Phoneme Background Model for Information Bottleneck based Speaker Diarization

Sree Harsha Yella[1,2], *Petr Motlicek*[1] *and Hervé Bourlard*[1,2]

[1] Idiap Research Institute, CH-1920 Martigny, Switzerland
[2] Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
sree.yella@idiap.ch, petr.motlicek@idiap.ch, herve.bourlard@idiap.ch

Abstract

Acoustic variability of speakers arises due to differences in their vocal tract characteristics. These individual speaker characteristics are reflected in a speech signal when speakers pronounce a given phoneme. The current work hypothesizes that clusters within a phoneme spoken by multiple speakers roughly correspond to different speakers. Based on this hypothesis, a Gaussian mixture model (GMM) based phoneme background model (PBM) is estimated. The components of such a PBM are used as a set of relevance variables in information bottleneck based speaker diarization system. Experiments are done using phone transcripts obtained from ground-truth and automatic speech recognition (ASR) system to estimate the PBM. The diarization experiments done on meeting recordings from AMI and NIST-RT corpora show that the proposed method achieves significant improvements over the system using a background model which ignores phoneme information.

Index Terms: speaker diarization, phoneme background model, information bottleneck, clustering

1. Introduction

Speaker diarization addresses the problem of "who spoke when" in a given multi-party conversation. It is an unsupervised task as there is no a-priori knowledge of the speakers or the number of speakers in a conversation [1, 2]. It has been studied in various domains such as broadcast news [3], telephone calls [4], with more recent focus on spontaneous meeting room conversations [2, 5, 6]. Methods of speaker diarization proposed in literature can be broadly placed in agglomerative (bottom-up) clustering framework [7, 8] or top-down splitting framework [9, 10]. Later works have tried to increase the robustness of the method by combining different diarization systems to exploit their complementary behavior [11, 12, 13]. More recent works have concentrated their effort in addressing the problems arising due to overlapping speech which contains multiple simultaneous speakers [14, 15, 16, 17].

Several methods have been proposed to make use of auxiliary information such as phone transcripts and non-speech segments of a given recording to help speaker diarization. In [18] a phonetic subspace mixture (PSM) model is proposed which uses phonetic information to make the Bayesian information criterion based distance measure (ΔBIC) used for agglomerative clustering more robust. In [19], phone adaptive training similar to speaker adaptive training performed in automatic speech recognition (ASR) system is proposed to remove the influence of phonetic content on the features so that they are more discriminative in speaker space. In [20], non-speech segments are provided as side-information to information bottleneck (IB) clustering to make the clustering algorithm more robust to background noise and errors made by automatic speech/non-speech detector.

In the current work, we use the information from phone transcript to estimate a Gaussian mixture model (GMM) based phoneme background model (PBM) that can be used for diarization in IB clustering framework. The PBM is estimated such that different modes of articulation of a phoneme are represented as different components of the GMM. This is based on our hypothesis that different modes of articulation of a phoneme arise due to different speakers pronouncing it. We report oracle experiments, where ground-truth phoneme and speaker information are used to estimate the PBM. We compare the resulting diarization output obtained by the oracle PBM with the background models obtained using just speaker information and a normal background model estimated without any knowledge of phoneme being spoken or the speaker. To use the PBM in a practical system, given a phoneme transcript (obtained from ASR system) of a meeting, we apply a simple clustering algorithm such as Ward's method [21] to identify clusters within each phoneme class. These clusters are represented by Gaussian components in the PBM. After the PBM is estimated, its components are used as a set of relevance variables to perform IB diarization [8]. We report diarization experiments on meetings from AMI and NIST-RT meeting corpora. These experiments suggest that a PBM based IB speaker diarization system gives lower error than the one using normal background model. The paper is organized as follows. Section 2 presents a brief overview of speaker diarization system based on IB clustering framework. Section 3 presents the motivation and details of the proposed method of estimation of PBM which is used in IB diarization system. Section 4 reports the experimental results on meetings from AMI and NIST-RT meeting corpora. Section 5 presents the conclusions and future directions.

2. Information bottleneck based speaker diarization system

This section briefly summarizes the agglomerative Information Bottleneck (aIB) speaker diarization system proposed in [8]. Information Bottleneck (IB) is a distributional clustering technique introduced in [22]. Consider a set of input variables $X = \{x_1, x_2, \ldots, x_n\}$ to be clustered into $C = \{c_1, c_2, \ldots, c_k\}$ clusters. The Information Bottleneck principle depends on a relevance variable set $Y = \{y_1, y_2, \ldots, y_m\}$ that carries important information about the problem. According to IB principle, any clustering C should be compact with respect to the input representation (minimum $I(X, C)$) and preserve as much mutual information as possible about relevance variables Y (maximum $I(C,Y)$). This corresponds to the maximization of:

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

where β is a Lagrange multiplier. The IB criterion is optimized w.r.t. the stochastic mapping $p(c_i|x_j)$ using iterative optimization techniques. The agglomerative Information Bottleneck clustering is a greedy way of optimizing the IB objective function [23]. The algorithm is initialized with each input element $x_i \in X$ as a separate cluster. At each step, two clusters are merged such that the reduction in mutual information w.r.t relevance variables is minimum. The distance measure which is dependent on the loss in mutual information w.r.t to relevance variables by merging two clusters c_i, c_j is obtained as:

$$\nabla \mathcal{F}_{\mathcal{IB}}(c_i, c_j) = [p(c_i) + p(c_j)] d_{ij}^{IB} \quad (2)$$

The distance d_{ij}^{IB} between two clusters c_i, c_j can be obtained in closed form by using Jensen-Shannon divergence as shown below, which arises naturally from the optimization of (1).

$$d_{ij}^{IB} = JS[p(Y|c_i), p(Y|c_j)] - \frac{1}{\beta} JS[p(X|c_i), p(X|c_j)] \quad (3)$$

The Jensen-Shannon divergence $JS[p(Y|c_i), p(Y|c_j)]$ is given by:

$$\pi_i D_{kl}\left[p(Y|c_i)||p(Y|c_{ij})\right] + \pi_j D_{kl}\left[p(Y|c_j)||p(Y|c_{ij})\right] \quad (4)$$

where $\pi_i = \frac{p(c_i)}{p(c_i)+p(c_j)}$, $p(Y|c_{ij})$ represents the distribution of relevance variables after the cluster merge and D_{kl} denotes the Kullback-Leibler divergence between two distributions. After each merge, $p(Y|c_i)$ and $p(Y|c_j)$ are averaged to get relevance variable distribution of the new cluster $p(Y|c_{ij})$. The number of clusters is determined by a model selection criterion based on a threshold on the normalized mutual information given by $\frac{I(C,Y)}{I(X,Y)}$ (see [8] for details).

To apply this method to speaker diarization, the set of relevance variables $Y = \{y_i\}$ is defined as the components of a background Gaussian Mixture Model (GMM) estimated from speech regions of a given recording [8]. The input to the clustering algorithm is uniformly segmented speech segments $X = \{x_j\}$ which represent the initial clusters with which the algorithm is initialized. The probability $p(y_i|x_j)$, i.e., the posterior probability of each Gaussian component conditioned on a speech segment can be computed using Bayes' rule. The speech segments with the smallest distance $\nabla \mathcal{F}_{\mathcal{IB}}$ given by (2) are then iteratively merged until the model selection criterion is satisfied. After the agglomerative clustering stops, Viterbi re-alignment is performed to smooth the arbitrary boundaries due to initialization by uniform segmentation.

3. Phoneme background model

Having the knowledge of what is being spoken has been shown to be a very useful information in speaker identification/verification tasks [24, 25]. This gives a chance to model individual variations in pronunciation of an acoustic class (phoneme/word) [26, 27]. Due to this, text constrained speaker identification/verification tasks usually have higher accuracies than their text independent counterparts. In the current work, we perform experiments to investigate whether the knowledge of what is being spoken helps to improve speaker diarization.

First of all, we perform an oracle speaker diarization experiment, with ground-truth phone transcription and speaker segmentation. In this experiment, we compare the IB diarization systems using different background models to perform speaker diarization. We compare plain background GMM (Plain-UBM) estimated from the speech regions of a given meeting recording, the background model estimated with the knowledge of speaker segmentation (Spkr-UBM) and the model estimated with the knowledge of both speaker and phoneme being spoken (Spkr-phone-UBM). In the Spkr-phone-UBM, each phoneme spoken by a speaker is represented by a Gaussian component in the background GMM. This is done by accumulating all the utterances of a phoneme by a speaker according to the ground-truth transcripts and approximating them with a Gaussian. In total there are 45 phonemes including silence class. To estimate the Spkr-UBM, speech segments belonging to a speaker are used to estimate n components per speaker in the background GMM, where n was varied to take different values from 5 to 45. The ground-truth phone transcripts are obtained by force-aligning the manual transcripts to individual head microphone channels which also produces speaker start and end times as a by product. Once the background model is obtained based on one of the approaches explained above (Plain-UBM/Spkr-UBM/Spkr-phone-UBM), IB speaker diarization is performed using the components of the background model as a relevance variable set as explained in the Sec. 2. We used 100 meetings from AMI corpus in this experiment. Fig. 1 plots the speaker error in IB diarization when using the three background models (Plain-UBM/Spkr-UBM/Spkr-phone-UBM). The lowest error for Spkr-UBM is at $n = 20$ which means 20 is the optimal number of components per speaker in the background GMM. It can be observed that Spkr-phone-UBM achieves the lowest error among all the background models which shows that having the knowledge of what is being spoken helps speaker diarization.

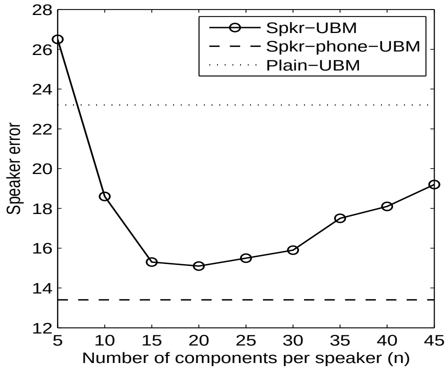

Figure 1: *Comparison of various background models: Speaker error for IB diarization using different types of UBMs.*

Motivated from the above oracle experiment, we propose a method to estimate a background model that captures different modes of articulation of a phoneme. The hypothesis, is that different modes of articulation of a phoneme arise due to differences in the individual speakers pronouncing it. Given a phone transcript of a meeting recording (obtained either from ground-truth or ASR system), we employ a simple clustering algorithm such as Ward's method to estimate the clusters in a phoneme and use the obtained clusters from all the phonemes to build a background GMM which is referred to as phoneme background model (PBM). Let $S = \{p_1, p_2, \ldots, p_i, \ldots, p_N\}$ denote the set of phonemes in the given transcript, where N is the number of unique phonemes in the transcription. Let $P_i = \{p_i^1, p_i^2, \ldots, p_i^{n_i}\}$ denote the set of all the occurrences of a phoneme p_i in the transcript, where n_i indicates the number of occurrences. Each occurrence p_i^j of phoneme p_i is approximated by a mean vector x_i^j computed from the feature vectors corresponding to that phoneme occurrence. This results in a set of points $X_i = \{x_i^1, x_i^2, \ldots, x_i^j, \ldots, x_i^{n_i}\}$ where, x_i^j represents the j^{th} occurrence of the phoneme p_i in the transcript. Agglomerative clustering is performed using Ward's method to cluster the occurrences of a given phoneme. The agglomerative clustering is initialized by a set of single-ton clusters represented by X_i. Ward's method is an greedy clustering method, where at each step, it merges two clusters that results in minimum increase of variance. The distance measure $\Delta(c_k, c_l)$ between two clusters c_k, c_l is given by:

$$\Delta(c_k, c_l) = \frac{n_{c_k} n_{c_l}}{n_{c_k} + n_{c_l}} \|m_k - m_l\|^2 \qquad (5)$$

where, $\|\|$ denotes Euclidean distance, n_{c_k} and n_{c_l} represent the number of samples in cluster c_k and c_l respectively and m_k, m_l represent the mean vectors (centroids) of clusters c_k and c_l respectively. The clustering continues until the desired number of clusters M is obtained. The clustering is performed for all the phonemes p_i in the set S and the final set of clusters $C = \{c_i^j\}\ \forall\ i \in \{1, \ldots, N\}$ and $\forall j \in \{1, \ldots, M\}$ is obtained. Each of the cluster in the final cluster set C is represented as a Gaussian component in the PBM. The mean of component is equal to the centroid of the respective cluster and variance is equal to the cluster variance. The weight of a component is obtained as the proportion of samples assigned to the respective cluster. The number of clusters for each phoneme class M is selected based on cross-validation on development set. Fig. 2 summarizes the procedure of estimating the PBM with the help of a block diagram.

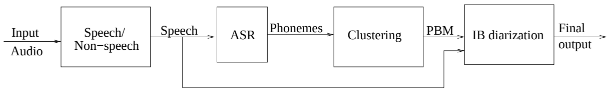

Figure 2: *Block diagram of the proposed diarization method using PBM.*

4. Experiments and Results

Speaker diarization experiments are conducted on meetings from AMI [28] and NIST-RT [29] meeting corpora. Out of 170 meetings present in the AMI corpus, 100 meetings are used in current experiments. The number of speakers in each meeting in AMI dataset varies between 3 to 5, but most of the meetings have 4 speakers. For experiments on NIST-RT corpus, we have used meetings from RT-05,06,07,09 datasets. The number of speakers in each meeting of NIST-RT corpus varies between 4 to 11. Both the corpora contain meetings recorded at multiple meeting room environments. The audio captured by the distant microphone array is enhanced by beamforming using *BeamformIt* [30] toolkit. 19 Mel-frequency cepstral coefficients (MFCC) are extracted for each frame of length 30 ms with a frame shift of 10 ms from this enhanced signal. These features are used for both speaker diarization and for clustering of data within each phoneme class to obtain a PBM. Prior to performing diarization, speech/non-speech detection is performed using SHOUT toolkit [31] and non-speech segments are ignored. The speaker diarization systems are evaluated using the metric diarization error rate (DER) used in NIST evaluation campaigns. DER is the sum of speech/non-speech error and the speaker error. Speech/non-speech error is the sum of miss and false alarm errors by the automatic speech/non-speech detection system and speaker error is the clustering error happening whenever speech segments of a speaker are attributed to a different one. Like the NIST evaluations, we used a forgiveness collar of ±0.25 seconds around the reference segment boundaries while scoring the automatic systems' output.

The ASR system used in the current study is a conventional HMM/GMM system [32]. The acoustic models are trained on 150 hours of labelled speech from AMI and ICSI [33] corpora. 13 dimensional MFCC along with their first and second order derivatives resulting in a 39 dimensional feature vector is extracted for every 10 ms from a speech frame of length 30 ms. The features are extracted from individual head microphone (IHM) channels worn by the speakers in the meetings. These features are used for training the models and for decoding. The 1-best word recognition output is subsequently transformed into a sequence of phonemes (i.e., top-down approach to deriving a phoneme sequence) which are later used for PBM estimation.

The optimal number of clusters M for each phoneme class in the PBM is decided based on diarization experiments on development data. Since the number of clusters is inherently dependent on the number of speakers in a given recording and since the number of speakers varies significantly between AMI and NIST-RT meetings, we performed separate development experiments for each corpus. For the development experiments on AMI corpus, we used 20 meetings that are not included in the 100 test meetings. For NIST-RT meetings, we used meetings from RT-05,06 as development set and used RT-07,09 as our test set. Development experiments on AMI data set revealed that $M = 6$ is optimal for meetings from AMI corpus. For NIST-RT meetings, the optimal value of M on development data was 15. Tab. 1 presents the speaker errors for baseline IB system (Bas-IB), and the system using PBMs estimated from phone transcripts obtained from ground-truth (GT-PBM) and ASR (ASR-PBM) systems on AMI and NIST-RT development sets.

Table 1: *Diarization experiments on development sets: Speaker error for different systems on development set of meetings from AMI (20 meetings) and NIST-RT (RT-05,06) corpora.*

Corpus	Bas-IB	GT-PBM	ASR-PBM
AMI	24.3	17.4	20
NIST-RT	16.8	13.4	14.8

To evaluate the performance of the proposed method on test set of meetings, we compare the performance of the baseline IB speaker diarization system presented in Sec. 2 with the IB diarization system using the PBMs estimated using ground-truth phone transcripts and ASR system. We also compare the IB speaker diarization system with the state of the art HMM/GMM based speaker diarization system [7]. In the HMM/GMM system, the states of the HMM represent speakers and the emission probability distributions of the states are modelled using GMMs. The system is initialized with uniform segmentation, resulting in 16 initial clusters (states). At each step of agglomerative clustering, the closest clusters obtained using modified ΔBIC [34] as distance measure are merged. After each merge, Viterbi re-alignment and re-estimation of the models is performed. The merging of clusters stops when there are no possible cluster merges according to the ΔBIC measure. Tab. 2 presents the evaluation results on AMI data set. Speech/non-speech error for all the systems is constant as speech activity detector output given by the SHOUT toolkit [31] is used for all the systems. It can be observed from the Tab. 2 that the baseline IB (Bas-IB) and HMM/GMM (Bas-HMM/GMM) systems have similar error rates on this set of 100 AMI meetings used as test set. Also, the PBM estimated using ground-truth phone transcripts (GT-PBM) gives the lowest DER among all the systems where it reduces the DER from 38.2 to 33.4. The PBM estimated from ASR transcripts also performs better than the baseline systems which reduces the DER from 38.2 to 34.8. This shows that the proposed method of using PBM for IB speaker diarization improves the performance of the baseline IB system. To check the fact that phone transcripts generated by ASR system are providing reliable information to PBM estimation, we performed an experiment where, the PBM is estimated by randomizing the phone transcript, where the phoneme class for a segment was chosen randomly out of the 45 phoneme classes. Since this randomization results in the loss of phoneme information, it is expected that the performance of this system will be similar to the baseline system which uses a background model estimated by ignoring the phoneme information. Tab. 2 also reports the performance of the system using the PBM estimated from randomized phone transcripts (Rand-PBM) which is similar to the baseline IB system(Bas-IB).

Table 2: *Diarization experiments on 100 test meetings from AMI corpus: Speech/non-speech error (SpNsp), Speaker error (Spkr) and total diarization error rate (DER) for different diarization systems.*

System	SpNsp	Spkr	DER
Bas-IB	15.0	23.2	38.2
Bas-HMM/GMM	15.0	23.6	38.6
GT-PBM	15.0	18.4	33.4
ASR-PBM	15.0	19.8	34.8
Rand-PBM	15.0	23.0	38.0

We also evaluate the proposed method on meetings from NIST-RT corpus using RT-07,09 data sets are our test set. Tab. 3 compares the performance of the various systems on NIST-RT test set of meetings from RT-07,09. It can be observed from the Tab. 3 that proposed method (ASR-PBM) reduces the DER of the baseline IB system from 14.5 to 13.6 on RT-07 data set and from 33.9 to 30.9 on RT-09 data set. Using ground-truth phone transcripts (GT-PBM) further reduces the error to 12 and 29.3 on RT-07, RT-09 data sets respectively. The lowest error on the two datasets (RT-07,09) is obtained by the baseline HMM/GMM system (Bas-HMM/GMM) which achieves DER of 10.1 and 27 respectively on RT-07 and RT-09. Even though the proposed method (ASR-PBM/GT-PBM) reduces the DER when compared to baseline IB system (Bas-IB), it has a higher error rate compared to baseline HMM/GMM system (Bas-HMM/GMM) on both RT-07,09 data sets. This might be due to the varying number of speakers in each meeting and higher rate of overlap in the meetings which is not the case with the meetings in AMI corpus.

Table 3: *Diarization experiments on NIST-RT test set: Speech/non-speech error (SpNsp), Speaker error (Spkr) and total diarization error rate (DER) for different diarization systems on NIST-RT 07, 09 data sets.*

Corpus	System	SpNsp	Spkr	DER
RT 07	Bas-IB	3.7	10.8	14.5
	Bas-HMM/GMM	3.7	6.4	10.1
	GT-PBM	3.7	8.3	12
	ASR-PBM	3.7	9.9	13.6
RT 09	Bas-IB	12.7	21.2	33.9
	Bas-HMM/GMM	12.7	14.3	27
	GT-PBM	12.7	16.6	29.3
	ASR-PBM	12.7	18.2	30.9

5. Conclusions and Future work

This paper proposed a method to incorporate information from “what is being spoken” (represented by phoneme transcripts) to improve the task of identifying “who is speaking when” (speaker diarization) in an information bottleneck based speaker diarization system. The information from phoneme transcript of a given audio recording was incorporated into the speaker diarization system by estimating a phoneme background model (PBM). The estimation of a PBM was motivated from the oracle experiment which showed that background model estimated from the knowledge of speaker and phoneme being spoken is more useful for diarization than background model estimated using only speaker information and a model estimated by ignoring both speaker and phoneme information. The PBM estimation was based on the hypothesis that clusters within a phoneme class roughly correspond to different speakers that have spoken it. The PBM was estimated by clustering the data within each phoneme class in a phoneme transcript of an audio file and representing each cluster with a Gaussian component in the PBM. The usefulness of such a PBM was evaluated by using it as a background model in IB speaker diarization. Experiments conducted on meetings from AMI and NIST-RT corpora showed that the PBMs estimated from ASR transcripts reduce the DER from 38.2 to 34.8 on AMI corpus, from 14.5 to 13.6 on RT-07 and 33.9 to 30.9 on RT-09 data sets.

As part of future work, we will explore methods to estimate the number of clusters M within each phoneme class based on a stopping criterion in the hope that it alleviates the need to do cross-validation based development studies whenever the nature of corpus changes drastically. Also to model the variations in phoneme pronunciations due to the context in which they occur, we will use syllable or tri-phone units to estimate the background model instead of using phonemes.

6. Acknowledgements

The authors would like to thank Dr. Andreas Stolcke for providing ground-truth force-aligned phone and word transcripts for NIST-RT meetings. This work was funded by the Swiss National Science Foundation (SNSF) through SNF-RODI grant.

7. References

[1] S.E. Tranter and D.A. Reynolds, "An overview of automatic speaker diarization systems," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 5, pp. 1557–1565, September 2006.

[2] X. Anguera Miro, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 2, pp. 356–370, feb. 2012.

[3] Daniel Moraru and al., "The ELISA consortium approaches in broadcast news speaker segmentation during the NIST 2003 rich transcription evaluation," in *ICASSP*, 2004, vol. 1, pp. 373–376.

[4] P. Kenny, D. Reynolds, and F. Castaldo, "Diarization of telephone conversations using factor analysis," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 4, no. 6, pp. 1059–1070, Dec 2010.

[5] Xavier Anguera, *Robust speaker diarization for meetings*, Ph.D. thesis, Universitat Politecnica de Catalunya, 2006.

[6] Deepu Vijayasenan, *An Information Theoretic Approach to Speaker Diarization of Meeting Recordings*, Ph.D. thesis, Ecole polytechnique fédérale de Lausanne, December 2010.

[7] Chuck Wooters and Marijn Huijbregts, "Multimodal technologies for perception of humans," chapter The ICSI RT07s Speaker Diarization System, pp. 509–519. Springer-Verlag, Berlin, Heidelberg, 2008.

[8] D. Vijayasenan, F. Valente, and H. Bourlard, "An information theoretic approach to speaker diarization of meeting data," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 7, pp. 1382–1393, 2009.

[9] S. Bozonnet, N. Evans, and C. Fredouille, "The lia-eurecom rt09 speaker diarization system: enhancements in speaker modelling and cluster purification," in *ICASSP*, Dallas, USA, 2010, pp. 4958–4961.

[10] N. Evans, S. Bozonnet, Dong Wang, C. Fredouille, and R. Troncy, "A comparative study of bottom-up and top-down approaches to speaker diarization," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 2, pp. 382–392, 2012.

[11] S. E. Tranter, "Two-way cluster voting to improve speaker diarisation performance," in *ICASSP*, 2005, pp. 753–756.

[12] Simon Bozonnet and al., "System output combination for improved speaker diarization," in *Interspeech*, 2010, pp. 2642–2645.

[13] Sree Harsha Yella and Fabio Valente, "Information bottleneck features for hmm/gmm speaker diarization of meetings recordings," in *Interspeech*, Florence, Italy, 2011, pp. 953–956.

[14] Kofi Boakye, Oriol Vinyals, and Gerald Friedland, "Improved overlapped speech handling for speaker diarization," in *Interspeech*, Florence, Italy, 2011, pp. 941–943.

[15] Jurgen Geiger, Ravichander Vipperla, Simon Bozonnet, Nicholas Evans, Bjorn Schuller, and Gerhard Rigoll, "Convolutive nonnegative sparse coding and new features for speech overlap handling in speaker diarization," in *Interspeech*, Portland, USA, 2012.

[16] M. Zelenák, C. Segura, J. Luque, and J. Hernando, "Simultaneous speech detection with spatial features for speaker diarization," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 2, pp. 436–446, 2012.

[17] Sree Harsha Yella and Hervé Bourlard, "Improved overlap speech diarization of meeting recordings using long-term conversational features," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, Vancouver, Canada, 2013.

[18] I-Fan Chen, Shih-Sian Cheng, and Hsin-Min Wang, "Phonetic subspace mixture model for speaker diarization," in *Interspeech*, Makuhari, Japan, 2010, pp. 2298–2301.

[19] Simon Bozonnet, Ravichander Vipperla, and Nicholas W. D. Evans, "Phone adaptive training for speaker diarization.," in *Interspeech*, Portland, USA, 2012.

[20] Sree Harsha Yella and Bourlard Hervé, "Information bottleneck based speaker diarization of meetings using non-speech as side information," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, Florence, Italy, 2014.

[21] Joe H. Ward, "Hierarchical grouping to optimize an objective function," *Journal of the American Statistical Association*, vol. 58, pp. 236–244, 1963.

[22] N Tishby, F Pereira, and W Bialek, "The information bottleneck method," in *NEC Research Institute TR*, 1998.

[23] N Slonim, N Friedman, and N Tishby, "Agglomerative information bottleneck," in *Advances in Neural Information Processing Systems*. 1999, pp. 617–623, MIT press.

[24] D.E. Sturim, D.A. Reynolds, R.B. Dunn, and T.F. Quatieri, "Speaker verification using text-constrained gaussian mixture models," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, Orlando, USA, May 2002, vol. 1, pp. 677–680.

[25] A. Stolcke, E. Shriberg, L. Ferrer, S. Kajarekar, K. Sonmez, and G. Tur, "Speech recognition as feature extraction for speaker recognition," in *Signal Processing Applications for Public Security and Forensics, 2007. SAFE '07. IEEE Workshop on*, 2007, pp. 1–5.

[26] Frederick Weber, Barbara Peskin, Michael Newman, Andrs Corrada-Emmanuel, and Larry Gillick, "Speaker recognition on single- and multispeaker data," *Digital Signal Processing*, vol. 10, no. 13, pp. 75 – 92, 2000.

[27] J. P. Eatock and J.S. Mason, "A quantitative assessment of the relative speaker discriminating properties of phonemes," in *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*, Adelaide, Australia, 1994, pp. 133–136 vol.1.

[28] "http://corpus.amiproject.org/," .

[29] "http://www.itl.nist.gov/iad/mig/tests/rt/," .

[30] "http://www.xavieranguera.com/beamformit/," .

[31] Marijn Huijbregts and Franciska de Jong, "Robust speech/non-speech classification in heterogeneous multimedia content," *Speech Communication*, vol. 53, no. 2, pp. 143–153, 2011.

[32] Petr Motlicek, Daniel Povey, and Martin Karafiat, "Feature and score level combination of subspace gaussians in lvcsr task," in *ICASSP*, Vancouver, Canada, 2013, pp. 7604–7608.

[33] Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, and Chuck Wooters, "The icsi meeting corpus," in *ICASSP*, Hong Kong, 2003, pp. 364–367.

[34] Jitendra Ajmera and Chuck Wooters, "A robust speaker clustering algorithm," in *IEEE Automatic Speech Recognition Understanding Workshop*, 2003, pp. 411–416.