INFORMATION BOTTLENECK BASED SPEAKER DIARIZATION OF MEETINGS USING NON-SPEECH AS SIDE INFORMATION

Sree Harsha Yella[1,2] *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, herve.bourlard@idiap.ch

ABSTRACT

Background noise and errors in speech/non-speech detection cause significant degradation to the output of a speaker diarization system. In a typical speaker diarization system, non-speech segments are excluded prior to unsupervised clustering. In the current study, we exploit the information present in the non-speech segments of a recording to improve the output of the speaker diarization system based on information bottleneck framework. This is achieved by providing information from non-speech segments as side (irrelevant) information to information bottleneck based clustering. Experiments on meeting recordings from RT 06, 07, 09, evaluation sets have shown that the proposed method decreases the diarization error rate by around 18% relative to the baseline speaker diarization system based on information bottleneck framework. Comparison with a state of the art system based on HMM/GMM framework shows that the proposed method significantly decreases the gap in performance between the information bottleneck system and HMM/GMM system.

Index Terms: speaker diarization, spontaneous meeting recordings, information bottleneck, clustering, side information.

1. INTRODUCTION

Speaker diarization task involves identifying "who spoke when" in a given multi-party speech recording. It involves inferring the number of speakers in a given audio recording along with attibuting each identified speaker with his/her utterances in the recording. It is an unsupervised task by definition as there is no a-priori information about the speakers in the recording [1, 2, 3]. In recent years, the focus has shifted towards performing this task on more natural data such as spontaneous meeting recordings and telephone conversations. Many approaches have been proposed in literature to accomplish this task such as parametric/non-parametric [4, 5, 6], top-down/bottom-up [7, 2, 6] frameworks and also different methods to combine these systems [8, 9, 10]. There are two main challenges for performing speaker diarization on spontaneous meeting recordings; they are, artefacts of spontaneous conversations such as simultaneous speakers (overlaps), short speaker turns and corruption of the audio captured by distant microphones due to room reverberation and background noise.

Several diagnostical studies were done to isolate the main sources of errors in speaker diarization systems [11, 12, 13]. These studies have shown that the significant sources of errors in a typical diarization system come from overlapping speech segments and errors in speech/non-speech detection. Recent works on speaker diarization have tried to address the issue of overlaps to improve diarization output [14, 15, 16, 17, 18]. However, less efforts have been put in addressing the errors caused by speech/non-speech detection and background noise. In [19], a method was proposed to improve the diarization output by detecting and eliminating non-speech frames that are included in clustering due to errors in speech/non-speech detection. In this work, it was observed that non-speech frames included in clustering have negative influence on the cluster merging decisions. Present work tries to address this issue in an information bottleneck based speaker diarization framework [6] by using information from the non-speech segments in the recording as side information for the clustering. This side information is provided in the form of irrelevant variable set to the information bottleneck with side information (IBSI) framework [20]. The IBSI framework [20] tries to cluster the input segments such that the resulting clusters maximize the mutual information with respect to relevant variables to clustering while minimizing the mutual information with respect to irrelevant variables which are provided as side information for clustering.

In the present work, the use of non-speech regions as side information (irrelevant variable) for the clustering is motivated by two reasons. In a typical diarization system, automatic speech/non-speech detection is performed before initializing the agglomerative clustering. Errors in speech/non-speech detection are a common artefact, which introduce some non-speech frames into clustering. Since a typical agglomerative speaker diarization system is initialized by uniform segmentation, segments of different speakers containing similar non-speech (background noise) might get merged due to their similarity in the non-speech regions rather than speech regions. The second reason is that, segments with low signal to noise ratio (SNR), belonging to different speakers, but are corrupted by similar background noise might get merged into a cluster due to the similar noise characteristics rather than their speech characteristics. In the present work, we hypothesize that providing the information from non-speech segments as irrelevant variable to the clustering penalizes such non desirable merges and improves the overall diarization output. To verify this hypothesis experiments are conducted on meetings from NIST RT 06, 07 and 09 evaluation sets [3] and the results obtained support our hypothesis. Since the method only uses the non-speech segments from a given recording, it does not require any pre-labelled non-speech data. It also has advantage that it uses the data that best represents the given recording. The remainder of this paper is organized as follows, section 2 presents the baseline speaker diarization system based on information bottleneck principles. Section 3 presents the proposed method of using non-speech as side information in agglomerative IBSI clustering framework. Section 4 presents the experiments and results obtained. Section 5 presents the conclusions and the future work.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

2. AGGLOMERATIVE INFORMATION BOTTLENECK

This section briefly summarizes the agglomerative Information Bottleneck (aIB) speaker diarization system proposed in [6]. Information Bottleneck (IB) is a distributional clustering technique introduced in [21]. 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:

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

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 [22]. 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:

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

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).

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

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

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

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 [6] 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) trained on speech regions of a given recording [6]. 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 posterior probability $p(y_i|x_j)$, i.e., the probability of each Gaussian component conditioned to the 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.

3. AGGLOMERATIVE INFORMATION BOTTLENECK WITH SIDE INFORMATION

The IBSI framework was proposed initially to identify relevant patterns among several conflicting patterns that might exist in the data [20]. The method has been successfully applied in document clustering, in processing neural spike train activity and in face recognition [20, 23]. The method incorporates information about irrelevant components of the data to better extract the relevant patterns' information. Given a set of input variables X that need to be clustered, a set of relevant variables Y^+ whose characteristics should be preserved in the final clustering, a set of irrelevant variables Y^-, and the joint distributions $P(X, Y^+)$ and $P(X, Y^-)$, the IBSI framework tries to cluster the input variable set X into clusters C such that the resulting clusters maximize mutual information w.r.t the relevant variable set Y^+ and minimize mutual information w.r.t the irrelevant variable set Y^-. This can be represented as maximization of the objective function below:

$$\mathcal{F}_{\mathcal{IBSI}} = I(Y^+, C) - \gamma I(Y^-, C) - \frac{1}{\beta} I(X, C) \tag{5}$$

where, γ and β are Lagrange multipliers.

Similar to the optimization of $\mathcal{F}_{\mathcal{IB}}$, $\mathcal{F}_{\mathcal{IBSI}}$ can also be optimized using various approaches [23] such as deterministic annealing, greedy agglomerative hard clustering and sequential K-means based clustering. To be compatible with the already existing diarization framework [6], in the current work, we used agglomerative hard clustering solution to the optimization problem. In this method, as with (2), loss in the objective function due to a merge of two clusters c_i and c_j can be obtained as:

$$\nabla\mathcal{F}_{\mathcal{IBSI}}(c_i, c_j) = [p(c_i) + p(c_j)]d_{ij}^{IBSI} \tag{6}$$

As in (3), the distance d_{ij}^{IBSI} between two clusters c_i and c_j can be obtained as:

$$\begin{aligned} JS[p(Y^+|c_i), p(Y^+|c_j)] \quad & -\gamma JS[p(Y^-|c_i), p(Y^-|c_j)] \\ & -\tfrac{1}{\beta} JS[p(X|c_i), p(X|c_j)] \end{aligned} \tag{7}$$

At each step of agglomerative clustering, the algorithm merges the two clusters that result in the lowest value of $\nabla\mathcal{F}_{\mathcal{IBSI}}$. By comparing the two distance measures d_{ij}^{IB} and d_{ij}^{IBSI}, respectively given by (3), (7) respectively, it can be observed that d_{ij}^{IBSI} incorporates an extra penalty term $\gamma JS[p(Y^-|c_i), p(Y^-|c_j)]$ which measures the similarity between two clusters in irrelevant variable domain Y^-. Due to this, the distance measure penalizes the merge of clusters with similar distribution over irrelevant variables. The whole method is summarized in Figure 1. The model selection criterion which gives the number of final clusters is based on a threshold on the normalized mutual information given by $\frac{I(C,Y^+)}{I(X,Y^+)}$.

To apply this method to speaker diarization, the set of relevant variables Y^+ is defined as the components of background GMM trained on the speech regions of a given recording similar to aIB framework. The set of irrelevant variables Y^- is defined as the components of background GMM trained on non-speech regions of a given recording. Similar to aIB, the clustering starts with uniformly segmented speech regions represented by X, and the posterior distributions of relevant and irrelevant variables $p(Y^+|X)$, $p(Y^-|X)$, are obtained using Bayes' rule. Clusters that have lowest distance measure (6) are merged at each step. The final number of clusters is obtained using the model selection criterion.

4. EXPERIMENTS AND RESULTS

Experiments are conducted on meetings from NIST RT 06, 07 and 09 evaluation data sets which contain meetings recorded in differ-

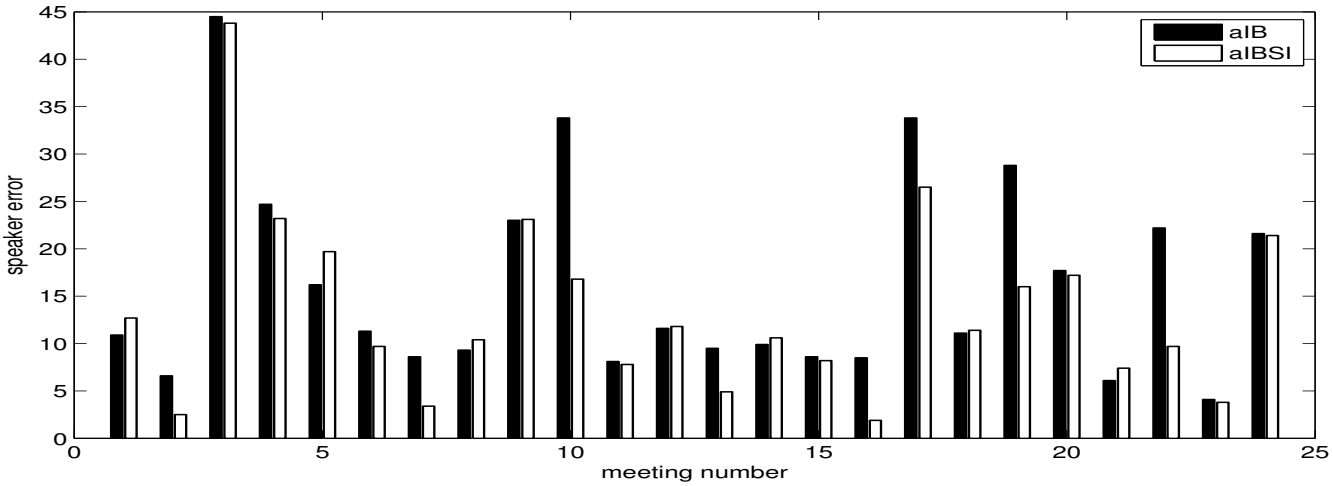

Fig. 2. *Meeting wise speaker error values for baseline aIB diarization system and aIBSI system.*

Input:

Joint Distribution $p(x, y^+), p(x, y^-)$

Trade-off parameters γ, β

Output:

C_m: m-partition of X, $1 \leq m \leq |X|$

Initialization:

$C \equiv X$

For $i = 1 \ldots N$

$c_i = \{x_i\}$

$p(c_i) = p(x_i)$

$p(y^+|c_i) = p(y^+|x_i) \forall y^+ \in Y^+$

$p(y^-|c_i) = p(y^-|x_i) \forall y^- \in Y^-$

$p(c_i|x_j) = 1$ if $j = i, 0$ otherwise

For $i, j = 1 \ldots N, i < j$

Find $\nabla \mathcal{F}_{\mathcal{IBSI}}(c_i, c_j)$

Main Loop:

While $|C| > 1$

$\{i, j\} = \arg\min_{i',j'} \nabla \mathcal{F}_{\mathcal{IBSI}}(c_{i'}, c_{j'})$

Merge $\{c_i, c_j\} \Rightarrow c_r$ in C

$p(c_r) = p(c_i) + p(c_j)$

$p(y^+|c_r) = \frac{[p(y^+|c_i)p(c_i) + p(y^+|c_j)p(c_j)]}{p(c_r)}$

$p(y^-|c_r) = \frac{[p(y^-|c_i)p(c_i) + p(y^-|c_j)p(c_j)]}{p(c_r)}$

$p(c_r|x) = 1, \forall x \in c_i, c_j$

Calculate $\nabla \mathcal{F}_{\mathcal{IBSI}}(c_r, c), \forall c \in C$

Fig. 1. Agglomerative IBSI algorithm

ent meeting room environments with varying number of participants [3]. The audio captured by multiple distant microphone channels is beamformed to get an enhanced signal using *BeamformIt* toolkit [24, 25]. 19 dimensional Mel frequency cepstral coefficients (MFCCs) are extracted per each frame with a frame rate of 10 ms and frame length of 30ms. These features are used as input features for the speaker diarization system. Speech/non-speech detection is performed using the SHOUT system [26].

It was observed that non-speech segments detected by the SHOUT system contained instances of laughter in the meeting conversations. Since the aim of using data from non-speech regions in the current method is to capture the background environment in the meeting, these laughter instances have to be eliminated from the non-speech regions before using them in clustering. Since the laughter segments detected as non-speech by the SHOUT system are usually very loud as they involve several people laughing together, they can be easily separated from silence/background noise in the recording. In the current study, we used a simple short-term spectral energy based system to detect the laughter segments in non-speech segments detected by the SHOUT system. The detected laughter segments are excluded from non-speech regions and the remaining data is used to train the non-speech background model.

The optimal value of γ in (5) for the aIBSI framework is obtained by picking the value of the parameter that minimized speaker error on RT 05 evaluation set of meetings which is used as a development set. The speaker error obtained for various values of γ on RT 05 evaluation set of meetings is plotted in Figure 3. These development experiments as reported in Figure 3 show that $\gamma = 0.1$ is optimal on the development set of meetings(RT 05). Therefore, the parameter value is fixed to 0.1 while testing on RT 06, 07 and 09 meetings. The value of β is fixed to 10 according to the prior work [6].

The performance of the two systems, the baseline aIB system and the proposed aIBSI system is measured in terms of Diarization Error Rate (DER) which is a standard metric used to evaluate speaker diarization systems in NIST RT evaluation campaigns [3] given a reference ground-truth segmentation. 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 clus-

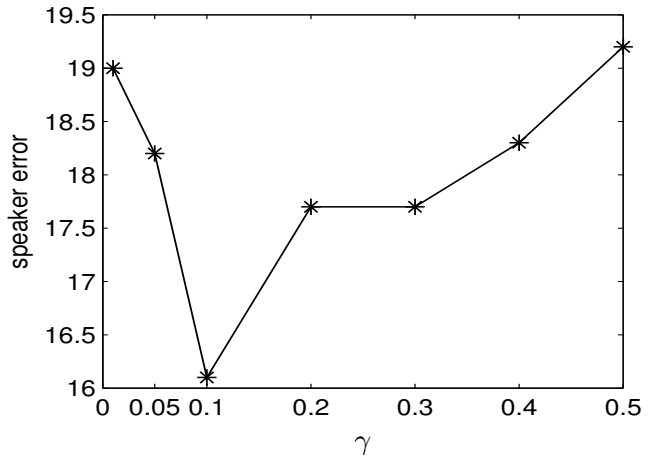

Fig. 3. *Speaker error for various values of γ on development set of meetings from RT05 eval set.*

tering 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 performance of the automatic speech/non-speech detector in terms of miss (Miss) and false-alarm (FA) rates on different test meeting sets is summarized in Table. 1. Since automatic speech/non-speech output is kept constant for all the meetings for both the baseline aIB framework and the proposed aIBSI framework, we compare the meeting wise speaker error or the error in the clustering for the two systems in Figure 2. It can be observed from Figure 2 that the proposed method either decreases or makes insignificant changes to DER for most of the meetings.

Table 1. *Speech/non-speech errors for RT 06, 07 and 09 sets.*

Data set	Miss	FA	TOTAL
RT-06	6.5	0.1	6.6
RT-07	3.7	0	3.7
RT-09	11.6	1.1	12.7
ALL	7.3	0.4	7.7

We compare the two methods aIB and aIBSI with a speaker diarization system based on HMM/GMM framework. The HMM/GMM system used in the current study was shown to give state-of-the-art performance in RT evaluation campaigns [5]. In this system the states of the HMM represent speakers and the emission probability distributions of the states are modelled using GMMs. The HMM/GMM system is initialized with uniform segmentation, resulting in 16 initial clusters (states). Then at each step of clustering, the closest clusters obtained using Bayesian information criterion (BIC) 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. In Table 2, we summarize the results at the data set level for the baseline aIB and the proposed aIBSI diarization systems and HMM/GMM system. It can be observed from table 2 that the proposed method decreases speaker error on all the meeting sets when compared to the baseline aIB system. The overall speaker error on all the three data sets is reduced from 16.6 to 13.3 by around 18.4% relative when compared to the baseline aIB system. Also, it can be observed that the proposed method reduces the gap between the state of the art HMM/GMM system and the aIB system by a significant margin. This is important because, diarization based on IB framework has been shown to be very fast [27] when compared to typical HMM/GMM based diarization framework, as it avoids multiple iterations of GMM re-estimation after each merge of clusters.

Table 2. *Speaker error for aIB, aIBSI diarization systems (with relative improvements over aIB baseline in parenthesis) on RT 06, 07 and 09 sets. We also report the performance of HMM/GMM system for comparison.*

Data set	aIB	aIBSI	HMM/GMM
RT-06	16.8	14.9 (+11.3)	13.6
RT-07	10.8	9.8 (+9.2)	6.4
RT-09	21.2	15.3 (+27.8%)	14.3
ALL	16.3	13.3 (+18.4%)	11.4

5. CONCLUSIONS AND FUTURE WORK

This paper proposed a method to improve information bottleneck based speaker diarization by incorporating information from non-speech regions of a given recording. The information was incorporated in the form of irrelevant variable set for clustering which is represented by a set of components of background GMM estimated over non-speech regions in the recording. Experimental results on meetings from RT 06, 07, 09 meeting sets have shown that, the proposed method decreases the speaker error or the clustering error on all the three data sets when compared to the baseline aIB diarization system. The combined speaker error on all the three data sets was reduced from 16.3% to 13.3% which is a reduction of around 18% relative to the baseline aIB system. Also, meeting level comparison between the two systems showed that the proposed method decreases the speaker error on most of the meetings.

As part of future work, we will run similar experiments to those conducted in this study on single channel distant microphone signals which have lower SNR than the enhanced signal used in the current study. We expect the relative improvements to be even better in this scenario. Also, in addition to non-speech based irrelevant variable we will also experiment on using lexical information as irrelevant variable to speaker clustering. A Multi Layer Perceptron (MLP) based phoneme posteriors can be used to represent the lexical information, where the set of phoneme posteriors can be used as irrelevant variable set for speaker clustering based on IBSI framework.

6. ACKNOWLEDGEMENTS

The authors would like to thank Dr. Fabio Valente for his useful comments on the work. This work was funded by the Swiss National Science Foundation through SNF-RODI grant and by the EU through FP7 SSPnet 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] "http://www.itl.nist.gov/iad/mig/tests/rt/," .

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.

[11] M. Huijbregts and C. Wooters, "The blame game: Performance analysis of speaker diarization system components," in *Interspeech*, Antwerp, Belgium, 2007, pp. 1857–1860.

[12] M. Huijbregts, D.A. van Leeuwen, and C. Wooters, "Speaker diarization error analysis using oracle components," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 2, pp. 393–403, feb. 2012.

[13] Mary Tai Knox, Nikki Mirghafori, and Gerald Friedland, "Where did i go wrong?: Identifying troublesome segments for speaker diarization systems," in *Interspeech*, Portland, USA, 2012.

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

[15] Martin Zelenák, Carlos Segura, and Javier Hernando, "Overlap detection for speaker diarization by fusing spectral and spatial features," in *Interspeech*, Makuhari, Japan, 2010, pp. 2302–2305.

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

[17] Sree Harsha Yella and Fabio Valente, "Speaker diarization of overlapping speech based on silence distribution in meeting recordings," in *Interspeech*, Portland, USA, 2012.

[18] 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.

[19] Xavier Anguera, Chuck Wooters, and Javier Hernando, "Purity algorithms for speaker diarization of meetings data," in *Acoustics, Speech and Signal Processing, IEEE Transactions on*, Toulouse, France, 2006.

[20] G Chechik and N Tishby, "Extracting relevant structures with side information," in *Advances in Neural Information Processing Systems*. 2003, pp. 857–864, MIT press.

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

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

[23] Gal Chechik, *Information theoretic approach to the study of auditory coding*, Ph.D. thesis, Hebrew University, July 2003.

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

[25] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 7, pp. 2011–2022, 2007.

[26] 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.

[27] Deepu Vijayasenan and Fabio Valente, "Diartk : An open source toolkit for research in multistream speaker diarization and its application to meetings recordings," in *Proceedings of Interspeech*, Portland, USA, 2012.