Diarizing Large Corpora using Multi-modal Speaker Linking

Marc Ferràs[1], *Stefano Masneri*[2], *Oliver Schreer*[2], *Hervé Bourlard*[1]

Idiap Research Institute, CH-1920 Martigny, Switzerland[1]
Fraunhofer Heinrich Hertz Institute, 10587 Berlin, Germany[2]

marc.ferras@idiap.ch, stefano.masneri@hhi.fraunhofer.de,
oliver.schreer@hhi.fraunhofer.de, bourlard@idiap.ch

Abstract

Speaker diarization of a collection of recordings with uniquely identified speakers is a challenging task. A system addressing such task must account for the inter-session variability present from recording to recording and it is asked to scale well to massive amounts of data. In this paper we use a two-stage approach to corpus-wide speaker diarization involving speaker diarization and speaker linking stages. The speaker linking system agglomeratively clusters speaker factor posterior distributions obtained via Joint Factor Analysis using the Ward method and the Hotteling t-square statistic as distance measure. We extend this framework to link speakers based on both speech and visual modalities to improve the robustness of the system. The system is evaluated using the data collected for the Augmented Multiparty Interaction (AMI) project, involving over one hundred meetings. We provide results in terms of within-recording and across-recording diarization error rates (DER) to support the effectiveness of multi-modal speaker linking to enable large scale speaker diarization.

Index Terms: Multi-modal speaker linking, hierarchical clustering, Ward method, speaker diarization, meeting corpus

1. Introduction

The explosion of video and audio content available in recent years has challenged current technologies to index and analyse huge amounts of data in addition to archiving them. On one side, large multimedia corpora can sample several dimensions of interest, such as multiple sources of variability, that can extend modeling possibilities. These corpora typically involve speech in a variety of scenarios including multiple speakers, multiple and variable acoustic conditions, multiple languages and even emotion or vocal effort variation. Video data poses similar challenges especially in terms of light variation, multiple participant tracking or simple but relevant issues such as processing small face sizes and low resolution images. On the other side, many of the algorithms used nowadays do not scale well to large data sets or are simply prohibitive to use in such conditions. This is the case of speaker diarization, a technology that is quite mature for reasonable sized recordings but it is not directly applicable to long recordings or processing many files at once. It also has some modeling flaws, such as performance being highly dependent on the recording conditions.

It is more and more common that recordings involve multiple modalities, typically video and audio. In the context of speaker diarization, the speaker's voice and face characteristics can be used together to improve the speaker segmentation of a recording. Multi-modal processing assumes that certain conditions are met, such as audio and video focusing on the same person at the same time instant or only frames with frontal faces being used for the video modality.

In this paper we aim at diarizing a large data set using speech and video recordings of meetings. This translates into finding unique speaker identifiers across the database as well as the start and end times for each of the speaker segments. A straight approach to solve this task, such as audio-visual diarization of the concatenated recordings, is currently feasible for a few hours of data only. As proposed in [1], we opt for a two-stage hierarchical approach using local and global speaker representations. A speaker diarization system finds a small set of speaker clusters, along with start and end times and a speaker identifier local to each recording. In the second stage, we further cluster the speaker clusters using a global reference to structure the speaker/face space of the whole data set. Each speaker cluster is represented as a speaker factor posterior distribution obtained using Joint Factor Analysis (JFA) [2, 3]. The resulting speaker clusters are then given a unique speaker identifier across the data set.

Large scale speaker diarization has been addressed in previous works. A speaker attribution system performing speaker linking after speaker diarization was proposed in [4]. This system uses a variety of linking methods and the Normalized Cross Likelihood Ratio (NCLR) as the distance measure. Another multi-stage approach [5] targetting large scale diarization diarizes chunks of speech data whose clusters are linked later. This system scales particularly well on large data sets but still offers variable performance depending on the chunk size. A system using a variational Bayes approach to diarization presented in [6] used speaker factor posterior distributions to perform soft clustering. All these studies focus on telephone speech conversations between two people or broadcasted data. In our work, we focus on meetings recorded using far-field microphones on multiple meeting rooms on the audio side, and video cameras recording each of the participants on the video side.

This paper is a multimodal extension to the speaker diarization and linking work in [1]. The paper is organized as follows: Section 2 describes the speaker diarization system we used in this work. Section 3 describes the features used for linking, focusing on the visual modality. Section 4 gives an overview of how speaker clusters are modeled prior to linking. Section 5 describes how speaker clusters are linked and the labeling process across the dataset. In Section 6 the data sets used for experimental evaluation as well as the details about the implemented

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.

systems are presented. Section 7 provides results that validate the proposed techniques and Section 8 gives some conclusions.

2. Speaker Diarization

Speaker diarization systems are asked to to partition a recording into acoustically homogeneous regions that were spoken by the same speaker while determining the actual number of speakers. In this work, we use a system based on the Information Bottleneck (IB) principle. This system uniformly splits an audio recording into short 1-2 second long segments that are then clustered using a greedy optimization of the IB objective function. The reader can refer to [7] for detailed discussion about this approach.

The final diarization solution is asked to maximize the IB objective function,

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

to preserve a set of relevance variables Y, Gaussian posterior probabilities of the initial segments w.r.t. a Universal Background Model (UBM), that represent the information to be preserved during clustering. A compressed representation C of the initial segments X is seeked such that as much mutual information with Y is preserved while keeping the representation as compact as possible. The parameter β balances the amount of information preserved versus the amount of compression in the representation.

We use the agglomerative IB (aIB) algorithm, 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 normalized mutual information, $NMI = I(Y,C)/I(X,Y)$, measuring the fraction of original mutual information captured by the clustering partition C, is used to infer the number of speakers. The optimal number of speakers is found after thresholding the NMI measure value across multiple partitions.

The boundaries of the clusters are finally refined using an ergodic HMM with a minimum duration constraint.

3. Feature Extraction

In this work we use speech and visual features to perform speaker linking. For the speech modality, the linking system is using spectral envelope features as in [1]. In this paper, we use slightly different features, including energy and longer temporal context (see Section 6 for details) as compared to our previous work.

The visual features are obtained in two steps. First, a face detection algorithm is run to detect frontal faces over which local features are extracted later. We use the Shore library [8], that applies the modified census transform to each input video frame prior to a cascade of classifiers trained using AdaBoost. The algorithm provides the position, size and rating, a measure of how likely the image is to be a frontal face. Face detection was run every 2 frames to retain a maximum amount of data and low-likelihood faces were discarded to minimize the false alarm error rate.

The most widespread algorithms for feature extraction are Principal Component Analysis, so-called eigenfaces method [9], Linear Discriminant Analysis [10], Elastic Graph Bunch Matching [11] and Local Binary Patterns (LBP) [12]. In the last years, Discrete Cosine Transform (DCT) features used together with generative modeling [13] allowing for inter-session variability compensation have obtained state-of-the-art performance on the face authentication task. After informal face recognition experiments using a nearest neighbour classifier, we limited the choice to LBP and DCT features, offering the best compromise in terms of performance, efficiency and extendibility.

LBP feature extraction assumes a face image to be a composition of micro-patterns that are invariant to monotonic gray scale transformations. A global description of the image is then obtained by combining these micro-patterns. The pixels of an image are labeled by thresholding the neighbourhood of each pixel with the centre value and considering the result as a binary number. After applying this operator on every pixel, the resulting image is divided into M different non-overlapping regions and the histogram of each region is then computed. A variant of the basic method using a larger neighbourhood size and uniform patterns only for histogram computation was used in this work. For speaker linking purposes, each of these histograms is used as a feature vector, ensuring that data are dense enough to be modeled by a generative model.

Regarding DCT features, each face image is split into the same number of regions of the same size as in the LBP case. For each region, we compute the DCT transform and we use as many coefficients as the histogram size in LBP feature extraction, so that the comparison across both approaches is fair. The low-frequency to high-frequency zig-zag scanning used for JPEG compression was used.

4. Speaker modeling

Each of the speaker clusters found during the diarization process is modeled using Joint Factor Analysis (JFA) [2, 3]. JFA adapts a Gaussian Mixture Model (GMM) to the features of a speaker cluster while disentagling speaker/face and session effects. The speaker factors provide a compact representation of speakers/faces that disregards session variability components. We use the simplified JFA model

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

where $\hat{\mathbf{m}}$ 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 UBM, 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 [3]. $\mathbf{y}$ 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. The factor loading matrices are retained and they are used for adaptation, where only the latent variables y and x are fit to the data. During this process, JFA provides the posterior mean and covariance matrix of the multivariate Gaussian random variable y, fully characterized by a mean vector $\mathbf{y}$ and covariance matrix $\mathbf{C}$. Please refer to [2, 1] for more details on how these parameters are estimated.

5. Speaker Linking

The speaker linking module is the second clustering stage that structures the speaker clusters of the database hierarchically. The speech data of each speaker cluster output by the diarization system is modeled as a single multivariate Gaussian with full covariance matrix, i.e. the speaker factor posterior distribution estimated via JFA. We use an agglomerative clustering approach as follows:

1. **Compute the distance matrix** for all pairs of initial speaker 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**.

We apply the Ward method [14] to merge the most compact clusters, i.e. with minimum within-cluster variance, at each merging step. Besides the merging criterion, Ward clustering also allows for a fast implementation, the so-called Lance-Williams recursion [15], enabling the distances between cluster pairs to be computed recursively from the initial distance matrix. In [1] we found that using the two-way Hotteling t-square statistic

$$d_{ttest}(p_i, p_j) = \frac{n_i n_j}{n_i + n_j}(\mathbf{y}_i - \mathbf{y}_j)^T \mathbf{C}_{pool}^{-1}(\mathbf{y}_i - \mathbf{y}_j) \quad (3)$$

$$\text{with} \quad \mathbf{C}_{pool} = \frac{(n_i - 1)\mathbf{C}_i + (n_j - 1)\mathbf{C}_j}{n_i + n_j - 2} \quad (4)$$

as the distance measure is especially well suited for this task while still being derived from the square Euclidean distance. The expression in (3) is used to test whether the means of two multivariate Gaussian distributions $p_i \sim \mathcal{N}(\mathbf{y}_i, \mathbf{C}_i)$ and $p_j \sim \mathcal{N}(\mathbf{y}_j, \mathbf{C}_j)$ are the same when the variances are assumed to be different. This is the multivariate equivalent of the two-way Student t statistic.

We assume that the speaker clusters can be found by thresholding the distance values in the clustering dendrogram. 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.

A full block diagram of the speaker diarization and linking system is depicted in Figure 1.

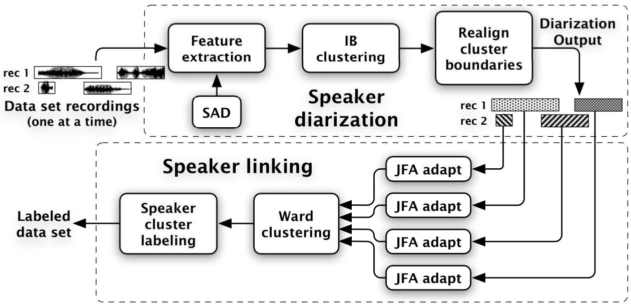

Figure 1: Block diagram of the speaker diarization and linking system.

6. Experimental Setup

The proposed speaker diarization and linking system was evaluated on the meeting data collected for the Augmented Multiparty Interaction (AMI) project. We compared the performance of speaker diarization alone versus the speaker diarization and linking approach with either speech or visual features, and also with a multi-modal fusion approach.

We used 19 Mel-Frequency Cepstral Coefficients (MFCC) extracted every 10ms using a 30ms window for the speaker diarization system. The initial segments were 2.5s long and the IB trade-off arameter was set to 10. The maximum number of speakers was 10 to encourage the system to undercluster, and the linking system can recluster to find a better partition. The NMI threshold was 0.3. All of these settings were optimized for the NIST RT'06 evaluation.

For the linking system using speech features, we used Perceptual Linear Prediction (PLP) features extracted every 10ms over 30ms windows, along with log energy and their delta and double delta coefficients (60 features). 50 hours of far-field AMI data were used to train a GMM-UBM with 512 Gaussian components as well as the JFA hyperparameters, using maximum likelihood estimation. The JFA factor loading matrices 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 24 different channels. We used decoupled estimation and 10 iterations of ML training to estimate the factor loading matrices. For adaptation we used joint estimation of speaker and session factors. All the available speaker factors, i.e. 132, and 50 session factors were used after informal optimization on preliminary experiments.

For the visual linking system we compute features on 80x80 pixel faces over 16 regions of 20x20 pixels. We extract either 59 LBP features or 59 DCT coefficients using the same number of regions and sizes. The individual-speaker video recordings were used for extracting the features alogn with the ground truth speaker segmentation. We used 128 Gaussians for the visual UBM to cope with the sparsity of visual data. All other settings for UBM and factor analysis hyperparameter training were kept the same as for speech features.

To prune the linking dendrogram we used a simple threshold optimizing the arDER on a development set and used on the linking tree of a evaluation set. Table 1 summarizes the data included in each of these data stes. For speech, only one recording per meeting was included using a microphone channel randomly chosen.

Combining speech and visual modalities consisted in fusing the corresponding distance matrices. Since clustering is unsupervised and no labels are available to perform calibration, min-max rescaling of the score ranges are performed prior to linearly combining pairs of scores. We used weights 0.7 and 0.3, optimized on the development set, if both modalities have scores available, otherwise only the available score was used.

6.1. Performance measures

We use the Diarization Error Rate (DER) as the main measure to evaluate the performance of the proposed systems. The DER assesses the impact of speaker linking on the diarization systems, using the references obtained by forced alignment of ASR transcripts with speakers labeled with unique identifiers within the recording. The within-recording DER (wrDER) assesses the effect of linking speakers within the recording. We use the across-recording DER (arDER) to assess the DER for the data

Data set	# meetings	# spk	# segments	DER (%)
Dev set	39	96	285	25.2
Eval set	88	130	667	24.7

Table 1: Details of the AMI data sets used for system development and evaluation. Only one single distant microphone channel was included for each meeting. The columns show the number of meetings, number of speakers, the number of speaker segments after per-recording speaker diarization.

set as a whole. We concatenated the references of all recordings in the data set with the within-recording speaker identifiers replaced by unique speaker identifiers across the data set. For all DER computations we use a collar of 250ms.

We also compute cluster purity and cluster coverage measures for systems using speaker linking. 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.

7. Experiments and Results

We first ran experiments comparing the speaker diarization performance of systems using speech and visual speaker linking. Table 2 shows the results for the dev and eval data sets. The speech system using PLP features outperformed the visual systems in both wrDER and arDER, although systems using DCT and LBP features still obtain reasonably small DER, given the complexity of the task. In absolute terms, all systems are able to diarize the whole data set at a DER comparable to single-recording DER, i.e. 25.8%, 29.7% and 32.3% versus 25.2% for the dev set. These results also confirm the effectiveness of the linking approach when compared to those obtained in [1]. The AMI data sets in this work involve all of the speakers in the database and a randomly chosen single-distant microphone recording per meeting, i.e. the most adverse scenario in terms of speaker and channel variability possible with the AMI data. Still, the arDER can be kept as low as those obtained for single-recording diarization. Regarding cluster purity and coverage, they are in the same range across all individual systems. Note that the initial DER obtained by the speaker diarization system is limiting the final DER, since the linking system is able to join speaker clusters but not to split them. This also translates into reducing linking performance as speaker representations are corrupted with around 25%, the DER, of feature vectors from wrong speakers in average. The number of estimated speakers is relatively accurate for the PLP system (78 vs. 96, 125 vs. 130) whereas it is slightly off for the DCT system, and not accurate for the LBP system. A non-negligeable fraction of the speaker clusters, 11% and 14% for the dev and eval sets respectively, are empty and they have no visual features available due to dropping non frontal faces. Although this is a structural limitation for the visual processing, it definitely has a negative effect on the DER of systems using visual features as well as the estimation of the number of speakers, as empty samples are kept as singleton clusters.

Table 3 shows results for experiments on the fusion of speaker linking systems. We explored fusing speech and visual systems using PLP+DCT or PLP+LBP features. For the development set, the fused systems obtain minor gains both in wrDER and arDER compared to the most performing individual system using PLP features. For the evaluation set, the fused systems keep the wrDER as low as for the PLP system but they slightly increase the arDER. These results suggest that the performances of the visual linking systems may not be good enough to provide gains after fusion. On the other side, fusing with the speech system is clearly effective in reducing DER from visual systems. The wrDER and arDER of the fused systems reduced largely after fusion, especially for the PLP+LBP over LBP systems, from 35.1% to 26.8%, ranking better that the PLP+DCT system. The estimated total number of speakers improves after fusion for the dev set, whereas the PLP system gave the closest number for the evaluation set.

Dev set			
System	#Spk	wr/ar DER(%)	Cp/Cc(%)
PLP	78	24.5/25.8	64.2/74.1
DCT	95	26.0/29.7	66.3/73.6
LBP	130	24.7/32.3	67.0/70.1

Eval set			
System	#Spk	wr/ar DER(%)	Cp/Cc(%)
PLP	125	24.0/26.2	64.2/71.3
DCT	187	25.3/33.8	68.7/64.4
LBP	263	25.1/35.1	69.7/62.3

Table 2: Diarization error rates after speaker linking experiments using speech (PLP) and face (DCT,LBP) features for the development and evaluation data sets. The remaining columns show the detected number of speakers, the DER, the within-recording and across-recording DERs and cluster purity and cluster coverage measures. The best results use bold typeface.

Dev set			
System	#Spk	wr/ar DER(%)	Cp/Cc(%)
PLP+DCT	91	24.2/25.3	61.9/73.6
PLP+LBP	88	24.3/25.4	62.6/73.8

Eval set			
System	#Spk	wr/ar DER(%)	Cp/Cc(%)
PLP+DCT	164	24.0/27.2	62.3/70.4
PLP+LBP	147	23.9/26.8	63.0/71.0

Table 3: Diarization error rates for speaker linking experiments using the score-level fusion of speech (PLP) and face (DCT,LBP) modalities for the development and evaluation data sets. The remaining columns show the detected number of speakers, the within-recording and across-recording DER and cluster purity and cluster coverage measures. The best results use bold typeface.

8. Conclusion

We have presented a multi-modal speaker linking approach that is able to diarize a whole database of highly interactive meetings with error rates as low as those obtained when diarizing a single meeting. This system aims at removing audio and visual inter-session variation in the linking phase. The linking system using speech features outperformed systems using visual features. This performance gap can be accounted for by the relative large amount of time, from 11% to 15% of the speech time, where visual features are not available. This has an negative impact in the visual and fused systems, the latter still obtaining gains in both within- and across-recording DER terms and especially for the development set.

9. References

[1] M. Ferras and H. Bourlard, "Speaker Diarization and Linking of Large Corpora," in *Proc. of the IEEE Workshop on Spoken Language Technology*, 2012.

[2] 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.

[3] 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.

[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] C. Kueblbeck and A. Ernst, "Face detection and tracking in video sequences using the modified census transformation," *Journal on Image and Vision Computing*, vol. 24(6), pp. 564–572, 2003.

[9] M. Turk and A. Pentland, "Face Recognition Using Eigenfaces," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 586–591.

[10] J. Lu and K. P, "Face Recognition Using LDA-Based Algorithms," *IEEE Trans. on Neural Networks*, vol. 14, pp. 195–200, 2003.

[11] L. Wiskott, J.-M. Fellous, N. Kruger, and C. Malsburg, "Face Recognition by Elastic Bunch Graph Matching," in *Intelligent Biometric Techniques in Fingerprint and Face Recognition*, 1999, pp. 355–396.

[12] T. Ahonen, "Face Recognition with Local Binary Patterns," in *Computer Vision - ECCV*, 2004, pp. 469–481.

[13] C. McCool, R. Wallace, M. McLaren, L. E. Shafey, and S. Marcel, "Session variability modelling for face authentication," *IET Biometrics*, vol. 2, no. 3, pp. 117–129, Sep 2013.

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

[15] 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.