

Speaker Recognition in Two-Wire Test Sessions

Hagai Aronowitz¹, Yosef A. Solewicz²

¹ IBM Haifa Research Labs, Haifa 31905, Israel

² Department of Computer Science, Bar-Ilan University, Ramat-Gan, Israel

hagaia@il.ibm.com, solewicz@013.net

Abstract

This paper deals with the task of speaker recognition in four-wire training and two-wire testing conditions. Instead of performing blind speaker diarization before the recognition stage, we directly perform the recognition on the non-segmented (or imperfectly diarized) speech. We present an analysis of the problem with respect to three different speaker recognition systems and propose improved recognition techniques both in the frame domain and in the model domain. The proposed techniques reduce error rate significantly. Furthermore, the developed techniques may be also beneficial in conjunction with an imperfect blind diarization stage.

Index Terms: speaker recognition, two-wire, summed channel

1. Introduction

Speaker recognition in two-wire (2w) test sessions is an important task that raises interesting research challenges. Recognition in 2w test sessions may be required when the audio is recorded in an open environment such as a meeting, or for processing of telephony where only a 2w recording is available.

Although the 2w testing (and training) condition is one of the conditions evaluated by the NIST speaker recognition evaluation [1], most participants choose not to address this condition. Those who do address it usually choose to apply a first stage of blind speaker diarization designed to segment the training/test sessions into two separate sides, each with a unique speaker. After the blind speaker diarization stage, speaker recognition techniques are applied as if the condition was four-wire (4w) [2].

The two-stage framework described above is not flawless. First, as no perfect speaker diarization algorithm is available [3], the sessions provided to the recognition phase are contaminated with audio from other speakers. The contamination may be more severe in noisy conditions where speaker diarization may perform poorly. Second, to achieve state-of-the-art diarization accuracy [4], significant computational resources must be allocated for the diarization task.

In this paper, we explore techniques to optimize the performance of state-of-the-art speaker recognition when the test session is either 2w or the outcome of an imperfect speaker diarization stage.

The remainder of this paper is organized as follows: Section 2 describes the experimental setup and the baseline systems. Section 3 presents an analysis of running the described speaker recognition systems on 2w test sessions. Section 4 presents techniques in the frame domain for tackling some of the challenges raised in Section 3. Section 5 describes techniques in the Gaussian Mixture Model (GMM)

supervector [6] domain. Finally, Section 6 concludes.

2. Experimental setup and systems

In this paper, we analyze and modify three different speaker recognition systems, which are described in Subsections 2.1-2.3. Subsection 2.4 describes the speaker diarization system we use for some of our experiments. Finally, Subsection 2.5 describes the datasets and protocol.

2.1. GMM-based system

The GMM baseline system was inspired by the GMM-based system described in [7]. The front end is based on Mel-frequency cepstrum coefficients (MFCC). An energy-based voice activity detector is used to locate and remove non-speech frames. The final feature set consists of 13 cepstral coefficients augmented by 13 delta cepstral coefficients extracted every 10ms using a 25ms window. Feature warping [8] is applied with a 300 frame window. GMMs of order 2048 with fixed diagonal covariance matrices are adapted from a universal background model (UBM). The GMM raw scores are normalized using TZ-norm [9].

2.2. Inter-session variability modeling based system

The GMM-supervector-based system with inter-session variability modeling is a modified version of the system described in [9]. The system is based on the parameterization of both training sessions and test sessions with GMMs and embedding the GMMs into a GMM-supervector space (equation (1)):

$$GMM = \{\bar{\mu}_g, \Sigma_g, w_g\} \rightarrow \left(\sqrt{\tilde{w}_1} \Sigma_1^{-\frac{1}{2}} \bar{\mu}_1, \dots, \sqrt{\tilde{w}_n} \Sigma_n^{-\frac{1}{2}} \bar{\mu}_n \right) \quad (1)$$

where $\bar{\mu}_g, \Sigma_g, w_g$ are the mean, covariance, and weight of the g^{th} Gaussian of the GMM and \tilde{w}_g is the weight of the g^{th}

Gaussian of the UBM. The above processing is done using the same front-end and GMM architecture as described in Subsection 2.1. Each target speaker is modeled as a multivariate normal distribution in the GMM-supervector space. The covariance matrix is shared among all speakers of the same gender, and estimated from a development set. Scoring is done by computing the likelihood of the supervector (extracted from the test session) given the probability density function of the target speaker. Final normalization is done as described in Subsection 2.1.

2.3. GMM-NAP-SVM based system

The GMM-NAP-SVM supervector system is quite similar to the one described in [10]. This system uses the same front-end processing as described above and similar GMM architecture, except for the GMM order, which is 512. The GMM-NAP-

SVM was introduced as an evaluation framework for the virtual 2w training method described in Section 5.

2.4. Speaker diarization system

For reference, a speaker diarization system based on Self Organizing Maps (SOM) [11] was used to segment the 2w test sessions. We denote this test condition as 2w+seg.

2.5. Datasets and protocol

The experiments reported in this paper consist of male target speakers only. The NIST-2004 SRE corpus [1] was used as a development dataset. The development dataset was used for training the gender independent UBM (240 sessions), inter-session variability modeling, NAP and SVM training (124 male speakers, 2507 sessions in total). 100 male speakers were used for T-norm, and 100 male sessions (from 100 speakers) were used for Z-norm. Note that some speakers were used for both inter-session variability modeling, NAP/SVM training and T-norm/Z-norm modeling.

The NIST-2005 SRE corpus [12] was used for training target speakers and for test data. The set of target speakers consists of 181 male speakers. For each target speaker, two 4w sessions were randomly selected from the official NIST training lists and used for training. The deviation from the official NIST protocol (using two sessions for training) is justified since we focus on noisy conditions and the recognition performance obtained with single-conversation models does not seem to be useful in such applications (for the 2w test scenario). For our 2w experiments, we artificially summed the two sides of the 4w conversations to get the corresponding 2w sessions. This methodology was chosen to enable the execution of controlled experiments (4w vs. 2w).

For testing, we used the male subset of the core 4w dataset. For 2w experiments, we again summed the two sides of the conversations. The total number of test sessions is 1013. Contrary to the NIST protocol, every target speaker is scored against every test session. For the 4w experiments, this results in male-male experiments only. For the 2w test condition, this results in experiments consisting of at least one male in each 2w test session.

3. Analysis

Training speaker models on 4w data and testing on 2w data using algorithms designed for pure 4w data is suboptimal and results in degraded accuracy. In this section, we examine different aspects of the recognition algorithms described in Section 2 with respect to 2w testing.

3.1. Baseline results

Table 1 presents the equal error rate (EER) and minimal DCF [12] for the baseline GMM, GMM-supervector with inter-session variability modeling, and GMM-supervector-NAP-SVM systems. EER and minimal DCF approximately double when using blind speaker diarization (2w+seg test), compared to 4w testing (except for the GMM system where degradation is more moderate). For 2w testing without a preceding speaker diarization phase, another doubling in EER (except for the GMM system where the degradation is again more moderate) and 50% degradation in minimal DCF is observed.

Table 1. Results for the baseline systems: GMM, GMM-supervector intersession modeling, and GMM-supervector-NAP-SVM. Training condition is 4w.

System	EER (%) 4w test (minDCF)	EER (%) 2w test (minDCF)	EER (%) 2w+seg test (minDCF)
GMM	8.8 (0.0314)	20.2 (0.0724)	14.1 (0.0462)
GMM + inter-session modeling	3.9 (0.0135)	13.9 (0.0478)	7.8 (0.0316)
GMM+NAP+ SVM	3.6 (0.0135)	14.4 (0.0459)	7.0 (0.0328)

3.2. Feature warping

Feature warping is the process of normalizing a frame with respect to its adjacent frames. For 2w sessions, some of the adjacent frames may belong to a different speaker, and therefore the normalization will be different than intended. To assess the degradation accounted to feature warping, we ran a cheating experiment by replacing each 2w session with a concatenation of its two 4w sides (in the audio domain). For the GMM system, we received an EER of 15.9%. Our conclusion is that roughly 40% of the observed degradation on 2w testing is due to inadequate feature normalization. Note that a similar problem is expected for other normalization methods such as cepstral mean removal and variance normalization.

3.3. GMM scoring

In the 2w condition, the GMM-based system scores frames that belong to more than a single speaker. Assuming one speaker is the target speaker (with α as the fraction of the frames spoken by the target speaker), the frames belonging to the second speaker modify the original score S_T (what we would get in a 4w scenario) by averaging it with a (usually lower) score S_N :

$$\tilde{S}_T = \alpha S_T + (1 - \alpha) S_N. \quad (2)$$

Given an estimate of α and S_N , one could estimate S_T from the calculated score \tilde{S}_T .

3.4. GMM supervector parameterization

Similarly to GMM scoring, the frames that do not belong to the target speaker contaminate the calculated supervector. The contamination is non-linear: let $\bar{\mu}_g^1, \bar{\mu}_g^2$ be the means of the g^{th} Gaussians of the GMMs corresponding to the first and second speakers (in a 4w framework), and w_g^1, w_g^2 be the corresponding weights. The observed mean vector in the 2w framework would therefore be $\frac{w_g^1}{w_g^1 + w_g^2} \bar{\mu}_g^1 + \left(1 - \frac{w_g^1}{w_g^1 + w_g^2}\right) \bar{\mu}_g^2$.

The mixing factor $\frac{w_g^1}{w_g^1 + w_g^2}$ is Gaussian dependent; therefore the overall mixing is non-linear.

3.5. Modeling in supervector domain

The non-linearity described in the previous subsection implies that inter-session variability modeling techniques such as those described in Section 2 will not perform as good as they perform in the 4w test condition. Indeed, as seen in Table 1, the EER reduction when using inter-session variability modeling is 56% for the 4w condition and only 31% for the 2w case. SVM classification in the supervector domain also suffers from the mismatch between training in the 4w condition and testing in the 2w condition.

4. Frame-based techniques

In this section, we examine frame-based techniques for improving the performance of the speaker recognition systems described in Section 2 for the 4w-2w (i.e., training in 4w, testing in 2w) case.

4.1. Top-F scoring

In the GMM scoring framework, equation (2) implies that knowledge of the fraction of frames spoken by the target speaker (α) and the average second speaker score (S_N) is sufficient for canceling the second speaker's effect on the GMM score. In practice, we set α to be a fixed constant. S_N is estimated by calculating the average score of the bottom- n scoring frames (n denotes the total number of frames). Note that the scores must be normalized by subtracting the UBM score before the selection of lowest scoring frames to get improved accuracy. Since removing the bottom scoring frames is equivalent to retaining the top scoring frames, we name this technique top-F scoring. The results are presented in Table 2. Contrary to the results in [5], trying to smooth the scores (with adjacent scores) before score selection did not give as good results as without any smoothing.

Table 2. Results for the GMM-based system on the 4w-2w task using top-F scoring.

System	EER (%)	minDCF
GMM baseline	20.2	0.0724
GMM top-30%	17.9	0.0592

4.2. Frame-weighted-based scoring

Given a GMM $G = \{\bar{\mu}_g, \Sigma_g, w_g\}$ trained for a target speaker, a UBM $\{\bar{\mu}_g, \Sigma_g, \tilde{w}_g\}$, and a 2w test session $X = \{x_1, \dots, x_F\}$ of length F , we approximate the GMM-based log-likelihood ratio (LLR) of frame x_f by the LLR obtained from the most likely Gaussian in G , as shown in equation (3).

$$LLR(x_f) \approx \log \frac{\Pr(x_f | \bar{\mu}_i, \Sigma_i, w_i)}{\Pr(x_f | \bar{\mu}_i, \Sigma_i, \tilde{w}_i)} \quad (3)$$

$$i = \arg \max_g \Pr(x_f | \bar{\mu}_g, \Sigma_g, w_g)$$

We model a target speaker by assuming that x_f was generated by the target speaker with probability $p_i = w_i / (w_i + \tilde{w}_i)$ and by the second speaker with the probability $1 - p_i$. The LLR for the first case is the same as in equation (3). For the second case, we assume the LLR is zero (as it reveals no relevant information). The expected LLR for frame f is therefore $p_i LLR(x_f)$.

A second approach is to optimize accuracy in the 2w testing condition by finding a set of weights $\{\alpha_1, \dots, \alpha_F\}$, $\sum \alpha_f = 1$, that will maximize the discrimination of the weighted LLR between the target speaker and imposters:

$$LLR'(X) = \sum_f \alpha_f LLR(x_f) \quad (4)$$

It can be shown with reasonable assumptions and approximations that $\alpha_f \propto \max\left\{\frac{w_i - \tilde{w}_i}{w_i + \tilde{w}_i}, 0\right\}$ is optimal [14]. Results for the two methods are shown in Table 3.

Table 3. Results for the GMM-based system using frame-weighted scoring optimized for 2w testing.

System	EER (%) (minDCF)
GMM baseline	20.2 (0.0724)
Expected LLR $\alpha_f = \frac{w_j}{w_j + \tilde{w}_j}$	19.0 (0.0675)
Max. discrimination $\alpha_f = \max\left\{\frac{w_j - \tilde{w}_j}{w_j + \tilde{w}_j}, 0\right\}$	17.6 (0.0589)

4.3. GMM supervector parameterization

The top-F scoring technique described for GMM scoring may also be used for GMM supervector parameterization. However, using this technique may partly reduce the computation efficiency of the supervector-based techniques. Given a target speaker and a 2w test session, the top-F scoring technique may be used in a first pass for frame selection; in a second pass, these frames may be used for GMM parameterization.

5. Model-based techniques

In this section, we examine model-based techniques for improving the performance of the speaker recognition systems described in Section 2 for the 4w-2w case.

5.1. Optimized weighting of supervector scoring

The essence of [6] is that the classic GMM scoring algorithm can be accurately approximated by extracting GMM supervectors from both training and test sessions and computing the following metric:

$$\Pr(X|Q) \cong -\frac{1}{2} \sum_g w_g^P (\mu_g^P - \mu_g^Q)^T (\Sigma_g^Q)^{-1} (\mu_g^P - \mu_g^Q) + C \quad (5)$$

In equation (5), Q and P denote the supervectors for the trained model and the test session respectively, and C is a constant. In a 2w testing condition, a different weighting scheme may be optimal. In order to find the optimal set of weights $\{\alpha_g\}$ for which $-\sum_g \alpha_g (\mu_g^P - \mu_g^Q)^T (\Sigma_g^Q)^{-1} (\mu_g^P - \mu_g^Q)$

is optimal (in the discriminative sense), we perform the following optimization: we estimate the expected value and variance of $S_g = -(\mu_g^P - \mu_g^Q)^T (\Sigma_g^Q)^{-1} (\mu_g^P - \mu_g^Q)$ for the target

speaker and for imposters (in the 2w scenario) and choose α_g to be proportional to $\frac{D}{V}$ where D is the difference between the expectations of S_g with respect to the target and the impostors, and V is the variance of S_g [14]. The results of two methods, assuming different assumptions and approximations (without proofs), are shown in Table 4. The above analysis may also be applied to the intra-speaker inter-session variability (ISIS) modeling framework. The results are shown also in Table 4.

Table 4. Results for GMM-supervector-based systems using a weighting scheme optimized for 2w testing.

System	EER (%) (minDCF) No ISIS	EER (%) (minDCF) ISIS
Baseline	20.4 (0.0751)	13.9 (0.0478)
$\alpha_g \propto \max \left\{ \frac{w_g - \tilde{w}_g}{w_g + \tilde{w}_g}, 0 \right\}$	16.6 (0.0589)	12.4 (0.0433)
$\alpha_g \propto \max \{w_g - \tilde{w}_g, 0\}$	16.0 (0.0540)	12.1 (0.0416)

5.2. SVM training – reducing training mismatch

In order to reduce the mismatch between the training condition (4w) and the testing condition (2w), we can train the SVM on GMM supervectors extracted from 2w conversations. However, this approach results in an additional degradation (EER of 22.3% for 2w-2w compared to the original 14.4% for the 4w-2w condition). This degradation is explained by the fact that important information is lost when the 4w training sessions are replaced by their 2w equivalents.

Instead, we propose to simulate a 2w training condition without losing information contained in the 4w training sessions. This is done by replacing every 4w training session with a set of virtual 2w sessions. Each virtual 2w session is derived from two 4w conversation sides. One side is the original 4w training session (the target speaker side), and the other side is extracted from the background training speakers. In practice, given GMM parameterizations for two 4w conversation sides, the virtual 2w GMM supervector $(\bar{m}_1, \dots, \bar{m}_n)$ is derived according to equation (6):

$$\bar{m}_g = \left(\frac{\sum_g^{-\frac{1}{2}} \alpha w_g^a \bar{\mu}_g^a + (1-\alpha) w_g^b \bar{\mu}_g^b}{\alpha w_g^a + (1-\alpha) w_g^b} \right) \quad (6)$$

where Σ_g is the covariance and $\bar{\mu}_g^a, w_g^a, \bar{\mu}_g^b, w_g^b$ are the means and weights of the g^{th} Gaussian of the GMM parameterizations of the two conversation sides. α is a weighting factor which simulates the target speaker conversation dominance (the fraction of the frames spoken by the target speaker). Preliminary results show a small improvement in EER and minDCF, but this method requires further exploration.

6. Conclusions

This paper focused on the task of speaker recognition in 2w sessions assuming a 4w training condition. The experiments conducted focused on instances where no blind speaker diarization is performed, but the analysis and techniques shown are also applicable where imperfect diarization is

available.

The analysis reported in Section 3 reveals that feature warping is a major source for degradation (for the GMM baseline it accounts for 40% of the overall degradation). Furthermore, inter-session variability modeling techniques suffer from the non-linear properties of the supervector representation with respect to 2w sessions.

We managed to improve the GMM based system significantly by 21% and bring its accuracy close to a corresponding system that performs blind speaker diarization before the GMM-based recognition (EER=16% compared to 14.1%).

We have also shown a significant improvement (EER=13.9% to 12.1%) for the GMM inter-session variability modeling-based system, and intend to apply it under the GMM-SVM-NAP framework.

Finally, for the SVM framework, we developed an approach for reducing the mismatch between the 4w training condition and the 2w testing condition.

7. Acknowledgments

The authors would like to thank Oshry Ben-Harush, Itshak Lapidot, and Hugo Guterman for providing automatic speaker diarization labels, and Reda Dehak, Najim Dehak, and David van Leeuwen for their assistance with the NAP-SVM system.

8. References

- [1] The NIST Year 2004 Speaker Recognition Evaluation Plan, <http://www.ist.gov/speech/tests/spk/2004/>.
- [2] C. Vair, D. Colibro, F. Castaldo, E. Dalmaso, and P. Laface, "Loquendo - Politecnico di Torino's 2006 NIST Speaker Recognition Evaluation System," in Proc. *Interspeech*, 2007.
- [3] D. A. Reynolds and P. Torres-Carrasquillo, "Approaches and Applications of Audio Diarization," in Proc. *ICASSP*, 2005.
- [4] H. Aronowitz, "Trainable Speaker Diarization," in Proc. *Interspeech*, 2007.
- [5] R. B. Dunn, D. A. Reynolds and T. F. Quatieri, "Approaches to Speaker Detection and Tracking in Conversational Speech," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 93-112, 2000.
- [6] H. Aronowitz and D. Burshtein, "Efficient Speaker Recognition Using Approximated Cross Entropy (ACE)," in IEEE Trans. on Audio, Speech & Language Processing, September 2007.
- [7] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," in *Digital Signal Processing*, Vol. 10, No.1-3, pp. 19-41, 2000.
- [8] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in Proc. *ISCA Odyssey Workshop*, 2001, pp. 213-218.
- [9] H. Aronowitz, D. Irony, D. Burshtein, "Modeling intra-speaker variability for speaker recognition," in Proc. *Interspeech*, 2005.
- [10] N. Brummer et al., "Fusion of Heterogeneous Speaker Recognition Systems in the STBU Submission for the NIST Speaker Recognition Evaluation 2006," in IEEE Trans. on Audio, Speech & Language Processing, September 2007.
- [11] I. Lapidot (Voitovetsky), H. Guterman, and A. Cohen, "Unsupervised Speaker Recognition Based on Competition Between Self-Organizing-Maps," in IEEE Trans. on Neural Networks, vol. 13, no.4, pp. 877-887, July 2002.
- [12] The NIST Year 2005 Speaker Recognition Evaluation Plan, http://www.nist.gov/speech/tests/spk/2005/sre-05_evalplan-v6.pdf
- [13] H. Aronowitz, D. Burshtein, and A. Amir, "A session-GMM generative model using test utterance Gaussian mixture modeling for speaker verification," in Proc. *ICASSP*, 2005.
- [14] H. Aronowitz, "Speaker Recognition in Two-Wire Test Sessions – Extended Version," [Online]. Available: <http://aronowitzh.googlepages.com/publicationlist>