding to a test utterance.

$$
\hat{n} = \arg min_n \{BhS(\Lambda_n, \Lambda)\} \qquad (7)
$$

BhS in (7) is the metric which was used to hypothesize the speaker in our SID. Thus, the broad contribution of our approach is to map features to models and then do classification in the space of models. This approach was guided by empirical studies where we found that the metric defined by us was more robust to noise than the corresponding metric in baseline GMM approach. We present more analysis indicating the robustness of our approach in Section 6.

4. EXPERIMENT

We did our SID experiments on a subset of YOHO database. We considered a total of 38 speakers in our task. We took 16 utterances from *Session*1 per speaker for training and 4 utterances from *Session*2 for testing. We extracted 13-dimensional Mel frequency cepstral coefficients (MFCC) from the speech utterances using [10]. The frame duration was kept at 25*msec* with 10*msec* of frame shift. We found that removing silence segments was extremely important for SID and silence was removed from training as well as test utterances. For silence detection we used an approach where we took the zeroth cepstral ($C0$) feature which is indicative of signal energy in that frame and obtained a threshold on $C0$ to segment speech frames into silence vs. non-silence (actual speech utterances). The thresholding was done by fitting a GMM with number of mixtures as 4, then assuming that noise is being modeled by the Gaussian mixture with the lowest mean, and rest 3 Gaussian mixtures represent actual speech, the threshold is obtained from the maximum a posterior criterion to separate the noise and speech under the above assumptions.

We also compared our classification results against those from Principle Component Analysis (PCA) and Support Vector Machine (SVM) [5]. PCA is widely used as a benchmark to compare the performance of novel algorithms. We used PCA to project MFCC features into a lower dimensional space and used nearest neighbor approach to do classification. In PCA, a weight was assigned to each class for each of the features. The weight was based on mean squared distance of the feature with the nearest feature belonging to the different classes. Weights are summed across features and the class with the smallest aggregate weight is taken to the result of SID. SVM developed by Vapnik [3][5] finds many practical application. SVM maps features with appropriate kernel function into a higher-dimensional feature space to make the features linearly separable in the mapped space.

In our experiments training was always done on clean training utterances and testing was done on clean test utterances as well on simulated noisy test utterances. Simulated data was generated from corresponding clean data by adding noise at various Signal to Noise Ratios (SNR). Training and testing on clean data indicates the performance of our approach under matched clean conditions. Training on clean and testing under noisy conditions indicates the noise robustness of our approach. In the next section we present our results comparing the performance of our approach against traditional GMM based approach and also against other important classifiers in pattern recognition.

5. RESULTS

We present and discuss our results in this section. Fig. 2 plots results with different approaches where the mismatch for test utterance was created in terms of additive white Gaussian noise. Training was done on clean features but testing on speech corrupted by white noise at different SNR levels. The number of mixture models for GMM was kept at 16. We see that under clean-matched conditions GMM provides highest SID accuracy but the SID accuracy drops significantly under noisy conditions. We also notice that results with PCA and SVM do not outperform GMM based SID except in the worst noise case. Next, we present results with our BhS based measure in the space of adapted GMM models as developed in Section 3. Those results are indicated as $GMM - BHS$ in the figures. For the GMM-BHS case, the number of Gaussian mixtures was kept as 1 as evaluating BHS in (6) becomes intractable for higher number of Gaussian mixtures. For single density Gaussian, BhS between two Gaussian distributions can be obtained in close form as in (8), where Λ_n and Λ are the two Gaussian models as obtained in Section 3 and indicated in Fig. 1.

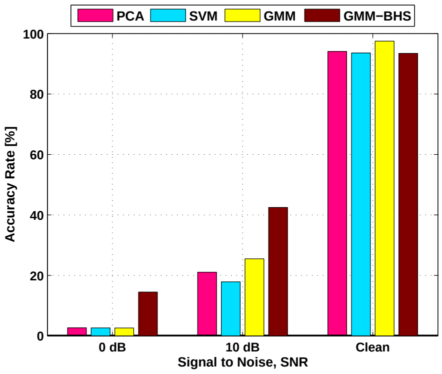

Figure 2: Accuracy Comparison under White Noise

$$BhS(\Lambda_n, \Lambda) = \frac{1}{2} ln \frac{|\frac{\Sigma_n + \Sigma}{2}|}{\sqrt{|\Sigma_n||\Sigma|}} \qquad (8)$$

We note that BhS in (8) is a function of only the covariance components, the mean component of the Gaussian mixtures does not affect the BhS. Next, in Fig. 2 we observe that as compared to GMM, $GMM - BHS$ provides an absolute improvement of approximately 18% at 10dB and 12% at 0dB. We also note that GMM outperforms $GMM - BHS$ in matched conditions. Lastly, Fig. 3 presents our results in presence of music noise. We see that $GMM - BhS$ approach provides big improvement in SID in low SNR conditions. As compared to GMM, by our approach in GMM-BHS, we obtained an absolute improvement of 22% in SID accuracy at music noise of 0dB, the relative improvement in SID accuracy for the same was 110%. We again note though the GMM system outperforms the GMM-BhS system at high SNR con-

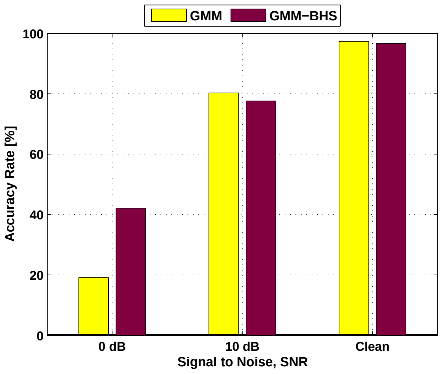

Figure 3: Accuracy Comparison under Music Noise

ditions, but the GMM-BhS is still very competitive in high SNR conditions too.

Thus, we find that we obtained important improvement in SID by working in the space of GMM models, we adapt training and test utterance against UBM and then do a distance comparison among the training and test GMM models to hypothesis the underlying speaker. The above approach works much better in low SNR conditions, and is very competitive for high SNR cases.

In our experiments, we restricted ourselves to SID tasks but our work can easily be extended to speaker verification tasks, where a user claims a particular identity and the system verifies whether the claim is true or not. Another point we wanted to note with respect to EM algorithm is that because EM algorithm can typically get stuck in local minima, so to robust against this issue, we performed EM algorithm with multiple starting points and selected the model with best likelihood across the different starting points of EM algorithms.

In our approach, we considered BhS metric for finding distance between two distributions. Studies done in [8][9] propose some other distance metric for the above comparison. Our future study can include a usage of those distance metrics in our application. In this work, for the for BhS measure, we restricted our evaluations to a single density Gaussian model. In future, we intend to evaluate a simplified BhS metric under some approximations else use Monte-Carlo based methods to extend our approach to multiple density Gaussians.

6. DISCUSSION

In this section we present an analysis behind the improved performance with GMM-BhS measure (8) against GMM measure (2). We begin by first developing the above equations in presence of noise in terms of parameters characterizing noise. Assuming a single density Gaussian model, the GMM parameters for the cepstral features of n^{th} speaker becomes $\Lambda_n = \{1, \mu_n, \Sigma_n\}$. Assuming that in presence of noise the corresponding features can be modeled by GMM parameters in $\Lambda = \{1, \mu = \mu_n, \Sigma = \Sigma_n + \sigma^2 I\}$. There, for our analysis we assumed that noise is zero mean with covariance $\sigma^2 I$ and additive in feature domain. We note that strictly speaking, noise is additive only in speech signal domain and not in cepstral feature domain where features have been obtained by non-linear processing on the signal. Even though, our analysis is not valid without above assumptions, we conjecture that our analysis does convey the message of robustness of BhS metric in an idealized setting. Next, we begin to write (2) and (8) as a function of the unknown in σ^2. Taking $\Sigma = \Sigma_n + \sigma^2 I$ and $\mu = \mu_n$ and assuming large number of independent observations O in (2), we derive an equivalent GMM metric for the log-likelihood score from (2). Under large number of observations the metric just becomes the expected value of log-likelihood score in (2), which can be shown to be as in (10),

$$\begin{aligned} GMM(\Lambda_n, \Lambda) &= tr(\Sigma_n^{-1}\Sigma) && (9) \\ &= tr(I + \sigma^2 \Sigma_n^{-1}) && (10) \end{aligned}$$

Constant terms involving additions or multiplications were ignored in (10). Using (8), we see the equivalent BhS metric in noise becomes (11).

$$BhS(\Lambda_n, \Lambda) = \frac{|I + \frac{\sigma^2}{2}\Sigma_n^{-1}|}{\sqrt{|I + \sigma^2 \Sigma_n^{-1}|}}. \qquad (11)$$

Thus we note that for large σ, GMM metric increases as a function of σ^2 but BhS metric increases as σ. Thus for large σ, the BhS metric is more robust to noise than GMM metric. We have therefore, provided an understanding of the results in Section 5, where from our experiments we found that BhS metric was more robust to noise than the baseline using GMM.

7. CONCLUSION

We studied the problem of robustness to noise in text independent speaker identification. We took the current state of art technology for SID and proposed algorithm for robustness with respect to additive noise. The novelty of our algorithm lies in doing classification in the space of GMM models for which we hypothesized, experimentally verified and analyzed that that under this approach, the SID is more robust to noise. We did experiments in presence of white noise as well as music noise and noted important gains for SID accuracy in low SNR conditions.

REFERENCES

[1] J. L. Gauvain, and C. H. Lee, ”Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains”, IEEE Trans. Speech Audio Process., vol. 2, pp. 291- 298, 1994.

[2] J. P. Campbell, ”Speaker recognition: a tutorial”, Proceedings of the IEEE, vol. 85, pp. 1437-1462, 1997.

[3] J. Platt, ”Probabilistic outputs for support vector machines and comparison to regularized likelihood methods”, *Advances in Large Margin Classifiers*, Cambridge, MA, 2000. MIT Press.

[4] D. A. Reynolds, T. F. Quatieri, R. B. Dunn, ”Speaker Verification Using Adapted Gaussian Mixture Models”, Digital Signal Processing. vol. 10, pp. 19-41., 2000.

[5] R. O. Duda, P. E. Hart, and D. G. Stork, ”Pattern Classification”, 2nd ed., John Wiley & Sons, Inc., 2001.

[6] J. F. Bonastre, F. Bimbot, L. J. Boe, J. Campbell, D. Reynolds, and I. M. Chagnolleau, ”Person authentication by voice: a need for caution”, European Conference on Speech Communication and Technology, 2003.

[7] F. Bimbot, J. F. Bonastre, C. Fredouille, G. Gravier, I. M. Chagnolleau, S. Meignier, T. Merlin, J. O. Garcia, D. P. Delacretaz, and D. A. Reynolds, ”A tutorial on text-independent speaker verification”, EURASIP J. Appl. Signal Process., vol. 4, pp. 430-451, 2004.

[8] G. Sfikas, C. Constantinopoulos, A. Likas, N. P. Galatsanos, ”An Analytic Distance Metric for Gaussian Mixture Models with Application in Image Retrieval”, Lecture Notes in Computer Science, Springer, vol. 3697, pp. 835-840, 2005.

[9] X. Peng, W. Xu, B. Wang, ”Speaker clustering via novel pseudo-divergence of Gaussian mixture models”, Proceeding of NLP-KE, pp. 111-114, 2005.

[10] CMU Sphinx Open Source Speech Recognition Engines, `http://cmusphinx.sourceforge.net/html/cmusphinx.php`.