Interspeech 2018
2-6 September 2018, Hyderabad

Estimation of the Number of Speakers with Variational Bayesian PLDA in the DIHARD Diarization 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 focuses on the estimation of the number of speakers for diarization in the context of the DIHARD Challenge at InterSpeech 2018. This evaluation seeks the improvement of the diarization task in challenging corpora (Youtube videos, meetings, court audios, etc), containing an undetermined number of speakers with different relevance in terms of speech contributions. Our proposal for the challenge is a system based on the i-vector PLDA paradigm: Given some initial segmentation of the input audio we extract i-vector representations for each acoustic fragment. These i-vectors are clustered with a Fully Bayesian PLDA. This model, a generative model with latent variables as speaker labels, produces the diarization labels by means of Variational Bayes iterations. The number of speakers is decided by comparing multiple hypotheses according to different information criteria. These criteria are developed around the Evidence Lower Bound (ELBO) provided by our PLDA.

Index Terms: DIHARD Challenge, Diarization, i-vectors, PLDA, Variational Bayes, number of speakers

1. Introduction

Diarization is the task of properly labeling some input audio according to the active speaker. These labels are just required to distinguish between the different speakers rather than providing a unique label per speaker (i.e. speaker identity). The recent growth of audiovisual resources has increased the necessity of these kind of systems for indexation purposes. For this reason, diarization has moved from the telephone channel environment to a wider range of scenarios, such as broadcast, meetings, etc.

Many solutions have been proposed for the diarization task. A popular solution is the bottom-up strategy, which consists of dividing the input audio into segments with only one speaker active and its posterior clustering. More detailed information is available in reviews such as [1][2]. A conceptual analysis let us divide the clustering task into three main subtasks or blocks: the segment representation, a similarity metric scoring and a clustering policy. The first two blocks have been deeply analyzed in the speaker identification state-of-the-art (JFA [3], i-vectors [4], PLDA [5], neural networks [6]). Regarding the clustering policy, many ideas have been proposed, such as BIC [7] [8], Variational Bayes [9], PCA [10], Mean-Shift [11] or Variational Bayes PLDA clustering[12].

Generally speaking, the diarization task does not assume the number of speakers to be known. To deal with this uncertainty diarization systems can generate several hypotheses, choosing the one which best fits the data. The impact of this choice in the diarization results is large. If this estimation is not correct, a significant loss in performance can be obtained, even if perfect labels were hypothesized too. Unfortunately, as described in [13], the total amount of possible hypotheses is too large to analyze all of them. Therefore, some sort of selection is needed, limiting the diarization performance to be as good as the best chosen hypothesis. Both decisions, with a noticeable influence in the diarization performance, become considerably more difficult when several scenarios are taken into consideration. This is because their criteria can differ from one scenario to each other.

The DIHARD Diarization Challenge is one of the latest evaluations about the diarization task. This evaluation seeks the application of the diarization technology to scenarios with high complexity, in which this task does not behave properly yet. Some of these scenarios include Youtube videos, meetings, court recordings, medical interviews, etc. In order to increase the complexity of the problem, no prior information about the scenarios is provided. Therefore, the development set is not guaranteed to be representative of the evaluation set. The evaluation includes two different modalities for the same audio. Track 1 considers manually annotated Voice Activity Detection (VAD) labels, distributed by the organization. Track 2 assumes that no VAD labels were released with the data, so it is up to the participants to obtain them.

In this work we have constructed our system around the standard i-vector PLDA architecture from speaker verification, including a Variational Bayes clustering for diarization purposes. After some initial segmentation an i-vector representation is extracted from each acoustic fragment and clustered by means of Variational Bayes Fully Bayesian PLDA[14][15]. Unsupervised model adaptation [15] is considered to mitigate domain mismatch between training and evaluation scenarios.

The paper is organized as follows. ViVoLab diarization system is introduced in Section 2. Section 3 describes our Voice Activity Detection approach. In section 4 the Variational Bayes clustering technique is explained in detail. The obtained results are presented in Section 5. Finally, Section 6 contains the conclusions.

2. System Description

The proposed system is an evolution of our diarization research in broadcast data [14][15]. A graphical representation is available in Fig. 1.

This approach combines a bottom-up approach with the speaker identification i-vector PLDA framework, state-of-the-art in speaker identification tasks. Given an audio, a feature extraction front-end is applied and a VAD estimation is obtained. Both types of information are taken into consideration to per-

This work has been supported by the Spanish Ministry of Economy and Competitiveness and the European Social Fund through the 2015 FPI fellowship, the project TIN2014-54288-C4-2-R and by the European Union FP7 Marie Curie action, IAPP under grant agreement no. 610986.

10.21437/Interspeech.2018-1841

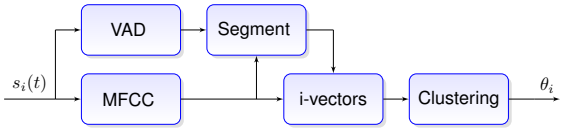


Figure 1: *Block diagram of the diarization system*

form the segmentation stage. This segmentation is performed by means of BIC [7], in which a sliding window analysis is carried out. A set of i-vectors is extracted from the obtained segments, assuming that a single speaker is present in each segment. The i-vector extraction includes centering, whitening and length normalization [16] to maximize its discriminative properties. The clustering step is performed by means of the Fully Bayesian PLDA and its Variational Bayes solution[12][14]. The clustering stage can be preceded by in-domain unsupervised adaptation [15] stage, applied to the PLDA model. This block is included to better fit the PLDA model to the evaluation audio conditions.

3. Voice Activity Detection

The proposed diarization system makes use of some VAD labels to perform the segmentation step. Whereas DIHARD track 1 includes perfect segmentation information provided by the organization, DIHARD track 2 does not, being up to each team the estimation of this information. In consequence, in track2 VAD becomes a new source of degradation, affecting the diarization performance.

The considered VAD solution is based on Recurrent Neural Networks (RNNs). This kind of neural network is specially designed to model sequential information, such as human speech. A popular RNN is the Long Short Term Memory (LSTM) [17], a RNN architecture which introduces the concept of memory cell. This cell is able to learn, retain and forget information along long sequences. This capability becomes very useful to carry out a long-term and short-term analysis simultaneously. LSTMs have been improved with the creation of Bidirectional LSTMs (BLSTMs or BiLSTMs). This network combines two LSTM networks processing the same sequence, but with opposite directions: one makes the forward analysis while the other performs the backward one. Therefore the network is capable of modeling causal and anti-causal dependencies for the same sequence.

The neural architecture proposed is shown in Fig. 2: the main component is a single layer Bidirectional Long Short Term Memory (BLTSM) with 128 neurons. Each output of the BLSTM layer is independently classified by a linear perceptron, which shares its values (weights and bias) for all time steps. Both the training and evaluation are performed with limited-length sequences (3 seconds, 300 frames), limiting the delay of dependencies to take into account.

The features for the neural network consist of a 32-component Mel filter bank and the log-energy. Features are computed using a window of 25 ms. length and an advance of 10 ms. Feature Mean and Variance Normalization is applied for each file in the database. The neural network has been trained with the DIHARD development dataset with Track 1 labels, assuming it is representative enough for the evaluation set.

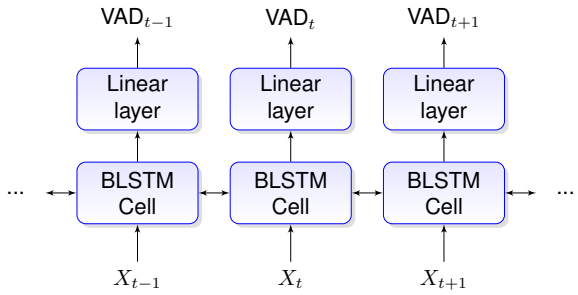


Figure 2: *BLSTM-based VAD architecture description. X_i represents the input features for frame i. VAD_i is the VAD label (speech, non-speech) for frame i*

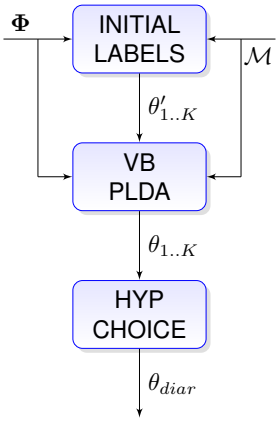


Figure 3: *Clustering stage: Inference of initial labels in terms of PLDA pairwise log-likelihood ratio. K possible hypotheses θ'_i are refined with our VB PLDA (θ_i). Finally hypothesis selection with BIC.*

4. Clustering

The clustering step is performed with the Fully Bayesian PLDA, originally described in [12]. This model, based on some initialization labels, is able to construct diarization hypotheses at segment level. These initialization hypotheses are obtained by means of pairwise log-likelihood ratio score as similarity metric. Different approaches have been tested to obtain the widest variety of hypotheses. For this purpose, we have tested Agglomerative Hierarchical Clustering and some processing with respect to the score matrix.

Unfortunately, these initialization methods do not include any selection criteria, that is, how to decide which hypothesis is more likely to properly represent the data. In the variational approach this problem is relevant due to the high dependency of its performance with respect to the initial labels. For this reason, our variational solution simultaneously analyzes multiple initializations, estimating multiple candidate labels and a quality metric, the Evidence Lower Bound (ELBO). This information feeds a model comparison stage, which chooses the best candidate as the final diarization labeling. The schematic of this system is represented in Fig. 3

4.1. Initialization with Image Processing

As described before, the initialization step must provide the widest range of initializations. The more different the initializations are, the more likely the global minimum error is reachable.

Despite the fact the original system [14][15] works with the Agglomerative Hierarchical Clustering (AHC), its hierarchical strategy becomes its main weakness. For an audio with N segments to cluster, only N possible initializations can be considered, and these hypotheses are very correlated. Besides, any mistaken decision committed during the hierarchical clustering is propagated along the posterior decisions. Our proposal is to identify some relational structure in a similarity matrix among the clustered elements, i. e., pairwise log-likelihood ratio score matrix. The relationships are obtained by assuming the score matrix to be an image, and applying basic image processing techniques. The image processing algorithm is inspired by [18], which also processes a score matrix to perform diarization. The algorithm, illustrated step by step in Fig. 4 is:

1. Construction of the score matrix. PLDA Pairwise log-likelihood ratio between all the segments is considered. For each row we substitute the element in the diagonal by the maximum of the remaining values in the row.
2. Gaussian blur of the image to mitigate erratic values.
3. Percentage value thresholding. Computed the histogram of the image, those values belonging to a threshold percentile are put to one and zero otherwise.
4. Conversion of the binary images to initialization speaker labels

This approach offers various advantages. The first one is independence among hypotheses, because hypotheses only depend on their specific percentile value. The other is the amount of combinations. While AHC with N elements is limited to N possible hypotheses, the image approach can provide as many hypotheses as desired, just by adjusting the percentile threshold.

4.2. Fully Bayesian PLDA

Originally described in [12], this model is a modification of PLDA [5] in which each i-vector $\boldsymbol{\phi}_j$ is produced by an unknown speaker i in a set of M possible candidates ($i = 1..M$), each one modeled by a hidden variable $\mathbf{y}_i$. This uncertainty about the speaker is modeled by substituting the fixed speaker label by a latent variable θ as follows:

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

The speaker latent variable θ follows a multinomial distribution with a Dirichlet prior π_θ.

$$P\left(\theta|\pi_\theta\right) = \prod_{i=1}^{M}\prod_{j=1}^{N_i} \pi_j^{\theta_{ij}} \quad (2)$$

$$P\left(\pi_\theta|\tau_0\right) = \mathbf{C}(\tau_0)\prod_{i=1}^{M} \pi_{\theta_i}^{\tau_0-1}; \mathbf{C}(\tau_0) = \frac{\Gamma(M\tau_o)}{\Gamma(\tau_0)^M} \quad (3)$$

Finally, in order to gain more robustness, the model parameters ($\boldsymbol{\mu}$, $\mathbf{V}$ and $\mathbf{W}$ and their prior $\boldsymbol{\alpha}$) are also considered to be latent variables rather than point estimates.

The high complexity of the proposed model makes its maximum likelihood solution unfeasible, so a Variational Bayes solution is proposed. The factor decomposition is:

$$P\left(\boldsymbol{\Phi}, \mathbf{Y}, \theta, \pi_\theta, \boldsymbol{\mu}, \mathbf{V}, \mathbf{W}, \boldsymbol{\alpha}\right) \approx$$
$$\approx q^*\left(\mathbf{Y}\right) q^*\left(\theta\right) q^*\left(\pi_\theta\right) q^*\left(\boldsymbol{\mu}\right) q^*\left(\mathbf{V}\right) q^*\left(\mathbf{W}\right) q^*\left(\boldsymbol{\alpha}\right) \quad (4)$$

The presented decomposition provides a useful tool $q^*\left(\theta\right)$, which describes how the i-vectors of each segment are distributed among M possible clusters or speakers.

4.3. In-domain Unsupervised Adaptation

DIHARD data is known to contain multiple scenarios or environments from which we have limited or unavailable data. Making use of in-domain unsupervised adaptation [15] we are able to adapt the PLDA model to the evaluation audio in absence of further in-domain information.

By simple techniques with no speaker knowledge we infer from scratch some pseudo-speaker labels, adapting the PLDA model afterwards. The PLDA model adaptation with the pseudo-speaker labels also takes advantage of the Fully Bayesian PLDA.

The employed technique is Mean-shift [19], using cosine distance as similarity metric [11].

4.4. Hypotheses comparison

The clustering by means of the Variational Bayes PLDA solution has already reported great results, significantly improving simpler clustering solutions such as Agglomerative Hierarchical Clustering. However, the fixing capabilities of the variational approach are limited by the initial state of the speaker labels,i. e., some mistakes can be too severe in the initialization to be fixed by the model.

To mitigate this effect, we can compare different hypotheses, generated with different initializations (i.e. different levels of the initial Agglomerative Hierarchical Clustering). The variational approach provides the Evidence Lower Bound (ELBO), a metric which represents how well the variational solution represents the data. However, direct comparison of ELBO metrics do not take into account the different complexity for each hypothesis, specially when a different number of speakers is assumed.

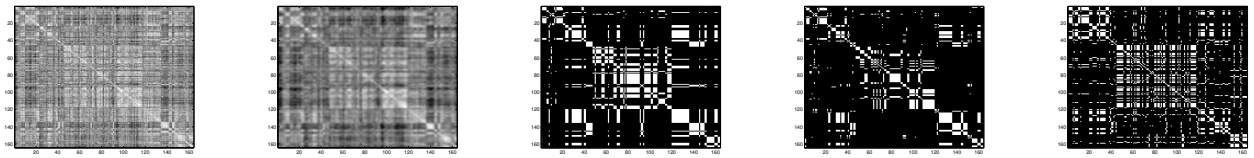
The proposed solution is a penalized score, inspired by the Bayesian Information Criterion (BIC) score [20]. This score, designed to compare how well some models fit some data, also includes a penalty term to compensate the different modeling capabilities or complexity. Consequently, the diarization labels Θ_{diar} are those which maximize the penalized ELBO like:

$$\Theta_{diar} = \arg\max_{\Theta_i}\left(ELBO(\Theta_i, \mathcal{M}_i) - \frac{1}{2}\lambda\log\Upsilon(i)\right) \quad (5)$$

where, for each hypothesis i, we have its speaker labels Θ_i, its model parameters and latent variable expectations $\mathcal{M}_i$. The penalty term $\frac{1}{2}\lambda\log\left(\Upsilon(i)\right)$ represents the extra modeling capabilities of the model for the hypothesis i, weighted by λ. In our work two different points of view have been applied: While in the first mode (Mode A), $\Upsilon(i)$ represents the number of total modeled speakers, in the second point of view (Mode B) $\Upsilon(i)$ represents the total number of free parameters the model uses.

5. Results

DIHARD challenge has moved forward in terms of diarization evaluation. Apart from the traditional Diarization Error Rate



a) b) c) d) e)

Figure 4: *Depiction of the Initialization by Image Processing. a) Initial scores with Diagonal Substitution. b) Blurred scores to soften noisy decisions. c,d) Two different hypotheses in terms of the threshold percentile. e) Reference pattern to recognize.*

(DER) metric, which analyzes the proportion of mistakenly labeled audio, another comparison metric has been proposed for the evaluation: The Mutual Information (MI).

Our submission consists of five systems, analyzing different lines of research. All of them consider a 20 MFCC frontend, without derivatives. The models (UBM, i-vector extractor and PLDA) have been trained with a combination of the Multi-Genre Broadcast 2015 [21] dataset, AMI Corpus [22], ICSI Meeting Corpus [23] and the Rich Transcription 2009 dataset. Their differences are:

- **Baseline**. This system applies unsupervised domain adaptation, AHC initialization and ELBO metric is penalized with mode A. Its main goal is the maximization of the new metric, the Mutual Information.
- **System 1**. This system includes unsupervised domain adaptation AHC initialization and considers Mode B as penalty term during hypothesis selection. This system is designed to compensate the errors of the baseline while maximizing Mutual Information.
- **System 2**. This system uses out-of-domain models without in-domain adaptation. Initialization with AHC. The hypothesis selection is done with speaker-penalized ELBO. This system is prepared to minimize DER metric.
- **System 3**. This system make use of non-adapted models. AHC Initialization is considered. Maximization of ELBO penalized in terms of the free parameters. This system is designed to fix error in System 2 while minimizing DER.
- **System 4**. This system make use of out-of-domain models. Initialization with Image Processing. The penalty term in the hypothesis selection depends on the number of speakers. This system is prepared to minimize DER.

Their results with DIHARD data with both development and evaluation set is included in Table 1.

The first detail we have come across is the high correlation between development and evaluation results, even though no common scenarios are guaranteed to be seen on each set.

The Baseline system has obtained the best results for Mutual Information with all configurations, suffering from a strong degradation in terms of DER. This degradation is fixed with System 1, significantly improving the baseline DER marks but obtaining the lowest MI results. Regarding System 2, 3 their performance is quite similar, improving the results of System 1 for both DER and MI, but they are surpassed by System 4. This last configuration leads the results in terms of DER (around 4% relative improvement) and obtains the second mark in terms of MI, with a 2% relative degradation compared with the best result. An especial mention for the BLSTM-based VAD, which has reported very good performance, just degrading an absolute 12% in terms of DER compared with the oracle VAD.

System	Development		Evaluation	
	DER(%)	MI	DER(%)	MI
Track 1				
Baseline	46.31	**8.48**	48.40	**8.52**
System 1	26.40	8.26	32.90	8.29
System 2	20.57	8.37	26.15	8.34
System 3	21.27	8.33	26.27	8.33
System 4	**20.42**	8.39	**26.02**	8.35
Track 2				
Baseline	47.13	**8.13**	51.78	**8.12**
System 1	37.16	7.96	44.91	7.89
System 2	31.16	8.04	39.20	7.97
System 3	31.51	8.00	39.15	7.96
System 4	**30.12**	8.07	**38.00**	7.99

Table 1: *Results in the development and evaluation sets*

6. Conclusions

All the systems have obtained satisfactory results, in terms of DER and MI, achieving their goals. Moreover, a better trade-off between DER and Mutual Information has been obtained when attempting to minimize DER. System 4, tuned to minimize DER, leads our results in terms of DER and has the second best Mutual Information score.

Regarding our contributions, the penalty term in terms of the number of free parameters has a larger impact compared with the number of speakers (Baseline vs system 1). Nevertheless, its larger value makes the number of speakers penalty much more suitable for finetuning. Besides, the new initialization with image processing has provided an extra improvement compared with AHC. Further work can obtain more robust structured relationships leading to a more oriented initializations.

7. References

[1] X. Anguera, 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. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6135543

[2] S. E. 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] P. Kenny, "Joint Factor Analysis of Speaker and Session Variability: Theory and Algorithms," *CRIM, Montreal,(Report) CRIM-06/08-13*, pp. 1–17, 2005.

[4] 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, 2011.

[5] 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.

[6] 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.

[7] S. Chen and P. Gopalakrishnan, "Speaker, Environment and Channel Change Detection and Clustering Via the Bayesian Information Criterion," *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, vol. 6, pp. 127–132, 1998.

[8] D. Reynolds and P. Torres-Carrasquillo, "Approaches and Applications of Audio Diarization," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. V, pp. 953–956, 2005.

[9] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, and C. Vair, "Stream-based speaker segmentation using speaker factors and eigenvoices," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2008*. Las Vegas, Nevada, USA: IEEE, 2008, pp. 4133–4136. [Online]. Available: http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=4518564

[10] C. Vaquero, A. Ortega, A. Miguel, and E. Lleida, "Quality Assessment of Speaker Diarization for Speaker Characterization," *IEEE Trans. on Acoustics, Speech and Language Processing*, vol. 21, no. 4, pp. 816–827, 2013.

[11] M. Senoussaoui, P. Kenny, P. Dumouchel, and T. Stafylakis, "Efficient Iterative Mean Shift Based Cosine Dissimilarity for Multi-Recording Speaker Clustering," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2013, pp. 7712–7715.

[12] 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.

[13] N. Brümmer and E. de Villiers, "The Speaker Partitioning Problem," *ODYSSEY 2012 The Speaker and Language Recognition Workshop*, no. July, pp. 194–201, 2010.

[14] 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.

[15] 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.

[16] 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.

[17] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1–32, 1997.

[18] Q. Wang, C. Downey, L. Wan, P. A. Mansfield, and I. L. Moreno, "Speaker Diarization with LSTM," 2017. [Online]. Available: http://arxiv.org/abs/1710.10468

[19] K. Fukunaga and L. Hostetler, "The Estimation of the Gradient of a Density Function, with Applications in Pattern Recognition," *IEEE Transactions on Information Theory*, vol. 21, no. 1, pp. 32–40, 1975.

[20] G. Schwarz, "Estimating the Dimension of a Model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978. [Online]. Available: http://projecteuclid.org/euclid.aos/1176344136

[21] 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," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2015Scottsdale, Arizona, USA, Dec. 2015, IEEE.*, vol. 1, no. 1, 2015.

[22] 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.

[23] 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