Odyssey 2016
21-24 Jun 2016, Bilbao, Spain

Short- and Long-Term Speech Features for Hybrid HMM-i-Vector based Speaker Diarization System

Abraham Woubie[1], *Jordi Luque*[2] *and Javier Hernando*[1]

[1] TALP Research Center, Department of Signal Theory and Communications,
Universitat Politecnica de Catalunya, Barcelona, Spain
[2] Telefonica Research, Edificio Telefonica-Diagonal, Barcelona, Spain

abraham.woubie.zewoudie@upc.edu, jls@tid.es, javier.hernando@upc.edu

Abstract

i-vectors have been successfully applied over the last years in speaker recognition tasks. This work aims at assessing the suitability of i-vector modeling within the frame of speaker diarization task. In such context, a weighted cosine-distance between two different sets of i-vectors is proposed for speaker clustering. Speech clusters generated by Viterbi segmentation are first modeled by two different i-vectors. Whilst the first i-vector represents the distribution of the commonly used short-term Mel Frequency Cepstral Coefficients, the second one depicts a selection of voice quality and prosodic features. In order to combine both the short- and long-term speech features, the cosine-distance scores of those two i-vectors are linearly weighted to obtain a unique similarity score. The final fused score is then used as speaker clustering distance. Our experimental results on two different evaluation sets of the Augmented Multi-party Interaction corpus show the suitability of combining both sources of information within the i-vector space. Our experimental results show that the use of i-vector based clustering technique provides a significant improvement, in terms of diarization error rate, than those based on Gaussian Mixture Modeling technique. Furthermore, this work also reports a significant speaker error reduction by augmenting i-vectors extracted from short-term spectral features with a second i-vector extracted from voice quality and prosody related speech features.

1. Introduction

Speaker diarization approaches segment and cluster a speech recording into homogeneous segments. While speaker segmentation partitions the audio data into acoustically homogeneous segments, speaker clustering groups speech segments of a particular speaker together [1].

The appropriate selection of speech features is one of the factors that affect the performance of speaker diarization system. Although short-term spectral features such as Mel Frequency Cepstral Coefficients (MFCCs) are the most widely used ones in speaker diarization [2], the works in [3, 4] show that long-term speech information can be employed to reveal individual differences which can not be captured by the short-term spectral features. Hence, these studies reveal that the fusion of short-term spectral features with long-term prosodic ones improves the performance of speaker diarization systems.

Jitter and shimmer voice quality features measure variations of the fundamental frequency and the amplitude of speakers voice, respectively. They have been applied in several speaker related tasks reporting successful results. It is reported in [5] that adding jitter and shimmer voice quality features to the baseline spectral ones improves the performance of a speaker recognition system. It is also shown in our work of [6] that the fusion of jitter and shimmer voice-quality features with with the spectral and prosodic ones improves the performance of a speaker diarization system. The fusion of voice-quality features with the spectral ones also provides better DER result than using only spectral features as reported in our work of [6].

Other factors that affect the performance of speaker diarization systems are the techniques employed to perform both speaker segmentation and speaker clustering. Speaker diarization systems mostly use Gaussian Mixture Modeling (GMM) based Bayesian Information Criterion (BIC) clustering technique to merge clusters within an Agglomerative Hierarchical Clustering (AHC) approach.

Factor analysis techniques which are the state of the art in speaker recognition have recently been successfully applied in speaker diarization experiments [7, 8, 9, 10, 11, 12, 13]. The speech clusters are first represented by i-vectors and the successive clustering stages are performed based on i-vector modeling. Representing the speech clusters by i-vectors enables to reduce the large-dimensional feature vector into a small dimensional one by retaining most of the relevant information. For instance, it is reported in [14] that modeling speech segments by i-vector and using cosine-distance clustering technique improves the performance of a diarization system more than GMM based BIC clustering technique. It is also shown in [7, 8, 9] that i-vector based cosine-distance clustering technique has been successfully applied in speaker clustering task.

Note that the above mentioned works extract i-vectors exclusively from short-term spectral features for speaker clustering task. Based on these studies, we propose the extraction of i-vectors from short-term spectral, and long-term voice-quality and prosodic features. The cosine distance scores of these i-vectors are then fused for speaker clustering task. The experiments have been conducted on a subset of AMI corpus [15], a multi-party and spontaneous speech set of recordings, and assessed in terms of speaker diarization error (DER). In order to validate the generalization of results, different parameters have first been tuned on a subset of AMI corpus. Then, we show how results of previous methodology generalize on held-out data both in single- and multiple-site scenarios of another subset of the AMI corpus.

The rest of this paper is organized as follows. The next sections give an overview of voice-quality and prosodic features followed by our speaker diarization system architecture. The fusion techniques are outlined in Section 4. Section 5 and 6 discuss about experimental results and conclusions, respectively.

doi: 10.21437/Odyssey.2016-58

2. Voice-quality and Prosodic Features

Although the most widely used features for speaker diarization are MFCC, it is shown in [3, 4, 6] that prosodic features can also be satisfactorily employed in speaker diarization systems. It is also reported in [16, 17] that prosodic features provides useful information for automatic speaker recognition. The work in [18] has also shown that prosodic features have also been successfully used in i-vector-based language identification task.

Jitter and shimmer, also known as voice quality features, measure variations of fundamental frequency and amplitude of speakers voice, respectively. Jitter and shimmer voice-quality features have been successfully used in speaker diarization experiments as reported in our work of [6]. They can be used to detect voice pathologies [19], speaking styles and can also be used to identify age and gender [20].

Although different estimations of jitter and shimmer measurements can be found in the literature, we have computed absolute jitter, absolute shimmer and shimmer apq3 measurements encouraged by the good results presented in [5]. The voice-quality features are estimated as described in [21].

- *Jitter (absolute):* It is a cycle-to-cycle perturbation in the fundamental frequency of the voice, i.e., the average absolute difference between consecutive periods.
- *Shimmer (absolute):* It is the average absolute logarithm of the ratio between amplitudes of consecutive periods.
- *Shimmer (apq3):* It is the three-point Amplitude Perturbation Quotient, the average absolute difference between the amplitude of a period and the average of the amplitudes of its neighbours, divided by the average amplitude.

Prosody is characterized by rhythm, intonation, stress and juncture of speech. Since these attributes cannot be measured directly, only their acoustic or perceptual correlates can be extracted from speech signal. We have extracted the following prosody-based features:

- *Fundamental Frequency:* It is influenced by age and gender. A typical adult male's fundamental frequency ranges from 100 to 150 Hz, and that of a typical adult female from 170 to 220 Hz. Therefore, fundamental frequency can be used to discriminate speakers.
- *Acoustic Intensity:* It can be used to mark stress and express emotions. Therefore, changes in loudness can be used as a potential speaker discriminant measure.
- *Formant Frequencies:* They occur only in voiced speech segments around frequencies that correspond to the speaker-specific resonances of the vocal tract. Therefore, they are suitable measures to help discriminate speakers.

Features related to the evolution in time of pitch, acoustic intensity and the first four formant frequencies have been extracted. Then, they are stacked with the three voice-quality features at the feature level, generating a nine dimensional feature vector. From now, for the sake of clarity, we shall refer to them as long-term features.

3. Speaker Diarization Architecture

Our speaker diarization system consists of three basic modules. The first module (Figure 1, Block A) performs mainly the feature extraction process. The second module (Figure 1, Block B) detects speaker change points and performs the Viterbi segmentation task. The third module (Figure 1, Block C) performs the bottom-up clustering and outputs the system hypothesis.

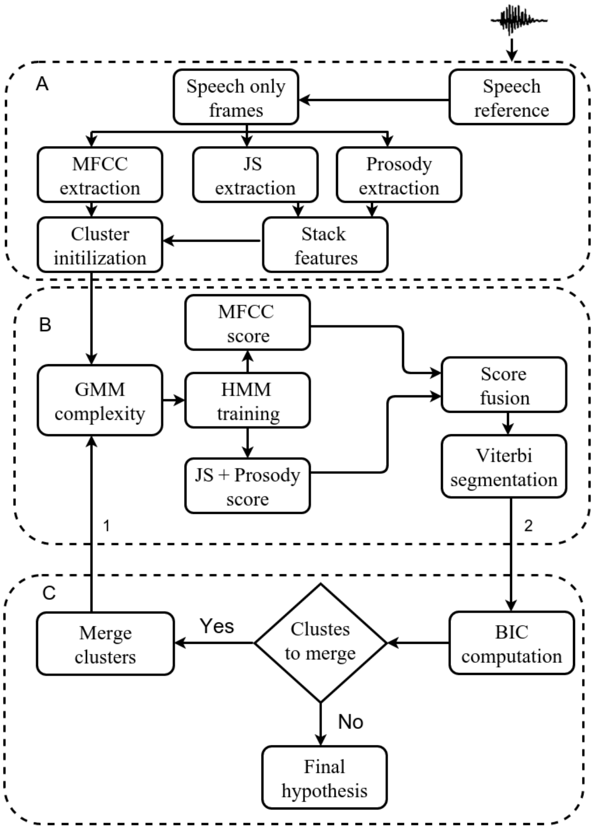

Figure 1: *Baseline speaker diarization architecture which draws a classical bottom-up agglomerative approach based on HMM/GMM modeling. The clustering technique is driven by Bayesian Information Criterion.*

3.1. Initialization (Figure 1, Block A)

The short-term and long-term speech features are masked by the use of speech/non-speech references. Our motivation resides in avoiding Speech Activity Detection (SAD) errors, thus focusing exclusively on speaker errors due to the diarization approach. The voice-quality and prosodic features are then stacked in the same feature vector. The speech signal is then equally partitioned to generate an initial number of clusters. The initial number automatically depends on meeting duration but it is constrained within the range $[10, 65]$. This method aims at tackling the common issues of AHC such as over-clustering and its high computational cost due to combinatorial explosion in pair-wise distance computation.

3.2. Speaker Segmentation (Figure 1, Block B)

The set of acoustic features are modeled using Hidden Markov Model(HMM)/GMM which is iteratively refined. Each state of the HMM is composed of a mixture of Gaussians, fitting the probability distribution of the features by the classical expectation-maximization (EM) algorithm. Note that for speaker segmentation, independent HMM models are estimated both for the short-term and long-term speech features and, at

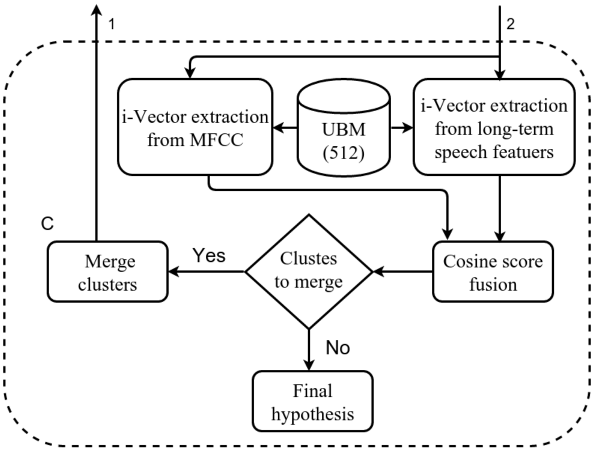

Figure 2: *Proposed i-vector based speaker clustering architecture based on a weighted cosine-distance among i-vectors.*

the end, their best paths obtained by Viterbi segmentation are fused. The number of mixtures is chosen as a function of available seconds of speech per cluster in the case of MFCC features. But, they are kept fixed for the long-term speech features. Finally, a time constraint, as in [22], is also imposed on the HMM topology. It constrains the minimum duration of the speaker turn to be greater than 3 seconds which is commonly used as mean value of a speaker intervention or speaker turn [22].

3.3. Speaker Clustering (Figure 2)

Our speaker diarization clustering system is based on agglomerative hierarchical clustering (AHC) technique. The baseline system of our speaker diarzation system uses GMM based BIC clustering technique.

Once the speech segments are generated by speaker segmentation, the speech segment clusters are first represented by a fixed dimensional i-vector. The successive clustering stages will group two acoustically similar segments per iteration based on their cosine distances among corresponding i-vectors. Factor analysis techniques, which provide an elegant way of obtaining a low dimensional fixed length representation, have been employed as proposed by [23].

Given speech segment clusters, the speaker and channel dependent GMM supervector (M) is represented as follows:

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

where *m* is a speaker and channel independent supervector from a Universal Background Model (UBM), *T* is a rectangular matrix of low rank that captures all kinds of variabilities, including speaker and session variabilities and *w* is a random vector having a standard normal distribution *N*(0,1). The components of the vector *w* are known as total factors. These new vectors, w, are called i-vectors.

As it is shown in Figure 2, two i-vectors have been extracted. While the first one represents the short-time spectral features, the second one represents the long-term ones. At each iteration, the Viterbi segmentation outputs a new clustering from which i-vectors are extracted. Then, the cosine-distance scores are computed among every pair of i-vectors representing each cluster and are linearly weighted.

At first, the similarity measure between all pairs of i-vectors is computed. Then, the two closest clusters are merged at each iteration, i.e., i-vector pairs with the highest cosine-distance scores. At the next iteration and after Viterbi segmentation, a new i-vector set is extracted from the new clustering and the similarity matrix between cluster pairs is updated. Note that i-vectors are only employed for speaker clustering task. The subsequent Viterbi segmentation and realignments stages employ short- and long-term speech feature as in our previous work [6].

The stopping criterion in the AHC is driven manually using a threshold λ on the matrix of distances. Once the cosine-distance scoring among all pair of clusters is less than λ, the merging process stops. Finally, the algorithm outputs the final speaker segmentation hypothesis, see Figure 3 for more details.

4. Fusion Techniques

The three voice-quality and the six prosodic features are first stacked in the same vector generating a nine dimensional feature vector. Note that this might be considered as simple fusion at the feature level. Two different score fusion techniques have been applied on speaker segmentation and clustering steps.

Independent HMM models are first estimated both for the short-term spectral and the long-term ones. The fusion of short-term spectral features with the long-term ones is carried out at the score level in speaker segmentation. It is done by fusing the log-likelihood scores corresponding to these feature sets.

Given a set of input features vectors, $\{\mathbf{x}\}$ and $\{\mathbf{y}\}$, the log-likelihood score is computed as a joint log-likelihood between features distributions as follows:

$$\begin{aligned}\log P(\mathbf{x},\mathbf{y}) &= \\ &\alpha \log P(\mathbf{x}|\theta_{ix}) + (1-\alpha)\log P(\mathbf{y}|\theta_{iy}),\end{aligned} \tag{2}$$

where $\log P(\mathbf{x},\mathbf{y})$ is the fused GMM score for cluster i, θ_{ix} is the model of cluster i from spectral feature vectors, and θ_{iy} is the model for the same cluster i using long-term features. The weight of the spectral feature vector is α and $(1-\alpha)$ is the weight of long-term speech features. The values of the α are tuned on development data set.

At the clustering step, once the speaker clusters are generated using Viterbi segmentation, the fused cosine-distance score is computed as follows:

$$\text{score}(i,j) = \beta.\frac{\mathbf{x_i}\cdot\mathbf{x_j}}{\|\mathbf{x_i}\|\|\mathbf{x_j}\|} + (1-\beta).\frac{\mathbf{y_i}\cdot\mathbf{y_j}}{\|\mathbf{y_i}\|\|\mathbf{y_j}\|}, \tag{3}$$

where score(i,j) is the fused cosine distance score between clusters i and j, $\mathbf{x_i}$ and $\mathbf{x_j}$ are the corresponding i-vectors extracted from short-term spectral features for clusters i and j, respectively. The vectors $\mathbf{y_i}$ and $\mathbf{y_j}$ stand for the i-vectors estimated using long-term speech features for same clusters i and j, respectively. Furthermore, two different weights are assigned to both cosine-distances. While β weights the cosine-distance of i-vectors extracted from short-term features, $(1-\beta)$ weights the cosine-distance of i-vectors extracted from the long-term features.

5. Experiments and Results

5.1. Databases and Experimental Setup

Wiener filtering technique is first applied on audio recordings. Then, speech references are used to mask the non-speech frames. This enables us to focus mainly on the speaker errors that occur due to segmentation and clustering. The short-term spectral-features are computed within a 30ms frame window at 10ms shift. The voice-quality and prosodic features are extracted over 30ms frame length and at 10ms shift using Praat software [24]. Then, each voice-quality and prosodic feature is estimated over a 500 ms window with 10ms shift. This is done to smooth out the feature estimation of the unvoiced frames. It is also done to synchronize the long-term features with the short-term ones.

The UBM and the T matrix are trained using 100 AMI shows which amount to 60 hours of audio. The AMI shows were recorded in English using three different setup rooms accounting for different acoustic properties. The recordings were performed at Idiap, Edinburgh, and TNO sites. Two universal background models (UBMs) of 512 Gaussians components have been trained. The first one is for the short-term spectral features and the second one is for the long-term ones. The UBM of short-term spectral features is trained on 20 cepstral co-efficients without the deltas. The UBM of long-term features is trained using the voice-quality and prosodic features. A 100 and 50 dimensional raw i-vector sizes are extracted from the short- and long-term speech features, respectively. The size of the total variability matrix is 100 for the short-term speech features and 50 for the long-term ones. The i-vector framework is carried out using the ALIZE open source software [25].

The experiments have been developed and tested on AMI corpus, a multi-party and spontaneous speech set of recordings [15].

- **Development database:** 10 shows from AMI corpus have been selected as a development set. These shows are used to tune the optimum parameters, i.e., stopping criterion threshold value, size of i-vectors and optimum set of weight values when score fusion is carried out. The total duration of the development set is 260.48 minutes. The development database is based on far-field microphone array channels sampled at 16kHz.
- **Test database:** In order to evaluate the performance of our proposed systems, two different experimental scenarios have been defined. The first one comprises 10 single-site audio recordings from Idiap site only. The second one is a multiple-site scenario. It includes 10 recordings from the Idiap, Edinburgh and TNO sites. The total duration of the single-site and multiple-site scenario dataset are 307.36 and 294.01 minutes, respectively. The test database is also based on far-field microphone array channels sampled at 16kHz.

Note that optimum parameters found through experimentation on the development database have been directly used on the test sets.

5.2. Performance Metrics

The performance metric employed for assessing speaker diarization systems is the Diarization Error Rate (DER). DER represents the sum of false alarm speech, missed speech and speaker error along time. Since speech references have been used, the rate of false alarms and missed speech have zero values in our experimental results. Hence, DER values reported in the following sections corresponds purely to speaker time confusion produced by the diarization system.

5.3. System Development

Experiments have been carried out first on the development set in order to find out the optimum parameters, i.e., the optimum α and β values, size of i-vectors and λ threshold value for stopping criterion.

The threshold value λ is selected manually as it is shown in Figure 3 in order to stop the iterative merging process. It is based on a data driven approach. The DER and corresponding cosine distance score values at each iteration are compared and λ that minimizes the DER value is selected. Thus, the system stops merging when the cosine distance score value among all pair of clusters is below this λ value. The optimum weight values for α and β are 0.975 for the short-term spectral features.

Table 1 depicts the results of the development dataset. The table shows that the baseline system of the development set has a DER of 30.09%. Note that the baseline system is based on GMM based BIC clustering technique exclusively on MFCC feature set. The table shows that replacing the BIC clustering of the development dataset by i-vector based cosine-distance speaker clustering technique on the same feature set decreases the DER to 27.03%. It represents more than 10% relative DER improvement more than the baseline system. The use of BIC clustering with MFCC and long-term features on the development dataset yields a DER of 25.98%. This corresponds to 13% relative DER improvement more than the baseline system. Finally, the table reports that the use of i-vector based cosine distance clustering technique with both short- and long-term speech features provides a DER of 25.42%. This DER value represents a 5.96% relative DER reduction more than the system based on same clustering technique but using only MFCC feature set.

Two main interpretations can be made from the development dataset results. The first one is that the results indicate the suitability of applying i-vector modeling technique within the clustering stage. The second one demonstrates that long-term speech features convey useful and complementary speaker discrimination more than the spectral features.

5.4. System Assessment

The tuned parameters, i.e., λ, β and size of i-vector(see Sections 4 and 5.3) have been used without modification on the single- and multiple-site scenario test sets.

As it is shown in Table 2, the baseline system system of the single-site scenario test set has a DER of 15.87%. The use of i-vector based cosine distance clustering using only MFCC feature set decreases the DER to 15.01%. This represents a 5.42% relative DER improvement more than the baseline system. Finally, the table reports that the use of i-vector based cosine distance clustering technique using short- and long-term speech features gives us the lowest DER, i.e., 13.37%. This represents a 10.99% relative DER improvement more than the system using same feature sets but applying GMM based BIC clustering technique. It also represents a 10.92% relative DER improvement more than the system using i-vector based clus-

Feature set	Clustering Technique	DER(%)
MFCC	GMM	30.09
MFCC + Long-term features	GMM	25.98
MFCC	i-vector	27.03
MFCC + Long-term features	i-vector	**25.42**

Table 1: *DER of Development dataset for GMM based BIC and i-vector based cosine distance clustering techniques using short- and long-term speech features.*

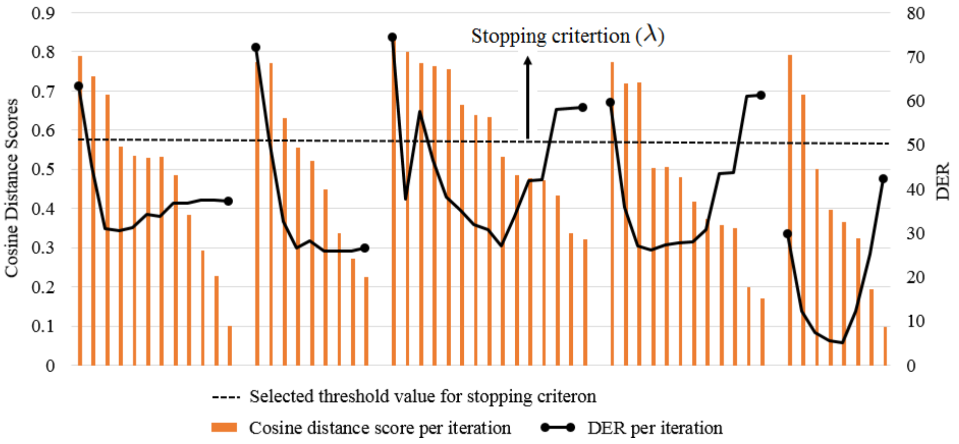

Figure 3: *DER and cosine-distance score per iteration for selected shows from the development set.*

Feature set	Clustering Technique	DER(%)
MFCC	GMM	15.87
MFCC + Long-term features	GMM	15.02
MFCC	i-vector	15.01
MFCC + Long-term features	i-vector	**13.37**

Table 2: *DER of Single-site scenario test set for GMM based BIC and i-vector based cosine distance clustering techniques using short- and long-term speech features.*

Feature set	Clustering Technique	DER(%)
MFCC	GMM	24.66
MFCC + Long-term features	GMM	22.96
MFCC	i-vector	22.79
MFCC + Long-term features	i-vector	**20.06**

Table 3: *DER of Multiple-site scenario test set for GMM based BIC and i-vector based cosine distance clustering techniques using short- and long-term speech features.*

tering technique and only MFCC feature set.

Finally, Table 3 reports results of the multiple-site scenario test set. The table shows that the baseline system of the multiple-site scenario test set provides a DER of 24.66%. The use of i-vector based based cosine distance clustering technique exclusively on MFCC feature set yields a DER of 22.79%. This represents a 7.58% relative DER improvement more than the baseline system. Finally, the table shows that the use of i-vector based clustering with short- and long-term speech features provides the best DER, i.e., 20.06%. It corresponds to a 12.63% relative DER improvement more than the system employing GMM clustering technique and using same feature sets. It also corresponds to a 11.98% relative DER improvement more than the system based on i-vector based clustering technique and only MFCC feature set.

The results reported on both single- and multiple-site conditions indicate the feasibility of using i-vector modeling for speaker clustering task. Moreover, the results show that the use of short- and long-term features enhance the performance of speaker diarization system by adding complementary speaker information, independently of the clustering approach.

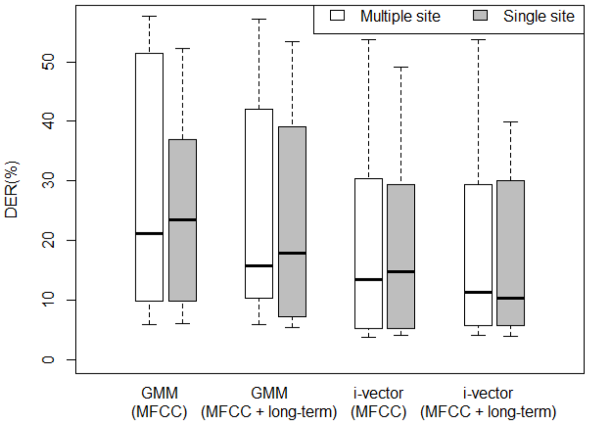

Figure 4: *Box plots of single- (grey) and multiple-site (white) scenario test recordings. Four different combinations are reported, resulting from the mixing of clustering approaches and feature sets.*

5.5. Discussion

The box plots in Figure 4 depict the DER distribution of the different recordings for both single- and multiple-site scenario test sets. The figure shows the minimum, lower quartile, median, upper quartile, and maximum DER performed by the GMM and i-vector clustering techniques. The figure shows that the proposed i-vector based clustering technique using short- and long-term features provides the minimum variance among all clustering techniques in terms of DER. Note that the DER variance is lower for single-site scenario test set than multiple-site one. As it is also shown in Figure 3, the DER values increase with the number of iterations because of over-clustering.

The DER values of different recordings show substantial differences as it is shown in Table 4. The duration of audio signals has lower correlation value with the DER for i-vector based clustering techniques than GMM based ones. This indicates that the use of i-vector based clustering reduces the DER range among the different recordings. As it shown in the table, the signal to noise (SNR) does not have a clear relationship with DER. Finally, the table shows that the final number of de-

	GMM	GMM	i-vector	i-vector			
Show	MFCC	MFCC+long-term	MFCC	MFCC+long-term	Duration	Real/Detected	SNR(db)
Name	DER(%)	DER(%)	DER(%)	DER(%)	(minutes)	# of speakers	
ES2015d	51.45	42.07	52.25	53.36	32.11	4/9	20.75
IS1006a	53.63	53.63	49.15	39.94	14.1	4/6	18.25
IS1004a	53.6	53.35	37.01	39.23	13.16	4/0	21.50
TS3009c	57.72	57.14	44.41	39.13	43	4/11	21.75
IS1006c	30.34	29.36	29.43	30.11	32.3	4/8	23
IS1009d	24	22.06	22.66	21.66	32.24	4/11	20
ES2008a	19.06	20.21	24.65	20.86	17.23	4/10	18.25
IS1008c	23.37	7.15	27.14	14.77	25.46	4/11	23.00
IS1002d	30.5	36.46	22.23	26.31	21.03	4/7	16.5
IS1000d	12.8	11.75	11.62	12.22	43.38	4/11	20.75
IS1002c	9.82	10.27	9.77	10.16	34.4	4/10	36.50
IS1004c	12.49	10.85	12.44	8.28	37.43	4/10	24.25
EN2003a	6.06	5.92	6.61	7.24	37.2	3/7	9
IS1008b	14	10.93	17.01	6.93	29.28	4/11	21.25
IS1000c	5.19	5.62	5.22	5.77	35.14	4/7	20.25
EN2009b	5.94	11.29	6.1	5.36	41.14	3/5	21.25
IS1009b	4.53	4.63	5.11	4.88	34.12	4/11	21.25
IS1004b	3.77	4.06	3.98	3.85	36.21	4/7	24

Table 4: DER for different shows of AMI recordings per each corresponding modeling technique and feature set combination. Recordings are sorted by DER of fourth column. Highlighted (grey) rows correspond to AMI multiple-site recordings. In addition, the recording duration and Signal to Noise (SNR) along with real/detected number of speakers are also reported. The real/detected number of speaker is for i-vector based clustering that uses MFCC and long-term features.

tected speakers is not optimum since a manual threshold value is used as a stopping criterion. Therefore, an automatic stopping criterion threshold value that varies per iteration and recording should be found.

Although the extraction of i-vectors and the use of i-vector based clustering techniques reduce the DER error for most of the recordings, the DER increases for few of them compared to the baseline system. The reasons for this issue are still unknown to us and needs further exploration.

In overall, our experimental results validate the usefulness of the proposed methodology. The use of the i-vector based clustering technique based on short- and long-term speech features increase the robustness and reliability of speaker diarization systems.

6. Conclusions

In this work, we have proposed the extraction of i-vectors from short- and long-term speech features and the fusion of their cosine-distance scores for speaker clustering task. Experimental results were carried out on two evaluation sets of subset of AMI corpus, i.e., single-site and multiple-site scenarios.

First of all, experimental results show that i-vector clustering technique based on short- and long-term features provides better DER than the same clustering technique using only short-term features. Secondly, the results show that i-vector clustering technique provides a substantial relative DER improvement for most of the recordings more than GMM one.

The results of our work manifest the usefulness of i-vector based clustering technique based on short- and long-term speech features within in the framework of speaker diarization.

7. Acknowledgment

This work has been supported by the Spanish Ministry of Economy and Competitiveness under contract PCIN-2013-067, the Generalitat de Catalunya under contract 2014-SGR-1660, and by the project TEC2015-69266-P (MINECO/FEDER, UE). This work has also received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 645323. This text reflects only the authors' view and the Commission is not responsible for any use that may be made of the information it contains.

8. References

[1] S.E. Tranter and D.A. Reynolds, "An Overview of Automatic Apeaker Diarization systems," *IEEE Transactions on Audio, Speech, and Language Processing,*, 2006.

[2] 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*, 2011.

[3] G. Friedland, O. Vinyals, Y. Huang, and C. Müller, "Prosodic and other Long-Term Features for Speaker Diarization," *IEEE Transactions on Audio, Speech, and Language Processing*, 2009.

[4] M. Zelenák and J. Hernando, "The Detection of Overlapping Speech with Prosodic Features for speaker diarization," in *INTERSPEECH*, 2011, pp. 1041–1044.

[5] M. Farrús, J. Hernando, and P. Ejarque, "Jitter and Shimmer Measurements for Speaker Recognition," in *INTERSPEECH*, 2007.

[6] A. Woubie, J. Luque, and J. Hernando, "Using Voice-quality Measurements with Prosodic and Spectral Features for Speaker Diarization," in *INTERSPEECH*, 2015.

[7] Patrick Kenny, Douglas Reynolds, and Fabio Castaldo, "Diarization of telephone conversations using factor analysis," *in IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 6, pp. 1059–1070, 2010.

[8] J. Franco-Pedroso, I. Lopez-Moreno, D.T. Toledano, and Joaquín González-Rodríguez, “ATVS-UAM system description for the audio segmentation and speaker diarization Albayzin 2010 evaluation,” in *FALA*, 2010.

[9] S. Shum, N. Dehak, E. Chuangsuwanich, D. A Reynolds, and J. R Glass, “Exploiting intra-conversation variability for speaker diarization,” in *Proceedings of Interspeech, Florence, Italy*, 2011.

[10] S. Shum, N. Dehak, and J. Glass, “On the use of spectral and iterative methods for speaker diarization,” in *Interspeech*, 2012.

[11] C. Vaquero Avils-Casco, *Robust diarization for speaker characterization (Diarizacin robusta para caracterizacin de locutores)*, Ph.D. thesis, University of Zaragoza, Zaragoza, 2011.

[12] M. Senoussaoui, P. Kenny, P. Dumouchel, and T. Stafylakis, “Efficient iterative mean shift based cosine dissimilarity for multi-recording speaker clustering,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013.

[13] Jesus Villalba, Alfonso Ortega, Antonio Miguel, and Eduardo Lleida, “Variational Bayesian PLDA for Speaker Diarization in the MGB Challenge,” in *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*. IEEE, 2015, pp. 667–674.

[14] J. Silovsky and J. Prazak, “Speaker diarization of broadcast streams using two-stage clustering based on i-vectors and cosine distance scoring,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012.

[15] “The Augumented Multi-party Interaction Project, AMI Meeting Corpus.,” Website, `http://corpus.amiproject.org`, 2011.

[16] Elizabeth Shriberg, Luciana Ferrer, Sachin Kajarekar, Anand Venkataraman, and Andreas Stolcke, “Modeling Prosodic Feature sequences for Speaker Recognition,” *Speech Communication*, vol. 46, no. 3, pp. 455–472, 2005.

[17] Najim Dehak, Pierre Dumouchel, and Patrick Kenny, “Modeling Prosodic Features with Joint Factor Analysis for Speaker Verification,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 7, pp. 2095–2103, 2007.

[18] David Martínez, Lukáš Burget, Luciana Ferrer, and Nicolas Scheffer, “ivector-based Prosodic System for Language Identification,” in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4861–4864.

[19] J. Kreiman and B. R. Gerratt, “Perception of aperiodicity in pathological voice,” *The Journal of the Acoustical Society of America*, vol. 117, no. 4, 2005.

[20] A. Sadeghi N. and M.M. Homayounpour, “Speaker age interval and sex identification based on Jitters, Shimmers and Mean MFCC using supervised and unsupervised discriminative classification methods,” in *8th International Conference on Signal Processing*, 2006.

[21] A. Woubie, J. Luque, and J. Hernando, “Jitter and Shimmer Measurements for Speaker Diarization,” in *VII Jornadas en Tecnología del Habla and III Iberian SLTech Workshop: Las Palmas de Gran Canaria, Spain*, 2014, pp. 21–30.

[22] J. Ajmera and C. Wooters, “A Robust Speaker Clustering Algorithm ,” in *Proceedings of IEEE Speech Recognition and Understanding Workshop*, St. Thomas, U.S. Virgin Islands, 2003.

[23] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front end factor analysis for speaker verification,” *in IEEE Transactions on Audio, Speech and Language Processing*, 2011.

[24] P. Boersma and D. Weenink, “Praat: doing phonetics by computer, version 5.3.69,” `http://www.praat.org/`.

[25] A. Larcher, J. F. Bonastre, B. G. B. Fauve, K. Lee, C. Lévy, H. Li, J. S. D. Mason, and J. Parfait, “ALIZE 3.0 - open source toolkit for state-of-the-art speaker recognition,” in *INTERSPEECH*, 2013, pp. 2768–2772.