*INTERSPEECH 2019*
September 15–19, 2019, Graz, Austria

# ViVoLAB Speaker Diarization System for the DIHARD 2019 Challenge

*Ignacio Viñals, Pablo Gimeno, Alfonso Ortega, Antonio Miguel, Eduardo Lleida*

ViVoLab, Aragón Institute for Engineering Research (I3A), University of Zaragoza, Spain
{ivinalsb, pablogj, ortega, amiguel, lleida}@unizar.es

## Abstract

This paper presents the latest improvements in Speaker Diarization obtained by ViVoLAB research group for the 2019 DIHARD Diarization Challenge. This evaluation seeks the improvement of the diarization task in adverse conditions. For this purpose, the audio recordings involve multiple scenarios with no restrictions in terms of speakers, overlapped speech nor quality of the audio. Our submission follows the traditional segmentation-clustering-resegmentation pipeline: Speaker embeddings are extracted from acoustic segments with a single speaker on them, later clustered by means of a PLDA. Our contribution in this work is focused on the clustering step. We present results with our Variational Bayes PLDA clustering and our tree-based clustering strategy, which sequentially assigns the different embeddings to its corresponding speaker according to a PLDA model. Both strategies compare multiple diarization hypotheses and choose their candidate one according to a generative criterion. We also analyze the impact of the different available embeddings in the state-of-the-art with both clustering approaches.

**Index Terms**: diarization, DIHARD Challenge, PLDA, Variational Bayes, Tree search, M-algorithm

## 1. Introduction

Speech signal is a very rich source of information with multiple levels of knowledge: from low level speech and speaker information to higher levels such as emotion or context. Some of them are worthy to be extracted. Therefore speech technologies have developed multiple tasks to process each sort of desired information. Diarization is the task focused on explaining a given audio in terms of the active speaker at each time. Thus diarization must identify the speaker every time somebody talks and differentiate his/her speech from the others.

Diarization is applied to multiple scenarios, such as telephone, meetings or broadcast. Each one of these scenarios presents its own particularities (number of speakers, noise, etc.) making solutions domain dependent. DIHARD 2019 evaluation seeks the improvement of diarization regardless of the scenario. For this reason the database counts with audios from multiple domains such as Youtube, court trials, meetings, etc. These domains are characterized to suffer from adverse conditions as unknown levels of noise or reverberation. Four conditions are presented, including single microphone (tracks 1 and 2) and multiple microphone (tracks 3 and 4) scenarios. Odd and even tracks only differ in the Voice Activity Detection (VAD), provided by the organization in tracks 1 and 3, and user made in tracks 2 and 4.

For DIHARD 2019 challenge ViVoLAB team has prepared a submission around the Bottom-Up diarization strategy: First we divide the input audio into segments with a single speaker by means of Bayesian Information Criterion [1] approaches. These segments, converted into a compact representation either by linear, e.g. i-vectors [2], or nonlinear solutions, i.e. x-vectors [3], are clustered afterwards. Regarding clustering, two strategies are considered: a diarization by means of Variational Bayes (VB) resegmentation [4, 5, 6] and a new diarization approach based on tree-modeling and helped by the M-algorithm [7]. The obtained labels are finally refined by means of HMMs with eigenvoice priors [8].

The paper is organised as follows. Our Variational Bayes clustering is explained in Section 2. Section 3 describes the tree search clustering approach. The experimental setup is commented in Section 4. In Section 5 contains our results. Finally, we include our conclusions in Section 6.

## 2. Clustering by means of Variational Bayes

The Bottom-Up approach in diarization consists of dividing the given audio into a set of of N segments $\mathbf{S} = \{s_1, s_2, ..., s_N\}$, compactly represented by the set of embeddings $\mathbf{\Phi} = \{\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, ..., \boldsymbol{\phi}_N\}$. These representations are clustered so elements from the same speaker are together. This is the same as labeling the embeddings with a partition $\Theta = \{\theta_1, \theta_2, ..., \theta_N\}$ where elements from the same cluster share a common label. Therefore clustering is an assignment task in which we must find those labels $\Theta$ which better explain the given data $\mathbf{\Phi}$. Mathematically:

$$\Theta_{\text{diar}} = \arg\max_{\Theta} P\left(\Theta|\mathbf{\Phi}\right) \tag{1}$$

In this clustering approach [9] we estimate the conditional distribution $P\left(\Theta|\mathbf{\Phi}\right)$ in terms of a PLDA [10] model and then carry out the maximization. This inference of the desired distribution is done by means of the Fully Bayesian PLDA [4] solved by Variational Bayes (VB).

The Fully Bayesian PLDA is a generalization of the PLDA model [10]. Given a set of $N$ elements from $M$ speakers, the original model assumes the assignment labels to be known. However, in our version a latent variable $\Theta$ is used instead, being in charge of the assignment. Hence, a set of $N$ embeddings is modeled by $M$ speakers as follows:

$$P\left(\mathbf{\Phi}|\Theta\right) = \prod_{j=1}^{N}\prod_{i=1}^{M}\mathcal{N}\left(\boldsymbol{\phi}_j|\boldsymbol{\mu} + \mathbf{V}\mathbf{y}_i, \mathbf{W}^{-1}\right)^{\theta_{ij}} \tag{2}$$

The hidden variable $\Theta$, representing the speaker labels, is modeled as a multinomial distribution, with a Dirichlet prior $\pi_\theta$. The Fully Bayesian approach also substitutes the model parameters ($\boldsymbol{\mu}$, $\mathbf{V}$ and $\mathbf{W}$), point estimates in the original, by latent variables, each one with its own prior $\boldsymbol{\alpha}$, to gain robustness. Its Bayesian Network is shown in Fig. 1 and further information about the model is in the original work [4].

The described model allows the definition of the joint conditional distribution $P\left(\mathbf{Y}, \Theta|\mathbf{\Phi}\right)$, which models the speaker variable $\mathbf{Y}$ and the speaker labels $\Theta$ given the embeddings $\mathbf{\Phi}$.



http://dx.doi.org/10.21437/Interspeech.2019-2462

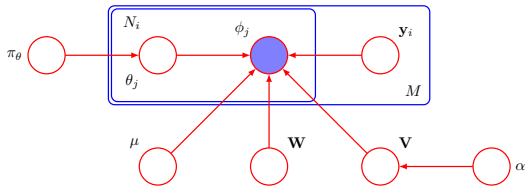


Figure 1: *Bayesian Network of Fully Bayesian PLDA*

However, the marginalization of the speaker latent variable $\mathbf{Y}$ is not straightforward.

In consequence we propose the approximation of the posterior by means of Variational Bayes. Its application approximates the joint posterior by a factorial of independent distributions, each one only dependent on a latent variable from our model. The considered decomposition is:

$$P(\mathbf{Y}, \Theta, \pi_\theta, \boldsymbol{\mu}, \mathbf{V}, \mathbf{W}, \boldsymbol{\alpha}|\boldsymbol{\Phi}) \approx \tag{3}$$
$$q(\mathbf{Y})\, q(\Theta)\, q(\pi_\theta)\, q(\boldsymbol{\mu})\, q(\mathbf{V})\, q(\mathbf{W})\, q(\boldsymbol{\alpha}) \tag{4}$$

Despite the closed formulation of the factor $q(\theta)$, its maximization requires an iterative reevaluation of those factors related to the labels ($q(\mathbf{Y})$, $q(\Theta)$ and $q(\pi_\theta)$) obtaining the labels by a cartesian ascent optimization. This solution requires an initial value for the distributions, with severe implications in the performance. Therefore multiple seeds are provided, opting for one of them according to a penalized Lower Bound [6].

## 3. Tree-based model for diarization purposes

Our alternative proposal reinterprets the maximization of the posterior probability $P(\Theta|\boldsymbol{\Phi})$. Instead of the estimation of the posterior distribution, we can carry out the maximization of the loglikelihood of the joint distribution as:

$$\Theta_{\text{diar}} = \arg\max_{\Theta} P(\Theta|\boldsymbol{\Phi}) = \arg\max_{\Theta} P(\boldsymbol{\Phi}|\Theta)\, P(\Theta) \tag{5}$$

### 3.1. The model

In [7] we propose a definition of $P(\boldsymbol{\Phi}|\Theta)$ where the set of embeddings $\boldsymbol{\Phi} = \{\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, ..., \boldsymbol{\phi}_N\}$ is arranged in a sequence so the turn $j$ is either assigned to an existing cluster or classified as the first element of a new cluster. This decision is made according to speaker models constructed with the previous embeddings in the sequence already classified. Mathematically

$$P(\boldsymbol{\Phi}|\Theta) = \prod_{j=1}^{N}\prod_{i=1}^{M} P\left(\boldsymbol{\phi}_j|\boldsymbol{\Phi}_i^{(j-1)}\right)^{\theta_{ij}} \tag{6}$$

where $\boldsymbol{\phi}_j$ represents the $jth$ embedding and $\theta_{ij}$ the label assigning the element $j$ to the speaker cluster $i$. For this purpose $\theta_{ij}$ only has a value of one if the embedding $j$ corresponds to the cluster $i$, being zero otherwise. The variable $\boldsymbol{\Phi}_i^{(j-1)}$ describes the previously assigned embeddings to cluster $i$.

Our definition of the conditional probability $P\left(\boldsymbol{\phi}_j|\boldsymbol{\Phi}_i^{(j-1)}\right)$ is done in terms of the PLDA model, which uses a latent speaker variable to model the embeddings. This latent variable is conditioned by the elements already assigned to the hypothesis cluster to be later marginalized. The imposition of Gaussian distributions makes our desired distribution Gaussian as

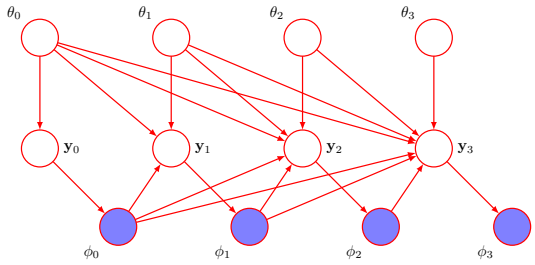


Figure 2: *Bayesian Network for out Tree Decoding model for a 4-element sequence*

$$P\left(\boldsymbol{\phi}_j|\boldsymbol{\Phi}_i^{(j-1)}\right) = \int P(\boldsymbol{\phi}_j|\mathbf{y}_{ij})\, P\left(\mathbf{y}_{ij}|, \boldsymbol{\Phi}_i^{(j-1)}\right) d\mathbf{y}_{ij} \tag{7}$$
$$= \mathcal{N}(\boldsymbol{\phi}_j|\boldsymbol{\mu}_{ij}, \Sigma_{ij}) \tag{8}$$

whose mean $\boldsymbol{\mu}_{ij}$ and variance $\Sigma_{ij}$ are defined in terms of PLDA parameters ($\boldsymbol{\mu}$, $\mathbf{V}$ and $\mathbf{W}$) by

$$\boldsymbol{\mu}_{ij} = \boldsymbol{\mu} + \mathbf{V}\boldsymbol{\mu}_{\mathbf{y}_{ij}} \tag{9}$$
$$\Sigma_{ij} = \mathbf{W}^{-1} + \mathbf{V}\Sigma_{\mathbf{y}_{ij}}\mathbf{V}^T \tag{10}$$

These terms depend on the posterior distribution of $\mathbf{y}_{ij}$, estimated in terms of the previous decisions $\boldsymbol{\Phi}_i^{(j-1)}$. Its mean $\boldsymbol{\mu}_{\mathbf{y}_{ij}}$ and variance $\Sigma_{\mathbf{y}_{ij}}$ are given by

$$\boldsymbol{\mu}_{\mathbf{y}_{ij}} = \Sigma_{\mathbf{y}_{ij}}\mathbf{V}^T\mathbf{W}\sum_{k=1}^{j-1}\theta_{ik}\,(\boldsymbol{\phi}_k - \boldsymbol{\mu}) \tag{11}$$

$$\Sigma_{\mathbf{y}_{ij}}^{-1} = \mathbf{I} + \mathbf{V}^T\sum_{k=1}^{j-1}\theta_{ik}\mathbf{W}\mathbf{V}. \tag{12}$$

The label prior $P(\Theta)$ is modeled based one the Distance-dependent Chinese Restaurant (DDCR) process [11, 12]. This process is a probability distribution over partitions, assigning a sequence of elements to different clusters. The element $\theta_j$, at the $jth$ turn, keeps in the last assigned cluster with probability $p_0$. Otherwise, it is either assigned to an existing cluster $k = 1..K$ proportionally to the occupation of clusters at time $j$ or assigned to a new empty cluster with a certain probability. Hence

$$P\left(\theta_j = k|\theta_1^{(j-1)}\right) \propto \begin{cases} p_0 & \text{if } k = \theta_{(j-1)} \\ n_k & \text{if } k \neq \theta_{(j-1)} \text{ and } k \leq K \\ \alpha & \text{if } k \neq \theta_{(j-1)} \text{ and } k = K+1 \end{cases} \tag{13}$$

where $n_k$ is the number of times in which variable $\theta$ was assigned to cluster $k$. The Bayesian network for the whole model can be seen in Fig. 2.

The model we have presented has the shape of a decision tree where in which every node represents an assignment decision. The number of branches depending on a node represent the number of decisions we can make. Therefore, each path from the root of the tree (the first element) to a leaf (the last embedding) is a possible partition of the set to cluster.

### 3.2. The maximization

The inference of the partition which maximizes the model is not straightforward. In fact, except for short sequences with very low number of speakers, a brute force approach comparing all possible partitions is unfeasible, as analyzed in [13]. Moreover, efficient search algorithms such as Viterbi cannot be applied here. In our model any decision is conditioned to previous choices, not allowing us to use the Markov assumption.

An efficient technique to deal with tree structures is the M algorithm, widely spread in Communications. Each time $j$ we propagate a handful of $L$ surviving paths along all possible branches from each node. Considering each node has $M$ branches, i.e. from every node we can transit to the $M$ candidate speakers, we take into account up to $LM$ candidate paths to progress through. This selection of paths is ranked in terms of likelihood, only considering the top $L$ paths as the most promising candidates to propagate to time $j+1$. This process is repeated until the end of the audio, when the most promising path alive, the one with higher log-likelihood, is chosen as the diarization labeling. By this technique, we limit an exponential number of partitions $O\left(N^M\right)$ from the brute force approach to a linear problem $(O\left(MNL\right))$.

## 4. Experimental setup

### 4.1. Data resources

DIHARD 2019 challenge imposes very low restrictions regarding the data. Except for a few datasets excluded due to be part in DIHARD subset, any other source of knowledge is allowed. In our case we have constructed a data pool trying to combine as many scenarios as possible. The MGB 2015 broadcast dataset [14] contributes with more than 1000 hours of labeled data from BBC in many scenarios. These data is combined with AMI [15] and ICSI [16], both meeting datasets with multiple speakers recorded with different microphones. Youtube domain is also included in the pool with Voxceleb I [17] and II [18]. Finally, DIHARD 2019 development set is also included in the data pool for adaptation purposes.

### 4.2. Speaker representation & clustering

Our diarization system works under the Bottom-Up approach and relying on the embedding-PLDA paradigm. Hence we need to convert the input audio into tractable representations with speaker awareness.

Any given audio is first converted into a stream of feature vectors. 20-coefficient MFCCs arrays are extracted from windows of 25ms, and considering window shifts of 10ms. These features are then used to carry out the initial segmentation with Bayesian Information Criterion [1]. Voice Activity Detection (VAD) information is obtained with a 2-layer BLSTM neural network, trained with DIHARD development set.

The obtained segments are then converted into compact representations. Two main representations are considered. On the one hand our baseline system follows our contribution in [6] using i-vectors [2]. A 512 Gaussian GMM is followed by a 200-dimension T matrix, both trained using MGB, AMI and ICSI. On the other hand, we also use x-vector [3] networks trained on VoxCeleb, AMI and ICSI. Multiple architectures varying both the size of the network and the exact training pool are considered. Before clustering both sorts of embeddings are centered, whitened and length-normalized [19].

Finally, clustering is done according to PLDA [10] models. While i-vectors work with 200-dimension PLDAs trained on AMI, MGB and ICSI, x-vectors are clustered in terms of a low dimension (50) PLDA, trained only on AMI and ICSI.

Table 1: *DER(%) results with the core systems. Results obtained with Variational Bayes (VB PLDA) and the tree-based sequential clustering (Tree Clustering) approaches. i-vectors and x-vectors are compared. Results obtained with eval. set.*

| Embedding | VB PLDA | Tree Clustering |
|---|---|---|
| Track1 | | |
| i-vector | 25.95 | **25.67** |
| x-vector | 25.70 | 27.50 |
| Track2 | | |
| i-vector | 38.99 | 38.49 |
| x-vector | **38.29** | 40.07 |

### 4.3. Resegmentation

Results provided by the clustering stage are later refined by means of resegmentation in order to improve the borders. To do so we make use of the resegmentation by HMMs and eigenvoices proposed in [8]. The same i-vector extractor (512 Gaussian and 200-dimension T matrix) from the speaker representation is considered here.

## 5. Results

Our experimentation is centered in the single microphone paradigm, i.e. tracks 1 and 2. For comparison reasons DIHARD evaluation considers DER (Diarization Error Rate) as the scoring metric for evaluation of performance and ranking.

Our submission can be divided into two stages. Our first step is the study of the core diarization system, that is, the segmentation and clustering steps. This part includes the analysis of speaker embeddings, representing the audio information, and clustering strategies. Table 1 shows the results with our candidate core systems.

According to the results in Table 1, the performance of our core systems is consistent between tracks 1 and 2. Two clustering setups, Tree-based and VBPLDA with i-vectors and x-vectors respectively, have the best performance in both tasks, with no significant differences between them. Regarding the other two systems, while VBPLDA with i-vectors performs slightly worse, the tree-based x-vector system obtains the worst score.

Surprisingly i-vectors are still competitive respect to x-vectors, many times outperforming them despite their well known performance capabilities. Regarding the clustering techniques, VBPLDA seems to be more robust to configuration changes than the Tree Clustering, which has obtained the best and worst results.

Furthermore, we have observed a high domain mismatch in the development set. As we can see in Table 2, trends of behaviour are completely different, with i-vectors always outperforming x-vectors.

Once we have analyzed the core diarization systems, we add the resegmentation stage, based on HMM with eigenvoices. The results are included in Table 3. They show a consistent improvement of the results obtained by all the core systems, with benefits around 0.5% in track 1 and 1.0% in track 2. Again our

Table 2: *DER(%) results with the core systems. Results obtained with Variational Bayes (VB PLDA) and the tree-based sequential clustering (Tree Clustering) approaches. i-vectors and x-vectors are compared. Results obtained with dev. set.*

| Embedding | VB PLDA | Tree Clustering |
|---|---|---|
| Track1 | | |
| i-vector | 24.85 | **24.65** |
| x-vector | 25.03 | 26.72 |
| Track2 | | |
| i-vector | 31.95 | **31.47** |
| x-vector | 32.55 | 33.20 |

Table 3: *DER(%) results with the core systems with resegmentation. Results obtained with Variational Bayes (VB PLDA) and the tree-based sequential clustering (Tree Clustering) approaches. i-vectors and x-vectors are compared. Results obtained with eval. set.*

| Embedding | VB PLDA | Tree Clustering |
|---|---|---|
| Track1 | | |
| i-vector | 25.55 | **25.02** |
| x-vector | 25.51 | 26.71 |
| Track2 | | |
| i-vector | 38.04 | **37.22** |
| x-vector | 37.63 | 39.03 |

tree based system is able to obtain the best and worst performance. However, this time the Variational Bayes PLDA works slightly worse with both embeddings, i-vectors and x-vectors.

# 6. Conclusions

Our work provides a comparative between our two clustering approaches, the well known Variational Bayes PLDA clustering and our new Tree-Based clustering. Both have been tested with two types of embeddings, i-vectors and x-vectors.

The direct comparison between the clustering methods has shown that our new sequential approach is capable to outperformn our Variational Bayes approach. However, those experiments carried out with x-vectors obtained the worst marks, illustrating a greater robustness of the VB strategy. This extra robustness of the VB approach can be caused by a lower number of tuneable hyperparameters (1 vs 3). Besides our results with i-vectors have demonstrated great performance with our two clustering strategies, many times outperforming the x-vectors.

Research carried out with development set has shown great domain mismatch issues, highly noticeable at PLDA level. This type of degradation is very adverse for our two clustering approaches, both PLDA based. However, this mismatch originated earlier in the embedding extraction step. While PLDA is commonly adapted in the state-of-the-art, DNN embedding extraction lacks from some method to adapt some trained network to unmatched data. Otherwise, these neural networks can include error terms too complex to mitigate afterwards. Further research should be done to provide methods to successfully adapt these relevant models.

# 7. Acknowledgments

This work has been supported by the Spanish Ministry of Economy and Competitiveness and the European Social Fund through the project TIN2017-85854-C4-1-R, Government of Aragón (Reference Group T36_17R) and co-financed with Feder 2014-2020 "Building Europe from Aragón". We gratefully acknowledge the support of NVIDIA Corporation with the donation of a Titan Xp GPU.

# 8. References

[1] S. S. Chen and P. Gopalakrishnam, "Speaker Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion," in *DARPA Broadcast News Worshop*, 1998, pp. 127–132.

[2] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-End Factor Analysis For Speaker Verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 788–798, may 2011.

[3] D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, Y. Carmiel, and S. Khudanpur, "Deep Neural Network-based Speaker Embeddings for End-to-end Speaker Verification," *iEEE Spoken Language Technology Workshop (SLT)*, pp. 165–170, 2016.

[4] J. Villalba and E. Lleida, "Unsupervised Adaptation of PLDA By Using Variational Bayes Methods," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2014, pp. 744–748.

[5] I. Viñals, A. Ortega, J. Villalba, A. Miguel, and E. Lleida, "Domain Adaptation of PLDA models in Broadcast Diarization by means of Unsupervised Speaker Clustering," *Interspeech*, pp. 2829–2833, 2017.

[6] I. Viñals, P. Gimeno, A. Ortega, A. Miguel, and E. Lleida, "Estimation of the Number of Speakers with Variational Bayesian PLDA in the DIHARD Diarization Challenge," *Interspeech*, no. September, pp. 2803–2807, 2018.

[7] I. Viñals, A. Ortega, A. Miguel, and E. Lleida, "Tree-based Search Strategy for Clustering in Speaker Diarization Using the M-Algorithm," *Interspeech 2019 (submitted)*, 2019.

[8] M. Diez, L. Burget, and P. Matejka, "Speaker Diarization based on Bayesian HMM with Eigenvoice Priors," *Proceedings of Odyssey 2018 - The Speaker and Language Recognition Workshop*, no. June, pp. 147–154, 2018.

[9] J. Villalba, A. Ortega, A. Miguel, and E. Lleida, "Variational Bayesian PLDA for Speaker Diarization in the MGB Challenge," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 667–674.

[10] S. J. D. Prince and J. H. Elder, "Probabilistic Linear Discriminant Analysis for Inferences About Identity," in *Proceedings of the IEEE International Conference on Computer Vision*, 2007.

[11] D. M. Blei and P. I. Frazier, "Distance Dependent Chinese Restaurant Processes," *Journal of Machine Learning Research (JMLR)*, vol. 12, pp. 2461–2488, 2011.

[12] A. Zhang, Q. Wang, Z. Zhu, J. Paisley, and C. Wang, "Fully Supervised Speaker Diarization," pp. 2–6, 2018.

[13] N. Brümmer and E. de Villiers, "The Speaker Partitioning Problem," *ODYSSEY The Speaker and Language Recognition Workshop*, no. July, pp. 194–201, 2010.

[14] P. Bell, M. J. F. Gales, T. Hain, J. Kilgour, P. Lanchantin, X. Liu, A. McParland, S. Renals, O. Saz, M. Wester, and P. C. Woodland, "The MGB Challenge: Evaluating Multi-Genre Broadcast Media Recognition," *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2015Scottsdale, Arizona, USA, Dec. 2015, IEEE.*, vol. 1, no. 1, 2015.

[15] 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, L. McCowan, W. Post, D. Reidsma, and P. Wellner, "The AMI Meeting Corpus: A pre-announcement," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 3869 LNCS, pp. 28–39, 2006.

[16] N. Morgan, D. Baron, J. Edwards, D. Ellis, D. Gelbart, A. Janin, T. Pfau, E. Shriberg, and A. Stolcke, "The meeting project at ICSI," *Proceedings of the first international conference on Human language technology research - HLT '01*, pp. 1–7, 2001. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1072133.1072203

[17] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2017-Augus, pp. 2616–2620, 2017.

[18] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxceleB2: Deep speaker recognition," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2018-Septe, no. ii, pp. 1086–1090, 2018.

[19] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of I-vector Length Normalization in Speaker Recognition Systems," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2011, pp. 249–252.