IDIAP RESEARCH REPORT

TWO-PASS IB BASED SPEAKER DIARIZATION SYSTEM USING MEETING-SPECIFIC ANN BASED FEATURES

Nauman Dawalatabad Srikanth Madikeri
C Chandra Sekhar Hema A Murthy

Idiap-RR-09-2018

JULY 2018

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Two-Pass IB based Speaker Diarization System using Meeting-Specific ANN based Features

Nauman Dawalatabad[1], *Srikanth Madikeri*[2], *C Chandra Sekhar*[1], *Hema A Murthy*[1]

[1]Indian Institute of Technology Madras, India
[2]Idiap Research Institute, CH-1920 Martigny, Switzerland

nauman@cse.iitm.ac.in, srikanth.madikeri@idiap.ch,
chandra@cse.iitm.ac.in, hema@cse.iitm.ac.in

Abstract

In this paper, we present a two-pass Information Bottleneck (IB) based system for speaker diarization which uses meeting-specific artificial neural network (ANN) based features. We first use IB based speaker diarization system to get the labelled speaker segments. These segments are re-segmented using Kullback-Leibler Hidden Markov Model (KL-HMM) based re-segmentation. The multi-layer ANN is then trained to discriminate these speakers using the re-segmented output labels and the spectral features. We then extract the bottleneck features from the trained ANN and perform principal component analysis (PCA) on these features. After performing PCA, these bottleneck features are used along with the different spectral features in the second pass using the same IB based system with KL-HMM re-segmentation. Our experiments on NIST RT and AMI datasets show that the proposed system performs better than the baseline IB system in terms of speaker error rate (SER) with a best case relative improvement of 28.6% amongst AMI datasets and 27.1% on NIST RT04eval dataset.

Index Terms: Speaker diarization, information bottleneck, ANN, bottleneck features.

1. Introduction

Speaker diarization is the task of identifying *"who spoke when?"* in an audio stream. The system does not assume any prior knowledge about the speakers or the number of speakers in a given audio [1, 2]. Speaker diarization has many applications, especially for tagging the audio in telephone conversations, broadcast news and meetings. Conversational meetings are spontaneous and therefore challenging. Diarization of meetings is a task that has received significant attention. The approaches to speaker diarization includes top-down, bottom-up, parametric and non-parametric clustering [1]. Bottom-up agglomerative clustering is the most popular approach. The state of the art speaker diarization systems include Hidden Markov Model/ Gaussian Mixture Model (HMM/GMM) [3, 4] and Information Bottleneck (IB) [5] systems.

Short term spectral features such as Mel Frequency Cepstral Coefficients (MFCC) are widely used for the task of speaker diarization [1]. Features like Mel Filterbank Slope (MFS) and Linear Filterbank Slope (LFS) have shown to be better at speaker discrimination compared with MFCC [6]. The i-vector based approaches have been applied to improve speaker models [7, 8]. Linear discriminant analysis (LDA) have been used for speaker diarization of telephone data [9]. Here, LDA is performed on MFCC after the initial system's output to obtain discriminative features which are then used in a fixed duration HMM for diarization. The diarization system itself is based on a 3-hyper-state HMM, for two speakers in the telephone conversation and a non-speech class. In [10], artificial neural network (ANN) based features have been found to be useful in adding discriminative information to the speaker diarization process. The ANN is trained on a large development set of meetings to determine whether two given speech segments came from the same or different speakers. The features extracted from the bottleneck layer of the ANN are then used along with the primary spectral features in a HMM/GMM system.

Speaker diarization involves segmentation and speaker clustering, where different speakers form different clusters. This requires features to be speaker discriminative. The discriminative information present in the output of the baseline IB system can be exploited. We, therefore propose a two-pass IB based speaker diarization system. The first pass includes the traditional IB system with KL-HMM re-segmentation [11]. This provides the first level segmentation of the meeting. Spectral features extracted from the segmented output are then used to train a meeting specific ANN. The bottleneck features (BNFs) are extracted from the penultimate layer of the trained ANN. These features are then used in the second pass, either independently or along with the spectral features. The second pass uses the same IB system with KL-HMM re-segmentation.

The rest of the paper is organized as follows: Section 2 discusses the IB based diarization system. Section 3 briefly describes the process of extracting ANN based features used in this paper. Section 4 presents the proposed two-pass IB based speaker diarization system. Section 5 presents experimental results. Finally, Section 6 concludes the paper.

2. Agglomerative Information Bottleneck

The agglomerative Information Bottleneck (aIB) [5] approach performs bottom-up agglomerative clustering based on the information bottleneck principle [12]. Let $\mathbf{X}$ represent a set of segments in an audio, $\mathbf{Y}$ represent a set of relevance variables that give meaningful information about the speaker in each segment, and $\mathbf{C}$ represent a clustering solution to $\mathbf{X}$. The aIB based approach converts the set of segments $\mathbf{X}$ into a set of clusters $\mathbf{C}$ that conveys as much information as possible about $\mathbf{Y}$. Each short segment is expected to contain only one speaker and is thus modelled by a Gaussian. According to the IB principle, any clustering $\mathbf{C}$ should be compact and should preserve as much information about the relevance variables $\mathbf{Y}$ as possible. Thus, the objective function is given by

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

where I denotes the mutual information and β is a Lagrange multiplier. The algorithm initializes each input segment $\mathbf{x_i} \in \mathbf{X}$ as a separate cluster and then iteratively merges the clusters such that the reduction in $\mathcal{F}$ is minimum. The change in the value of objective function can be represented as

$$\Delta\mathcal{F}(\mathbf{c_i}, \mathbf{c_j}) = (p(\mathbf{c_i}) + p(\mathbf{c_j})).d_{ij} \quad (2)$$

where the distance d_{ij} between the clusters $\mathbf{c_i}$ and $\mathbf{c_j}$ is given by

$$d_{ij} = JS[p(\mathbf{y}|\mathbf{c_i}), p(\mathbf{y}|\mathbf{c_j})] - \frac{1}{\beta} JS[p(\mathbf{x}|\mathbf{c_i}), p(\mathbf{x}|\mathbf{c_j})] \quad (3)$$

The Jensen-Shannon (JS) divergence for $JS[p(\mathbf{y}|\mathbf{c_i}), p(\mathbf{y}|\mathbf{c_j})]$ is given by

$$\pi_i D_{KL}[p(\mathbf{y}|\mathbf{c_i})||q_Y(\mathbf{y})] + \pi_j D_{KL}[p(\mathbf{y}|\mathbf{c_j})||q_Y(\mathbf{y})] \quad (4)$$

where D_{KL} denotes the Kullback-Leibler divergence, $q_Y(\mathbf{y})$ = π_i $p(\mathbf{y}|\mathbf{c_i})$ + π_j $p(\mathbf{y}|\mathbf{c_j})$ and $\pi_i = p(\mathbf{c_i})/(p(\mathbf{c_i}) + p(\mathbf{c_j}))$. The second term in (3) can be calculated in a similar manner. A pair of clusters with the minimum $\Delta\mathcal{F}$ are considered for merging and this is done in an iterative fashion. The new cluster $\mathbf{c_r}$ obtained by merging $\mathbf{c_i}$ and $\mathbf{c_j}$ is characterized by

$$p(\mathbf{c_r}) = p(\mathbf{c_i}) + p(\mathbf{c_j}) \quad (5)$$

$$p(\mathbf{y}|\mathbf{c_r}) = \frac{p(\mathbf{y}|\mathbf{c_i})p(\mathbf{c_i}) + p(\mathbf{y}|\mathbf{c_j})p(\mathbf{c_j})}{p(\mathbf{c_r})} \quad (6)$$

A threshold on the normalized mutual information (NMI) given by $\frac{\mathrm{I}(\mathbf{Y},\mathbf{C})}{\mathrm{I}(\mathbf{C},\mathbf{X})}$ is used to terminate the iterative method [5]. Once clustering terminates, the output is re-segmented using HMM/GMM or KL-HMM.

The IB approach provides a convenient way to combine different feature streams by fusing their respective posteriors [13]

$$p(\mathbf{y}|\mathbf{s_t^a}, \mathbf{s_t^b}) = p(\mathbf{y}|\mathbf{s_t^a})P_a + p(\mathbf{y}|\mathbf{s_t^b})P_b \quad (7)$$

where $\mathbf{s_t^a}$ and $\mathbf{s_t^b}$ are feature vectors at time t from feature streams a and b, respectively. Here, P_a and P_b are the weights assigned to the feature streams a and b respectively, such that P_a + P_b = 1. When multiple feature streams are used, the KL-HMM approach has been shown to perform better than the HMM/GMM approach for re-segmentation [11].

3. Extraction of ANN based features

ANNs with multiple layers have been used for speech, speaker and language recognition tasks [14, 15, 16]. ANNs can be used for both classification and feature extraction [16]. In this work, we have used ANNs for the purpose of feature extraction.

A multi-layer feed-forward neural network with two hidden layers is used. Neurons in the first hidden layer have a logistic sigmoid activation function while those in the second hidden layer have a linear activation function. The output layer neurons have a softmax activation function. Features are extracted from the second hidden layer. The ANN is trained to discriminate the speakers in the input meeting. Training is done using the output labels of the baseline IB system and the spectral feature stream. Unlike conventional system, features are not concatenated during training, as this would increase the complexity of ANN, thus increasing the parameters, the training time and the amount of data required for training.

The spectral features are input to the trained ANN to get the output of the second hidden layer which are referred as the bottleneck features (BNFs)[1] in this paper. The aim is to emphasize the discriminative information present in the input feature vectors, which is also similar to projecting the input feature vectors into a space where they are better separated than in the input space. As the ANN is trained with the speaker labels, the hidden layer learns the discriminative information. Bottleneck features are also known to capture information that is complementary to the input feature [16]. It is important to note that the ANN is specifically trained for a particular audio meeting, and thus does not require any development dataset for training.

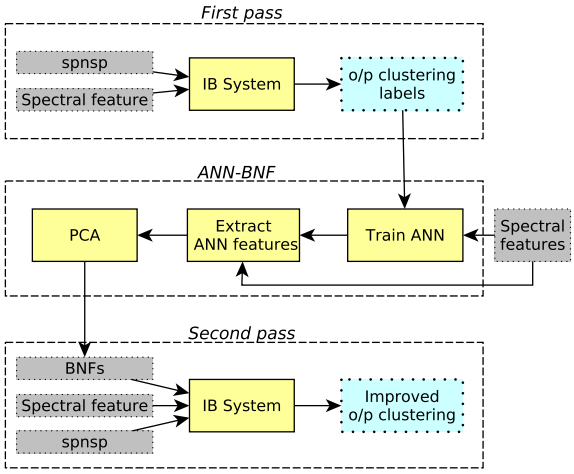

Figure 1: *Block diagram of a two-pass IB based speaker diarization system. The first pass is the conventional IB system whose output is used to train an ANN with the speaker labels obtained. The bottleneck features from the ANN are then used in the second pass independently or complementary to the spectral features.*

4. Two-pass IB diarization system

In this section, the proposed two-pass IB diarization system is described. The architecture of the system is given first. This is followed by a description of the different types of bottleneck features used.

4.1. Proposed system

The block diagram of the proposed two-pass IB based system is shown in Figure 1. The two-pass system refines the output of the diarization system to train an ANN and subsequently obtain additional discriminative information that is otherwise unavailable. The proposed algorithm consists of 3 steps:

1. *First pass*: The aim of the first pass is to obtain an initial set of labels to train the ANN. In this stage, speech and non-speech (spnsp) boundary details and the spectral features are given as input to the baseline IB system. The IB system initializes short segments of speech as individual clusters. The initial clusters are merged according to the IB criterion until the NMI threshold is satisfied. After clustering the initial segments, KL-HMM

[1]The term "bottleneck features" used in this paper denotes the output of the hidden layer of ANN and should not be confused with the feature compression, as in the case of auto-encoders.

Table 1: List of AMI meeting datasets.

AMI-1	ES2008c, ES2013a, ES2013c, ES2014d, ES2015a, IS1001c, IS1007a, IS1008c, IS1008d, IS1009c
AMI-2	ES2010b, ES2013b, ES2014c, ES2015b, ES2015c, IS1004b, IS1006c, IS1007c, IS1008a, IS1009d

re-segmentation is applied to realign the boundaries. The IB system diarizes the audio and outputs labelled speaker segments.

2. *ANN training and BNF extraction*: The output labels from the *first pass* along with the spectral feature stream of the segmented output are used to train a meeting-specific ANN. Bottleneck features corresponding to the feature stream used for training are extracted. Speaker clusters which contribute less than 3 seconds are discarded while training, as these are possibly spurious clusters. PCA is then applied on the bottleneck features to ensure that the features are whitened. This is required as the IB system models the speech segments as Gaussians. All the principal components are retained.

3. *Second pass*: The projected bottleneck features are used either independently or along with the spectral features using the same IB system with KL-HMM re-segmentation. When multiple features are used, the different feature streams are combined at the posterior level as given in (7). The output segments from the second pass is the final diarization output.

4.2. Discriminative bottleneck features

In this sub-section we describe different types of bottleneck features used in this paper. After obtaining the first pass output for an audio meeting, we make use of these labels to train the ANN. The bottleneck features depend on the spectral features used to train the ANN. In this paper, we have experimented with two different spectral features, namely, MFCC and MFS. The bottleneck feature extracted from the ANN trained on MFCC features is denoted by BN_{mfcc} while that extracted from ANN trained on MFS features is denoted as BN_{mfs}. These bottleneck features are used in the second pass after applying PCA.

5. Experiments and results

Experiments are performed on NIST RT04dev, RT04eval and RT05eval datasets from Linguistic Data Consortium (LDC) [17]. We also used 2 sets of meetings which are sub-sets of the Augmented Multi-Party Interaction (AMI) corpus [18]. Each set contains 5 randomly selected meetings each from the IDIAP (IS) and the Edinburgh (ES) groups. The list of meetings in each of the AMI sub-sets are shown in Table 1. RT04dev was used as development dataset to tune parameters of all models while the rest were used for testing. All the results reported in this paper are on the best performing parameters tuned on the development dataset. Multiple distant microphone (MDM) data was used for each meeting after beamforming using *BeamformIt* [19].

5.1. Systems and parameters

The aim of our experiments is to check whether the extracted bottleneck features helps in improving the performance in the second pass. We used the IB diarization toolkit [20] in all our experiments. Matlab's Neural Network toolbox was used to train the ANN and extract bottleneck features. The HTK toolkit [21] was used to extract MFCC and mel-filter bank energies. These features were extracted from the beamformed audio at a frame rate of 10ms with an analysis window size of 25ms. MFS was calculated from mel-filter bank energies [6]. After DCT, 19 coefficients were retained for both MFCC and MFS. Based on the performance on the development set, the values of β and NMI threshold were set to 10 and 0.3, respectively. Maximum segment duration for IB system was limited to 2.5 seconds. The hidden layers of the ANN (h1-h2) were configured to 30-19, where h1 and h2 represent the number of neurons in the respective hidden layers. Speech and non-speech (spnsp) details were obtained from the ground-truth. The speaker diarization performance is reported in terms of the speaker error rate (SER). Real time factor (RTF), which is defined as the ratio of the run-time to the length of an input audio, is used to evaluate the run-times of the systems presented.

5.2. Results

The results of the evaluation for different datasets are reported in Table 2. The SER after the second pass is compared with the baseline IB system. Based on the input features, three baseline systems (I, II and III) are used for comparison. The input spectral features to the baseline IB systems I, II and III are MFCC, MFS and MFCC+MFS, respectively. Similarly, depending on the input feature combinations used in the second pass of the proposed system, three different systems (A, B and C) are used for comparison. The first pass for A, B and C are the baseline IB systems I, II and III, respectively. For example, in the case of the MFCC+BN_{mfs} under system B, the BN_{mfs} features were extracted by training the ANN with the output labels of system II, which is then used in the second pass along with the MFCC features. The weights used for feature fusion are indicated in parentheses. The performance (SER) of the systems A, B and C is compared with the baseline IB systems I, II and III, respectively. For each dataset, the SER of the proposed system with the best case relative improvement (R.I) is indicated using a bold font.

It can be inferred from Table 2 that the proposed system performs better than the baseline IB system in most cases. On the development set when MFCC+BN_{mfcc} was used in the second pass, a relative improvement (R.I) of 15.9% was observed compared to the baseline system I. Inclusion of MFS with BN_{mfcc} has shown R.I of 16.6%. This was expected as MFS features are better at speaker discrimination than MFCC [6]. The R.I of 28.6% for system A is observed for the AMI-1 set when MFS+BN_{mfcc} were used in the second pass. On the other hand for AMI-2, a best case R.I of 16.3% compared to system III is observed when only BN_{mfs} was used in the second pass. A best case R.I of 21.9% compared to system I was observed for RT05eval when MFCC+BN_{mfcc} were used in the second pass. For RT04eval, R.I of 27.1% was observed when MFCC+MFS+BN_{mfcc} were used. In general, even the standalone bottleneck features were observed to improve the performance of the baseline IB system. In combination with the spectral features, the diarization performance improved further. This confirms the conjecture that the bottleneck features included in the second pass are indeed discriminative.

Figure 2 shows the performance of the proposed system for different values of feature fusing weights on the development dataset. Figures 2(a) and 2(b) shows the performance after the second pass, with first pass being the baseline systems I and II,

Table 2: Results of experiments are reported in Speaker Error Rate (SER). Performance of systems A, B and C are compared with the baseline systems I, II and III, respectively. For each dataset, the SERs with the best case relative improvements are indicated with bold font. The feature fusing weights are mentioned in parentheses.

	System/Dataset	Dev. set	Test set			
		RT04dev	RT04eval	RT05eval	AMI-1	AMI-2
	Baseline IB system					
I	MFCC	14.5	13.8	19.2	22	23.4
II	MFS	14.5	13.5	16.4	16.9	21.3
III	MFCC+MFS (0.7, 0.3)	13.3	15.5	17.9	19.2	24.6
	Proposed two-pass system					
A	BN_{mfcc}	13.1	14.3	15.7	19.2	21.6
	MFCC+BN_{mfcc}(0.2, 0.8)	12.2	12	**15**	19.1	21.4
	MFS+BN_{mfcc}(0.7, 0.3)	**12.1**	13.3	17	**15.7**	21.8
B	BN_{mfs}	14	13	14.1	13.7	20.5
	MFS+BN_{mfs}(0.3, 0.7)	12.9	10.9	14.4	13.6	21.3
	MFCC+BN_{mfs}(0.3, 0.7)	12.9	11.1	14.5	13.7	21
C	BN_{mfcc}	16.4	12.7	18.5	16.5	22.2
	BN_{mfs}	13.9	12.9	15.7	16.4	**20.6**
	MFCC+MFS+BN_{mfcc} (0.2, 0.3, 0.5)	12.9	**11.3**	16.5	16.8	22.3
	MFCC+MFS+BN_{mfs} (0.2, 0.4, 0.4)	12.3	13.8	16.1	15.4	21.7

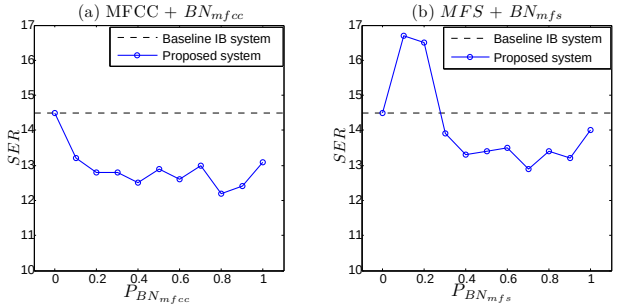

Figure 2: *SER comparison with baseline for different systems on development set. With P_a = 1-P_b, where, a and b are 2 different feature streams used in second pass of the proposed system.*

respectively. For most of the fusion weight combinations, the proposed system shows improved performance. On the development set, the observed average error in estimating accurately the number of speakers by the baseline system is 2.0 while that for the proposed system is 1.13.

The advantage of keeping the size of ANN small is that, it avoids over-fitting to the training data. The training time is also small, which enables the use of meeting specific ANNs. The performance of the proposed system is evaluated on a machine with 4 cores. The maximum number of ANN training epochs were set to 1000. The RTF on the development data for the baseline IB system is 0.02 while the RTF is 0.17 for the proposed system.

To check the effectiveness of the discriminative nature of the bottleneck features, we conducted an experiment where instead of using output labels from the baseline (first pass), the labels were taken from the ground-truth for training the ANN. ANNs were trained using MFCC and MFS features and corresponding bottleneck features were extracted. In the second pass, bottleneck features (BN_{mfcc} / BN_{mfs}) alone were used for diarization using IB system. The SER obtained in this experiment is shown in Table 3. As labels were considered from the ground-truth, this also provides a lower bound on SER when bottleneck features alone are used for diarization using the IB system with this framework.

Table 3: Lower bounds on SER when only the bottleneck features are used in the second pass of the proposed system.

Sys./Dataset	NIST RT			AMI	
	04dev	04eval	05eval	AMI1	AMI2
BN_{mfcc}	10.2	9.2	9.6	8.3	11
BN_{mfs}	10.5	9.1	7.4	8.7	12.9

6. Conclusion and Future work

A two-pass IB based speaker diarization system is proposed. The system is tested on NIST RT and AMI datasets. Inclusion of discriminative information in the second pass in the form of bottleneck features improves the performance of the baseline IB system. With a trade-off in terms of increased running time, the proposed system performs significantly better than the baseline system. The best case relative improvements of 27.1% (RT04eval) amongst NIST RT datasets and 28.6% (AMI-1) amongst AMI datasets supports our hypothesis.

In future, we plan to explore the framework by continuously looping through both the passes. This may be helpful in further refining the speaker boundaries. We also plan to investigate the output of the baseline, so as to provide more precise labelling information for training the ANN.

7. Acknowledgement

This work was supported by the Defence Research and Development Organisation (DRDO), India under the project CSE1314142DRDOHEMA.

8. References

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

[2] S. Tranter and D. Reynolds, "An Overview of Automatic Speaker Diarization Systems," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1557–1565, 2006.

[3] J. Ajmera and C. Wooters, "A Robust Speaker Clustering Algorithm," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2003, pp. 411–416.

[4] C. Wooters and M. Huijbregts, *Multimodal Technologies for Perception of Humans*. Springer Berlin Heidelberg, 2008, ch. The ICSI RT07s Speaker Diarization System, pp. 509–519.

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

[6] S. Madikeri and H. Bourlard, "Filterbank Slope based Features for Speaker Diarization," *IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP)*, 2014.

[7] S. Shum, N. Dehak, E. Chuangsuwanich, D. Reynolds, and J. Glass, "Exploiting Intra-Conversation Variability for Speaker Diarization," in *Proceedings of Interspeech*, 2011.

[8] S. Madikeri, P. Motlicek, and H. Bourlard, "Combining SGMM Speaker Vectors and KL-HMM Approach for Speaker Diarization," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[9] I. Lapidot and J. F. Bonastre, "Integration of LDA into a Telephone Conversation Speaker Diarization System," in *IEEE 27th Convention of Electrical Electronics Engineers in Israel (IEEEI)*, 2012.

[10] S. H. Yella, A. Stolcke, and M. Slaney, "Artificial Neural Network Features for Speaker Diarization," in *IEEE Spoken Language Technology Workshop (SLT)*, 2014, pp. 402–406.

[11] S. Madikeri and H. Bourlard, "KL-HMM based Speaker Diarization System for Meetings," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4435–4439.

[12] N. Tishby, F. C. Pereira, and W. Bialek, "The Information Bottleneck Method," *in NEC Research Institute TR*, 1998.

[13] D. Vijayasenan, F. Valente, and H. Bourlard, "An Information Theoretic Combination of MFCC and TDOA Features for Speaker Diarization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 2, pp. 431–438, 2011.

[14] G. Hinton, L. Deng, D. Yu, G. Dahl, A. R. Mohamed, N. Jaitly, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep Neural Networks for Acoustic Modeling in Speech Recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[15] F. Richardson, S. Member, D. Reynolds, and N. Dehak, "Deep Neural Network Approaches to Speaker and Language Recognition," *IEEE Signal Processing Letters*, vol. 22, no. 10, pp. 1671–1675, 2015.

[16] D. Yu and M. Seltzer, "Improved Bottleneck Features Using Pretrained Deep Neural Networks," in *Proceedings of Interspeech*, 2011.

[17] "Linguistic Data Consortium," https://catalog.ldc.upenn.edu/.

[18] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The AMI Meeting Corpus: A Pre-announcement," in *Proceedings of the Second International Conference on Machine Learning for Multimodal Interaction*, ser. MLMI'05, 2006, pp. 28–39.

[19] X. Anguera, C. Wooters, and J. Hernando, "Acoustic Beamforming for Speaker Diarization of Meetings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2011–2022, 2007.

[20] D. Vijayasenan and F. Valente, "DiarTk: An Open Source Toolkit for Research in Multistream Speaker Diarization and its Application to Meetings Recordings," in *Proceedings of Interspeech*, 2012.

[21] "HTK toolkit," http://htk.eng.cam.ac.uk/.