Zewoudie *et al. EURASIP Journal on Audio, Speech, and Music Processing* (2018) 2018:14
https://doi.org/10.1186/s13636-018-0140-x

EURASIP Journal on Audio,
Speech, and Music Processing

RESEARCH Open Access

The use of long-term features for GMM- and i-vector-based speaker diarization systems

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

Abstract

Several factors contribute to the performance of speaker diarization systems. For instance, the appropriate selection of speech features is one of the key aspects that affect speaker diarization systems. The other factors include the techniques employed to perform both segmentation and clustering. While the static mel frequency cepstral coefficients are the most widely used features in speech-related tasks including speaker diarization, several studies have shown the benefits of augmenting regular speech features with the static ones.

In this work, we have proposed and assessed the use of voice-quality features (i.e., jitter, shimmer, and Glottal-to-Noise Excitation ratio) within the framework of speaker diarization. These acoustic attributes are employed together with the state-of-the-art short-term cepstral and long-term prosodic features. Additionally, the use of delta dynamic features is also explored separately both for segmentation and bottom-up clustering sub-tasks. The combination of the different feature sets is carried out at several levels. At the feature level, the long-term speech features are stacked in the same feature vector. At the score level, the short- and long-term speech features are independently modeled and fused at the score likelihood level.

Various feature combinations have been applied both for Gaussian mixture modeling and i-vector-based speaker diarization systems. The experiments have been carried out on Augmented Multi-party Interaction meeting corpus. The best result, in terms of diarization error rate, is reported by using i-vector-based cosine-distance clustering together with a signal parameterization consisting of a combination of static cepstral coefficients, delta, voice-quality, and prosodic features. The best result shows about 24% relative diarization error rate improvement compared to the baseline system which is based on Gaussian mixture modeling and short-term static cepstral coefficients.

Keywords: Cosine-distance, Fusion, GNE, i-Vector, Jitter, PLDA, Prosody, Segmentation, Clustering, Shimmer

1 Introduction

An audio recording normally consists of different speakers, music segments, noises, etc. Speaker diarization needs to first classify the speech and non-speech parts of the audio signal. Then, it marks the speaker changes in the detected speech. Finally, it clusters speech segments which belong to the same speaker [1].

One of the factors that affect the performance of speaker diarization systems is the extraction of relevant speaker features. Mel frequency cepstral coefficients (MFCCs) are the most widely used short-term speech features for speaker diarization [2]. Despite its broad employment in speaker diarization, it is described in [3, 4] that augmenting short-term speech characteristics with long-term ones improves the performance of speaker diarization systems. These results manifest that the long-term features provide some complementary and discriminative information about different speakers not captured by the classical MFCCs.

One of the main contributions of this work is the use of jitter and shimmer voice-quality features both for GMM- and i-vector-based speaker diarization systems. Jitter and shimmer voice-quality measurements discern variations of fundamental frequency and amplitude, respectively. Studies show that these measurements can be used to detect voice pathologies [5], speaking styles, and emotions [6] and also identify age and gender [7]. For example, the authors in [8] report that fusing jitter and shimmer voice-quality measurements with the baseline cepstral features improves the performance of speaker

* Correspondence: abraham.woubie.zewoudie@upc.edu
[1] TALP Research Center, Department of Signal Theory and Communications, Universitat Politecnica de Catalunya, Barcelona, Spain
Full list of author information is available at the end of the article

recognition systems. It is also described in [6] that the use of jitter and shimmer measurements together with cepstral ones improves the classification accuracy of different speaking styles. These measurements have also been successfully used in speaker diarization in our previous works of [9, 10].

In addition to speech features, other factors that affect the performance of speaker diarization systems are the statistical techniques employed to carry out speaker segmentation and speaker clustering. Speaker diarization approaches typically employ Gaussian mixture modeling (GMM)-based Bayesian information criterion (BIC) clustering technique to merge clusters.

Factor analysis techniques which are the state of the art in speaker recognition have recently been successfully applied in speaker diarization experiments [11–17]. In these works, the speech clusters are first represented by i-vectors and the successive clustering stages are carried out using i-vector modeling techniques. The representation of the speech clusters by i-vectors enables to reduce the large-dimensional feature vector into a small-dimensional one while retaining most of the relevant information. For instance, it is reported in [18] that modeling speech segments by i-vector and using cosine-distance clustering technique improves the performance of a diarization system compared to GMM-based BIC clustering technique. It is also shown in [19, 20] that use of i-vector-based probabilistic linear discriminant analysis (PLDA) clustering technique provides better diarization error rate (DER) result compared to both GMM-based BIC and i-vector-based cosine-distance clustering techniques.

Note that the above mentioned works extract i-vectors exclusively from the short-term cepstral features for speaker clustering. In our work, we have proposed the extraction of i-vectors from the short-term cepstral and long-term speech features and the fusion of their cosine-distance and PLDA scores. These results have already been published in our previous works of [21, 22].

Based on these studies, we have proposed the use of jitter and shimmer voice-quality features both for GMM- and i-vector-based speaker diarization systems. The voice-quality features are used together with other long-term features (i.e., pitch, intensity, formants, and Glottal-to-Noise Excitation ratio) and short-term cepstral features. The fusion of the voice-quality features with the other long-term and short-term cepstral features is carried out both at the feature and score levels.

The other contribution of this work is the use of delta dynamic features for speaker clustering both for GMM- and i-vector-based speaker diarization systems. The first time derivative of the instantaneous cepstral features (i.e., deltas) have been successfully used in speaker recognition [23] and speech recognition [24]. However, they are not widely used in speaker diarization systems. For example, it is reported in [25] that since the delta features deteriorate the diarization results, only the static MFCC features are used in speaker diarization. It is also reported in [26] that delta features are not used in speaker diarization systems.

Since delta features provide dynamic information to the static cepstral coefficients, we have also analyzed the impact of delta features both for GMM- and i-vector-based diarization systems. The delta features have been used only in speaker clustering because of their limited temporal resolution in segmentation. The delta features are used together with the static cepstral coefficients in speaker clustering.

In all of our previous works [9, 10, 21, 22], only the static MFCCs were used. The deltas were not used in these works. We have also analyzed the impact of using formants together with MFCC both for GMM- and i-vector-based speaker diarization systems. The test experiments have been carried out on 112 AMI shows. In our previous works, the test experiments were carried out on only 20 AMI shows [27].

The rest of this paper is organized as follows. The next section gives an overview of the proposed long-term speech features followed by the proposed speaker diarization systems. Finally, Section 4.2 and Section 6 are presented.

2 Long-term speech features for speaker diarization

Mel frequency cepstral coefficients (MFCCs) are the most widely used short-term speech features in speaker diarization [2]. While short-term features are extracted from a single speech frame, long-term features are extracted from portions of speech longer than one frame. Long-term features capture phonetic, prosodic, lexical, syntactic, semantic, and pragmatic information. Although short-term spectral features are the most widely used ones in different speech applications, the authors in [3, 4, 28] show that long-term features can be employed to reveal individual differences which cannot be captured by short-term spectral features.

Since long-term features add complimentary information to the classical MFCC features, fusion of short-term spectral features with long-term features has been applied on speaker diarization experiments [3, 4, 29]. Fusion techniques also increase the reliability and robustness of a system [30].

The proposed long-term speech features proposed in this work are described as follows:

2.1 Dynamic features

It is possible to obtain more detailed speech features by using a derivation on the MFCC acoustic vectors. This permits the computation of the dynamic MFCCs, as the first-order derivatives of the MFCC. The speech features which are the time derivatives of the spectrum-based speech features are known as dynamic speech features. The delta dynamic features can complement the static information obtained by the MFCC.

Most of the state-of-the-art speaker diarization systems use only the static MFCC for diarization [2]. The static MFCC features cannot accurately capture the transitional characteristics of speech signal which contains speaker-specific information. The delta features can be used to extract more detailed speech features using the time derivation of static MFCC acoustic vectors. Hence, they provide dynamic information to the static MFCC features [31]. The dynamic features are not also adversely affected by convolution noise (i.e., channel effect) like the static MFCCs.

The MFCC feature vector describes only the power spectral envelope of a single frame. However, a speech signal has also information in the dynamics (i.e., what are the trajectories of the MFCC coefficients over time). The delta features are computed as the time differences between a set of consecutive feature vectors. They are usually appended to the static coefficients at the frame level. The extraction of the MFCC trajectories and appending them to the static MFCC features improve the performance of different speech applications. The delta features have been shown to improve the performance of speaker recognition [23], speech recognition [24], and speaker classification [32] systems.

The delta features are computed by the weighted sum of feature vector differences between two consecutive static coefficients as follows:

$$d_t = \frac{\sum_{\theta=1}^{\omega}\theta(C_{i+\theta}-C_{i-\theta})}{2\sum_{\theta=1}^{\omega}\theta^2} \tag{1}$$

where d_t is the delta coefficient at time t computed in terms of the corresponding static coefficients from $C_{i-\theta}$ to $C_{i+\theta}$. The delta window size is represented by ω.

The delta dynamic features are used together with the static and other long-term ones. While the static MFCC and other long-term features are used both in segmentation and clustering, the deltas are used only in speaker clustering.

2.2 Voice-quality features

Jitter and shimmer voice-quality measurements measure variations of the fundamental frequency and the amplitude of speaker's voice, respectively. They have been applied in several speaker-related tasks reporting successful results. For instance, it is reported in [8] that adding jitter and shimmer voice-quality features to the baseline spectral ones improves the performance of a speaker recognition system. By using Praat [33], five different jitter and six different shimmer measurement estimations can be extracted. But we have extracted only absolute jitter, absolute shimmer, and shimmer apq3 as they are used in [8]. It is reported in [8] that absolute jitter (Fig. 1), absolute shimmer (Fig. 2), and shimmer apq3 measurements provide better results for speaker recognition compared to the other jitter and shimmer measurements.

- 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, expressed as:

$$\text{litter}(\text{absolute}) = \frac{1}{N-1}\sum_{i=1}^{N-1}|T_i-T_{i+1}| \tag{2}$$

where T_i are the extracted pitch period lengths and N is the number of extracted pitch periods.

- Shimmer (absolute): It is the average absolute logarithm of the ratio between amplitudes of consecutive periods, expressed as:

$$\text{Shimmer}(\text{absolute}) = \frac{1}{N-1}\sum_{i=1}^{N-1}\left|20\log\left(\frac{A_i+1}{A_i}\right)\right| \tag{3}$$

where A_i are the extracted peak-to-peak amplitude data and N is the number of extracted pitch 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 neighbors, divided by the average amplitude. It is expressed as:

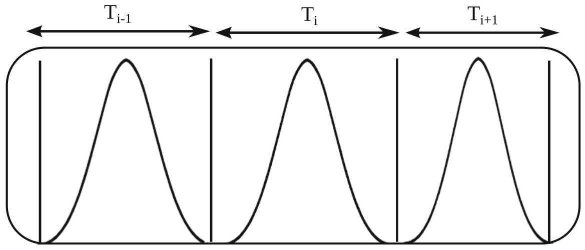

Fig. 1 Absolute jitter measurement for $N = 3$ pitch periods

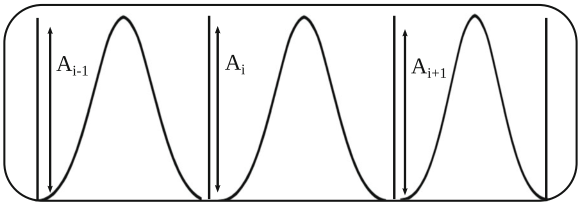

Fig. 2 Absolute shimmer measurement for $N = 3$ pitch periods

$$\mathrm{apq3} = \frac{\frac{1}{N-1}\sum_{i=2}^{N-2}\left|A_i - \frac{A_{i-1}+A_i+A_{i+1}}{3}\right|}{\frac{1}{N-1}\sum_{i=1}^{N}A_i} \tag{4}$$

where A_i are the extracted peak-to-peak amplitude data and N is the number of extracted pitch period.

2.3 Prosodic features

Prosody studies those aspects of speech that typically apply to a level above that of the individual phoneme.

It is expressed using intonation, rhythm, and stress which are perceived by listeners as changes in fundamental frequency, sound duration, and loudness, respectively [34].

Encouraged by work of [4], we have extracted features related to the evolution in time of pitch, acoustic intensity, and the first four formant frequencies to validate their performance in this work.

- Pitch

Pitch is the most important prosodic property of speech. It contains speaker-specific information. The default pitch value and range of a speaker is influenced by the length and mass of the vocal folds in the larynx [35]. The pitch values of different speakers vary in relation to their age and gender. Pitch can be used as an important acoustic cue for tone, lexical stress, and intonation.

- Acoustic intensity

It shows changes in loudness or energy of a speech signal. It is used to mark stress and express emotions. Therefore, changes in loudness can be used as a potential speaker discriminant measure.

- Formant frequencies

They are concentrations of acoustic energy around particular frequencies at roughly 1000-Hz intervals. 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.

2.4 Glottal-to-Noise features

In addition to jitter and shimmer acoustic parameters that are used to measure perturbations of speech signals, noise parameters can also be used to assess voice quality of a speaker [36]. Noise parameters can be used to assess the noise content of the signal and can be used in the evaluation of voice quality [36].

Glottal-to-Noise (GNE) is an acoustic measure that can be used to assess the amount of voice excitation by vocal-fold oscillations versus excitation by turbulent noise. It indicates whether a given voice signal originates from vibrations of the vocal folds or from turbulent noise generated in the vocal tract [37]. The main advantage of GNE is its computation is independent of variations of fundamental frequency and amplitude [36, 38]. GNE is closely related to breathiness and is considered as a reliable measure for the relative noise level even in the presence of strong amplitude and frequency perturbations.

The process of extracting GNE (see Fig. 3) is described in [37] as follows:

1. Down-sample the audio signal to 10 kHz.
2. Do inverse filtering of the speech signal.
3. Calculate the Hilbert envelopes of different frequency bands with fixed bandwidth and different center frequencies.

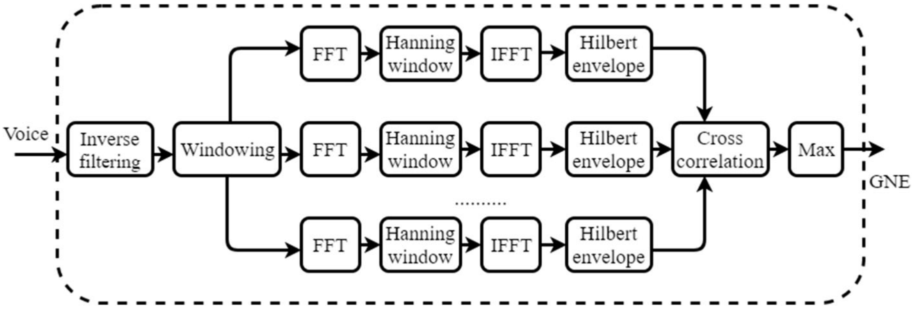

Fig. 3 Process of GNE extraction

4. Consider every pair of envelopes for which the difference of their center frequencies is equal or greater than half the bandwidth: calculate the cross correlation function between such envelopes.
5. Pick the maximum of each correlation.
6. Pick the maximum from the maxima.

It is shown in the work of [36] that GNE parameter has a significant potential to screen voices since it quantifies the amount of voice excitation and turbulent noise. It is also reported in [39] that GNE provides reliable measurements in terms of discrimination among normal and pathological voices compared to other classical long-term noise measurements, such as normalized noise energy and harmonics-to-noise ratio. It has also been used successfully to screen voice disorders in [39].

3 Proposed speaker diarization systems

The baseline or reference speaker diarization system used in this work is depicted in Fig. 4. Conceptually, the system performs three tasks: the first task performing the feature extraction (Fig. 4, block A), the second detecting speaker changes to segment the speech data (Fig. 4, block B), and the third one grouping the segmented regions together into speaker-homogeneous clusters and displaying the system hypothesis (Fig. 4, block C).

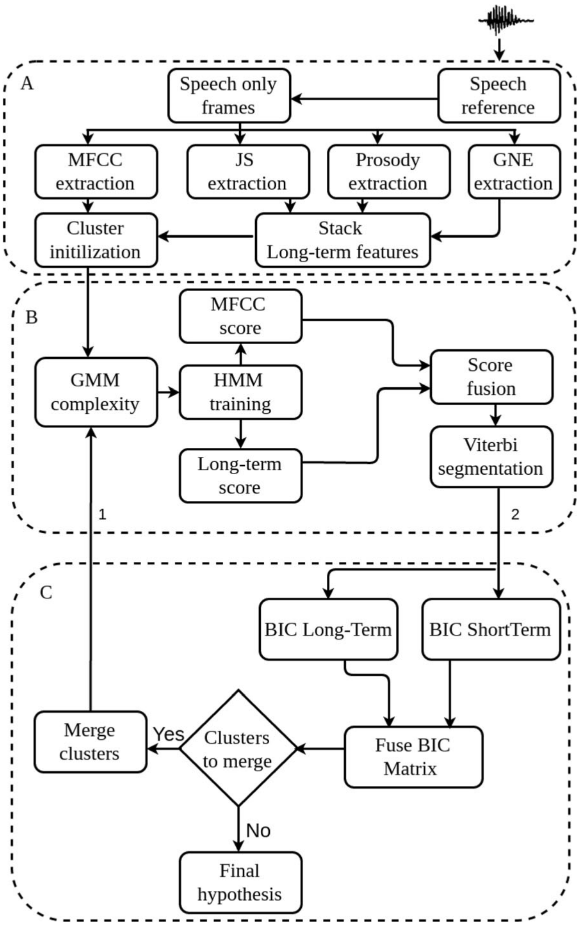

Fig. 4 Proposed HMM-GMM speaker diarization system. The dotted rectangles are the ones proposed in the proposed system. The undotted ones are the baseline system

In the proposed speaker diarization systems, the short-term cepstral coefficients are augmented with long-term feature information. Firstly, the non-cepstral features (i.e., prosodic and GNE) are combined together at the feature level. Then, they are fused with MFCCs at the score level to assess the contribution of each set of feature representation. Depending on the proposed system, the fused score is interpolated either by using the likelihood produced by GMM models or by the corresponding i-vector scoring method.

3.1 Proposed GMM-based speaker diarization system

One of the main differences of the proposed GMM speaker diarization system is the inclusion of jitter and shimmer voice-quality information and its combination with formants, pitch, intensity, GNE, and MFCC characteristics.

In addition to the long-term information, the proposed HMM/GMM speaker diarization system is also augmented with the extraction of the delta dynamic features. In order to capture several speaker properties, two feature representations are explored. While the first one is the combination of static cepstral coefficients and delta features, the second speaker vector consists of the voice-quality, prosodic, and GNE information. It is worth to mention that both sets of feature representations are employed both for speaker segmentation and clustering. However, delta features are only employed in the clustering task and discarded from the audio segmentation one.

After the computation of short- and long-term speech features, the audio signal is equally partitioned to generate the initial number of clusters (see Fig. 4, block A). The initial number of clusters depends on meeting duration, but it is constrained between 10 and 65. This is done to avoid the common issues of agglomerative hierarchical clustering (AHC) such as over-clustering and its high computational cost of pair-wise distance computation. Independent HMM/GMM models are estimated for each feature set. The fusion of short- and long-term information in speaker segmentation is carried out at the score level (see Fig. 4 block B). It is done by a weighted interpolation of the two log-likelihoods resulting from each GMM distribution, i.e., cepstral and long-term feature distributions.

Given a set of input feature vectors {x} and {y}, the log-likelihood score in the proposed segmentation is computed as a joint log-likelihood between feature distributions as follows:

$$\log\,\Pr\left(x,y|\theta_i^x,\theta_i^y\right)=\alpha\,\log\,\Pr\left(x|\theta_i^x\right)+(1-\alpha)\,\log\,\Pr\left(y|\theta_i^y\right)\qquad(5)$$

where $\log\,\Pr\left(x,y|\theta_i^x,\theta_i^y\right)$ is the logarithm of the fused emission probabilities for cluster i. While the model of cluster i using cepstral feature vectors is represented by θ_i^x, the model for the same cluster i using long-term features is denoted by θ_i^y. The interpolation weight which controls the individual contribution of each feature set is α. The values of α are tuned on development data set.

Once the audio data is segmented by the Viterbi algorithm, the homogeneous segments are grouped together based on Bayesian information criterion (BIC) distance among clusters (Fig. 5 block C).

Given two speech segments i and j, the BIC distance computation is carried out as follows:

$$BIC(i,j)=\beta.BIC_{ij}^x+(1-\beta).BIC_{ij}^y\qquad(6)$$

where BIC_{ij}^x and BIC_{ij}^y are the BIC distances between clusters i and j generated using short- and long-term speech features. The BIC distances computed using the short- and long-term feature sets are weighted by β and $(1-\beta)$, respectively, driving the contribution of each distance to the final fused BIC distance.

3.2 Proposed i-vector-based speaker diarization system

The extraction of i-vectors from audio segments and its usage for speaker clustering are the key components of our proposed i-vector system. The extraction is independently performed from the short-term static cepstral coefficients and from the long-term speech features. Therefore, two sets of i-vectors are extracted for each segmentation, representing different speaker traits. While the first i-vector is extracted from short-term features, the second i-vector is computed from the long-term ones, i.e., the combination of voice-quality, pitch, intensity, formants, and GNE.

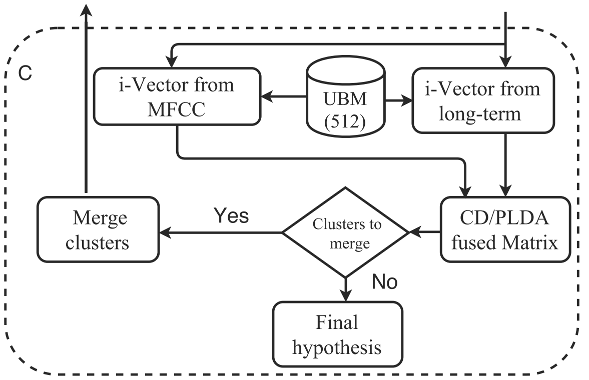

Fig. 5 Proposed i-vector-based clustering architecture. The proposed i-vector-based cosine-distance and PLDA clustering are based on short- and long-term speech features

We have also investigated the use of delta features for speaker clustering. The delta features are combined together with the static cepstral coefficients at the feature level, yielding a vector of size 40. Once the i-vectors are extracted from the audio segments, the fused distance between two segments is computed as interpolation of two i-vector distances. We have carried out both cosine and PLDA scoring techniques to explore the better one. It is worth to highlight that i-vectors are uniquely employed for speaker clustering task. In fact, it is the main difference of the proposed i-vector system compared to the previous HMM/GMM system. Note that the feature extraction and speaker segmentation modules are the same both in the proposed GMM system (see previous section) and in the proposed i-vector system. Results reported in the literature [20] suggest that i-vector modeling is not well suited for the speaker segmentation task since it is difficult to reliably compute an estimate of i-vector from segments of short duration. The extraction of i-vectors from short duration has shown to degrade the performance of speaker recognition systems [20]. Several works have paid attention to this issue during the last few years within the speaker recognition community [40, 41]. We have carried out some experiments to explore the impact of applying i-vectors on our segmentation module. Similar to results of speaker recognition in [40], the experimental results show that the use of i-vectors for speaker segmentation results in performance degradation in terms of diarization error rate.

3.2.1 Cosine-distance scoring

In order to perform the agglomerative cluster procedure, the similarity measure among all pairs of i-vectors is needed to be computed leading to a symmetric matrix of distances. Then, at each iteration of the agglomerative clustering, the two closest clusters are merged, i.e., i-vector pairs with the highest cosine-distance score. After merging the two closest clusters, the Viterbi segmentation is carried out, and this process iterates by extracting a new set of i-vector from the newly hypothesized clustering. The similarity matrix between cluster pairs is also updated. The previous steps iterate until the speaker diarization system reaches the stopping criterion and provides the final segmentation.

The cosine-distances from both sets of i-vectors are linearly weighted and applied as a fused distance metric for speaker clustering. The fused cosine-distance score is computed as follows:

$$\mathrm{CDS} = \beta \cdot \frac{w_i \cdot w_j}{\|w_i\|\|w_j\|} + (1 - \beta)\frac{w_i^{'} \cdot w_j^{'}}{\left\|w_i^{'}\right\|\left\|w_j^{'}\right\|} \quad (7)$$

where CDS stands for the fused distance score between clusters i and j. The corresponding i-vectors extracted from short-term cepstral coefficients and for clusters i and j are denoted by w_i and w_j, respectively. Similarly, the vectors $w_i^{'}$ and $w_j^{'}$ represent the i-vectors extracted by modeling long-term speech features for the 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.

3.3 PLDA scoring

The main contribution of this section comprises the replacement of the i-vector-based cosine-distance by an i-vector-based PLDA scoring within the speaker clustering. Since PLDA is the state of the art in speaker verification [42] and speaker diarization systems [21], we are also motivated to use the i-vector-based PLDA clustering technique in our proposed speaker diarization system.

Firstly, two sets of i-vectors are extracted from the short- and long-term speech features. Then, the PLDA scores of these two i-vectors are linearly weighted to obtain a fused distance score for further use on the speaker clustering task.

Finally, the fused PLDA score is computed as follows:

$$\begin{aligned} & y.\ \log\frac{p\left(w_i, w_j|H_1\right)}{p(w_i|H_0)p\left(w_j|H_0\right)} \\ & +(1 - \gamma)\frac{p\left(w_i^{'}, w_j^{'}|H_1\right)}{p(w_i|H_0)p\left(w_j|H_0\right)} \end{aligned} \quad (8)$$

where w_i and w_j are the i-vectors extracted from the short-term cepstral coefficients for cluster i and cluster j, respectively. Similarly, $w_i^{'}$ and $w_j^{'}$ represent the i-vectors extracted from long-term speech features for the same clusters i and j, respectively. Moreover, hypothesis H_1 and H_0 assume that the two i-vectors belong to the same and different speakers, respectively. Finally, the PLDA scores of i-vectors extracted from the short- and long-term features are weighted by γ and $(1 - \gamma)$, respectively.

Note that the long-term features in Eqs. 7 and 8 refer to two possibilities: voice quality with prosodic features and voice-quality with prosodic and GNE features.

4 Experimental setups and results

4.1 Databases and experimental setup

The experiments are developed and tested on AMI corpus, a multi-party and spontaneous speech set of recordings [27]. The AMI is a meeting corpus consisting 100 h of audio in 171 shows which use a range of signals synchronized to a common timeline. The shows were recorded using close-talking and far-field microphones. The meetings were recorded in English using three different rooms with different acoustic properties in IDIAP, Edinburgh, and TNO sites.

The development and test set are based on the far-field microphone array channels sampled at 16 kHz.

Development set: 10 shows are selected from IDIAP, Edinburgh, and TNO sites as a development set. These shows are used to tune the optimum parameters (i.e., optimum set of weight values for the short- and long-term speech features and optimum stopping threshold value). The total and average duration of the development set is 284 and 28.4 min, respectively.

Test set: In order to evaluate the performance of the proposed systems, the test experiments are carried out on 112 AMI shows selected from the IDIAP, Edinburgh, and TNO sites. The total and average duration of the test sets of the whole recording are 4075 min (about 69 h) and 36.38 min, respectively.

Manually annotated speech references are used to extract the speech frames and discard the non-speech ones. The main reason for using the speech references, instead of speech activity detector (SAD), is that we are motivated to focus exclusively on speaker errors that occur to the diarization approach. But we have carried out a few set of experiments with SAD to compare the results of the systems with and without using SAD.

The short-term cepstral coefficients are computed within a 30-ms frame window at a 10-ms shift. The dimension of the cepstral coefficients is 20. The voice-quality, prosodic, and GNE features are extracted over a 30-ms frame length and at a 10-ms shift using Praat software [43]. Then, each of the voice-quality, prosodic, and GNE features is estimated over a 500-ms window with a 10-ms shift. This is done to smooth out the feature estimation of the unvoiced frames. It is also done to synchronize them with the cepstral coefficients.

The UBM and the T matrix are trained using 100 AMI shows which have a duration of 60 h. Three universal background models (UBMs) of 512 Gaussian components have been trained. While the first UBM is trained using only the static cepstral coefficients, the second UBM is trained using both the static cepstral coefficients and delta dynamic features. The third UBM is trained using long-term features (i.e., voice-quality, pitch, intensity, formants, and GNE).

One hundred- and 50-dimensional raw i-vectors are extracted from the short- and long-term speech features, respectively. The sizes of the total variability matrix are 100 and 50 for the short- and long-term speech features, respectively. The i-vector extraction is carried out using ALIZE open-source toolkit [44].

The PLDA of the short-term and long-term speech features is trained on 40 and 20 dimensional speaker spaces, respectively. The PLDA is trained on the same data used to train the UBM and T matrix. The i-vectors are length normalized before the PLDA training.

The selection of threshold value for stopping criterion for the proposed i-vector-based speaker diarization systems is carried out as it is shown in Fig. 6. It is based on a data-driven approach. The DER and corresponding cosine-distance/PLDA score values at each iteration are compared, and λ value that minimizes the DER is selected. Thus, the system stops merging when the highest cosine-distance/PLDA score value among all pair of clusters is less than λ. As it is shown in Fig. 6, the DER values decrease for some iterations. However, its values start to increase after some number of iterations because of over-clustering.

Note that optimum parameters found through experimentation on the development set are directly applied on the test set.

The conventional 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 are used, the rate of false alarms and missed speech have zero values in the experimental results. Hence, DER values reported in the following sections correspond purely to speaker time confusion produced by the diarization system. We have used a collar of 250 ms around every speaker segment to discard any inaccuracies in the reference annotation.[1]

4.2 Experimental results

As it is shown in Table 1, the baseline system of the test set has a DER of 23.97%. Note that the baseline system is based on BIC clustering and static MFCC feature set both for segmentation and clustering. The table shows that the addition of long-term features (i.e., voice-quality, pitch, intensity, formants, GNE, and delta) to the static MFCC features improves the DER both for GMM- and i-vector-based speaker diarization systems.

For the GMM-based speaker diarization system, the table depicts that the fusion of the concatenated or individual long-term speech features with the cepstral coefficients provides better DER compared to using only the later feature set. For example, the use of formants together with MFCC provides a DER of 21.11%. This corresponds to a 11.93% relative DER improvement compared to the same system using static MFCC feature set. Similarly, the table shows that the fusion of voice-quality features together with the cepstral coefficients provides a 4.76% relative DER improvement compared to the system based only on cepstral coefficients.

Furthermore, the use of the concatenated long-term features (i.e., jitter, shimmer, pitch, intensity, and formants) together with the static cepstral coefficients improves the performance of both i-vector-based cosine-distance and PLDA clustering techniques. The table reports that the use of these concatenated long-term feature sets with the MFCC ones provides a DER of 20.13% and 20.03% for i-vector-based cosine-distance and PLDA clustering techniques, respectively. These represent a 12.33% and 8.75% relative DER improvement compared to the system based on GMM-based BIC clustering technique and using the same feature sets.

The table also shows that the use of delta dynamic features only in clustering improves the DER compared to using the static cepstral coefficients both in segmentation and clustering. The DER improvements are both for GMM- and i-vector-based speaker diarization systems.

Hence, the use of delta dynamic features in GMM-based BIC clustering reduces the DER to 21.57%. This represents a 10.01% relative DER improvement

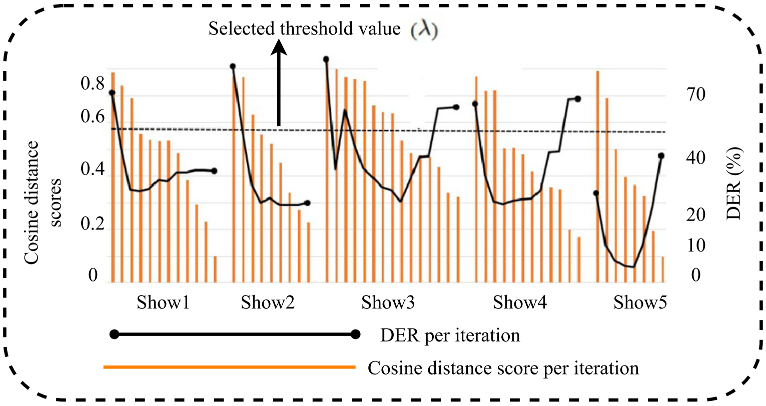

Fig. 6 DER and cosine-distance score per iteration for selected shows from the development set

Table 1 DER of the test set using the proposed long-term speech features and proposed speaker diarization systems

Features for segmentation	Features for clustering	Speaker diarization system		
		GMM/BIC	i-Vector/ CD	i-Vector/ PLDA
MFCC	MFCC	23.97	22.96	21.05
MFCC	MFCC + delta	21.57	19.34	19.47
MFCC + JS	MFCC + JS	22.83	–	–
MFCC + formants	MFCC + formants	21.11	20.26	20.71
MFCC + (formants + pitch + intensity)	MFCC + (formants + pitch + intensity)	23.45	–	–
MFCC + (JS + formants + pitch + intensity)	MFCC + (JS + formants + pitch + intensity)	21.68	20.13	20.03
MFCC + (JS + formants + pitch + intensity + GNE)	MFCC + (JS + formants + pitch + intensity + GNE)	21.91	20.44	19.46
MFCC + (JS + formants + pitch + intensity)	MFCC + delta + (JS + formants + pitch + intensity)	21.76	18.2	19.37
MFCC + (JS + formants + GNE)	MFCC + delta + (JS + formants + GNE)	21.52	18.87	19.2
MFCC + (JS + formants + pitch + intensity + GNE)	MFCC + delta + (JS + formants + pitch + intensity + GNE)	22.68	18.68	18.95

JS jitter and shimmer, *CD* cosine-distance

compared to the baseline system. Similarly, the use of delta dynamic features in clustering improves the DER both for i-vector-based cosine-distance and PLDA clustering techniques. It provides a DER of 19.34% and 19.47%. These results correspond to a 15.77% and 11.3% relative DER improvement compared to the system using the same clustering technique and apply static cepstral coefficients both for segmentation and clustering.

Note that the delta features are used only in speaker clustering. The motivation for using deltas only in clustering is because they capture the transitional characteristics of the speech signal which contains speaker-specific information. Since they don't have temporal resolution, they are not used in speaker segmentation. In our work, i-vectors are also applied only on the clustering stage. They are not applied in speaker segmentation since it is difficult to reliably estimate i-vectors from segments of short duration [21].

We have carried out some experiments on subset of AMI shows to explore the impact of applying deltas and i-vectors both in segmentation and clustering. The experimental results show that the use of delta and i-vectors in segmentation degrades the performance of speaker diarization system (i.e., the DER values are high).

Table 1 reveals that the best results are found when i-vector-based cosine-distance clustering is used together with the short- and long-term speech features (i.e., MFCC, delta, voice-quality, pitch, intensity, and formants). It provides a DER of 18.2%. This provides a 24.07% relative DER improvement compared to the baseline system. It also corresponds to a 9.59% relative DER improvement compared to the system that uses the same clustering technique and uses static MFCC, voice-quality, pitch, intensity, and formant feature sets. It also represents a 5.89% relative DER improvement compared to the system that uses the same clustering technique and uses only the static and delta dynamic features.

We have also carried out experiments using speech activity detector (SAD) for the baseline (i.e., GMM-based speaker diarization system using only static MFCC feature set) and the best performing system (i-vector-based cosine-distance system using MFCC, delta, jitter, shimmer, pitch, intensity, and formants). Table 2 shows that the use of SAD on these experiments also exhibit the use of long-term features improves the performance of speaker diarization system both for the GMM- and i-vector-based speaker diarization systems. While the DER of the baseline system using SAD is 48.91% (miss speech = 19.7%, false alarm = 11.5%, and speaker error = 17.7%), the DER of the best performing system is 44.2% (miss speech = 19.7%, false alarm = 11.5%, and speaker error = 13%). The best performing system provides a 9.63% relative DER improvement compared to the baseline system. If we compare only the speaker errors of the

Table 2 Speech activity detector (SAD) experiments for the baseline and the best performing system

Features	Speaker diarization system	
	GMM/BIC	i-Vector/CD
MFCC	48.91	–
MFCC + delta + (JS + formants + pitch + intensity)	–	44.2

two systems, the best performing system provides a 26.55% relative speaker error improvement compared to the baseline system.

Since we have used a simple SAD based on energy thresholding and the AMI database corresponds to a complex scenario (i.e., there are reverberations, background noises, echos), the percentage of miss speech is relatively high.

5 Discussion

The box plots in Fig. 7 depict the DER distribution of the different recordings for the test set. It 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. As it is also shown in Fig. 6, the DER values increase with the number of iterations because of over-clustering.

Although the use of long-term features in GMM- and i-vector-based systems reduces the DER for most of the recordings, the DER values increase for few of them compared to the baseline system. One of the possible reasons is the threshold value used for the stopping criterion. In this work, we have proposed the same threshold value for all shows. It is also worth to investigate the impact of an automatic threshold value that varies per iterations and recordings in the proposed systems.

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 increases the robustness and reliability of speaker diarization systems.

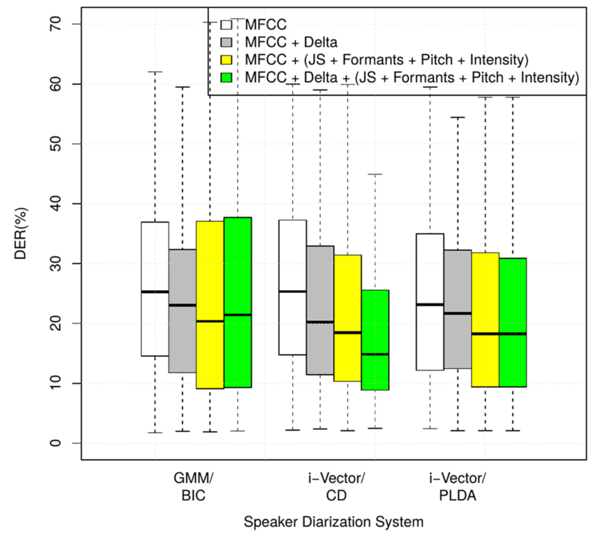

Fig. 7 Box plots of the test shows. It shows the DER variations of different shows for the baseline and proposed systems

6 Conclusions

This work has proposed the use of voice-quality features for GMM- and i-vector-based speaker diarization systems. The proposed voice-quality features are used together with the short-term cepstral and other long-term features (i.e., pitch, intensity, formants, and GNE).

The work has also analyzed the use of delta dynamic features for speaker clustering since they capture the transitional characteristics of the speech signal which contains speaker-specific information. The delta dynamic features are not used in speaker segmentation because of their limited temporal resolution in segmentation. But, they are used in speaker clustering since they capture the transitional characteristics of the speech signal which contains speaker-specific information. The delta features are used together with the short-term static cepstral coefficients and other long-term speech features (i.e., voice-quality, pitch, intensity, formants, and GNE) for GMM- and i-vector-based speaker clustering techniques.

The experimental results show that the use of voice-quality features together with the other long-term and short-term spectral features improves the performance of both GMM- and i-vector-based speaker diarization systems. The experimental results also show that i-vector clustering techniques based on short- and long-term features provides better DER compared to the same clustering technique using only short-term features. Moreover, the results show that i-vector-based cosine-distance and PLDA clustering techniques provide a substantial relative DER improvement compared to GMM-based BIC clustering. Finally, the experimental results manifest that the use of delta dynamic features in clustering improves the DER both for GMM- and i-vector-based speaker diarization systems.

The results of our work manifest the usefulness of long-term speech features both for GMM- and i-vector-based speaker diarization systems.

7 Endnotes

[1] The scoring tool is the NIST RT scoring used as ./md-eval-v21.pl -1 -nafc -c 0.25 -o -R reference.rttm-S hypothesis.rttm

Acknowledgements

This work has also received funding from the EU's Horizon 2020 research and innovation program under grant agreement No 645323. The text reflects only the author's view and the Commission is not responsible for any use that may be made of the information it contains.

Author's contributions

All authors are responsible for proposing the algorithm and the manuscript. All authors have read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

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

Received: 16 November 2017 Accepted: 10 September 2018
Published online: 26 September 2018

References

1. Meignier, S., Moraru, D., Fredouille, C., Bonastre, J.-F., & Besacier, L. (2006). Step-by-step and integrated approaches in broadcast news speaker diarization. *Comput. Speech Lang., 20*(2), 303–330.
2. Anguera, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., & Vinyals, O. (2011). Speaker diarization: a review of recent research. *IEEE Trans. Audio Speech Lang. Process.*
3. Friedland, G., Vinyals, O., Huang, Y., & Muller, C. (2009). Prosodic and other long-term features for speaker diarization. *IEEE Trans. Audio Speech Lang. Process., 17*(5), 985–993.
4. Zelenák, M., & Hernando, J. (2011). The detection of overlapping speech with prosodic features for speaker diarization. In *INTERSPEECH* (pp. 1041–1044).
5. Kreiman, J., & Gerratt, B. R. (2005). Perception of aperiodicity in pathological voice. *J Acoust Soc Am, 117*(4).
6. Li, X., Tao, J., Johnson, M. T., Soltis, J., Savage, A., Leong, K. M., & Newman, J. D. (2007). Stress and emotion classification using jitter and shimmer features. In *IEEE International Conference on Acoustics, Speech and Signal Processing*.
7. Sadeghi N., A. Homayounpour, M.M.: 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).
8. Farru's, M., Hernando, J., & Ejarque, P. (2007). Jitter and shimmer measurements for speaker recognition. In *INTERSPEECH*.
9. Woubie, A., Luque, J., & Hernando, J. (2014). Jitter and shimmer measurements for speaker diarization. In *VII Jornadas en Tecnolog´ıa del Habla and III Iberian SLTech Workshop* (pp. 21–30).
10. Woubie, A., Luque, J., & Hernando, J. (2015). Using voice-quality measurements with prosodic and spectral features for speaker diarization. In *INTERSPEECH*.
11. Kenny, P., Reynolds, D., & Castaldo, F. (2010). Diarization of telephone conversations using factor analysis. IEEE Journal of Selected Topics in Signal Processing, 4(6), 1059.
12. Franco-Pedroso, J., Lopez-Moreno, I., Toledano, D. T., & González-Rodríguez, J. (2010). ATVS-UAM system description for the audio segmentation and speaker diarization Albayzin 2010 evaluation. In *FALA*.
13. Shum, S., Dehak, N., Chuangsuwanich, E., Reynolds, D. A., & Glass, J. R. (2011). Exploiting intra-conversation variability for speaker diarization. In *Proceedings of Interspeech, Florence, Italy*.
14. Shum, S., Dehak, N., & Glass, J. (2012). On the use of spectral and iterative methods for speaker diarization. In *Interspeech*.
15. Vaquero Avilés-Casco, C. (2011). *Robust diarization for speaker characterization (diarización robusta para caracterización de locutores)*. University of Zaragoza, Zaragoza: PhD thesis.
16. Senoussaoui, M., Kenny, P., Dumouchel, P., & Stafylakis, T. (2013). Efficient iterative mean shift based cosine dissimilarity for multi-recording speaker clustering. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.
17. Villalba, J., Ortega, A., Miguel, A., & Lleida, E. (2015). Variational Bayesian PLDA for speaker diarization in the MGB challenge. In *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop On* (pp. 667–674) IEEE.
18. Silovsky, J., & Prazak, J. (2012). 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)*.
19. Prazak, J., & Silovsky, J. (2011). Speaker diarization using PLDA-based speaker clustering. In *6th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems (IDAACS), vol. 1*.
20. Sell, G., & Garcia-Romero, D. (2014). Speaker diarization with PLDA i-vector scoring and unsupervised calibration. In *Proceedings of SLT*.
21. Woubie, A., Luque, J., & Hernando, J. (2016). Short- and long-term speech features for hybrid HMM-i-vector based speaker diarization system. In *Odyssey 2016-The Speaker and Language Recognition Workshop* Paper Accepted.
22. Woubie, A., Luque, J., & Hernando, J. (2016). Improving i-vector and PLDA based speaker clustering with long-term features. In *INTERSPEECH* (pp. 372–376).
23. Furui, S. (1981). Cepstral analysis technique for automatic speaker verification. *IEEE Trans. Acoust. Speech Signal Process., 29*(2), 254–272.
24. Kumar, K., Kim, C., & Stern, R. M. (2011). Delta-spectral cepstral coefficients for robust speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference On* (pp. 4784–4787) IEEE.
25. Luque, J. (2012). *Speaker diarization and tracking in multiple-sensor environments*. Barcelona: PhD thesis, Universitat Polit'ecnica de Catalunya.
26. Yella, S. H. (2015). *Speaker diarization of spontaneous meeting room conversations*. Switzerland: PhD thesis, EPFL, Lausanne.
27. The Augmented Multi-party Interaction Project, AMI Meeting Corpus (2011). Website, http://corpus.amiproject.org.
28. Farrús, M., Garde, A., Ejarque, P., Luque, J., Hernando, J. (2006). On the fusion of prosody, voice spectrum and face features for multimodal person verification. In *9th International Conference on Spoken Language Processing, ICSLP, pp. 2106–2109*.
29. Pardo, J.M., Anguera, X., Wooters, C.: Speaker diarization for multiple-distant-microphone meetings using several sources of information. in IEEE Transactions on Computers (2007).
30. Wang, X.-G., & Shen, H. C. (1999). Multiple hypothesis testing fusion method for multisensor systems. In *IEEE International Conference on Intelligent Robots and Systems, vol. 2*.
31. Memon, S., Lech, M., & Maddage, N. (2009). Speaker verification based on different vector quantization techniques with Gaussian mixture models. In *Third International Conference on Network and System Security* (pp. 403–408).
32. Nguyen, P.T.: Automatic speaker classification based on voice characteristics. University of Canberra, ??? (2010).
33. Boersma, P., Weenink, D.: Praat: doing phonetics by computer (2009).
34. Adami, A. G. (2007). Modeling prosodic differences for speaker recognition. *Speech Comm., 49*(4), 277–291.
35. Dellwo, V., Huckvale, M., & Ashby, M. (2007). How is individuality expressed in voice? An introduction to speech production and description for speaker classification. In Speaker Classification I (pp. 1-20). Springer, Berlin, Heidelberg.
36. Sáenz Lechón N., Osma Ruiz V., Fraile Muňoz R., Godino Llorente J.I., Gómez Vilda P.: Screening voice disorders with the glottal to noise excitation ratio (2009).
37. Michaelis, D., Gramss, T., & Strube, H. W. (1997). Glottal-to-noise excitation ratio–a new measure for describing pathological voices. *Acta Acustica united with Acustica, 83*(4), 700–706.
38. Michaelis, D., Frohlich, M., & Strube, H. W. (1998). Selection and combination of acoustic features for the description of pathologic voices. *J Acoust Soc Am, 103*(3), 1628–1639.
39. Godino-Llorente, J. I., Osma-Ruiz, V., Sáenz-Lechón, N., Gómez-Vilda, P., Blanco-Velasco, M., & Cruz-Roldán, F. (J Voice, 2010). The effectiveness of the glottal to noise excitation ratio for the screening of voice disorders*., 24*(1), 47–56.
40. Kanagasundaram, A., Vogt, R., Dean, D. B., Sridharan, S., & Mason, M. W. (2011). i-vector based speaker recognition on short utterances. In *Proceedings of the 12th Annual Conference of the International Speech Communication Association, pp. 2341–2344* International Speech Communication Association (ISCA).
41. Poddar, A., Sahidullah, M., & Saha, G. (2015). Performance comparison of speaker recognition systems in presence of duration variability. In *India Conference (INDICON), 2015 Annual IEEE* (pp. 1–6) IEEE.
42. Kenny, P. (2010). Bayesian speaker verification with heavy-tailed priors. In *Odyssey, p. 14*.
43. Boersma, P., Weenink, D.: Praat: doing phonetics by computer, Version 5.3. 69. http://www.praat.org/
44. Larcher, A., Bonastre, J. F., Fauve, B. G. B., Lee, K., L'evy, C., Li, H., Mason, J. S. D., & Parfait, J. (2013). ALIZE 3.0 - open source toolkit for state-of-the-art speaker recognition. In *INTERSPEECH* (pp. 2768–2772).