SPEAKER DIARIZATION AND LINKING OF LARGE CORPORA

Marc Ferràs[1] *and Hervé Bourlard*[1,2]

[1]Idiap Research Institute, Martigny, Switzerland
[2]Ecole Polytechnique Fédérale, Lausanne (EPFL),Switzerland
marc.ferras@idiap.ch, herve.bourlard@idiap.ch

ABSTRACT

Performing speaker diarization of a collection of recordings, where speakers are uniquely identified across the database, is a challenging task. In this context, inter-session variability compensation and reasonable computation times are essential to be addressed. In this paper we propose a two-stage system composed of speaker diarization and speaker linking modules that are able to perform data set wide speaker diarization and that handle both large volumes of data and inter-session variability compensation. The speaker linking system agglomeratively clusters speaker factor posterior distributions, obtained within the Joint Factor Analysis framework, that model the speaker clusters output by a standard speaker diarization system. Therefore, the technique inherently compensates the channel variability effects from recording to recording within the database. A threshold is used to obtain meaningful speaker clusters by cutting the dendrogram obtained by the agglomerative clustering. We show how the Hotteling t-square statistic is an interesting distance measure for this task and input data, obtaining the best results and stability. The system is evaluated using three subsets of the AMI corpus involving different speaker and channel variabilities. We use the within-recording and across-recording diarization error rates (DER), cluster purity and cluster coverage to measure the performance of the proposed system. Across-recording DER as low as within-recording DER are obtained for some system setups.

Index Terms— speaker diarization, speaker linking, agglomerative clustering, joint factor analysis, ward method

1. INTRODUCTION

In the last decades, speech technologies have been faced up to the challenge of dealing with large collections of multimedia data. These corpora typically involve speech in a variety of scenarios including multiple speakers, multiple acoustic conditions, multiple languages, and even emotion or vocal effort variation. For some other corpora, the data being captured is only limited by the users' imagination and the available technology. In practice, the size and variety of recording conditions end up posing new challenges for the speech processing techniques, whilst they are asked to keep an adequate computation time. The speaker diarization task is a technology that is currently quite mature. However, changes in acoustic conditions still can result in a performance drop and the computational cost can become prohitive for long recordings or collections of recordings.

Fortunately, the availability of lots of data can also turn into a valuable source for modeling new phenomena. In speaker recognition technology it is now common to model the inter-session variability of speakers in addition of the speaker variability itself so that more robust recognition systems can be built. In techniques such as Joint Factor Analysis (JFA)[1, 2], large databases involving multiple speakers and multiple sessions per speaker are analyzed to separate the speaker and session effects in the speaker Gaussian Mixture Models (GMM) obtained after adaptation.

In this paper we are interested in performing speaker diarization of a data set involving many recordings, that is, uniquely identify the speakers across the data set and find the set of segments for each recording where each of the speakers is speaking. This task could be solved by simply concatenating all the recordings of the data set and then running a standard speaker diarization system, but this is not a practical or even feasible approach at this time for the volume of data we are targetting. We opt instead for a two-stage approach where the amount of data processed at each stage is compressed. First, a standard speaker diarization system obtains within-recording speaker clusters using a agglomerative clustering at the acoustic observation level. The speaker clusters are given a set of start and end times and a unique speaker identifier within each recording. In the second stage, another agglomerative clustering algorithm whose input are the speaker clusters output by the diarization system is run to structure the speaker space of the data set. Each input speaker cluster is represented as a speaker factor posterior distribution obtained after adaptation of a Universal Background Model (UBM) to the speech data using JFA. The resulting speaker clusters, i.e. clusters of speaker factor posteriors distributions, are then given a unique speaker identifier across the data set. If unique identifiers are correctly assigned, there should be a corresponding improvement of diarization performance, both within-recording, when the number of speakers within the recording has not been correctly determined and across-recording as well. Such a system benefits from adequate processing for each stage. The first stage, within recording, benefits from a UBM fitted to the recording conditions to finely detect speaker differences as well as dealing with a tractable number of speakers. The second stage benefits from a global UBM involving multiple acoustic conditions, JFA inter-session compensation and more data per speaker to obtain the speaker models. We assess the impact of using the unique speaker labels on the Diarization Error Rate (DER) within and across recordings of the full data set as well as on the cluster purity and coverage measures.

Some work related to large scale speaker diarization and speaker linking can be found in the literature. Cluster impurity and cluster entropy measures are proposed in [3] for speaker linking evaluation, although not focusing on system development. In [4] a so-called speaker attribution system that performs speaker linking of the speaker clusters found by a speaker diarization system is proposed. This system is similar to our proposal in that it clusters the speaker clusters agglomeratively. The system uses the complete linking method to compute distances between clusters and the Nor-

This work was supported by the European Union under the FP7 Integrated Project inEvent (Accessing Dynamic Networked Multimedia Events), grant agreement 287872. The authors gratefully thank the EU for their financial support and all project partners for a fruitful collaboration.

malized Cross Likelihood Ratio (NCLR) as the distances between pairs of initial speaker clusters. Cluster and speaker purity measures are given to compare MAP and JFA approaches to adaptation. Targetting large scale speaker diarization is [5], which proposes a multi-stage system involving speaker diarization followed by speaker linking of chunks of speech data. Although splitting the database in small chunks increases diarization error rates, this system scales particurlarly well on large data sets. Only the work in [5] focuses on interview and meeting data, the others targetting telephone speech conversations between two people. In our work, we target a challenging scenario with meetings of 4 participants each recorded using various types of far-field microphones and several recording rooms. Regarding speaker factor posterior distributions, the system proposed in [6] includes the posterior distributions into the Variational Bayes method to perform soft clustering diarization on telephone speech.

The paper is organized as follows: Section 2 describes the speaker diarization system, based on the Information Bottleneck (IB) clustering framework, that we use as a black box in this work. Section 3 gives an overview of JFA, focusing on how the speaker factor posterior distributions are estimated. Section 4 describes how the JFA framework is used to model the speaker segments output by the diarization system, how they are clustered and how unique speaker identifiers across the data set are obtained. In Section 5 the data sets used for experimental evaluation as well as the details about the implemented systems are presented. Section 6 gives some results to validate the proposed techniques and Section 7 gives some conclusions.

2. SPEAKER DIARIZATION

The goal of the speaker diarization task is to split a recording into acoustically homogeneous regions that were spoken by the same speaker, while also determining the number of speakers. After feature extraction and speech activity detection, these systems typically detect boundaries between speaker turns, so-called speaker change detection, and then cluster these segments into speaker clusters across the recording, so-called speaker clustering. Since speaker change detection is straightforward in our system, the recording is uniformly split into 1-2 second long segments that are considered homogeneous due to its short length, we only detail the speaker clustering stage here. Please refer to [7] for more details about the speaker diarization system.

The speaker clustering stage uses an Agglomerative Information Bottleneck (aIB) approach based on information theoretic principles. The IB framework defines a set of relevance variables Y, posterior probabilities of the initial segments with respect to a GMM-UBM in our case, that represent the information to be preserved in the clustering process. If C is a compressed representation of the inital segments X, then the IB principle states that C should preserve as much information as possible about the relevance variables Y. This objective function can be formalized in terms of mutual information as

$$\mathcal{F} = I(Y,C) - \frac{1}{\beta} I(C,X) \quad , \tag{1}$$

where β is a trade-off between the amount of information preserved and the compression from the initial representaion.

The aIB algorithm is a greedy approach to optimize Eq. 1 where the initial segments are iteratively merged by pairs so that the decrease in the objective function is minimum at each merging step. The distance measure between two clusters is a combination of Jensen-Shannon divergences, a measure naturally arising from the maximization of Eq. 1. To infer the number of speakers the system uses the Normalized Mutual Information criterion, $NMI = I(Y,C)/I(X,Y)$, measuring the fraction of original mutual information I(X,Y) captured by the current cluster representation C. The optimal number of speakers is found when the NMI measure is larger than a specified threshold.

Once the clusters have been found, their boundaries are refined using an ergodic HMM with duration constaints.

3. JOINT FACTOR ANALYSIS

Joint Factor Analysis (JFA) [1, 2] is a technique for adaptation of Gaussian Mixture Models (GMM) based on Maximum-A-Posteriori (MAP) estimation that allows for disentangling the speaker and session effects. Assuming the simplified JFA model[1]

$$\hat{\mathbf{m}} = \mathbf{m} + \mathbf{V}\mathbf{y} + \mathbf{U}\mathbf{x} \quad , \tag{2}$$

$\hat{\mathbf{m}}$ and $\mathbf{m}$ are the speaker-adapted and speaker-independent Gaussian mean supervectors of a GMM, i.e. the concatenation of the mean vectors into a single vector. The speaker-independent supervector $\mathbf{m}$ is formed by the mean vectors of a Universal Background Model (UBM) typically trained with data from many speakers. $\mathbf{Vy}$ is a speaker-dependent low-rank term assumed to model speaker variation. $\mathbf{Ux}$ is a session-dependent low-rank term modeling session variation. The factor loading matrices $\mathbf{V}$ and $\mathbf{U}$ are speaker-independent and they are trained off-line using data from many speakers and several session per speaker [2]. $\mathbf{y}$ and $\mathbf{x}$ are the so-called speaker and session factors, assumed to be a priori i.i.d following a normal distribution with zero mean and unit variance. The number of speaker and session factors affects the quality of the adaptation, the more factors the higher the dimensionality of the adapted subspaces.

Training a JFA model consists of fitting the factor loading matrices $\mathbf{V}$ and $\mathbf{U}$ and the latent variables $\mathbf{y}$ and $\mathbf{x}$ to the speech of a database in the maximum-likelihood sense, typically by alternating the estimation of latent variables and loading matrices until convergence. The factor loading matrices are retained and they are used for adaptation, where only latent variables are fit to the adaptation data. Note that as few as the number of speaker and session factors need be estimated to adapt a GMM, whatever the number of Gaussian mixtures. Once all the variables are available, a session-compensated speaker model could be synthesized as $\mathbf{m} + \mathbf{Vy}$.

In the training or adaptation phases, JFA estimates the posterior distribution of the speaker factors. Since they are assumed to be multivariate Gaussian[1] we can characterize the posterior distribution with a mean vector $\mathbf{y}$ and a covariance matrix $\mathbf{C}$, computed as

$$\begin{aligned} \mathbf{y} &= \mathbf{C}\mathbf{b} && (3) \\ \mathbf{C} &= \big(\mathbf{I} + \sum_{g=1}^{G} N_s^g \mathbf{V}^{g,T} \mathbf{\Sigma}^{g,-1} \mathbf{V}^g\big)^{-1} && (4) \\ \mathbf{b} &= \sum_{g=1}^{G} \mathbf{V}^{g,T} \mathbf{\Sigma}^{g,-1} \overline{\mathbf{X}}_s^g && (5) \end{aligned}$$

with G being the number of Gaussian components of the GMM-UBM. For Gaussian mixture g, $\mathbf{V}^g$ is the corresponding submatrix of $\mathbf{V}$ and Σ^g the corresponding covariance matrix. $\overline{\mathbf{X}}_s^g$ are the first

[1] We dropped the diagonal speaker term $\mathbf{Dz}$ typically used in JFA.

order statistics that account for the term $\mathbf{Vy}$. They are computed by removing the UBM and session effects from the first order statistics $\mathbf{X}_s^g$ as

$$\overline{\mathbf{X}}_s^g = \mathbf{X}_s^g - N_s^g \mathbf{m}^g - \sum_{h \in s} N_{h,s}^g \mathbf{U}^g \mathbf{x} \qquad (6)$$

where N_s^g and $N_{h,s}^g$ are the expected number of frames assigned to Gaussian g for speaker s and session h and $\mathbf{U}^g$ is the corresponding submatrix of $\mathbf{U}$. When only one session is available for adaptation we use $N_s^g = N_{h,s}^g$. Equations analogous to Eq. 3 are used to estimate session factors.

4. SPEAKER LINKING

The goal of the speaker linking system is to uniquely identify the speakers output by the speaker diarization system for all the recordings in the data set. The agglomerative clustering and labeling steps are discussed in the following:

4.1. Agglomerative clustering

We use an agglomerative clustering algorithm to group similar speaker clusters from the output of the speaker diarization system. Linking speaker clusters from different recordings is a challenging task, but it also benefits from two advantages over wihtin-meeting speaker diarization: (a) The speech data in the data set can be analysed as a whole. In particular, the adapted models can be compensated for session variation via JFA. (b) The amount of data used for speaker modeling can be significantly larger if the speaker appears in more than one recording. This surely has a positive impact in the quality of the adapted models.

The speech data of each speaker cluster is modeled as a single multivariate Gaussian with full covariance matrix, which is indeed the speaker factor posterior distribution estimated by JFA given the speech data and a GMM-UBM (see Section 3). These are the objects that the speaker linking algorithm is clustering. We follow a standard approach to agglomerative clustering: each initial cluster is assigned one speaker cluster. The two closest clusters are then successively merged. To keep the whole clustering dendrogram we stop the merging process when only one cluster remains:

1. **Compute the distance matrix** for all pairs of speaker clusters, that become the initial clusters.
2. **Merge** the two closest clusters.
3. **Update the distance matrix**, from the merged cluster to all other clusters.
4. **Go to 2.** If only one cluster remains, **stop**.

A key point of agglomerative clustering algorithms is the linking method, or how to measure the distance between two clusters at some stage in the clustering process. We use Ward's method[8], which merges the two clusters that result in the minimum increase of the total within-cluster variance after merging, i.e. it aims at obtaining compact clusters. Ward's method is typically implemented in a recursive manner using the Lance-Williams algorithm[9]. When two clusters c_i and c_j are to be merged, the distances between the merged cluster c_{ij} and all other clusters c_k are updated using the recursion

$$d_{(ij)k} = \alpha_i d_{ik} + \alpha_j d_{jk} + \beta d_{ij} \qquad (7)$$

with d_{ij} being the distance from cluster i to cluster j, $\alpha_i = \frac{n_i+n_k}{n_i+n_j+n_k}$, $\alpha_j = \frac{n_j+n_k}{n_i+n_j+n_k}$, $\beta = \frac{n_k}{n_i+n_j+n_k}$ and n_i is the number of samples in cluster c_i. The number of samples for the initial clusters is taken as the number of feature vectors used in the estimation of speaker factor posterior distributions.

4.2. Cluster Dissimilarity

Although the Lance-Williams recursion is stricly valid for initial distances that are proportional to the squared Euclidean distance, we use it with other dissimilarity measures as well. Assuming we are comparing two F-dimensional[2] multivariate Gaussian distributions $p_i \sim \mathcal{N}(\mathbf{y}_i, \mathbf{C}_i)$ and $p_j \sim \mathcal{N}(\mathbf{y}_j, \mathbf{C}_j)$:

- The **cosine distance** is a widely use metric in the speaker recognition community to compare speaker and total factor mean vectors estimated via JFA or Eigenvoice MAP estimation. It has been noted in [10] that the resuting scores are stable to the point that the derived recognition systems do not require any score normalization. The distance measure is taken from the normalized projection of two vectors, $\mathbf{y}_i$ and $\mathbf{y}_j$ as

$$d_{cos}(\mathbf{y}_i, \mathbf{y}_j) = 1 - \frac{\mathbf{y}_i^T \mathbf{y}_j}{||\mathbf{y}_i||\,||\mathbf{y}_j||} \qquad (8)$$

- The **symmetrised Kullback-Leibler divergence** is an information theoretic measure of dissimilarity between two distributions. The KL divergence measures the amount of information required to encode samples of a distribution using a code based on another distribution. By using the symmetrised KL divergence, $d_{skl}(p_i, p_j) = d_{kl}(p_i, p_j) + d_{kl}(p_j, p_i)$, which is non-negative and symmetric, only half of the elements of the distance matrix need to be computed. The closed form of the KL divergence for multivariate Gaussian distributions can be written by

$$\begin{aligned} d_{kl}(p_i, p_j) =& \frac{1}{2}\Big(\mathrm{tr}(\mathbf{C}_j^{-1}\mathbf{C}_i) + (\mathbf{y}_i - \mathbf{y}_j)^T \mathbf{C}_j^{-1} (\mathbf{y}_i - \mathbf{y}_j) \\ & - \ln \frac{|\mathbf{C}_i|}{|\mathbf{C}_j|} - F\Big) \end{aligned} \qquad (9)$$

- The **two-way Hotteling t-square statistic** is the multivariate equivalent of the two-way Student t statistic. It is used for testing the hypothesis that the means of two samples assumed to be Gaussian distributed with equal covariance matrices are different. The statistic is written by

$$d_{ttest}(p_i, p_j) = \frac{n_i n_j}{n_i + n_j} (\mathbf{y}_i - \mathbf{y}_j)^T \mathbf{C}_{pool}^{-1} (\mathbf{y}_i - \mathbf{y}_j) \qquad (10)$$

with

$$\mathbf{C}_{pool} = \frac{(n_i - 1)\mathbf{C}_i + (n_j - 1)\mathbf{C}_j}{n_i + n_j - 2} \qquad (11)$$

In the hypothesis test, Eq. 10 is typically transformed into an F statistic that is evaluated against an F distribution to obtain a p-value, the probability of rejecting the hypothesis that the two mean vectors are the same. Since the p-values vanish when computed with large n_i and n_j, we use the statistic of Eq. 10 as the dissimilarity measure between two clusters. Under the assumption that both Gaussian distributions share the same covariance matrix, this measure has the form of the Euclidean distance between spherified Gaussian distributions, therefore matching the assumptions of the Lance-Williams recursion.

[2] F is the number of speaker factors.

4.3. Speaker labelling

It is expected that speaker clusters naturally arise during the agglomerative clustering process. As shown in [10], speaker factor mean vectors exhibit very good discrimination amongst speakers. They are also inherently normalized by the priors, which is likely to render them more comparable amongst speakers. In this work, we assume the speaker clusters can be simply found by thresholding the distance values in the clustering dendrogram obtained as described in Section 4.1. For parent node p and child node c in the dendrogram, if $d_p > th$ and $d_c < th$, all descendants including node c are assigned the same global speaker identifer.

5. EXPERIMENTAL SETUP

To evaluate the proposed system, we ran experiments to compare the performance of speaker diarization system alone versus the speaker diarization plus linking system, what we call *full* speaker diarization system.

We use the same frontend for the whole system, extracting 19 MFCC features every 10ms using a 30ms window. No delta or acceleration coefficients are used.

The speaker diarization system relies solely on the data of each recording to do the speaker partitioning. No training data other than the recording itself is required. We used 2.5s long segments for the initial segmentation. The IB trade-off parameter β was 10. Since the number of participants in the AMI meetings is 4 and the maximum number of speaker clusters is set to 10, the system tends to under-cluster, i.e. find more speaker clusters than actual speakers. The NMI threshold for speaker detection was 0.3. These settings were optimized for meeting data from the NIST RT'06 evaluation.

We use the speech data collected for the Augmented Multiparty Interaction (AMI) project for training the speaker linking system. JFA adaptation requires a GMM-UBM that we trained using around 50 hours of far-field array data from the ES, IS and TS meetings in the AMI corpus. We use a gender-independent 512 Gaussian mixture UBM and Maximum Likelihood (ML) estimation.

The JFA factor loading matrices $\mathbf{V}$ and $\mathbf{U}$ were trained using speech data involving 132 speakers from 4 far-field microphone channels per meeting, using the ES, IS and TS meetings. These meetings are recorded in different rooms using different microphones, with a total of 12 different channels is present in this data set, plus the speaker-to-microphone placement which is unknown as well as microphone placement changes. To estimate the loading matrices we used a decoupled estimation scheme with 10 iterations of ML estimation for training the JFA model. For adaptation, the speaker and session factor posterior distributions were jointly estimated for each speaker cluster hypothesized by the speaker diarization system. All the available speaker factors, i.e. 132, and 20 session factors were used after informal optimization on the AMI8 data set, described below.

The speaker diarization systems were evaluated on the three following data sets:

- **AMI8**: involves 8 speakers, 18 meetings, 4 acoustic channels, 1 room, 135 speaker clusters output by the diarization system to be linked. This is a small development data set used to analyze the behaviour of the system and to tune the system parameters.
- **AMI56**: involves 56 speakers, 146 recordings, 56 meetings, 4 channels, 1 room, 1044 speaker clusters to be linked. This is an evaluation data set with larger speaker variability but recorded on the same room as AMI8.
- **AMI56CH**: involves 56 speakers, 181 recordings, 85 meetings, 12 channels, 3 rooms, 1262 speaker clusters to be linked. This is the evaluation data set with the largest speaker and session variability.

5.1. Performance measures

We use several measures to evaluate the performance of the speaker diarization and the full speaker diarization systems. The Diarization Error Rate (DER) assesses the within-recording and across-recording impact of speaker linking on the diarization systems. When speaker diarization systems detect more speakers than the actual number of speakers, the within-recording DER (wrDER) assesses the effect of grouping speakers within the recording that were considered the same by the speaker linking system. wrDER uses the references obtained by forced alignment of ASR transcripts with speakers labeled with unique identifiers within the recording. The wrDER also allows us to directly compare the output of the diarization and full diarization systems, although only the within-recording improvement can be observed. We use the across-recording DER (arDER) to assess the DER for the data set as a whole. For this purpose, we concatenate the references of all recordings in the data set as if it were a single recording with the within-recording speaker identifiers replaced by unique speaker identifiers across the data set. To compute the DER, a one-to-one mapping between the set of reference and system speaker identifiers is performed first. Next, the ratio of the number of frames with reference speaker not matching the mapped system speaker to the total number of frames is computed. For all DER computations we use a collar of 250ms.

For the full diarization systems we also compute cluster purity and cluster coverage measures. Given a particular cluster, the cluster purity is defined as the ratio of the number of frames assigned to the dominant speaker over the total number of frames of the cluster. Conversely, for a given speaker, the cluster coverage is computed as the ratio of the number of frames of the dominant cluster to the total number of frames of that speaker. We give average values over the data set for both measures.

Note that, since the speaker linking task is indeed an identification task, the performance is dependent on the number of speakers that are being identified, the more speakers the higher the arDER.

6. EXPERIMENTS AND RESULTS

The first set of experiments is aimed at exploring the behaviour of the clustering and labelling steps of the speaker linking system and assess their impact on their performance. Table 1 gives the results for these experiments. Two types of systems, Dia and FullDia are shown corresponding to the standard and full speaker diarization systems. For the FullDia systems we tested the discussed dissimilarity measures above. We use the optimal a posteriori threshold that minimizes the wrDER.

All FullDia systems obtain lower wrDER than the Dia baseline. These gains are due to speaker segments in each recording being clustered as the same global speaker. Note, however, that the number of speakers detected is far from 8 for the systems using the cosine and skl dissimilarity measures. For these two systems, raising the threshold led to a smaller number of speakers, but for the right number of speakers a bunch of recordings obtained much worse wrDER compared to the baseline system. Conversely, lowering the threshold resulted in a slight wrDER gain although only few recordings were given a different speaker assignment compared to the baseline system. In contrast, the system using the ttest distance measure exhibited more stability, being able to smoothly balance the trade-off between wrDER and number of speakers. For the optimal a posteriori threshold, this system finds the right number of speakers and it simultaneously optimizes the wrDER, reaching over 40% of DER relative improvement over the baseline. This large gain is probably due to the fact that the ttest distance is proportional to an Euclidean distance once the pooled covariance has been spherified. In this case, the distance measure is matched to the Gaussian input data as well as Ward's method and the Lance-Williams algorithm assumptions.

Fig. 1 shows dendrograms for the FullDia systems of Table 1. For the FullDia ttest system of Fig. 1(right), a big distance gap between merged nodes is present after a speaker cluster has been found. For the cosine distance measure of Fig. 1(left) these gaps are much more gradual and a clear cutting threshold can not be visually identified. For the FullDia skl system of Fig. 1(center), there seems to be a set of meetings on the left part of the graph for which the threshold seems to work whereas the distance explodes for another set of meetings on the right part of the graph. In this case, further tuning of the threshold only increased the wrDER.

For the FullDia cos and FullDia skl systems, the threshold that optimizes the wrDER, even though it results in improvements, it turns out to be rather bad in terms of full diarization, with arDER becoming three times worse than wrDER. This supports the idea that these systems are not able to identify natural clusters, at least using a single threshold. arDER for the FullDia ttest system stays the same as wrDER, meaning that the optimal speaker assignment was found in this case. Cluster purity and especially cluster coverage measures are significantly larger for the FullDia ttest system.

System	Th.	#Spk	wr/ar DER(%)	Cp/Cc(%)
Dia	—	—	14.5/	—
FullDia cos	0.6	21	11.2/36.9	69.6/53.2
FullDia skl	3e4	17	12.6/33.6	69.2/53.3
FullDia ttest	0.25	**8**	**8.5**/**8.5**	**75.0**/**74.2**

Table 1. Speaker diarization experiments on the AMI8 data set involving 8 speakers. The type of diarization system and the dissimilarity measure are shown in the first column. The optimal a posteriori threshold, the detected number of speakers, the within-recording/across-recording DER, the average cluster purity and average cluster coverage are shown in the remaining columns. The best results use bold typeface.

Table 2 shows results for the AMI56 data set which has more speaker variability but around the same session variability. Patterns similar to those found in the previous experiments can be identified here too. The use of cosine and skl dissimilarity measures result in some wrDER gains compared to the baseline system, but not as much as using the ttest distance. In terms of full diarization, the arDER for the FullDia cos and FullDia skl systems is about 50% larger than their corresponding wrDER. However, the FullDia ttest system obtains comparable wrDER and arDER figures, with the arDER being slightly better than the baseline wrDER. Although wrDER and arDER use different speaker assignments and comparing them is not strictly correct, these similar figures suggest that the speaker linking system is doing a good job considering that full speaker diarization is a considerably more difficult task. Cluster purity and coverage measures are in the line of the previous experiments with a clear improvement for the FullDia ttest system.

Figs. 2 show the histograms of the absolute improvement of wrDER across all the meetings in the AMI56 data set for the three FullDia systems of Table 2. All the systems resulted in mostly improvement, as revelead by the asymmetry of the distributions towards the right side and around gain 0. For the cosine and skl dissimilarity measures, a large proportion of the meetings did not change the speaker assignment. In this case, this led to better wrDER than further assigning the wrong speakers. Note that losses from -10% to -27% wrDER are observed in three recordings for the FullDia cosine system. For the FullDia ttest system, out of the 146 recordings, only 5 recordings are left with the original speaker assignment, 131 recordings obtain gains and for 10 recordings the new speaker assignment results in some loss. However, these losses are bounded at around -2.5% absolute.

System	Th.	#Spk	wr/ar DER(%)	Cp/Cc(%)
Dia	—	—	24.5/	—
FullDia cos	0.6	165	23.4/36.9	63.0/55.1
FullDia skl	3e4	90	23.6/33.4	63.3/59.0
FullDia ttest	0.25	**58**	**21.7**/**23.6**	**69.8**/**70.2**

Table 2. Speaker diarization experiments on the AMI56 data set involving 56 speakers. The same columns of Table 1 are shown.

System	Th.	#Spk	wr/ar DER(%)	Cp/Cc(%)
Dia	—	—	27.6/	—
FullDia cos	0.6	247	**26.0**/38.9	62.3/56.4
FullDia skl	3e4	121	27.3/33.0	60.7/64.0
FullDia ttest	0.2	**86**	26.8/**28.0**	**67.5**/**72.8**

Table 3. Speaker diarization experiments on the AMI56CH data set involving 56 speakers. The same columns of Table 1 are shown.

Table 3 shows the results for the AMI56CH data set, which has the largest channel variability of the three data sets. Although this data set is more difficult than AMI56 we noticed that the same thresholds in the AMI56 experiments roughly optimized the wrDER here too. Absolute wrDER are overall higher than those obtained for the AMI56 data set, e.g. 27.60% versus 24.50% for the baseline Dia system. This data set involves recoding in different sites that use different recording setups as well as slightly different structure of the meeting. The FullDia systems give some slight improvement over the baseline, around 5% relative for the FullDia cos system, meaning that the speaker clusters within the meeting are still properly linked to some extent. However, the number of detected speakers is much larger than the actual number of speakers, 56. This has a negative impact on the arDER rate. The FullDia cos and FullDia skl detect twice and five times the actual number of speakers respectively. Regarding the FullDia ttest system, 86 speakers were detected, i.e. around 50% more than the actual number of speakers. However, note that the arDER for this system is still very close to the baseline wrDER, 28% vs. 27.6%, whereas the full speaker diarization task is considerably more challenging than within-recording speaker diarization. Note that, with a similar 50% increase of the number of speakers detected, the FullDia skl system performance decreased by around 40% the arDER in the AMI56 experiments. This suggests that the ttest distance is robust to estimation errors in the number of natural speaker clusters besides obtaining better figures in absolute terms.

7. CONCLUSION

We proposed a two stage system for performing speaker diarization on a full data set via speaker linking of the speaker clusters output by a speaker diarization system. According to the results obtained for three data sets taken from the AMI corpus, using a dis-

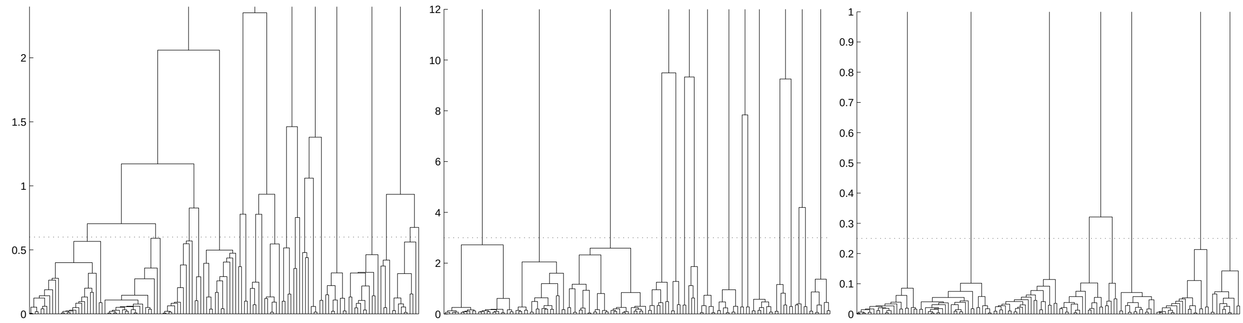

Fig. 1. From left to right, dendrograms obtained by the FullDia cos, FullDia skl and FullDia ttest systems on the AMI8 data set. On the x axis are the speaker clusters output by the standard speaker diarization system. The height of the nodes represents the merging distance. The dotted line is the threshold that minimized the wrDER.

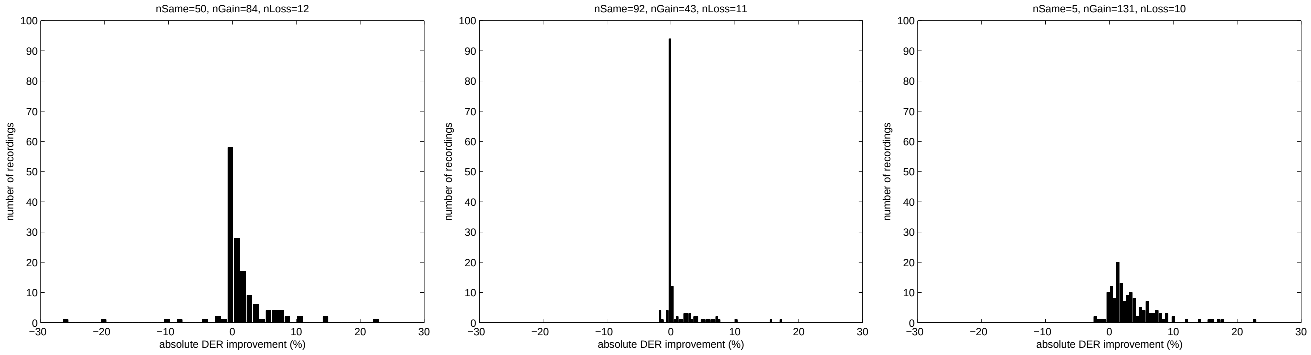

Fig. 2. Histograms of the absolute wrDER improvement across all the meetings in the AMI56 data set. The three graphs correspond to the FullDia cosine, FullDia skl and FullDia ttest systems from left to right. On top of the graphs, the number of meetings whose assignment was not changed after speaker linking is shown along with the number of meetings that resulted in a gain or loss of wrDER.

tance derived from the Hotteling t-square statistic in the agglomerative speaker clustering stage greatly helps in producing meaningful speaker clusters across the database. Although the systems using the cosine distance and the symmetric KL divergence obtained some within-recording DER improvements, diarization error rates across the whole data set were much higher than those obtained with the ttest distance. The optimal thresholds for speaker labelling in the speaker linking task were stable across different data sets with different amount of acoustic channel variability, although the threshold was decisive in predicting the number of speakers of the data set, a parameter that remains critical. The systems using the ttest distance were also much more precise at guessing the right number of speakers. Even the sensitivity of the diarization error rate across the data set to these errors was found to be low when using the ttest distance. These conclusions are supported by the cluster purity and cluster coverage measures as well. For the data set with largest speaker and channel variability diarization error rates across the data set were kept almost as low as diarization error rates performed recording by recording.

8. REFERENCES

[1] P. Kenny, G. Boulianne, and P. Dumouchel, “Eigenvoice modeling with sparse training data,” *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 3, pp. 345–354, 2005.

[2] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, “A study of inter-speaker variability in speaker verification,” *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2008.

[3] D.A. van Leeuwen, “Speaker Linking in Large Data Sets,” in *Proc. of the IEEE Speaker Odyssey Workshop*, 2010.

[4] H. Ghaemmaghami, D. Dean, R. Vogt, and S. Sridharan, “Extending the Task of Diarization to Speaker Attribution,” in *Proc. INTERSPEECH*, 2011, pp. 1049–1052.

[5] M. Huijbregts and D. van Leeuwen, “Large Scale Speaker Diarization for Long Recordings and Small Collections,” *IEEE Trans. on Audio, Speech and Language Processing*, pp. 404–413, 2012.

[6] P. Kenny, D. Reynolds, and F. Castaldo, “Diarization of telephone conversations using factor analysis,” *IEEE Journal of Selected Topics in Signal Processing*, December 2010.

[7] D. Vijayasenan, F. Valente, and H. Bourlard, “Information Theoretic Approach to Speaker Diarization of Meeting Data,” *IEEE Trans. on Audio, Speech and Language Processing*, vol. 17, no. 7, pp. 1382–1393, 2009.

[8] J. H. Ward, “Hierarchical Grouping to Optimize an Objective Function,” *American Statistical Association*, vol. 58, no. 301, pp. 236–244, 1963.

[9] G. N. Lance and W. T. Williams, “A General Theory of Classificatory Sorting Strategies. 1. Hierarchical Systems,” *Computer Journal*, vol. 9, pp. 373–380, 1967.

[10] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-End Factor Analysis for Speaker Verification,” *IEEE Trans. on Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 788–798, 2009.