INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Domain Adaptation of PLDA models in Broadcast Diarization by means of Unsupervised Speaker Clustering

Ignacio Viñals[1], *Alfonso Ortega*[1], *Jesús Villalba*[2], *Antonio Miguel*[1] *and Eduardo Lleida*[1]

[1]ViVoLAB, Aragón Institute for Engineering Research (I3A), University of Zaragoza, Spain
[2]Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, USA

{ivinalsb,ortega,amiguel,lleida}@unizar.es
jvillal7@jhu.edu

Abstract

This work presents a new strategy to perform diarization dealing with high variability data, such as multimedia information in broadcast. This variability is highly noticeable among domains (inter-domain variability among chapters, shows, genres, etc.). Therefore, each domain requires its own specific model to obtain the optimal results. We propose to adapt the PLDA models of our diarization system with in-domain unlabeled data. To do it, we estimate pseudo-speaker labels by unsupervised speaker clustering. This new method has been included in a PLDA-based diarization system and evaluated on the Multi-Genre Broadcast 2015 Challenge data. Given an audio, the system computes short-time i-vectors and clusters them using a variational Bayesian PLDA model with hidden labels. The proposed method improves 25.41% relative w.r.t. the system without PLDA adaptation.

Index Terms: diarization, clustering, adaptation

1. Introduction

The goal of diarization is to identify who speaks when in an audio file. Historically, diarization has been based on dividing the input data into homogeneous acoustic segments and clustering them afterwards. Some successful diarization systems [1] use Bayesian Information Criterion [2] for segmentation and Agglomerative Hierarchical Clustering (AHC) for clustering. Other systems [3] make use of JFA [4] to create streams of speaker-factors clustered later on according to different metrics. Variational Bayes is considered as well [5], proceeding to the joint estimation of the speaker-factors and the speaker labels, both latent variables, to maximize the log-likelihood lower bound. Some of the newest approaches diarize according to the i-vector paradigm [6], and make decisions in terms of PLDA [7], such as the system described in [8].

Therefore, diarization is closely linked to speaker recognition, inheriting its state-of-the-art techniques while assuming an important drawback: training data must be adapted to the operating conditions. In diarization with broadcast data, e.g., the 2015 Multi Genre Broadcast diarization evaluation [9], we find large performance differences across domains. This variability is interpretable as a mismatch among scenarios characteristics. In consequence, we could benefit from adapting our models to the target domain. Unfortunately, in-domain labeled data can be limited so adaptation strategies should work with unlabeled data.

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.

In our experiments we adapted our diarization system [8], based on i-vectors and PLDA, with in-domain unlabeled data. Four stages were candidates for the adaptation: the GMM-UBM, the i-vector extractor, the i-vector normalization and the PLDA model. We focused our efforts on the last two of them, the i-vector normalization and the PLDA model, leaving the first two steps unmodified. In order to obtain the speaker labels, necessary for the PLDA model adaptation, we inferred them by unsupervised clustering techniques such as Variational Bayes GMMs [10], Agglomerative Hierarchical Clustering (AHC) on different metrics, or Mean-Shift [11] with cosine distance kernel [12].

In our analysis we employed the data provided for the 2015 Multi-Genre Broadcast (MGB2015) diarization challenge [9]. The challenge, evaluated on BBC TV recordings from different shows and genres, analyzed speaker diarization in a longitudinal setting, across multiple recordings from the same show.

The paper is organized as follows. Section 2 presents the reference diarization system. Our proposal about unsupervised speaker clustering and PLDA adaptation is available in section 3. The experimental procedure is included in section 4. Finally, section 5 summarizes the conclusions.

2. Baseline Diarization System

Our reference diarization system [8] is based on the i-vector paradigm [6]. Its schematic is shown in Fig. 1. Given an episode to diarize, we perform a MFCC feature extraction and an initial segmentation based on Bayesian Information Criterion (BIC) [2], isolating homogeneous acoustic segments. I-vectors are extracted for each segment. These resultant i-vectors are clustered to obtain their speaker labels. Consecutive segments with a common speaker label are merged afterwards. We repeat twice the process of i-vector extraction, clustering and segment merging. With this procedure we compute i-vectors on longer segments, which contain more discriminative information.

2.1. I-vector extraction

The i-vector concept [6] assumes each utterance s can be represented by a GMM whose supervector $\mathbf{M}_s$, the result of concatenating the means for each component, is mapped on the total variability subspace by an affine transformation like:

$$\mathbf{M}_s = \mathbf{m} + \mathbf{T}\boldsymbol{\phi}_s \tag{1}$$

where $\mathbf{m}$ is the speaker- and channel- independent supervector. $\mathbf{T}$ is a low rank matrix describing the total variability space and $\boldsymbol{\phi}_s$ a latent variable with a standard normal prior, reflecting the projection of the utterance on the subspace. The posterior distribution of $\boldsymbol{\phi}_s$ given the utterance s, $P(\boldsymbol{\phi}_s|s)$, is Gaussian dis-

http://dx.doi.org/10.21437/Interspeech.2017-84

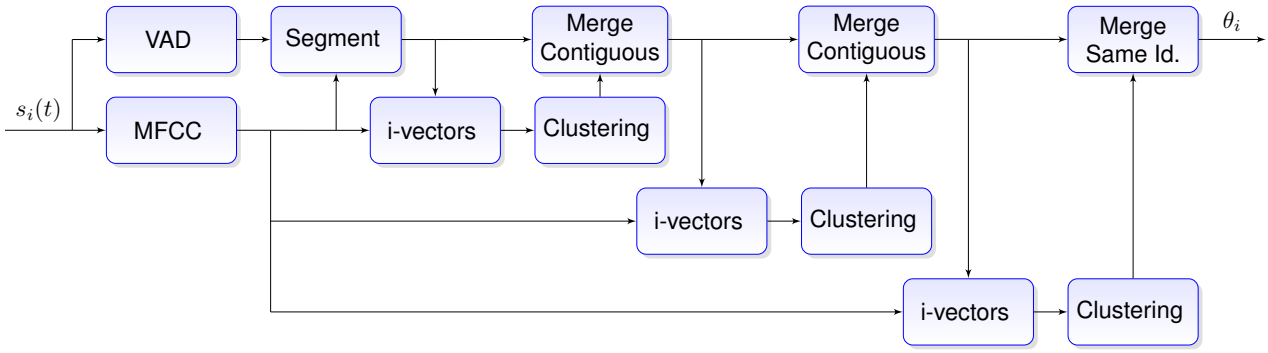

Figure 1: *Block Diagram for the proposed Baseline Diarization System*

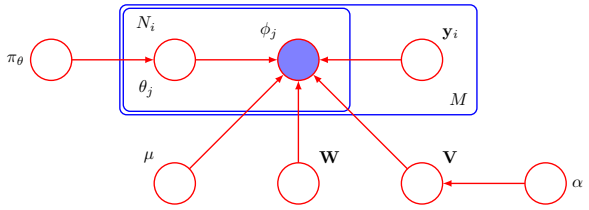

Figure 2: *Bayesian Network of Fully Bayesian Simplified PLDA with speaker label priors*

tributed. The maximum of the posterior is referred as i-vector in the bibliography.

The obtained i-vectors are length normalized following [13]. Before length normalization, centering and whitening are performed on i-vectors to evenly distribute them in the unit hypersphere.

2.2. Fully Bayesian PLDA with speaker label priors

Our clustering procedure has been constructed around a Fully Bayesian PLDA model [14], which assumes each utterance *j* of the speaker *i* can be written as

$$\boldsymbol{\phi}_{ij} = \boldsymbol{\mu} + \mathbf{V}\mathbf{y}_i + \boldsymbol{\epsilon}_{ij} \tag{2}$$

where $\mathbf{V}$ is a low rank matrix, describing the eigenvoices subspace, $\mathbf{y}_i$ the speaker-factor latent variable, ϵ_{ij} the within class variability term and $\boldsymbol{\mu}$ the speaker independent term. $\mathbf{y}_i$ and $\boldsymbol{\epsilon}_{ij}$ have been defined with normal distributions, the former a standard normal and the latter a normal distribution with zero mean and $\mathbf{W}$ precision.

This fully Bayesian PLDA approach considers the model parameters ($\boldsymbol{\mu}$, $\mathbf{V}$ and $\mathbf{W}$) to be hidden variables. Moreover, our model also defines the speaker labels $\boldsymbol{\theta}$ as hidden variables with a prior $\boldsymbol{\pi}_{\boldsymbol{\theta}}$, dividing N i-vectors into M speakers. All the prior distributions, for both the model parameters ($P(\boldsymbol{\mu})$, $P(\mathbf{V}|\boldsymbol{\alpha})$, $P(\boldsymbol{\alpha})$ and $P(\mathbf{W})$) and the speaker labels ($P(\boldsymbol{\theta}|\boldsymbol{\pi}_{\boldsymbol{\theta}})$ and $P(\boldsymbol{\pi}_{\boldsymbol{\theta}})$) are described in detail in [14]. The Bayesian network is shown in Fig. 2

The high complexity of the model makes impossible its closed form solution, so a Variational Bayes approximation has been carried out. Thus we have approximated the joint posterior to a product of factor distributions like:

$$P(\mathbf{Y}, \boldsymbol{\theta}, \boldsymbol{\pi}_{\boldsymbol{\theta}}, \boldsymbol{\mu}, \mathbf{V}, \mathbf{W}, \boldsymbol{\alpha}|\boldsymbol{\Phi}) \approx$$
$$q(\mathbf{Y})\, q(\boldsymbol{\theta})\, q(\boldsymbol{\pi}_\theta) \prod_{r=1}^{d} q\left(\tilde{\mathbf{v}}'_r\right) q(\mathbf{W})\, q(\boldsymbol{\alpha}) \tag{3}$$

During training, the variational factors for the parameters and its priors ($q(\mathbf{V})$, $q(\mathbf{W})$ and $q(\boldsymbol{\alpha})$) as well as the speaker-factors $q(\mathbf{Y})$ are iteratively updated to maximize the lower bound.

At evaluation, a clustering process is performed. Given an initial seed for the speaker labels, the maximization of the lower bound is carried out in terms of the joint estimation of the speaker label factors $q(\boldsymbol{\theta})$, $q(\boldsymbol{\pi}_{\boldsymbol{\theta}})$ and the speaker-factor $q(\mathbf{Y})$. In this task, the model parameters and its priors remain unmodified. In the model description [14] the initial labels are estimated by an agglomerative hierarchical clustering (AHC), employing the pairwise PLDA log-likelihood ratio of i-vectors as metric. After the optimization of the latent variables, the final speaker labels $\boldsymbol{\theta}_{\text{diar}}$ are the expected values for $\boldsymbol{\theta}$ given $q(\boldsymbol{\theta})$.

3. PLDA Domain Adaptation with Unsupervised Speaker Clustering

Our assumption of mismatch between scenarios in broadcast data makes domain adaptation a suitable solution. This idea fits with the complex statistical models in our system, which strongly depend on their training data. Considering the implications (complexity, duration, necessity of further adaptations,...) for adapting our statistical models to each domain, we opted for the adaptation of the PLDA model, leaving the GMM and the i-vector extractor unmodified. This decision was motivated by its relatively fast and low-memory-consuming adaptation in comparison to the other two options. Thus speaker labels are required for the PLDA adaptation, but in most cases labeled in-domain data is not available. Therefore some pseudo speaker labels must be inferred.

In our proposal we performed this inference with Unsupervised Speaker Clustering strategies, exploiting the similarity between i-vectors from the same speaker. We analyzed the following strategies:

- **Agglomerative Hierarchical Clustering (AHC)**. This hierarchical clustering sequentially estimates the pairwise proximity of clusters according to a metric, and combines the closest. In our experiments we tested two metrics: cosine distance on the i-vectors, as well as a

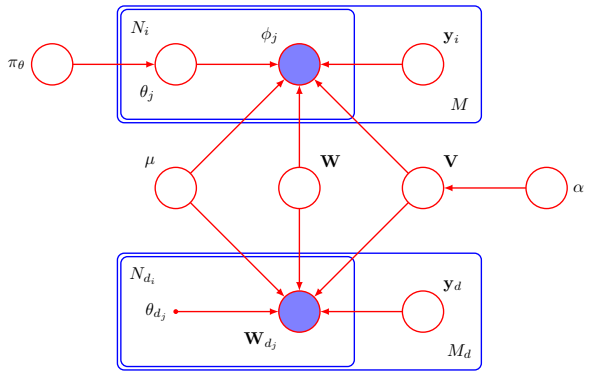

Figure 3: *Bayesian Network of Fully Bayesian Simplified PLDA with speaker label priors for two different domains. This situation is considered during our model adaptation.*

pairwise PLDA log-likelihood ratio based metric [8][14], estimated with a general domain model.

- **Variational Bayesian GMM (VBGMM)**. This clustering method models the i-vectors with a GMM distribution, associating each component to a speaker. By using a Variational Bayesian approach [10], our distribution is able to infer the number of speakers.

- **Mean-Shift**. In this technique [11], each i-vector is attracted to the densest position in a local vicinity, finishing when all i-vectors converge to their local highest density location. The number of clusters is equal to the amount of local maximums. Making use of [12] we opted for cosine distance as kernel.

While some of the previous techniques are able to infer the number of speakers (VBGMM and Mean-Shift), the agglomerative hierarchical clustering does not, requiring a tunable hyperparameter adjusted during development. The tuning range was limited, intending to properly cluster the most relevant speakers for each domain.

The obtained label estimations were used in the PLDA adaptation making use of the method described in [14]. This work proposes the adaptation of PLDA models, specifically those described in the section 2.2, by their optimization for two weighted domains. This methodology can be represented by a Bayesian network, illustrated in Fig. 3.

Whereas in the first scenario the PLDA model must approximate the general domain data with reliable labels, in the second scenario the PLDA models the in-domain data in terms of speaker label latent variables. The adaptation consists of the optimization of the joint likelihood lower bound for both domains by the reestimation of the model parameters and the speaker labels hidden variables.

In our strategy the output of the Unsupervised Clustering assumes the role of initialization for the speaker labels hidden variables. This role is critical because the given labels are refined iteratively during the adaptation. However, depending on the initial seed quality, this tuning of the speaker labels can converge to a local maximum rather than the global one. Therefore, the more accurate our initial clustering is, the more effective our adaptations are.

4. Experiments

We chose to carry out our experiments with the 2015 Multi Genre Broadcast (MGB2015) dataset [9]. This set consists of, approximately, 1600 hours of broadcasted audio from British Broadcasting Corporation (BBC), divided in three subsets, train, development and evaluation. The available metadata for the training set is formed by the emitted subtitles, refined with lightly supervised alignment systems. Two quality scores (Phone Matched Error Rate and Word Matched Error Rate) are also released for this set to facilitate data selection. Regarding the development set, hand-transcribed metadata was provided, as well as baseline labels for speech/non-speech and speaker segmentation. Similar baseline labels are available for the evaluation set too.

The rules for the MGB2015 challenge limit the use of data to those provided by the organization, so no other source of information is allowed. All our experiments followed the mentioned rules for comparison reasons.

4.1. Diarization system setup

The previously described diarization system was configured in the following way: from the audio data we have obtained 20 ETSI [15] standard MFCCs, and performed short time cepstral mean and variance normalization. With the obtained features we carried out the offline training as follows. A 256 Gaussian UBM-GMM and a 100-dimension i-vector extractor were trained on the training set. For the PLDA, the human-transcribed set (development) was employed, training a 50-dimension PLDA model. This PLDA model has worked as general domain model.

Regarding the diarization process, two configurations were analyzed. The first one considered a ground truth segmentation in addition to a ground truth VAD, to study the benefits of domain adaptation in optimal conditions. The other analysis opted for a segmentation based on BIC, modeling with a Gaussian distribution, and the baseline VAD provided by the organization, making our results comparable with those presented in MGB2015 [9].

4.2. Adaptation

Broadcast data is known to present high variability, specially considering different shows and genres. Due to this dissimilarity a loss in performance is expected to appear.

We propose domain adaptation as a solution for this mismatch, applied to the i-vector normalization and the PLDA models. We have opted for adapting our global-domain Fully Bayesian PLDA with the episode to diarize, using pseudo speaker labels inferred with unsupervised clustering methods. The finetuned PLDA model estimates the final diarization labels afterwards. By limiting our experiments to episode adaptation, adapting and carrying out diarization with only information from the episode, we fulfill the MGB2015 rules for longitudinal diarization. In the following lines we describe the different evaluated strategies:

4.2.1. Centering and whitening

The first sort of adaptation is applied to the i-vector normalization stage. Instead of using global information to perform this task, we have substituted it by a local version estimated per episode of each show. In fact, our experiments have primarily focused on the centering, maintaining the global whitening information. This decision was made due to the reduced

amount of available in-domain data to properly estimate a reliable whitening matrix. The results for this modification can be seen in Table 1

Table 1: *In-domain i-vector normalization adaptation*

Centering	VAD	Segmentation.	DER(%)
Not adapted	Ground Truth	Ground Truth	32.18
Adapted	Ground Truth	Ground Truth	**30.47**
Not adapted	Baseline	BIC	42.77
Adapted	Baseline	BIC	**38.54**

The obtained results indicate domain adaptation is effective and reports benefits in oracle and real conditions, despite its simplicity. Thus any further experiment will follow this procedure.

4.2.2. PLDA adaptation

The other experimented adaptation was applied to the PLDA model. This option should obtain a more significant improvement than the i-vector normalization, due to the PLDA modeling capabilities. However, any adaptation to be performed on PLDA needs speaker labels. In consequence, we have made use of the strategy proposed in section 3 for the PLDA adaptation with unsupervised labels.

In our experiments we have analyzed different methods to create the initial labels by using unsupervised clustering on the normalized i-vectors: Agglomerative Hierarchical Clustering with cosine distance and PLDA log-likelihood ratio, Variational Bayesian GMM and Mean-Shift with cosine distance kernel. Besides, experiments with a ground truth clustering, VAD and segmentation were also carried out, to measure the maximum achievable improvement.

Considering a simplified experiment with ground truth VAD (GTVAD) and ground truth segmentation (GTS), the obtained results are shown in Table 2

Table 2: *Initialization for PLDA domain adaptation. Ground Truth VAD and Segmentation*

Initialization Method	DER(%)
Non-adapted Baseline	32.18
I-vector Normalization Adaptation	30.47
*Ground Truth	11.14
PLDA LLR AHC	32.39
Cosine Distance AHC	24.85
Variational Bayes GMM	30.33
Mean-shift Cosine Kernel	**23.49**

*Adaptation weights in the same range of analysis

The results indicate that agglomerative clustering with cosine distance obtains a 22.77% relative improvement compared to non-adapted systems. This gain is increased by Mean-Shift (27.00% relative improvement). However, PLDA log-likelihood ratio gets degraded, probably due to domain mismatch. Variational Bayes GMM seems to be ineffective.

The same experiments, now performed on the same data but considering a real VAD, provided by the organization and a real segmentation, BIC, give the results included in Table 3.

Table 3: *Initialization for PLDA domain adaptation. Real VAD and BIC segmentation*

Initialization Method	DER (%)
Non-adapted Baseline	42.77
I-vector Normalization Adaptation	38.54
PLDA LLR AHC	39.66
Cosine AHC	35.06
Variational Bayes GMM	37.82
Mean-shift Cosine Kernel	**31.90**

In real conditions, the trend of results remains. PLDA log-likelihood ratio still degrades and Variational Bayes GMM slightly improves its performance. The other two techniques are consistent and robust to noisy segmentations and VADs (18.00% and 25.41% relative improvement compared to non-adaptation for cosine distance AHC and Mean-Shift respectively).

5. Conclusions

According to the results, domain adaptation has demonstrated its potential and usefulness, specially on the PLDA models. The inference of speaker labels by unsupervised methods and its posterior usage in model adaptation (PLDA) leads to more accurate models for each episode, obtaining important improvements in performance (up to 25.41% relative improvement). Moreover, this strategy is beneficial despite the simplicity of the considered unsupervised clustering techniques (agglomerative hierarchical clustering with cosine distance and the Mean-Shift with cosine distance kernel have a 18.00% and 25.41% relative improvement respectively).

Besides, our results are consistent in optimal (ground truth conditions) and real conditions, preserving the obtained relative improvement when oracle VAD and segmentation (22.47% relative improvement with agglomerative clustering and 27.00% by using Mean-Shift) are substituted by real estimates.

Comparing the results obtained with the different clustering techniques, simple techniques (Cosine Distance AHC and Mean Shift) only considering in-domain information are our best option. Data scarcity (only data from one episode is clustered each time) and domain mismatch can explain the poor performance of Variational Bayes GMM and PLDA respectively. Our GMM approach can also be degraded by our clustering assumption, one Gaussian per speaker, specially when the number of speakers rises.

Finally, the ground truth results show the influence of the initial seed accuracy for the PLDA model adaptation. Any improvement in the unsupervised clustering techniques, applied to our initialization, implies a reduction of the diarization error.

6. References

[1] 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.

[2] 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.

[3] C. Vaquero, A. Ortega, A. Miguel, and E. Lleida, "Quality Assess-

ment of Speaker Diarization for Speaker Characterization," *IEEE Trans. on Acoustics, Speech and Language Processing*, vol. 21, no. 4, pp. 816–827, 2013.

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

[5] D. Reynolds, P. Kenny, and F. Castaldo, "A study of New Approaches to Speaker Diarization." in *Interspeech*, 2009, pp. 1047–1050.

[6] 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.

[7] 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.

[8] 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.

[9] P. Bell, M. Gales, T. Hain, J. Kilgour, P. Lanchantin, X. Liu, A. McParland, S. Renals, O. Saz, M. Wester, and P. 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.

[10] H. Attias, "Inferring Parameters and Structure of Latent Variable Models by Variational Bayes," in *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence 1999 Stockholm, Sweden*, vol. 1, no. 1, 1999, pp. 21–30.

[11] 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.

[12] M. Senoussaoui, P. Kenny, T. Stafylakis, and P. Dumouchel, "A study of the Cosine Distance-Based Mean Shift for Telephone Speech Diarization," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 22, no. 1, pp. 217–227, 2014.

[13] 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.

[14] 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.

[15] ETSI, "ETSI ES 202 050 Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Advanced Front-end Feature Extraction Algorithm; Compression," pp. 1–45, 2002.