Novel Approaches for One- and Two-speaker Detection

Sachin S. Kajarekar[1], *André G. Adami*[1], *Hynek Hermansky*[1,2]

[1]OGI School of Science and Engineering, Oregon Health and Science University, Portland, USA
[2]International Computer Science Institute, Berkeley, California, USA
{sachin, adami, hynek}@asp.ogi.edu

Abstract

The paper reviews OGI submission for NIST 2002 speaker recognition evaluation. It describes the systems submitted for one- and two-speaker detection tasks and the post-evaluation improvements. In one-speaker detection system, we present a new design of a data-driven temporal filter. We show that using few broad phonetic categories improves the performance of speaker recognition system. In post evaluation experiments, we show that combinations with complementary features and modeling techniques significantly improve the performance of the GMM-based system. In two-speaker detection system, we present a structured approach to detect speaker in the conversations.

1. Introduction

Speaker recognition is the task of recognizing the identity of the speaker based on his or her voice. Performance of the systems on this task is known to be sensitive to communication channel variations. Over the years, we have addressed this issue by investigating into robust feature extraction techniques and alternatives to commonly used Gaussian mixture modeling (GMM) techniques [1]. In this paper, we describe robust feature extraction using a new data-driven temporal filter. We describe the advantage of alternative modeling techniques by combining our results with IIT Madras system that uses complementary features and modeling techniques. We apply the novel speaker segmentation technique on the two-speaker detection task. The results show that improvements in the speaker segmentation algorithm are directly related to the two-speaker detection algorithm.

The systems based on our techniques are compared to the state-of-the-art using NIST 2002 Speaker Recognition Evaluation (SRE). They are evaluated on two tasks: one-speaker detection and two-speaker detection. The one-speaker detection task is to determine whether a specified speaker is speaking during a given one-side of the conversation. The two-speaker detection task is to determine if an unspecified speaker is speaking during the entire conversation. The data for both tasks is from the second release of LDC's cellular switchboard corpus (Switchboard Cellular – Part 2) [2].

The paper is organized in two parts. Sections 2 describes the task and the submitted system. Section 3 describes improvements in the systems after the evaluation. Sections 4 and 5 describe the systems submitted for two-speaker detection task and improvements in the systems after the evaluations. The paper concludes with summary in Section 6.

2. One-speaker Detection Submission

NIST 2002 SRE one-speaker detection task has 330 speakers with approximately 2 minutes of training data per speaker. It has 3570 tests utterances where each utterance is scored against 11 punitive speakers [2].

OGI systems are based on universal background model-Gaussian mixture model (UBM-GMM) framework [3]. In this framework, we apply the robust feature extraction techniques. The basic features are 13 Mel frequency cepstral coefficients (MFCCs). The temporal trajectories of these features are filtered using a data-driven temporal filter (see Section 2.1) and are normalized using cepstral mean subtraction (CMS). Delta and double delta features are appended to MFCCs. The combined features are selected either using the phone label or using the speech-silence segmentation decision (see Section 2.2). Finally the selected features vectors are gaussianized using a medium-term gaussianizer [4].

We submitted two systems for one-speaker detection task. Both systems used a 256-component GMM. In feature extraction part, both used the data-driven filter. They differed in the feature selection.

2.1. Data-driven temporal RASTA filter

Data-driven temporal filter for speaker recognition has been investigated before by different researchers, such as van Vuuren et. al. [5] and Malayath et. al. [1]. van Vuuren et. al. [5] studied the importance of the modulation frequency components for speaker recognition using systematic filtering of the modulation spectrum. They proposed a band-pass filter that was a combination of a 10 Hz low-pass filter and CMS. Malayath et. al. [1] used oriented principal component analysis (OPCA) to derive a data-driven filter for speaker verification. This filter was shown to significantly outperform the filter proposed by van Vuuren.

This work is an extension of [1] where we assume that phone and speaker variabilities are the useful variabilities, and channel and residual variabilities are harmful variabilities for speaker recognition. The phone, speaker+channel, and residual variabilities are estimated using OGI Stories database [6]. The channel variability is estimated using OPCA as described in [1]. The temporal filter is derived using linear discriminant analysis (LDA) [7]. Across-class covariance is computed as sum of phone and speaker+channel variabilities, and within-class covariance is computed as the sum of channel and residual variabilities. The impulse and frequency responses of the filter are similar to those of symmetric data-driven RASTA filter [5, 8, 9] used in speech recognition.

The performance of the filter on NIST 2002 SRE is shown in Figure 1. The feature extraction is same as the one described above without the Gaussianizer. The frame selection is performed using energy-based speech-silence segmentation (see Section 2.2). Result shows that the filter improves the speaker verification performance at all points and reduces the equal error rate (EER) from 10.6% to 9.8% (This is consistent with the improvements on NIST 2001 SRE tasks).

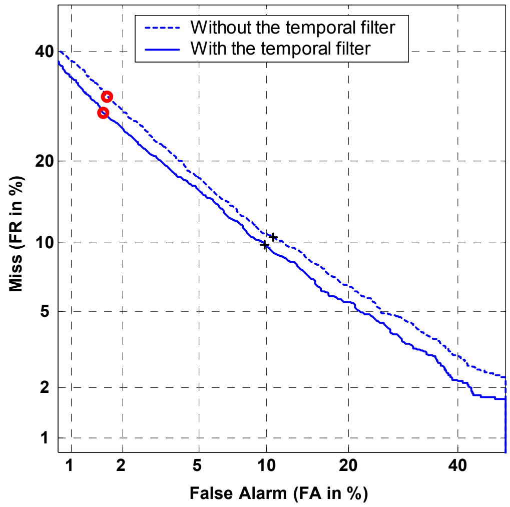

Figure 1: Effect of data-driven temporal filter on the performance

2.2. Feature-selection

Two different methods for selecting feature vectors were used in the submission. First method, a commonly used method, performs speech-silence segmentation using the estimate of frame-energy (or C0 coefficient). The system based on this method was the primary system (OGI1). Second method assumes that all the phone classes are not equally important for speaker recognition. Thus, speaker recognition performance can be improved by selecting the most sensitive classes to the speaker changes. In this approach, each utterance is transcribed into sequence of four broad categories – vowels+diphthongs (VD), glides+nasals (GN), fricatives (F), and silence+stops (SS). The class labels are obtained using automatic speech recognition system trained using CTIMIT database. As shown in [10], the frames labeled as VD and GN are used in the statistical modeling. The system based on this method was a secondary system (OGI2).

The EER of the system without any feature selection is 9.8%. It improves to 8.8% after applying any one of the feature selection methods described above. This suggests that the high-energy frames, selected using energy-based segmentation, are related to the frames corresponding to VD and GN broad phone categories. The performance degradation without the energy-based segmentation or by adding the other phone categories is an important result. Theoretically, the performance of a well-trained classifier can not degrade by adding less-useful data. In this case, our hypothesis is that by adding more data, we reduce the relative assignment of Gaussian components to the useful categories. More investigation is needed to understand this issue and to understand how to use the unused data to improve the performance of the baseline system.

3. Post-evaluation Improvements in One-speaker Detection Task

Scores from the primary system are compensated for difference in test durations using TNORM. Test part of NIST 2001 SRE is used to model the distribution of the impostor scores. The score compensation improves the performance in the low false-alarm area of the ROC curve but it does not change EER significantly.

3.1. Combination with IIT Madras Systems

In a joint effort, we are working with IIT Madras on alternative techniques for modeling the distribution of features. For NIST 2002 SRE, IITM submitted two systems: 1) IITM1: used linear predictive cepstral coefficients (LPCCs) with auto-associative neural networks (AANNs) and 2) IITM2: used residual from LPC analysis as features with AANNs. We combined the scores from these systems with those from OGI1 system (with TNORM). Table 1 shows the system and combination results for the primary condition and for three recording conditions -- 1) inside building (IN), outside building (OUT), and inside vehicle (VEH).

Table 1: Performance of OGI and IITM systems under different conditions

System	%EER			
	Primary	IN	OUT	VEH
OGI1	8.6	9.1	8.5	10.8
IITM1	17.2	18.6	14.8	17.6
IITM2	23.8	25.8	19.9	21.6
OGI+IITM1	8.0	8.6	7.7	10.8
OGI+IITM2	7.8	8.3	8.1	10.8
OGI+IITM1+ IITM2	7.1	8.1	6.9	9.0

Table 1 shows that although IITM systems perform worse than OGI1 system, the combination of scores from these systems significantly outperforms OGI1 system. Figure 2 that combination of scores improves performance at all the points of ROC curve. Results for different recording conditions also show that IITM systems improve the performance of OGI1 system in all the recording conditions. In terms of the combinations, note that the worst performing IITM2 system, which uses LPC residual, gives the best results after the combination. This clearly shows the advantage of using complementary features in speaker recognition.

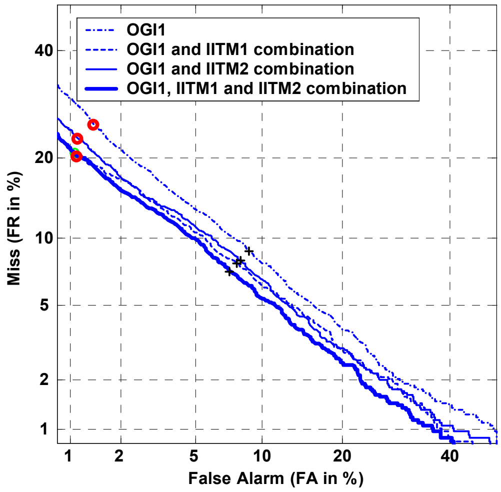

Figure 2: Combination of scores from OGI1 and IITM systems

The complementary information in IITM system is also due to different modeling technique. This can be shown by modeling LPCCs using GMM, and combining the scores with the primary system. Results show only moderate improvement over the primary system. In conclusion, we show that alternative modeling approaches may have complementary information that can be used to improve the performance of the conventional systems.

4. Two-speaker Detection System

NIST 2002 SRE two-speaker detection task has 309 speakers with 3 complete conversations with each speaker as the training data. There are 1460 test segments, which are tested against 22 punitive speakers [2].

The two-speaker detection is divided into three tasks: 1) separate speakers in the conversation (two-speaker segmentation), 2) select appropriate speaker(s) for training or testing (speaker selection), and 3) perform one-speaker detection (speaker detection). The two-speaker segmentation and speaker selection use 16 line spectral pairs (LSP) coefficients computed every 10 ms from the speech using a 32ms-window. The LSP coefficients are computed from the same order Linear Predictive Coding [11]. Speaker detection was performed using the primary system for one-speaker detection

4.1. Two-speaker Segmentation

The two-speaker segmentation can be divided into three steps: speaker-change detection, clustering, and re-segmentation. The speaker-change detection detects when a change of speaker occurs. The clustering merges all the segments that belong to the same speaker. The re-segmentation refines segmentation using a speaker detection approach.

Since the performance of the segmentation depends on the accuracy of the speaker-change detection, we investigated different approaches for speaker-change detection – one energy-based and two distance-based.

In the primary system (referred as OGI1), the energy-based approach uses the silence+stops labels, from the broad categories described in Section 2.2, to detect the speaker changes. The approach splits the conversation into nominally 1-second long segments (the duration can varies from 0.8 to 1.5 seconds) by removing the longest silence+stops regions.

The secondary systems (OGI2 and OGI3) use a distance-based method to detect a speaker change [12]. It assumes that one of the speakers initiates the conversation. So one second of speech at beginning of the conversation is used to train a model for one of the speakers. This model is used to calculate the likelihood of the speaker given all the frames in the conversation. The second speaker model is estimated using the regions that give low likelihood (or largest distance). A coarse segmentation is performed by computing likelihood with respect to the speaker models and then more data (3 seconds) is obtained to improve the estimation of the speaker models. Two different distance measures are used to compute the likelihood. OGI2 uses the Generalized Likelihood Ratio (GLR) distance, which is computed using a 16-component GMM adapted from the entire conversation. OGI3 uses a 16-component GMM for each speaker to detect when that speaker is more likely speaking. OGI3 follows the speaker detection GMM/UBM framework.

The clustering step uses a hierarchical clustering algorithm. The algorithm is initialized using segments produced by the speaker-change detection step. It used GLR distance to merge clusters. It was iterated until two clusters are obtained. The GLR distance is computed using a 16-component GMM, which is mean-adapted from the entire conversation.

The re-segmentation step uses the GMM/UBM framework to produce a more accurate segmentation using the data from the two clusters. First, it uses the entire conversation to train a 32-component GMM as the background model. Second, it uses the data from each cluster to train the speaker models, which are mean-adapted from the background model. Then, the conversation is re-scored at frame level using the two models. Finally, the sequence of scores is smoothed using a 1.5 second hamming window and, the frames are assigned to the speaker with maximum likelihood.

4.2. Speaker Selection

Two different approaches were used for training and testing. For training, three complete conversations were provided with the target speaker common among them. The common speaker was selected using an automatic gender detection system and the GLR distance. First, the gender detection algorithm classifies the gender of each speaker data produced by the two-speaker detection step. For the conversations that have different gender speakers, the speaker data that matches the gender of the target speaker is elected. For the remaining conversations (same gender speakers), the GLR distance is computed between every pair across different conversations. The pair with smallest distance is selected as the common speaker data.

From the testing, a single conversation was provided. We tested both the hypothesized speaker sides against the target model. The highest likelihood of a side was used as a likelihood of the conversation.

4.3. Results

Figure 3 presents the results of the two-speaker detection submission. The primary system OGI1 has 16.2% EER, OGI2 has 20.2% EER and OGI3 has 19.7% EER.

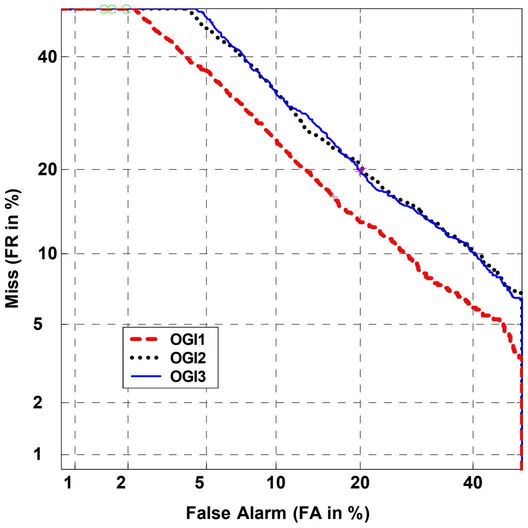

Figure 3: two-speaker Detection Performance: OGI1 – Energy-based; OGI2 and OGI3 – distance-based

The OGI2 and OGI3 systems give similar performance. In addition, the model adaptation from the entire conversation in the two-segmentation part (speaker-change detection step in section 4.1) affects the speaker model estimation. It is due to small (one second) amount of training data, and due to the background model, which is trained, using both the speakers.

5. Post-evaluation Improvements in Two-speaker Detection Task

In addition to TNORM (described in Section 3), we created the speaker model (in speaker-change detection step in OGI2) from the initial one-second segment (instead of adapting it from background model). Figure 4 shows the performance of the primary system and improved versions of the secondary system OGI2. The EER of OGI2 system with the new model estimation (referred as OGI2-POST) is 15.6% (23% relative improvement over the submitted OGI2 system). As the performance of OGI2-POST is better than OGI1, we conclude that a better estimation of change of speaker also improves the performance of the segmentation. Finally, TNORM reduces EER of OGI2-POST to 14.6% (referred as OGI2-POST+TNORM in Figure 4) and reduces EER of OGI1 to 15.2%.

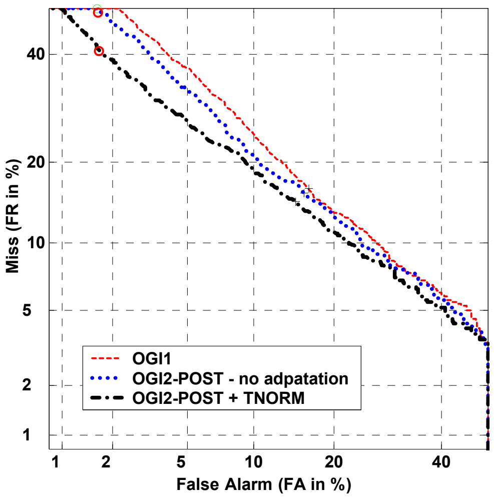

Figure 4: Post-evaluation improvements: TNORM and no model adaptation

6. Summary

We presented recent improvements in the OGI Speaker recognition system. A new method of deriving data-driven filter for speaker recognition was proposed. Results using the filter showed significant improvements in the performance. The importance of different broad-phonetic categories for speaker recognition was re-examined using transcriptions generated by ASR system. Results showed that vowels, diphthongs, glides and nasals are the most important categories for speaker recognition. This also agrees with our previous work using multivariate analysis of variance (MANOVA). Overall, the system submitted to NIST 2002 SRE was one of the best systems in the speaker recognition tasks.

We also presented results by combining the scores from our primary system with those from IITM systems. IITM systems were chosen due to their complementary features and modeling techniques. The combinations significantly improved the performance of our primary system. This was related to the complimentary information in the features and the modeling techniques used in IITM systems.

In two-speaker recognition task, we presented a structured approach to deal with two-speaker detection by using 3 different systems: two-speaker segmentation, speaker selection (gender detection), and one-speaker detection. The post-evaluation improvements in the two-speaker segmentation improved the performance of the two-speaker detection system. The improvement show that the performance of the two-speaker detection depends on the performance of the two-speaker segmentation

7. References

[1] N. Malayath, H. Hermansky, S. Kajarekar, and B. Yegnanarayana, "Data-Driven Temporal Filters and Alternatives to GMM in Speaker Verification," *Digital Signal Processing*, vol. 10 55-74, 2000.

[2] A. Martin, "NIST 2002 Speaker Recognition Evaluation Plan," 2002.

[3] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification Using Adapted Mixture Models," *Digital Signal Processing*, vol. 10 181-202, 2000.

[4] S. S. J. Pelecanos, "Feature Warping for Robust Speaker Verification," In Proc. of 2001: A Speaker Odyssey: The Speaker Recognition Workshop, Crete, Greece, pp. 213-218, 2001.

[5] S. v. Vuuren and H. Hermansky, "Data-driven design of RASTA-like filters," In Proc. of Proc of Eurospeech, Rhodes, Greece, pp. 409-412, 1997.

[6] S. Kajarekar, N. Malayath, and H. Hermansky, "Analysis of Speaker and Channel Variability in Speech," In Proc. of Workshop on Automatic Speech Recognition and Understanding, Colorado, 1999.

[7] K. Fukunaga, *Statistical Pattern Recognition*, Second ed: Academic Press, 1990.

[8] H. Hermansky, A. Bayya, N. Morgan, and P. Kohn, "RASTA-PLP Speech Analysis Technique," In Proc. of ICASSP'92, San Francisco, pp. 121-124, 1992.

[9] S. Kajarekar and H. Hermansky, "Analysis of Information in Speech and Its Application in Speech Recognition," In Proc. of Proc. of TSD, Brno, Czech Republic, pp. 283-288, 2000.

[10] S. K. a. H. Hermansky, "Speaker Verification Based on Broad Phonetic Categories," In Proc. of Proc. of 2001: A Speaker Odyssey, Crete, Greece, 2001.

[11] F. Itakura, "Line Spectrum Representation of Linear Predictive Coefficients of Speech Signal," *The Journal of The Acoustical Society of America*, vol. 57, S35 1975.

[12] A. Adami, S. Kajarekar, and H. Hermansky, "A New Speaker Change Detection Method for Two-speaker Segmentation," In Proc. of Proc. of ICASSP, Orlando, Florida, pp. 3908-3911, 2002.