SYSTEM FUSION AND SPEAKER LINKING FOR LONGITUDINAL DIARIZATION OF TV SHOWS

Marc Ferràs[1]*, Srikanth Madikeri*[1]*, Petr Motlicek*[1] *and Hervé Bourlard*[2]

[1]Idiap Research Institute, Martigny, Switzerland
[2]Ecole Polytechnique Fédérale, Lausanne (EPFL),Switzerland

{mferras,msrikanth,pmotlicek,bourlard}@idiap.ch

ABSTRACT

Performing speaker diarization while uniquely identifying the speakers in a collection of audio recordings is a challenging task. Based on our previous work on speaker diarization and linking, we developed a system for diarizing longitudinal TV show data sets based on the fusion of speaker diarization system outputs and speaker linking. Agreement between multiple diarization outputs is found prior to speaker linking, largely reducing the diarization error rate at the expense of keeping some speech data unlabelled. To deal with noisy clusters, a linear prediction based technique was used to label speakers after linking. Considerable gains for both fusion and labelling are reported. Despite the challenges of the longitudinal diarization task, this system obtained similar performance for linked and non-linked tasks under moderate session variability, highlighting the viability of a linking approach to longitudinal diarization of speech in the presence of noise, music and special audio effects.

Index Terms— speaker diarization, linking, longitudinal , fusion, clustering, i-vector, ward

1. INTRODUCTION

Automatically structuring multimedia archives containing large amounts of data is a difficult task. These data sets typically involve speech with a large variety of speakers, acoustic environment conditions, languages and expressive states. While speaker diarization systems are currently quite mature, changes in speech expressiveness and environmental noise typically result in performance drops while the computational cost can become prohibitive for long recordings or collections of recordings.

The Multi-Genre Broadcast (MGB) Challenge [1] is an international evaluation campaign of speech technologies on TV recordings from the British Broadcasting Corporation (BBC). This data is especially challenging for speech technologies as it involves multi-genre data from the whole range of TV shows of the BBC. For speaker diarization, this data set faces systems to frequent overlapping audio, such as voice over music or applause, and a large variability in expressive speech, e.g. in soap operas. On the other side, speaker turn structure is highly variable from show to show, making it difficult for systems to be tuned to specific speaker interaction patterns. Task 4 in the MGB Challenge is a longitudinal diarization task, asking to diarize speakers across different recordings of the same show. This translates into finding start and end times for every speaker while labelling the speakers globally within the show.

This work was supported by Speaker Identification Integrated Project (SIIP), funded by the European Unions Seventh Framework Programme for research, technological development and demonstration under grant agreement no 607784.

Large scale speaker diarization has been addressed by previous work on telephone speech and broadcast news data. A speaker linking system using cross-likelihood ratio (CLR) and normalized CLR (NCLR) scores on Joint Factor Analysis (JFA) compensated models was proposed in [2, 3], for linking telephone speech and brodcast news data. Another multi-stage approach [4] targeting large scale diarization diarizes chunks of speech data whose clusters are linked in a later stage. This system scales particurlarly well on large data sets but still offers variable performance depending on the chunk size.

In this paper, we further develop Idiap's speaker linking system for the longitudinal task of the MGB evaluation. We addressed specific challenges posed by the MGB data by fusing speaker diarization outputs in order to increase speaker cluster purity and prevent error propagation to the linking system. A variable threshold pruning approach to speaker labelling was applied to the speaker linking dendrogram to cope with noisy dendrograms after linking.

The paper is structured as follows: Section 2 gives a system overview. Sections 3 and 4 describe Idiap's diarization strategy to process the MGB Challenge data. Speaker linking, clustering speaker clusters output in the diarization stage, is described in Sections 5 and 6. The experimental setup and results are given in Sections 8 and 9. Section 10 gives conclusion about this work.

2. SYSTEM OVERVIEW

The system submitted to the MGB Challenge was based on the speaker diarization and linking approaches developed in [5, 6] for far-field meeting speech data. This approach is well-suited to processing a large amount of speech data while modeling speakers within-recording and across-recording. *Diarization* is able to precisely find cluster boundaries within a recording at the expense of underclustering, while *linking* is able to link speaker clusters using knowledge from a large speaker population. If a training data set with speaker labels is available, speaker models can be compensated for session variability, e.g. using Joint Factor Analysis (JFA) [7, 8] or Probabilistic Linear Discriminant Analysis (PLDA). Unfortunately, although the MGB Challenge provides a large training data set, it does not provide global speaker labels that can directly be used for training such session variability compensation models.

In this work, we focused on two topics addressing a) the fusion of diarization system outputs and b) improving the pruning strategy for speaker labelling after linking. Figure 1 shows a block diagram of the overall system. The diarization stage uses fused outputs from two speaker diarization systems. Fusion has the objective of finding agreement between diarization outputs expecting the purity of the resulting speaker clusters to increase and hopefully reduce errors in the linking stage. Therefore, only a portion of the total speech time is passed on to the linking module. The linking stage hierarchically

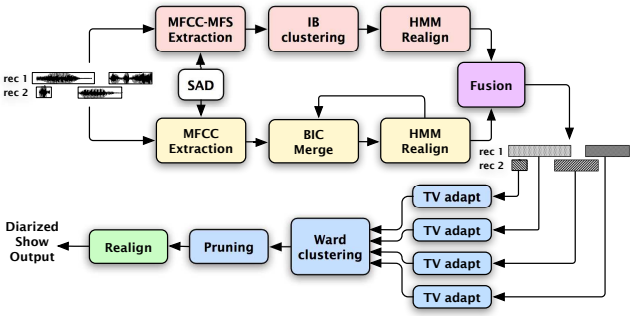

Fig. 1. Block diagram of the full speaker diarization and linking system. In red and yellow colors are the IB and HMM/GMM speaker diarization system modules respectively. In purple, the fusion of their outputs based on agreement. In blue, the speaker linking module. In green, the final realignment module.

clusters the agreed segments using an i-vector representation along with the i-vector covariance matrices obtained during i-vector extraction, as in [5]. Finally, the speech data for each linked speaker cluster is used to train HMM/GMM speaker models that are used to realign the non-labeled speech portions of the recordings.

3. SPEAKER DIARIZATION

The goal of the speaker diarization task is to split a recording into acoustically homogeneous regions that were spoken by the same speaker. After feature extraction and speech activity detection, these systems typically detect boundaries between speaker turns and then cluster these segments into speaker clusters across the recording, so-called speaker clustering. Two diarization systems were used:

- **IB diarization**: This is a fast agglomerative clustering algorithm based on the information bottleneck (IB) principle [9]. After uniformly segmenting the audio recording into short segments, the IB framework iteratively merges pairs of clusters using the Jensen-Shannon divergence, resulting in a minimum decrease of the objective function $\mathcal{F} = I(Y,C) - \frac{1}{\beta} I(C,X)$, where Y are a set of relevance variables, frame posteriors over the mixtures of a Gaussian Mixture model (GMM), C is the clustering solution and X are the initial segments. β is a trade-off between the amount of information preserved and the compression from the initial representation. The stopping criterion is given by setting a theshold on the Normalized Mutual Information criterion, $NMI = I(Y,C)/I(X,Y)$, measuring the fraction of original mutual information I(X,Y) captured by the current cluster representation C. Once the clustering has stopped, cluster boundaries are refined using Viterbi decoding on an ergodic HMM with a minimum duration constaint.
- **HMM/GMM diarization**: This is a traditional approach to speaker segmentation using GMM to characterize each speaker cluster. Starting from a set of uniform initial segments, the Bayesian Information Criterion (BIC) [10] is evaluated for all cluster pairs to determine the best merge candidate. After each merge, Viterbi decoding on an ergodic HMM with a minimum duration constraint is used to resegment the data to refine the speaker cluster boundaries. The process is iterated until the ΔBIC values for merging are under a given threshold.

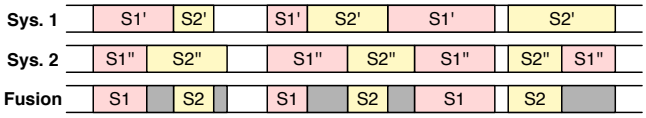

Fig. 2. Illustration of the fusion process. Only segments for which both systems agree upon speaker labels are output after fusion. A new set of labels is created for the fused output. Regions of non-agreement are marked as such and processed in a later stage.

4. FUSION OF SPEAKER DIARIZATION OUTPUTS

System combination of speaker diarization systems is a poorly studied topic. Some researchers [11] have cascaded two IB based diarization systems, the second algorithm refining the first system output. The same authors combined multiple feature streams in [12]. Frame-by-frame fusion of diarization outputs has been explored in [13, 14]. In [15], system combination is addressed by finding common segments in the outputs of two diarization systems while reclassifying the rest. We used a similar approach using two different front-ends together with the two speaker diarization systems described in Section 3.

From the outputs of two diarization systems, a new output is obtained by considering the segment boundaries of both diarization system outputs simultaneously. As illustrated in Figure 2, for each cluster of the first system, the cluster with the largest number of frames in common is found, and the pair of speakers becomes a new speaker in the output. This process is iterated over all speakers, outputting as many new speakers as speakers in the reference diarization system. Segments that were not assigned to any new speaker are given a "non-labelled" speaker name and processed separately.

This approach tends to purify the speaker clusters at the price of missing speaker labels for non-agreed regions. Bootstrapping the speaker linking system with purer speaker cluster is expected to prevent error propagation across diarization and linking modules.

5. SPEAKER CLUSTER MODELING

In our prior work, speaker clusters were modeled using JFA [7, 8], a parametric GMM adaptation technique allowing for the disentanglement of speaker and session effects. Although this approach is effective for observed linear distortion effects due to channel variability, its effectiveness on additive noise may be questioned. Furthermore, training data provided in the MGB challenge does not provide global speaker labels that can be used to train such models.

A single-factor eigenmodeling approach, such as a total variability (TV) or i-vector approach [7, 16], was used instead. The total variability model

$$\hat{\mathbf{m}} = \mathbf{m} + \mathbf{T}\mathbf{w} \tag{1}$$

is a parametric map from Gaussian supervectors, $\hat{\mathbf{m}}$ down to low-dimensional vectors, $\mathbf{w}$, that characterize the acoustics of the segment. The speaker-independent supervector $\mathbf{m}$ is formed by the mean vectors of a Universal Background Model (UBM) trained using data from many speakers. $\mathbf{T}\mathbf{w}$ is a low-rank term that models the acoustic variation. After $\mathbf{T}$ has been estimated in the maximum-likelihood sense using a large speech database, only latent variables are fit to the test, i.e. finding parameter estimates for the posterior distribution of $\mathbf{w}$, the latter being the objects used in the speaker linking phase.

6. SPEAKER LINKING

The goal of the speaker linking module is to assign unique identifiers to the clusters output by the speaker diarization output for all recordings in a TV show, i.e. considering longitudinal speaker linking within a show. Two major steps, agglomerative clustering and labelling, are discussed in the following:

6.1. Agglomerative clustering

The speech data of each cluster is modeled as a single multivariate Gaussian with a full covariance matrix, which is indeed a total factor posterior distribution. Initially, each initial cluster is assigned one speaker cluster output by the diarization system. The two closest clusters are then successively merged, until only one cluster remains:

1. **Compute the distance matrix** for all pairs of speaker clusters, that become the initial clusters.
2. **Merge** the two closest clusters.
3. **Update the distance matrix**, from the merged cluster to all other clusters.
4. **Go to 2.** If only one cluster remains, **stop**.

We use Ward's method [17], merging the two clusters that result in the minimum increase of the total within-cluster variance after merging, i.e. it aims at obtaining compact clusters. Ward's method is implemented in a recursive manner using the Lance-Williams algorithm [18, 5]. When two clusters c_i and c_j are to be merged, the distances between the merged cluster c_{ij} and all other clusters c_k are updated using the formula $d_{(ij)k} = \alpha_i d_{ik} + \alpha_j d_{jk} + \beta d_{ij}$. The values for α_i, α_j and β can be found in [18]. In [5], we found that the two-way Hotteling t-square statistic, the multivariate equivalent of the two-way Student-t statistic, outperformed other distance measures such as cosine distance or Kullback-Leibler divergence. We use the squared Euclidean distance term, $(\mathbf{w}_i - \mathbf{w}_j)^T \mathbf{C}_{ij}^{-1} (\mathbf{w}_i - \mathbf{w}_j)$, for the initial inter-cluster distances, which spherifies the distance between cluster means with the pooled covariance matrix of both clusters.

6.2. Speaker labelling

A binary tree can be obtained from all the merging steps performed during clustering. The tree structure indicates the order in which merges occurred and each merge node in the tree stores the merge cost. While we found in [5, 6] that pruning this tree using a single threshold for all series was enough to give satisfactory results on meeting data, it was difficult to find an optimal threshold that is valid for all TV series in the MGB development set.

In a first approach, a TV series dependent threshold was set, namely a fraction of the largest initial inter-cluster distance between all speaker clusters from all recordings of that series, making the threshold larger as more spread the initial data are.

A second approach combined the TV series dependent threshold with an approach aiming at predicting merging costs at a node in the tree from merging costs in surrounding nodes, labelling a node as a speaker if a jump in predictability is found. For each node, a linear prediction model [19] is fit to each of four possible subtree trajectories, predicting the parent cost from current, child, and child of child node costs. If the minimum of the prediction errors, normalized by the central node distance, is larger than a threshold, the node is labelled as a new speaker. This threshold and the TV series dependent threshold are combined with an AND operation. The sibling nodes of already labelled nodes are labeled as new speakers if they were not already labeled as speakers. This ensures that all initial samples are given a speaker label.

7. REALIGNMENT

After speaker linking, a new set of speaker labels is obtained for those segments for which agreement was found during system fusion. The rest of the segments have no speaker labels assigned. The purpose of realignment is to give "agreed" speaker identifiers to "non-agreed" speech segments. For this purpose, an ergodic HMM/GMM is trained on the "agreed" speaker speech of each recording and decoded using the Viterbi algorithm (with a minimum duration constraint) on the "non-agreed" speech data.

8. EXPERIMENTAL SETUP

The presented speaker diarization and linking system was partially tuned using the development MGB Challenge data set before evaluation. The linked Diarization Error Rate (DER) performance of this system was compared to systems using no speaker diarization fusion.

Only audio labelled as speech in the segmentation provided during the MGB Challenge evaluation was used, with ground-truth segmentation being available for development and an automatic segmentation being available for evaluation.

The IB diarization system uses a front-end with 19 Mel-Frequency Cepstral Coefficients (MFCC) and 19 Mel-Filterbank Slope features (MFS) [20]. This setup used 2.5s uniform initial segmentation, $\beta = 7$, a stopping threshold of 0.3 and a maximum number of speaker clusters of 10 were used. The minimum duration for Viterbi decoding was set to 2.5s. The HMM/GMM diarization system used 19 MFCC features. The initial number of speaker clusters and Gaussian components was set to 10 and 5, respectively, the BIC threshold was set to 0.7 and the minimum duration constraint for Viterbi decoding was 2.5s. Only the number of speakers and the minimum duration for Viterbi decoding were tuned to the MGB development data, the rest being optimized for meeting data.

For speaker cluster modeling, we used the speech data from the "train.full" condition of the challenge, over 2000 hours, to train a gender-independent GMM-UBM with 512 Gaussian components as well as the the total variability matrix $\mathbf{T}$ using 5 and 10 EM iterations of maximum likelihood estimation respectively. The optimal i-vector dimensionality was found to be 100, after optimization on the development set. The speaker linking module thresholds were optimized on the development data set as well, obtaining an optimal linear prediction threshold of 4, and a maximum absolute threshold of 0.2 times the maximum intercluster initial distance.

9. RESULTS

We ran longitudinal diarization experiments on the MGB Challenge data, using the non-linked and linked DER performance measures, i.e. the amount of time that system output does not agree with the non-linked and linked references, as performance measures.

Table 1 shows DER for the speaker diarization and linking system. In the first block, DER of 43.6% are shown for both IB and HMM/GMM diarization. Fusing these two systems results in a relative drop of 33.4% DER, from 43.6% to 29.0%, at the price of introducing 20% of non-labeled speech. This illustrates a trade-off in the diarization fusion approach: the more reliable the fused speaker clusters the more speech is not labelled, and viceversa.

Regarding linking, a performance gap can be observed from non-linked to linked DER due to the increased difficulty of the linking task. The increase is 6% to 16% DER absolute for IB and HMM/GMM diarization, from 43.6% to 50.4% and 43.6% to 60.6%, respectively. For fused diarization, the increase stays around 5%, from 43.6% to 48.9%, after labelling the non-labeled regions using Viterbi decoding. The latter is able to label 5.4% out of the non-labeled 20% with correct speaker labels, assuming the same linked speakers are present in the non-labelled speech obtained after fusion, which is far from being true in practice for the MGB data. Indeed, we noticed that the number of speakers correctly labeled was far below the actual number of speakers. We believe this is due to optimizing DER in very noisy data, with tuning resulting in correctly labelling major speakers while dismissing the rest.

All of these systems used the same pruning thresholds during speaker labelling. Table 2 shows fixed, show-adaptive and linear prediction based threshold optimizations for the fusion system. Using an adaptive threshold, relative to the maximum intercluster distance, results in 13% relative improvement over using a fixed threshold (41.3% vs. 35.9%). Using the linear prediction based thresholding together with an adaptive threshold brings a 4% additional relative gain for an absolute linked DER of 34.3%. Non-linked DER is not optimal at this operating point, but it is slightly better for a fixed threshold. We believe this might be a byproduct of having corrupted speech in the speaker clusters being linked.

We consider speaker diarization and linking results as satisfactory given the challenging data and the use of an i-vector front-end with no session compensation. These results are in line with our previous work on speaker linking for far-field meeting data [5, 6], where the increase of difficulty of the longitudinal diarization task barely affected the DER. However, for the MGB challenge, the starting diarization error rates are roughly twice those reported for far-field meeting data and no JFA session compensation, or PLDA, was used.

These systems were run on the evaluation data, with the difference that the provided speech/non-speech segmentation is output by an automatic speech/non-speech detection system. Table 3 shows the linking system DER as well as missed speech and speaker and false alarm times. In this case, missed speech gathers non-labelled speaker time and non-speech. The non-linked DER for the Fusion system remains in the same range as for the development data set, 30.8%, with a large proportion of the 25.5% missed speech time to be realigned. However, bootstrapping the linking system with these diarization outputs results in significantly larger linked DER when compared to the development data set. Absolute increases in linked DER are indeed almost 12% (from 30.8% to 42.6%) range compared to 5% in the development set, suggesting a deviation in behavior of the linking system, but not the diarization system. The development set has an average of 3.8 recordings per show whereas the evaluation data set has 9.5, more than twice as many recordings per show. The reported results suggest that the linking system is vulnerable to the large session variability in the evaluation data set while the DER increase was acceptable for the development data set. Such decrease in linking performance results in more incorrectly labelled data overall compared to the development data set. For the evaluation set, out of the 72.1%=100%-25.5%-2.4% of labelled speaker time, 42.6%, i.e. more than half of it, are incorrectly labeled. These figures are lower for the development set.

System	Not-Linked DER (%)	Linked DER (%)	Missed Spkr. (%)
Diarization			
IB	43.6	-	0.0
HMM	43.6	-	0.0
Fusion	29.0	-	20.0
Linking			
IB	50.4	56.3	0.0
HMM	60.6	69.1	0.0
Fusion	30.6	34.3	20.0
Fusion + Realign	44.6	**48.9**	0.00

Table 1. Diarization Error Rates (DER) for the speaker diarization and linking system. DER on not-linked and linked references are reported as well as the percentage of missed speaker time.

System	Not-Linked DER (%)	Linked DER (%)	Missed Spkr. (%)
Linking			
Fusion th=5e4	**29.2**	41.3	20.0
Fusion th=0.2max	29.7	35.9	20.0
Fusion th=0.2max lp=4	30.6	**34.3**	20.0

Table 2. Diarization Error Rates (DER) for the speaker diarization and linking system using the labelling strategies described in Section 6.2. DER on not-linked and linked references are reported as well as the percentage of missed speaker and speech time.

System	Not-Linked DER (%)	Linked DER(%)	Missed Spch+Spkr (%)	False Spch. (%)
Linking				
Fusion	30.8	42.6	25.5	2.4
Fusion + Realign	45.1	**58.0**	6.0	4.0

Table 3. Diarization Error Rates (DER) for the speaker diarization and linking system on the evaluation data. DER on not-linked and linked references are reported as well as the percentage of missed speaker and speech time and false speech time.

10. CONCLUSION

This paper presented and analyzed the speaker linking and diarization system used for the longitudinal diarization task of the MGB Challenge 2015. An agreement-based approach to diarization fusion reduced the DER by 33% at the price of missing labels for 20% of the data. The agreed speaker clusters were linked using Ward clustering and a linear prediction based strategy to pruning the clustering dendrogram. Together with a show-adaptive threshold this approach resulted in 13% of relative improvement in linked DER with respect to a fixed threshold strategy, and 4% with respect to a show-adaptive threshold. Finally, the non-labelled speech was realigned using speaker models trained on the agreed speaker clusters.

This work shows that, while the task of longitudinal diarization is more complex than diarizing one recording at a time, the DER for both tasks can be kept in the same range. This has been possible on two very noisy data sets and bootstrapping the linking system using incorrect speaker labels in around half of the initial diarization output.

11. REFERENCES

[1] 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 Transcription," in *ASRU 2015*, 2015.

[2] H. Ghaemmaghami, D. Dean, R. Vogt, and S. Sridharan, "Extending the Task of Diarization to Speaker Attribution," in *Proc. INTERSPEECH*, 2011, pp. 1049–1052.

[3] H. Ghaemmaghami, D. Dean, and S. Sridharan, "Speaker Attribution of Australian Broadcast News Data," in *Proceedings of the First Workshop on Speech, Language and Audio in Multimedia (SLAM), Marseille, August 22-23*, 2013.

[4] M. Huijbregts and D. van Leeuwen, "Large Scale Speaker Diarization for Long Recordings and Small Collections," *IEEE Trans. on Audio, Speech and Language Processing*, pp. 404–413, 2012.

[5] M. Ferras and Hervé Bourlard, "Speaker Diarization and Linking of Large Corpora," in *Proc. of the IEEE Workshop on Spoken Language Technology*, 2012.

[6] O.Schreer M. Ferras, S. Masneri and Hervé Bourlard, "Diarizing Large Corpora using Multi-modal Speaker Linking," in *Proc. INTERSPEECH*, 2014.

[7] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 3, pp. 345–354, 2005.

[8] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of inter-speaker variability in speaker verification," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2008.

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

[10] J. Ajmera and C. Wooters, "A robust speaker clustering algorithm," in *Proc. IEEE ASRU*, 2003.

[11] F. Valente D. Vijayasenan and H. Bourlard, "Combination of agglomerative and sequential clustering for speaker diarization," in *Proc. IEEE ICASSP*, 2008, p. 43614364.

[12] F. Valente D. Vijayasenan and P. Motlicek, "Multistream speaker diarization through Information Bottleneck system outputs combination," .

[13] Sylvain Meignier, Daniel Moraru, Corinne Fredouille, Jean-François Bonastre, and Laurent Besacier, "Step-by-step and integrated approaches in broadcast news speaker diarization," *Computer Speech & Language*, vol. 20, no. 2, pp. 303–330, 2006.

[14] Simon Bozonnet, Nicholas Evans, Xavier Anguera, Oriol Vinyals, Gerald Friedland, and Corinne Fredouille, "System output combination for improved speaker diarization," in *Interspeech 2010, September 26-30, Makuhari, Japan*, 2010, pp. Interspeech–2010.

[15] V. Gupta, P. Kenny, P. Ouellet, G. Boulianne, and P. Dumouchel, "Combining gaussianized/non-gaussianized features to improve speaker diarization of telephone conversation," *IEEE Signal Processing Letters*, pp. 1040–1043, 2007.

[16] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 788–798, 2009.

[17] J. H. Ward, "Hierarchical Grouping to Optimize an Objective Function," *American Statistical Association*, vol. 58, no. 301, pp. 236–244, 1963.

[18] G. N. Lance and W. T. Williams, "A General Theory of Classificatory Sorting Strategies. 1. Hierarchical Systems," *Computer Journal*, vol. 9, pp. 373–380, 1967.

[19] J. Makhoul, "Linear prediction: A tutorial review," in *Proceedings of the IEEE*, 1975.

[20] S. Madikeri and H. Murthy, "Mel Filter Bank energy-based Slope feature and its application to speaker recognition," in *National Conference on Communications (NCC), Bangalore*, 2011, pp. 1–4.