Odyssey 2010
The Speaker and Language Recognition Workshop
28 June – 1 July 2010, Brno, Czech Republic

Unsupervised Compensation of Intra-Session Intra-Speaker Variability for Speaker Diarization

Hagai Aronowitz

IBM Research - Haifa
Haifa University Mount Carmel, Haifa 31905, Israel
hagaia@il.ibm.com

Abstract

This paper presents a novel framework for unsupervised compensation of intra-session intra-speaker variability in the context of speaker diarization. Audio files are parameterized by sequences of GMM-supervectors representing overlapping short segments of speech. Session-dependent intra-session intra-speaker variability is estimated in an unsupervised manner, and is compensated using the nuisance attribute projection (NAP) method. The proposed compensation method is evaluated in the context of speaker diarization in two-speaker conversations. A simple and effective two-speaker diarization algorithm is introduced in which speaker diarization is performed in the compensated supervector-space. The proposed diarization algorithm was evaluated on summed telephone conversations and achieved a speaker error rate of 2.8% which is a 54% relative error reduction compared to a baseline BIC-based system. Finally, we evaluate the proposed system on a speaker recognition task in the summed-speech condition where improvement in speaker recognition accuracy is observed using the proposed diarization system.

1. Introduction

In recent years, two major approaches have been proven to be very effective for automatic speaker recognition. The first approach is inter-speaker variability modeling [1-4], and the second approach is inter-session intra-speaker variability modeling [4-8]. Inter-speaker variability modeling (modeling the speaker-space) is used by eigenvoice based systems [1], anchor-modeling [2-3] and joint factor analysis [4]. This approach and has led to improvements in both efficiency and accuracy of automatic speaker recognition. The inter-session intra-speaker variability modeling approach (often referred to as modeling the channel-space), has been proven to be even more effective for improving the accuracy of speaker recognition systems. This approach is used by joint factor analysis which jointly models inter-session intra-speaker variability and inter-speaker variability [4], eigen-channel MAP adaptation [5-6], explicit statistical modeling in GMM-supervector space [7], and nuisance attribute projection (NAP) in the GMM-supervector space [8].

The apparent success of these techniques for speaker recognition has drawn attention to these techniques in the scope of other speech classification tasks such as language identification [9].

In the context of speaker diarization, prior knowledge about the distribution of speaker population (inter-speaker variability) has been exploited for speaker diarization by the anchor modeling approach where spoken segments are projected into a space of reference speaker models named anchor-space [10-11]. In [12-13], factor analysis was successfully used to model the speaker-space and parameterize the processed speech in the speaker-factor space. Overall, the speaker diarization methods in [10-13] do not make use of intra-speaker variability modeling.

Modeling the inter-session component of intra-speaker variability seems to actually degrade accuracy of speaker diarization systems as channel-related information may be beneficial for discriminating between different speakers in a conversation [13]. On the contrary, modeling the intra-session component of intra-speaker variability can potentially improve accuracy for speaker diarization. Such variability can be accounted to the following types: phonetic content, energy level, speech rate, acoustic (speaker intrinsic), and non-speech rate (due to voice activity detection errors).

Indeed, we have found intra-session intra-speaker variability modeling to be useful for speaker diarization in [14] where a framework for explicit modeling of the intra-session component of intra-speaker variability has been introduced. Intra-session intra-speaker variability was modeled using a manually speaker-segmented training corpus and was used to learn an appropriate distance function between speech segments. The modeling was done by embedding segments into a segment-space using kernel-PCA (principal component analysis), followed by explicit statistical modeling of intra-speaker variability in the segment-space. The framework described in [14] leaded to a significant improvement in diarization accuracy in the broadcast domain. However, using such a framework requires the availability of manual speaker-segmentation of a training dataset. A significant channel mismatch between the training data and the test data may reduce the effectiveness of the framework.

In this paper, we propose to model intra-session intra-speaker variability without the need of any training datasets. Instead, intra-speaker variability is modeled on-the-fly in an unsupervised manner. Contrary to [14] where intra-speaker variability was estimated globally (speaker-independently), in this paper we estimate intra-speaker variability separately for each audio session. Unsupervised estimation of intra-speaker variability is possible by exploiting the following assumption: we assume that the characteristics of a speaker (phonetic content, energy level, etc.) change typically faster that the typical rate of speaker identity change (speaker turns).

Finally, we propose a new algorithm for speaker diarization in two-speaker conversations based on GMM-supervectors and using the proposed intra-session intra-speaker compensation technique.

The remainder of this paper is organized as follows: Section 2 describes the proposed technique for unsupervised estimation and compensation of intra-session intra-speaker

variability. In section 3 we describe the proposed algorithm for two-speaker diarization. In section 4 we describe the experimental setup, datasets and results. In section 5 we present speaker recognition results in the summed speech condition using the proposed speaker diarization system. Finally, we conclude in section 6.

2. Intra-Session Intra-Speaker Variability Modeling and Compensation

2.1. Generative Model

We model a speaker *S* in a particular session with a dynamic generative process represented by a time-dependent probability density function (PDF) S_t. S_t represents the PDF (GMM in practice) used to generate the observation at time *t*. We further assume a memory-less process. Therefore, a speaker in a given session is modeled by a single PDF over the GMM space. This is contrary to advanced speaker-recognition approaches which model a speaker in a given session with a single GMM [1-8]. Recently, a similar model has been proposed in the context of speaker recognition in short test sessions [15].

2.2. GMM-supervector parameterization as a front-end

We follow the GMM-supervector parameterization framework taken in our previous work [16-17, 7] and in SVM GMM-supervector-based speaker recognition [8]. According to this framework, both training and test sessions are mapped into a supervector-space using classical MAP (maximum a posteriori) adaptation [18] of a universal background model (UBM) and concatenating the adapted GMM-means in a fixed order. The actual modeling and classification is done in the supervector space. The reason we take this approach (contrary to using factor analysis) is that we do not assume prior information such as the inter-speaker variability covariance matrix or the channel variability covariance matrix which require a proper development dataset for estimation.

In order to adapt the GMM-supervector parameterization approach to the speaker diarization framework, we parameterize the speech signal with a time series of supervectors. The speech signal is divided into evenly spaced overlapping superframes (sequences of frames) of length 1s and with an offset of 100ms (superframe rate is 10/s). We estimate a supervector for each superframe using classical MAP. The parameterization procedure is outlined as following:

GMM-supervector parameterization

1. Define evenly spaced overlapping superframes of length 1s with an offset of 100ms.
2. Estimate a GMM for each superframe by adapting the UBM to the frames of the superframe using classical MAP.
3. Parameterize each superframe with the supervector created by concatenating the means of its estimated GMM.

2.3. Estimating intra-session intra-speaker variability

Definitions:

s_i – Original uncompensated supervector at superframe *i*.
δ_i – Delta supervector between two consecutive supervectors: $\delta_i=s_{i+1}-s_i$
I_i – Intra-session intra-speaker variability at superframe *i*
m_j – Mean supervector for speaker *j*
k_i – Speaker at superframe *i*
r – Ratio between superframe length and superframe offset
p – Probability of speaker identity change (between two consecutive superframes)

We first analyze the simpler setup with no overlap between superframes and assume that there is no speaker change within a superframe. We assume that supervector s_i is a sum of two independent random supervectors as shown in Equation (1).

$$s_i = m_{k_i} + I_i \tag{1}$$

Supervector m_{k_i} is the mean supervector of the supervectors corresponding to speaker k_i, and supervector I_i is the intra-session intra-speaker variability component at superframe *i*.

Given a sequence of supervectors $\{s_i,\}$ δ_i is the difference between two consecutive supervectors. Taking in mind that *p* is the probability of speaker identity change (between two consecutive superframes) we get:

$$\delta_i = \begin{cases} I_i - I_{i-1} & \text{with } \Pr(p) \\ I_i - I_{i-1} + m_{k_i} - m_{k_{i-1}} & \text{with } \Pr(1-p) \end{cases} \tag{2}$$

from which we derive the following expression for estimation of the covariance matrix of the intra-speaker variability:

$$Cov(I_i) = \tfrac{1}{2} Cov(\delta_i) - pCov(m_{k_i}) \tag{3}$$

where $Cov(m_{k_i})$ is the covariance of the speaker-dependent mean supervectors (inter-speaker variability). Updating Equation (3) to handle overlapping superframes with an offset ratio of r (r=10 in our implementation), we get the following modification for Equation (3):

$$Cov(I_i) = \tfrac{r}{2} Cov(\delta_i) - pCov(m_{k_i}). \tag{4}$$

We approximate the covariance matrix of the intra-speaker variability term I_i by assuming that the length of speaker turns is much larger than the superframe rate ($p<<1$) and discard the second term in the RHS of Equation 4:

$$Cov(I_i) \cong \tfrac{r}{2} Cov(\delta_i). \tag{5}$$

In practice, on our conversational dataset (with average speaker turn length equal to 3s) p=1/3. The empirical results in section 4 indicate that the approximation in Equation (5), though not very accurate, is of value.

2.4. Intra-session intra-speaker variability compensation

Similarly to the NAP [8] technique, we assume that most of the intra-speaker variability is confined to a low dimensional affine subspace of the supervector space. We denote by *d* the

dimension of the low-dimensional subspace. PCA is applied to the estimated intra-speaker variability covariance matrix. The eigenvectors corresponding to the d largest eigenvalues are stacked to form matrix U. Projection T defined by $T=(I-UU^t)$ can be now used to compensate the estimated intra-speaker supervector affine subspace.

Furthermore, the original feature space may be compensated using the feature-domain version of NAP compensation (fNAP) [19]. For a given superframe i, the nuisance supervector η_i for supervector s_i can be calculated as following:

$$\eta_i = UU^t s_i \ . \tag{6}$$

Nuisance supervector η_i may be effectively removed from the frames corresponding to superframe i by splitting η_i back to its individual Gaussian components $\{\eta_{i,1},\ldots,\eta_{i,G}\}$ (G denotes the order of the GMM), and subtracting a weighted average of these components from each of the original feature vectors. The weights are set according to the Gaussian occupation probabilities. For a given feature vector o_t in superframe i, the compensated feature vector is:

$$\hat{o}_t = o_t - \sum_g \Pr(g|o_t)\eta_{i,g} \tag{7}$$

where $\Pr(g|o_t)$ is the Gaussian occupation probability of Gaussian g in frame t.

3. Supervector-Based Speaker Diarization in Two-Speaker Conversations

In this section we propose a new algorithm for two-speaker diarization. Two-speaker diarization is a special case of the general speaker diarization task in which it is known a-priori that the number of speakers in each session is exactly two. Two-speaker diarization may be applicable for summed telephone calls in which the availability of only a summed channel is usually due to operational constraints (such as in some eavesdropping scenarios). Another important scenario is when a conversation between two speakers is recorded using a far-field microphone. A comprehensive review of available two-speaker diarization algorithms can be found in [12-13].

3.1. Motivation

Let x and y denote two multivariate normally distributed random variables. Given a non-labeled mixed sample from both x and y, the goal is to classify each single sample to either population x or population y. A 2-dimensional example of this setup is illustrated in Figure 1. In this example, the samples drawn from $x\sim N(\mu_x=(-1,1), \sum)$ are marked by asterisks and the samples drawn from $y\sim N(\mu_y=(1,-1), \sum)$ are marked by circles. The shared covariance matrix $\sum$ is diagonal with $\sum_{1,1}=9$ and $\sum_{2,2}=1$.

It is clear from Figure 1 that without prior knowledge about the distributions of the two classes, it is hard to separate them. For instance, trying to classify using PCA by projecting all samples on the eigenvector corresponding to the largest eigenvalue of the sample covariance matrix is not very successful (Figure 2) due to large intra-class variability along the x-axis. However, if a model of intra-class variability (covariance matrix $\sum$) is given, it may be used to compensate part of the intra-class variability with the hope of not removing most of the inter-class variability. This may be done by removing only a low-dimensional subspace estimated using PCA applied on the covariance matrix $\sum$. In Figure 3, the subspace spanned by the eigenvector corresponding to the largest eigenvalue of $\sum$ is removed. Applying PCA to the covariance matrix of the compensated samples will now lead to more accurate classification.

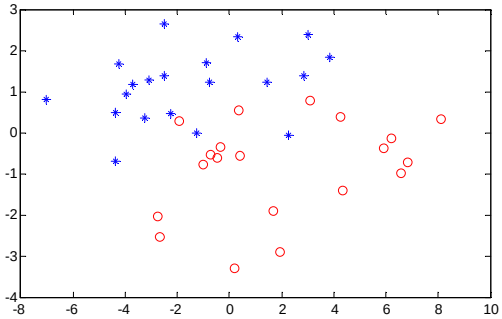

***Figure 1*:** A random sample of two normally distributed 2-dimensional random variables.

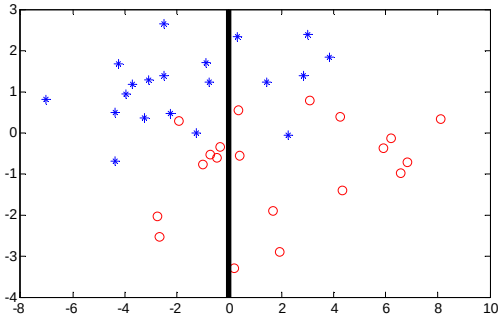

***Figure 2*:** Classification is done using the eigenvector corresponding to the largest eigenvalue of the sample covariance matrix (13 errors).

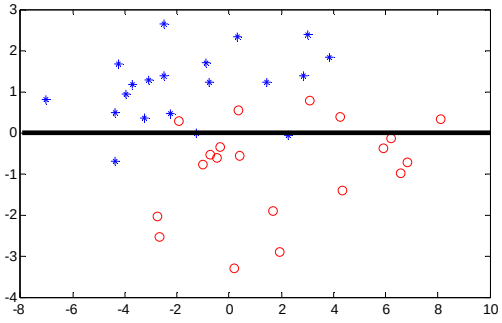

Figure 3: Classification is done using the eigenvector corresponding to the largest eigenvalue of the sample covariance matrix after compensation of the intra-class (7 errors).

3.2. Two-speaker diarization using PCA

Given an audio file we apply the framework outlined as following:

Two-speaker diarization using PCA

1. Compute standard frame-based features without channel normalization.
2. Detect and remove non-speech frames.
3. Estimate a session-dependent UBM (trained from scratch on the current session).
4. Divide the audio session into superframes and estimate a GMM-supervector for each superframe (subsection 2.2).
5. Estimate and compensate intra-session intra-speaker variability (subsections 2.3 and 2.4).
6. Compute the covariance matrix of the compensated supervectors.
7. Apply PCA to find the eigenvector corresponding to the largest eigenvalue of the covariance matrix from step 6.
8. Project each compensated supervector onto the eigenvector found in step 7.
9. The outcome of step 8 is converted to a LLR (log-likelihood ratio) with respect to the two speakers (see subsection 3.3 below) using a linear transformation.
10. Viterbi segmentation is used to convert the superframe-based LLRs into a smoothed segmentation (see subsection 3.4 below).
11. Optionally, perform a few iterations of adaptation and Viterbi re-segmentation in the original feature space (see subsection 3.5 below).

3.3. Converting a projected compensated supervector into a LLR

Definitions:

- c_i – Compensated supervector at superframe i: $c_i=T(s_i)$
- v – The eigenvector corresponding to the largest eigenvalue of the covariance matrix of the compensated supervectors (step 7 in subsection 3.2)
- μ – Mean of compensated supervectors over entire session
- p_i – Compensated supervector at superframe i projected onto eigenvector v: $p_i = v^t(c_i - \mu)$
- m_j – Mean supervector for speaker j
- k_i – Speaker at superframe i
- α – The fraction of speech frames spoken by the first speaker
- Γ – Residual intra-session intra-speaker variability after compensation. A diagonal form $\Gamma=\eta^2 I$ is assumed
- D - Dimension of supervector space

We claim that the outcome of step 8 in the two-speaker diarization algorithm (p_i) is approximately equal to a scaled and shifted LLR with respect to the two speakers. This claim is expressed in Equation (8):

$$p_i \cong a\log\frac{\Pr(c_i|spk_1)}{\Pr(c_i|spk_2)}+b \qquad (8)$$

The correctness of Equation (8) results from an assumption that the covariance matrix of the compensated supervectors is dominated by the speaker mean-supervectors m_{k_i}. This is a reasonable assumption due to the fact that most intra-speaker variability is already compensated, and that the covariance of the compensated vectors is a sum of the contribution of the inter-speaker variability and the residual intra-speaker variability:

$$Cov(c_i)=\alpha(1-\alpha)(m_1-m_2)(m_1-m_2)^t+\Gamma\ . \qquad (9)$$

The eigenvector corresponding to the largest eigenvalue of $Cov(c_i)$ is therefore approximately proportional to supervector m_1-m_2. Consequently we get:

$$p_i \cong \frac{(m_1-m_2)^t}{\|m_1-m_2\|}(c_i-\mu)\ . \qquad (10)$$

with $\mu=\alpha m_1+(1-\alpha)m_2$.

The likelihood of a compensated supervector c_i given speaker j using a multivariate normal distribution with diagonal covariance matrix $\Gamma=\eta^2 I$ is expressed in Equation (11):

$$\Pr(c_i|spk_j)=(2\pi\eta^2)^{\frac{D}{2}}e^{-\frac{(c_i-m_j)^t(c_i-m_j)}{2\eta^2}}\ . \qquad (11)$$

The LLR we are trying to approximate can be expressed by Equation (12) which follows from Equation (11) using simple mathematical manipulation:

$$\log\frac{\Pr(c_i|spk_1)}{\Pr(c_i|spk_2)}=\frac{(m_1-m_2)^t c_i}{\eta^2}-\frac{m_1{}^t m_1-m_2{}^t m_2}{2\eta^2} \qquad (12)$$

Combining Equations (10) and (12) leads to the claim in Equation (8), namely that $p_i \cong a\log\frac{\Pr(c_i|spk_1)}{\Pr(c_i|spk_2)}+b$ with $a=\frac{\eta^2}{\|m_1-m_2\|}$ and $b=\frac{\left(\frac{1}{2}-\alpha\right)\left(m_1{}^t m_1-m_2{}^t m_2\right)}{\|m_1-m_2\|}$.

We conclude that it is possible to estimate the LLR without explicitly estimating the mean supervectors m_1 and m_2. In order to estimate the LLR we need to estimate factors a and b in Equation (8). We assume a balanced speaker distribution (α=0.5) which results in b=0. We further assume that parameter a is fixed across audio sessions. These assumptions may be refined using iterative EM estimation.

3.4. Viterbi segmentation

The LLR described in subsection 3.3 is used by a standard Viterbi segmentation algorithm that models each speaker by a hidden Markov model (HMM). The transition probabilities are derived from a prior estimate of the average speaker turn length, and minimal speaker length is enforced using an appropriate HMM topology. The Viterbi algorithm is used to find a maximum likelihood (ML) segmentation using the

HMM topology, the transition probabilities and the estimated LLRs.

3.5. Viterbi re-segmentation

The first Viterbi pass described in the previous subsection may be optionally refined by a second pass using Viterbi re-segmentation [10]. The first-pass segmentation is used to MAP-adapt a single GMM for each speaker using the original frame-based feature vectors (using the fNAP compensation technique described in subsection 2.4 resulted in an insignificant improvement). The adapted GMMs are used to calculate updated LLRs which are used by the same Viterbi-based segmenter described in the previous subsection. The adaptation-segmentation scheme is iterated for several iterations

4. Experiments and Results

4.1. Datasets and protocol

A subset of the NIST-2005 SRE [20] core dataset was used as an evaluation set (630 sessions), and a disjoint part of the NIST-2005 SRE was used to tune the HMM transition parameters and the parameter *a* required by the LLR calibration method. We artificially convert the stereo datasets to mono by summing both channels. The ground truth was derived from the automatically produced transcripts provided by NIST.

Speech/non-speech segmentation is not the main focus of this work. Therefore, use the standard speaker error rate (SER) measure and do not include speech/non-speech errors. SER is computed according to the standard NIST protocol for evaluation of a two-speaker segmentation task, which is available in [21].

4.2. Baseline BIC-based diarization system

The baseline system is based on the Bayesian Information Criterion (BIC) which is perhaps the most common approach nowadays [12]. Our implementation is inspired partly by the system described in [22]. The outline of the system is as following:

Baseline speaker diarization system

1. Compute standard frame-based features without channel normalization.
2. Detect and remove non-speech frames.
3. Detect speaker change points using BIC [23].
4. Initialize each cluster with a single detected speaker turn.
5. Iterative Viterbi re-segmentation / agglomerative BIC clustering:
   a) Viterbi re-segmentation:
      i. Estimate a 16-component GMM for each cluster.
      ii. Compute a ML segmentation using Viterbi.
   b) Agglomerative BIC clustering:
      i. Estimate a single full covariance Gaussian for each cluster.
      ii. Compute pair-wise distances between each cluster.
      iii. Merge closest clusters.
      iv. Update distances of remaining clusters to new cluster.
      v. Iterate steps ii-iv until a BIC stopping criterion is met.
   c) Iterate step a-b until the number of clusters reaches two.
6. Final Viterbi re-segmentation:
   a) Estimate a 64-component GMM for each cluster.
   b) Compute a ML segmentation using Viterbi.

4.3. Front-end

The front-end used in all the diarization experiments we report is based on of Mel-frequency cepstrum coefficients (MFCC). An adaptive energy based voice activity detector is used to locate and remove non-speech frames. The final feature set consists of 13 cepstral coefficients extracted every 10ms using a 25ms window. The use of feature warping with a 300 frame window tuned to single speaker sessions [26]) and delta MFCC features was also investigated.

The output of the front-end is passed to the speaker segmentation systems. Alternatively, the reference speech/non-speech segmentation from the ground truth was used to assess the sensitivity of the diarization algorithm to speech/non-speech segmentation errors.

4.4. Selected results

Table 1 presents results for three selected systems. The first system is the baseline BIC-based system. The second system is the proposed system with intra-speaker compensation disabled. The third system is the full proposed system. For Viterbi re-segmentation, each speaker is modeled by a 50-state HMM. Transition probabilities are tuned for an average speaker turn of 3 seconds. The GMM order used in these experiments is 64.

Table 1. *SER for the proposed system compared to the baseline.*

System	SER (%)
Baseline BIC-based diarization system	6.1
Proposed system Intra-session intra-speaker compensation disabled	4.8
Proposed system	2.8

The results in Table 1 show a clear advantage to the proposed system compared to the baseline even without compensation of intra-speaker variability (21% relative reduction in SER). The use of intra-speaker variability compensation results in an additional relative reduction of 33% in SER (54% in total).

4.5. Detailed results

In this subsection we analyze the sensitivity of the proposed system to various configurations.

4.5.1. Front-End

Table 2 presents results for varying front-end configurations. The compensation order is 25 and the GMM order is 64. We can conclude from Tables 2 that plain MFCC features give best results. The degradation observed when using feature warping is in-line with the results in [13]. The degradation due to imperfect speech/non-speech segmentation was found to be modest (0.2% absolute).

Table 2. *SER for the proposed system using various front-end configurations (GMM order is 64, NAP order is 25).*

System	SER (%)
MFCC c0-c12	2.8
MFCC c0-c12 + feature warping	3.9
MFCC c0-c12 + delta c0-c12	3.0
MFCC c0-c12 reference speech/non-speech segmentation	2.6

4.5.2. GMM order

Table 3 presents results for various GMM orders. The NAP compensation order is 25. According to these results we choose to use a GMM order of 64 for the rest of our experiments.

Table 3. *SER for the proposed system using various GMM orders (NAP compensation order is 25).*

GMM order	8	16	32	64	128
SER (%)	5.1	3.6	3.3	2.8	2.8

4.5.3. NAP compensation order

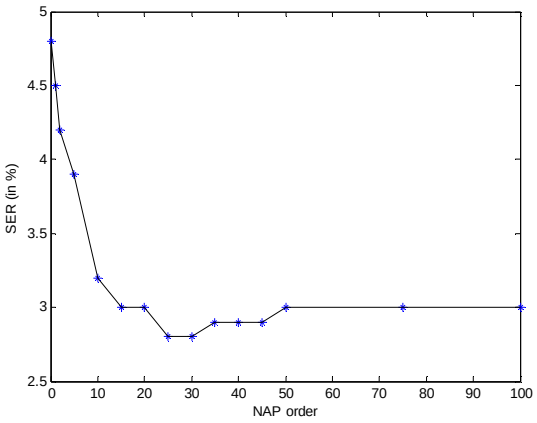

Figure 4: SER for the proposed system using various NAP compensation orders (GMM order is 64).

Figure 4 presents results for various NAP compensation orders. According to these results we choose to use a NAP order of 25. Note that SER is almost insensitive to the NAP order in the range 25-45.

4.5.4. Viterbi re-segmentation

Table 4 presents an analysis of the contribution of the Viterbi re-segmentation step to the overall system. For the proposed system, the Viterbi re-segmentation step reduces SER by 43%. When Intra-speaker variability compensation is disabled, Viteribi re-segmentation reduces SER by 47%.

Table 4. *An analysis of the contribution of the Viterbi re-segmentation step to the overall system.*

System	SER (%)
Proposed system	2.8
Proposed system without Viterbi re-segmentation	4.9
Proposed system Intra-session intra-speaker compensation disabled	4.8
Proposed system without Viterbi re-segmentation Intra-session intra-speaker compensation disabled	9.1

5. Speaker recognition in summed conversations

We compare three speaker diarization sources in the context of speaker recognition where either the training condition or the test condition is summed speech. The diarization sources are manual segmentation obtained from the available automatic transcription, an automatic diarization obtained from our baseline diarization system (SER=6.1%), and an automatic diarization obtained from our proposed diarization system (SER=2.8%). A detailed overview of our speaker recognition system and setup is presented in [24]. In short, our speaker recognition system evaluated in this paper is based on GMM-supervectors trained on warped MFCC features compensated with standard NAP and scored using a linear kernel, followed by standard score normalization.

The experiments reported in this paper were performed on the female core set of the NIST-2005 SRE protocol [20]. We artificially sum the two sides of each original conversation in order to produce a summed conversation (this enables comparison of the results to the original stereo experiments). For each trial, the training conversation is segmented and the target cluster is selected using an automatic comparison to the manual segmentation (best matching cluster is selected). The test conversations are handled in a standard manner by testing on both clusters and selecting the maximal score. Note that the non-standard framework we use is motivated by the characteristics of certain security-related usage scenarios.

Table 5 presents the speaker recognition accuracy measured in equal error rate (EER) and in minimal detection cost function (minDCF) which are defined in [25]. For the manual diarization training condition, the EER is not significantly sensitive to the diarization scheme, which is in-line to the findings in [12]. However, for minDCF we do observe a difference between the schemes. For the condition where both training and testing is performed using automatic diarization, we do observe a significant degradation using the baseline diarization system (EER goes up from 7.0% to 8.9%, and minDCF goes up from 28×10^{4} to 37×10^{3}). This degradation is roughly cut by 50% using the proposed diarization system.

Table 5. *Speaker recognition accuracy on the NIST-2005 SRE core dataset (females only).*

Train diarization	Test diarization	ERR (%)	minDCF x10^3
Manual	Manual	7.0	28
Manual	Baseline	7.1	33
Manual	Proposed	7.0	29
Baseline	Baseline	8.9	37
Proposed	Proposed	7.7	33

6. Conclusions

In this paper a novel approach for speaker diarization is introduced. The main novelty is on-the-fly unsupervised estimation and compensation of intra-session intra-speaker variability. Unsupervised estimation is possible by exploiting the fact that speaker turns, though possibly being short, are still long enough to enable modeling intra-speaker variability which results in a decrease of 42% in SER. In addition, we propose to carry out segmentation using PCA in the GMM-supervector space followed by Viterbi smoothing. Applying these techniques with a final standard Viterbi re-segmentation pass we manage to reduce SER by 54% compared to a conventional approach (BIC-segmentation, bottom-up clustering). The proposed system requires the tuning of only a few parameters and seems to be not very sensitive to these parameters. Finally, we get a significant accuracy improvement in the summed-speech speaker recognition condition using proposed diarization system compared to using the baseline diarization system.

Possible future work is to generalize and evaluate the proposed techniques on more general diarization tasks. A simple approach would be using fNAP (see subsection 2.4) to compensate intra-session intra-speaker variability in the feature domain, and to use this as a preprocessing step before applying standard diarization systems.

7. Acknowledgment

The author wishes to acknowledge the financial support under the HERMES project funded by European FP7 Programme with Contract IST-216709.

8. References

[1] Thyes, O., Kuhn, R., Nguyen, P., and Junqua, J.-C., “Speaker identification and verification using eigenvoices”, in Proc. *ICSLP*, 2000.

[2] Sturim, D. E., Reynolds, D. A., Singer, E. and Campbell, J. P., "Speaker indexing in large audio databases using anchor models," in Proc. *ICASSP*, 2001.

[3] Collet, M., Mami, Y., Charlet, D. and Bimbot, F., “Probabilistic anchor models approach for speaker verification,” in Proc. *Interspeech*, 2005.

[4] Kenny, P., Dumouchel P., "Disentangling Speaker and Channel Effects in Speaker Verification”, In *Proc. ICASSP*, 2004.

[5] Brummer N., “Spescom DataVoice NIST 2004 system description,” in Proc. NIST Speaker Recognition Evaluation, 2004.

[6] Vogt, R. J., Baker, B. J. and Sridharan, S., “Modeling session variability in text independent speaker verification”, in Proc. *Interspeech*, 2005.

[7] Aronowitz H., Irony D., Burshtein D., "Modeling Intra-Speaker Variability for Speaker Recognition", in Proc. *Interspeech*, 2005.

[8] Campbell, W.M. Sturim, D.E. Reynolds, D.A. Solomonoff, A., “SVM Based Speaker Verification using a GMM Supervector Kernel and NAP Variability Compensation”, in Proc. *ICASSP*, 2006.

[9] Noor E. and Aronowitz H., "Efficient language Identification using Anchor Models and Support Vector Machines", in Proc. *Odyssey*, 2006.

[10] D. A. Reynolds and P. Torres-Carrasquillo, “The MIT Lincoln Laboratory RT-04F Diarization Systems: Applications to Broadcast Audio and Telephone Conversations,” in Proc. *DARPA RT04*, 2004.

[11] M. Collet, D. Charlet, F. Bimbot, “Speaker tracking by anchor models using speaker segment cluster information,” in Proc. *ICASSP*, 2006.

[12] Reynolds, D., Kenny, P., and Castaldo, F., “A Study of New Approaches to Speaker Diarization”, in Proc *Interspeech*, 2009.

[13] Kenny, P., Reynolds, D., and Castaldo, F., “Diarization of Telephone Conversations using Factor Analysis”, submitted to IEEE Journal of Selected Topics in Signal Processing, August 2009.

[14] H. Aronowitz, “Trainable speaker diarization”, in *Proc. Interspeech*, 2007.

[15] Vogt, R., Pelecanos, J., Scheffer, N., Kajarekar, S., Sridharan, S., “Within-Session Variability Modeling for Factor Analysis Speaker Verification”, in Proc. *Interspeech*, 2009.

[16] Aronowitz H., Burshtein D., Amir A., "Speaker indexing in audio archives using test utterance Gaussian mixture modeling", in Proc. *ICSLP*, 2004.

[17] Aronowitz H. and Burshtein D., “Efficient Speaker Recognition Using Approximated Cross Entropy (ACE)”, in IEEE Trans. on Audio, Speech & Language Processing, September 2007.

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

[19] Vair, C., Colibro, D., Castaldo, F., Dalmasso, E., and Laface, P., “Channel factors compensation in model and feature domain for speaker recognition,” in Proc. *Odyssey*, 2006.

[20] The NIST Year 2005 Speaker Recognition Evaluation Plan, http://www.nist.gov/speech/tests/spk/2005/sre-05_evalplan-v6.pdf

[21] NIST segmentation scoring script, Available online: “http://www.itl.nist.gov/iad/mig/tests/sre/2002/SpkrSegEval-v07.pl”, 2002.

[22] Zhu, X., Barras, C., Meignier, S. and Gauvain, J.-L., ”Combining speaker identification and BIC for speaker diarization”, in Proc. *Interspeech*, 2005.

[23] S. S. Chen and P. S. Gopalakrishnam, “Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion,” in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, 1998.

[24] Solewicz, Y.A., Aronowitz, H., "Two-Wire Nuisance Attribute Projection", in Proc. *Interspeech*, 2009.
[25] Available online: http://www.ist.gov/speech/tests/sre/.
[26] J. Pelecanos and S. Sridharan, “Feature Warping for Robust Speaker Verification”, in Proc. of *Odyssey*, 2001.