Speaker Diarization Using a priori Acoustic Information

Hagai Aronowitz

IBM Research – Haifa, Haifa, Israel

Hagaia@il.ibm.com

Abstract

Speaker diarization is usually performed in a blind manner without using a priori knowledge about the identity or acoustic characteristics of the participating speakers. In this paper we propose a novel framework for incorporating available a priori knowledge such as potential participating speakers, channels, background noise and gender, and integrating these knowledge sources into blind speaker diarization-type algorithms. We demonstrate this framework on two tasks. The first task is agent-customer speaker diarization for call-center phone calls and the second task is speaker-diarization for a PDA recorder which is part of an assistive living system for the elderly. For both of these tasks, incorporating the a priori information into our blind speaker diarization systems significantly improves diarization accuracy.

Index Terms: speaker diarization, a priori knowledge, speaker segmentation, speaker clustering, within session variability, intra session variability

1. Introduction

Speaker diarization is the task of automatic segmentation of audio sessions into speaker coherent segments (speaker change detection) and associating these segments according to speaker identity (speaker clustering). Most of the published speaker diarization research in the past years has been done under the assumption that no a priori information is available about the identity of the speakers. A currently popular algorithm for blind speaker diarization has been developed by LIMSI [1] and is based on the Bayesian Information Criterion (BIC) [2]. Recently, modern techniques based on speaker recognition innovation, namely joint factor analysis (JFA) and nuisance attribute projection (NAP) were reported to obtain very accurate results for two-speaker blind speaker diarization on conversational telephony [3-5].

However, when dealing with harder tasks such as diarization of broadcast audio and diarization of recorded meetings with multiple participants, accuracy is still an issue. Furthermore, techniques such as JFA require enormous amount of development data and are not easy to port to other languages and channel characteristics.

One of the possible sources for improved accuracy in speaker diarization is exploiting a priori speaker information which may be available. An extreme case is when all the speakers in the audio session are known in advance, and training data is available for each of the speakers. A less extreme case is when a set of possible pre-trained speakers is given and an assumption is made that one or some of the speakers may be speaking in a given audio session. This work is motivated by two specific tasks. The first task is speaker diarization in summed (two-wire) conversations between a known agent and an unknown customer. The second task is speaker diarization for a personal assistance system for the elderly in which conversational audio is recorded using a PDA. These tasks are described in detail in section 3.

Exploitation of a priori speaker knowledge in speaker diarization has been previously addressed in [6] where two cases were explored. For the first case it was assumed that training data was available for all the speakers in a given audio session. For the second case training data was assumed to be available for only a single speaker in a given session. Both cases were addressed by first applying standard techniques to segment the audio into speaker coherent segments, and then by detecting independently for each segment whether it should be associated to one of the pre-trained speakers (using a likelihood ratio test). The resulting segments which were not associated to any pre-known speaker were then handled separately by a clustering stage. However, the authors of [6] have concluded that for the case when training data was available for only one speaker in a given session, only a small gain has been obtained using the a priori information.

In this paper we propose a novel method for integrating a priori acoustic information into standard blind speaker diarization systems which take as input a time series of feature vectors (usually Mel frequency cepstral coefficients). The a priori information is used to modify the standard feature vectors in a way that both information sources (standard features and a priori information) are combined optimally.

The remainder of this paper is organized as follows: Section 2 describes in detail the theoretical model and analysis. Section 3 describes the experimental setup, datasets and results. Finally, Section 4 concludes.

2. Theoretical model and analysis

Speaker diarization algorithms are often based on change detection and clustering algorithms which use one or several scoring functions. In this work we focus on the scoring function component. A scoring function is defined as following: given two segments X and Y (we use the terms segment and cluster interchangeability), a good scoring function $f(X, Y)$ gives a high score when X and Y belong to the same speaker, and a low when they do not belong to the same speaker. In subsection 2.1 we define a scoring function that is motivated by the speaker recognition framework and investigate its properties. In subsection 2.2 we show how to incorporate a priori acoustic information into the speaker recognition-based scoring function. In subsection 2.3 we describe approximations to the function presented in subsection 2.2. Finally, in subsection 2.4 we present how to encode a priori acoustic information in the feature domain. Note that the encoding method developed in subsection 2.4 enables the use of a priori acoustic information under diarization frameworks that do not necessarily use the speaker recognition-based scoring function.

2.1. Speaker recognition-based scoring

A commonly used scoring function is based on the speaker recognition framework. According to this framework, the hypothesis of associating segment X to segment Y is scored by estimating a model M_X from segment X and computing the log-likelihood ratio (LLR) of Y with respect to M_X and with respect to a universal background model (UBM) denoted by *U*. We denote this scoring function by f_{SID}:

$$f_{SID}(X,Y)=\log\frac{\Pr(Y|M_X)}{\Pr(Y|U)}. \tag{1}$$

In [5] we show that segments *X* and *Y* can be appropriately parameterized by GMM-supervectors which we denote by *x* and *y* respectively. Similarly to [5] we assume that the supervectors that correspond to segments of a given speaker distribute normally with a speaker dependent mean and a common (speaker-independent) intra-speaker within-session covariance matrix denoted by Σ. Consequently, we estimate the model for the speaker of segment *X* to be the normal distribution $N(x, \Sigma)$.

ZT-Score normalization [9] is a common practice in speaker recognition for getting improved accuracy. In [7] it has been shown that score normalization is also beneficial for speaker diarization. We therefore normalize the scoring function f_{SID} and denote the normalized function by f_{SIDzt}. For the sake of mathematical simplicity we only normalize the mean Z- and T-statistics. In [9] we have shown that under the above assumptions (normal distribution in supervector space and mean ZT-normalization) the scoring function reduces to

$$f_{SIDzt}(x,y)=(x-u)^t\Sigma^{-1}(y-u) \tag{2}$$

where *u* denotes the UBM supervector.

2.2 Exploiting a priori acoustic information

Let $C_1,\ldots,C_k$ be a set of disjoint acoustic classes (such as distinct speakers). We model each class C_i by a normal probability density function (PDF) P_i over the GMM-supervector space. PDF P_i is parameterized by $\{w_i, \mu_i, \Lambda_i\}$ where w_i denotes the prior probability of a segment to originate from class *i*, μ_i denotes the mean supervector of class *i*, and Λ_i denotes the full covariance matrix (in supervector space) for class *i*. Note that classes $C_1,\ldots,C_k$ do not necessarily cover the whole acoustic space as unknown speakers are expected. We therefore define C_0 as the complementary of the union of classes $C_1,\ldots,C_k$. and approximate its PDF with a universal class PDF denote by P_0.

By taking into account the a priori acoustic classes, the speaker recognition-based scoring function f_{SID} from Eq. (1) turns into the following expression:

$$f_{SID}=\log\frac{\sum_i \Pr(C_i|X)\Pr(Y|X,C_i)}{\sum_i \Pr(C_i)\Pr(Y|C_i)} \tag{3}$$

In the appendix we show that $\Pr(C_i|X)\Pr(Y|X,C_i)$ is equal to $\Pr(\hat{C}_i|X)N(y;\mu_{i,x},\Sigma_{i,x}+\Sigma)$ where $\hat{C}_i$ is a class with a prior probability w_i and a normal PDF $N(x;\mu_{i,},\Sigma+\Lambda_i)$, $\mu_{i,x}=(\Sigma^{-1}+\Lambda_i^{-1})^{-1}(\Sigma^{-1}x+\Lambda_i^{-1}\mu_i)$ and $\Sigma_{i,x}=(\Sigma^{-1}+\Lambda_i^{-1})^{-1}+\Sigma$. The expression $\sum_i \Pr(C_i)\Pr(Y|C_i)$ is dependent on *Y* only (and not on *X*) and therefore it cancels out after ZT-score normalization. Eq. (3) reduced therefore to

$$f_{SID'}=\log\sum_i \Pr(\hat{C}_i|X)N(y;\mu_{i,x},\Sigma_{i,x}) \tag{4}$$

where f_{SID} and $f_{SID'}$ are equal after applying ZT-score normalization.

2.3 Approximations

The expression derived in Eq. (4) may be lower bounded using Jensen's inequality:

$$f_{SID'}\geq\sum_i \Pr(\hat{C}_i|X)\log N(y;\mu_{i,x},\Sigma_{i,x}). \tag{5}$$

We use the lower bound as an approximation to $f_{SID'}$ and denote it by f_{SID*}.

We further assume that all audio classes share the same covariance matrix ($\Lambda_i=\Lambda$) and that the shared covariance matrix is proportional to the intra-speaker within-session covariance matrix: $\Lambda=(\alpha-1)\Sigma$. We therefore obtain:

$$f_{SID*}=\sum_i \Pr(\hat{C}_i|X)\log N(y;\mu_{i,x},\Sigma_\alpha) \tag{6}$$

with $\mu_{i,,x}$=(1-$\frac{1}{\alpha}$)x+$\frac{1}{\alpha}\mu_i$ and Σ_α= (2-$\frac{1}{\alpha}$)Σ. We now compute the ZT-normalized scoring function which we denote by f_{SIDzt*}. We remove from f_{SIDzt} additive terms that are either constant or dependent on either *X* only or *Y* only. Similarly to the arguments borrowed from [9] that lead to our Eq. (2) we get:

$$\begin{aligned} f_{SIDzt*}&=\sum_i \Pr(\hat{C}_i|X)(\mu_{i,x}-E(\mu_{i,x}|\hat{C}_i))^t\Sigma_\alpha{}^{-1}(y-Ey)\\ &=(\hat{x}-u)^t\Sigma_\alpha{}^{-1}(y-u) \end{aligned} \tag{7}$$

with $\hat{x}=(1-\frac{1}{\alpha})x+\frac{1}{\alpha}\sum_i \Pr(\hat{C}_i|X)\mu_i$.

Note that our expression for f_{SIDzt*} results from an approximation (using Jensen's inequality). Due to symmetry, we can get another approximation by exchanging the roles of segments *X* and *Y* in Eq. (7). Furthermore, we can take the average of both approximations and get:

$$f_{SIDzt**}=\tfrac{1}{2}(\hat{x}-u)^t\Sigma_\alpha{}^{-1}(y-u)+\tfrac{1}{2}(x-u)^t\Sigma_\alpha{}^{-1}(\hat{y}-u). \tag{8}$$

We now aim at representing the expression in Eq. (8) as a single inner-product. We define $\widetilde{x}=\frac{1}{2}(x+\hat{x})$ and $\widetilde{y}=\frac{1}{2}(y+\hat{y})$ and reformulate Eq. (8) as follows:

$$\begin{aligned} f_{SIDzt**}&=(\widetilde{x}-u)^t\Sigma_\alpha{}^{-1}(\widetilde{y}-u)-(\hat{x}-\widetilde{x})^t\Sigma_\alpha{}^{-1}(\hat{y}-\widetilde{y})\\ &=(\widetilde{x}-u)^t\Sigma_\alpha{}^{-1}(\widetilde{y}-u)-\tfrac{1}{4}(\hat{x}-x)^t\Sigma_\alpha{}^{-1}(\hat{y}-y) \end{aligned}. \tag{9}$$

We can reformulate the expression in Eq. (9) as a single dot product by extending the adjusted supervectors $\widetilde{x}$ and $\widetilde{y}$ to be $\vec{x}=[\widetilde{x},\frac{1}{2}(\hat{x}-x])$ and $\vec{y}=[\widetilde{y},\frac{1}{2}(\hat{y}-y])$ respectively, and then extending *u* to be $\vec{u}=[u,0]$. $f_{SIDzt**}$ then turns out to be

$$f_{SIDzt^{**}} = (\vec{y} - \vec{u})^t \Sigma_\alpha^{-1} (\vec{x} - \vec{u}). \tag{10}$$

An alternative approximation to Eq. (9) is achieved by discarding the second inner product in the RHS of Eq. (9). We then get

$$f_{SIDzt^{**}} \cong (\tilde{x} - u)^t {\Sigma_\alpha}^{-1} (\tilde{y} - u) \ . \tag{11}$$

2.4 Encoding a priori acoustic information into the feature domain

Following the approximation described in Eq. (11), an observed feature vector at frame *t* denoted by *O(t)* is modified according to the recipe described in Eq. (12), where *x(t)* denotes the supervector extracted for an audio segment centered at time *t*, $\gamma_m(t)$ denotes the UBM Gaussian occupation probability at frame *t*, and $\mu_{i,m}$ denotes the subvector of supervector μ_i corresponding to the *m*-th Gaussian.

$$O(t)' = \left(1 - \tfrac{1}{2\alpha}\right) O(t) + \tfrac{1}{2\alpha} \sum_i \Pr\left(\hat{C}_i \middle| x(t)\right) \sum_m \gamma_m(t) \mu_{i,m} \tag{12}$$

The rational behind Eq. (12) is similar to the one underlying the feature-domain NAP (fNAP) method [8] used to effectively compensate a GMM-supervector indirectly in the feature domain. In our case we want to modify the feature vectors in such a way that the supervector extracted for a segment centered at time *t* (*x(t)*) effectively transforms into $\tilde{x}$. We obtain this by effectively compensating the supervector $x(t) - \tilde{x}(t)$ (which is equal to $\tfrac{1}{2\alpha} \sum_i \Pr\left(\hat{C}_i \middle| x(t)\right)\left(x(t) - \mu_i\right)$) using the fNAP method.

3. Experimental setup

In this section we report results on using the method described in subsection 2.4 for encoding a priori acoustic information into the feature domain. The method is evaluated as a pre-processing stage for two different diarization algorithms which are described in subsection 3.1. We report results on two tasks which are briefly described in subsection 3.2. The comparative diarization results are reported in subsection 3.3.

3.1 Baseline diarization algorithms

The first baseline system is BIC-based and is inspired partly by the system developed by LIMSI [1]. A detailed description of the system can be found in [5]. The system first detects speaker change points using BIC. The resulting segments are clustered using iterations of agglomerative BIC-based clustering and Viterbi re-segmentation. Finally, the diarization is refined using frame-based Viterbi re-segmentation. The system achieved an SER (speaker error rate) of 6.1% on NIST 2005 summed telephone conversations.

The second baseline system is based on unsupervised compensation of intra-speaker within-session variability followed by PCA-based clustering and is described in detail in [5]. The algorithm parameterizes the audio using GMM-supervectors extracted for evenly overlapping 1 second segments. Intra-speaker within-session variability is estimated in an unsupervised manner and removed from the GMM-supervectors using the NAP method. The algorithm exploits the assumption that only two speakers exist in a session by applying PCA to the compensated GMM-supervectors scatter matrix and distinguishing between the two speakers by taking each GMM-supervector and classifying it according to the sign of its projection on the first eigenvector (the one corresponding to the largest eigenvalue). The segmentation is smoothed using Viterbi decoding, and refined by applying Viterbi re-segmentation using the original frame-based features. The algorithm achieves a SER (speaker error rate) of 2.8% on NIST 2005 summed telephone conversations.

3.2 Tasks

This work is motivated by two specific tasks. The first task is speaker diarization in summed (two-wire) telephone conversations between a known agent and an unknown customer. For each agent we have 100 summed conversations for training an acoustic model. These conversations were segmented using a blind speaker diarization algorithm (BIC-based) prior to training. The test dataset consists of 400 sessions, of which 25% are shorter than 30 sec. For each test session the identity of the agent is available for the diarization algorithm. Note that the reference segmentation provided for this dataset is far from perfect. Furthermore, some of the sessions contain a third speaker which is not reflected in the reference segmentation.

The second task is speaker diarization for a personal assistance system for the elderly, in which conversational audio is recorded using a PDA. We use 15 sessions for evaluation. All sessions are about 5 minutes long and contain two speakers only (which differ between sessions).

3.3 Results

Given a test session, we have three testing conditions for each baseline system: no a priori information (baseline), using a single acoustic class equal to the agent/user, and using a single acoustic class equal to a random speaker not present in the conversation. The single parameter (α) was estimated in an unsupervised manner independently for each session. The results are presented in Table 1 for the call-center task and in Table 2 for the personal assistance system. Note that both tasks are much harder than the NIST task. The results in Tables 1, 2 indicate a significant improvement using the a priori speaker information when the one of the speakers is known in advance. In the case where a third speaker is used as an a priori information source, a small improvement is

Table 1. *SER for the call-center task.*

System	Baseline SER (in %)	Agent model SER (in %)	Random speaker SER (in %)
BIC	16.3	12.4	16.2
Supervector-based [5]	14.1	10.2	13.9

Table 2. *SER for the personal assistance system.*

System	Baseline SER (in %)	User model SER (in %)	Random speaker SER (in %)
BIC	28.1	21.5	27.6
Supervector-based [5]	23.9	18.8	23.2

observed, probably due to cases where the third speaker is close (in speaker space) to one of the speakers in the conversation.

4. Conclusions

In this paper a novel approach for using a priori acoustic information for speaker diarization is introduced. We have shown how to adjust the speaker recognition-based scoring function using a priori acoustic information and derived several approximations for the adjusted functions, including one which results with a simple inner product. The inner product approximation was used to develop a method for encoding the a priori information in the feature domain by adjusting the standard frame-based features. The developed method was evaluated for the case of two-speaker diarization when the identity of one of the speakers is known and was found to yield a significant error reduction on two different tasks.

Possible future work is to use acoustic classes other than speakers. Such classes may be channels (mobile / landline), gender and noise characteristics (noisy / quiet). Another important future work is modifying the model to handle audio classes which are not disjoint. Furthermore, it would be interesting to evaluate the system on other setups such as broadcast and meetings.

5. Acknowledgments

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

The author wishes to thank Omer Ziv and Ran Achituv for providing the call-center data.

6. Appendix

Let segments *X* and *Y* correspond to a speaker associated with a normal distribution over the supervector space with an unknown mean μ and a shared within-session covariance matrix Σ. Eq. (13) describes the posterior probability of *Y* given *X* and a prior class C_i.

$$\Pr(Y|X,C_i) = \int \Pr(y|\mu)\Pr(\mu|X,C_i)d\mu \tag{13}$$

Eq. (14) follows from Baye's rule and is used to obtain Eq. (15).

$$\Pr(\mu|X,C_i) = \frac{\Pr(x|\mu)\Pr(\mu|C_i)}{\Pr(x|C_i)} \tag{14}$$

$$\Pr(Y|X,C_i) = \frac{1}{\Pr(x|C_i)} \int \Pr(y|\mu)\Pr(x|\mu)\Pr(\mu|C_i)d\mu \tag{15}$$

Lemma: Given three multivariate normal distributions $f \sim N(\mu_f, \Sigma_f)$, $g \sim N(\mu_g, \Sigma_g)$, and $h \sim N(\mu_h, \Sigma_h)$ the integral of the product of the three distributions is:

$$\int f(x)g(x)h(x)dx = N(\mu_f;\mu_{g,},\Sigma_f+\Sigma_g)N\left(\mu_h;\mu_{fg,},\left(\Sigma_f^{-1}+\Sigma_g^{-1}\right)^{-1}+\Sigma_h\right) \tag{16}$$

A proof of lemma is given in [10]. The lemma is used to obtain the following expression:

$$\Pr(Y|X,C_i) = \frac{1}{\Pr(X|C_i)} N(x;\mu_{i,},\Sigma+\Lambda_i)N(y;\mu_{i,x},\Sigma_{i,x}) \tag{17}$$

with $\mu_{i,x} = \left(\Sigma^{-1}+\Lambda_i^{-1}\right)^{-1}\left(\Sigma^{-1}x+\Lambda_i^{-1}\mu_i\right)$ and $\Sigma_{i,x} = \left(\Sigma^{-1}+\Lambda_i^{-1}\right)^{-1}+\Sigma$.

Consequently, by multiplying both sides of Eq. (17) with the term Pr(C_i |*X*) and applying Bayes' rule once more we get

$$\Pr(C_i|X)\Pr(Y|X,C_i) = \Pr(\hat{C}_i|X)N(y;\mu_{i,x},\Sigma_{i,x}) \tag{18}$$

with $\hat{C}_i \sim N(x;\mu_{i,},\Sigma+\Lambda_i)$ and $\Pr(\hat{C}_i) = \Pr(C_i)$.

7. References

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

[2] 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.

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

[4] P. Kenny, D. Reynolds, and F. Castaldo, "Diarization of Telephone Conversations using Factor Analysis", in IEEE Journal of Selected Topics in Signal Processing, December 2010.

[5] H. Aronowitz, "Unsupervised Compensation of Intra-Session Intra-Speaker Variability for Speaker Diarization", in Proc. Speaker Odyssey, 2010.

[6] D. Moraru, L. Besacier, and E. Castelli, "Using a-priori information for speaker diarization ", Proc. *Speaker Odyssey*, 2004.

[7] H. Aronowitz, "Trainable speaker diarization", in *Proc. Interspeech*, 2007.

[8] C. Vair et al.., "Channel factors compensation in model and feature domain for speaker recognition," in Proc. *Odyssey*, 2006.

[9] H. Aronowitz, V. Aronowitz, "Efficient score normalization for speaker recognition", in *Proc.* ICASSP, 2010.

[10] H. Aronowitz, "The integral of a product of three Gaussians", 2011. [Online]. Available: http://sites.google.com/site/aronowitzh/publicationlist