Robust Speaker Identification in a Meeting with Short Audio Segments

Giorgio Biagetti, Paolo Crippa, Laura Falaschetti, Simone Orcioni and Claudio Turchetti

Abstract The paper proposes a speaker identification scheme for a meeting scenario, that is able to answer the question "is somebody currently talking?", if yes, "who is it?". The suggested system has been designed to identify during a meeting conversation the current speaker from a set of pre-trained speaker models. Experimental results on two databases show the robustness of the approach to the overlapping phenomena and the ability of the algorithm to correctly identify a speaker with short audio segments.

Keywords Speaker identification · Meeting conversation · Speaker diarization · Overlapping speech

1 Introduction

Speaker identification aims at detecting which speaker a given pool the unknown speech is derived from, and can be considered as a particular case of the more general problem of speaker recognition which is addressed to recognize, indentify or verify individuals using speech [13]. The main tasks involved in a speaker identification system are feature extraction, speaker modeling, and speaker classification.

G. Biagetti · P. Crippa · L. Falaschetti (✉) · S. Orcioni · C. Turchetti
DII – Dipartimento di Ingegneria dell'Informazione, Università Politecnica
delle Marche, Via Brecce Bianche 12, 60131 Ancona, Italy
e-mail: l.falaschetti@univpm.it

G. Biagetti
e-mail: g.biagetti@univpm.it

P. Crippa
e-mail: p.crippa@univpm.it

S. Orcioni
e-mail: s.orcioni@univpm.it

C. Turchetti
e-mail: c.turchetti@univpm.it

© Springer International Publishing Switzerland 2016

I. Czarnowski et al. (eds.), *Intelligent Decision Technologies 2016*,
Smart Innovation, Systems and Technologies 57,
DOI 10.1007/978-3-319-39627-9_41

The usually adopted features in speaker identification are the same adopted in speech recognition, namely Mel frequency cepstral coefficients (MFCCs), perceptual linear prediction coefficients, etc. [10]. Among these, MFCCs have shown the best performance due to their particular robustness to the environment and flexibility.

As far as speaker modeling is concerned, assuming an utterance from a speaker is a random sequence of frames, the Gaussian mixture model (GMM) is widely used in speaker identification [11]. Since this model falls into the family of statistical model, it requires training data sampled from the class of speakers to be identified.

For the task of speaker classification, the optimal Bayesian classifier guarantees the minimum classification error by identifying the speaker model which exhibits the maximum GMM a posteriori probability [13].

In the classic speaker indentification scenario it is required that the identification system be able to identify a person when one speaker alone is speaking for a time interval. In the different scenario of a meeting [3, 8, 14] speech from one speaker can abruptly change to, or can be overlapped with, speech from another speaker. In particular overlapping speech can greatly degrade the performance of speaker identification. These problems are common in Speaker Diarization whose main goal is to segment audio into speaker-homogeneous regions with the goal of answering the question "who spoke when?".

However in diarization system the output is limited to labeling speaker region with number or letters, without detecting the speaker's indentity. This goal is performed without prior training of specific models, as many of such systems work completely unsupervised. The main operational tasks to be carried out in a speaker diarization system are: *speech activity detection* (to separate speech from non-speech), *segmentation* (to detect speaker changes to segment the audio data), *clustering* (to group the segmented regions together into spoken-homogeneous clusters).

The aim of this paper is to derive a robust speaker identification scheme for a meeting scenario that is able to answer the question "is somebody currently talking?", if yes, "who is it?". Thus this task is performed using results from both the field of speaker identification and speaker diarization.

The suggested system has been designed to identify during a meeting conversation the current speaker from a set of pre-trained speaker models. In particular a robust speaker identification algorithm has been adopted in order to mitigate the problem of the overlapping speech.

The paper is organized as follows. Section 2 provides a brief overview of the speaker identification algorithm. Section 3 presents the experimental results carried out on a data base properly designed to simulate a true meeting including overlapping phenomenon and on the AMI Meeting Corpus. Section 4 summarizes the conclusions of the present work.

2 Speaker Classification Algorithm

The speaker classification algorithm used in this work is based on the approach [1] used in a classic speaker identification scenario, and it is summarized in the following.

2.1 Single Frame Classification

We denote with $y[n]$, $n = 0, \ldots, N-1$, a frame representing the power spectrum of the speech signal, extracted from the time domain waveform of the utterance under consideration, through a pre-processing algorithm including pre-emphasis, framing and log-spectrum. Typical duration values for frames range from 20 to 30 ms (usually 25 ms) and a frame is generated every 10 ms (thus consecutive 25 ms frames generated every 10 ms will overlap by 15 ms).

In a Bayesian speaker identification scheme, a group of S speakers is represented by the probability density functions (pdfs)

$$p_s(\mathrm{y}) = p(\mathrm{y} \mid \theta_s), \qquad s = 1, 2, \cdots, S \tag{1}$$

where θ_s are the parameters to be estimated using the training set $\mathcal{W}$.

The objective of classification is to find the speaker model θ_s which has the maximum a posteriori probability for a given frame y belonging to the testing set $\mathcal{Z}$. Using Bayes' theorem and assuming that $p(\theta_s)$ and $p(\mathrm{y})$ are independent of S, it results:

$$\hat{s}(\mathrm{y}) = \underset{1 \le s \le S}{\operatorname{argmax}} \left\{ p(\theta_s \mid \mathrm{y}) \right\} = \underset{1 \le s \le S}{\operatorname{argmax}} \left\{ p_s(\mathrm{y}) \right\} \ . \tag{2}$$

The main issue in Bayesian classification is to accurately estimate the pdf $p_s(\mathrm{y})$. To this end the most generic statistical speaker modeling one can adopt for the single speaker is the GMM [11], is given by the equation

$$p(\mathrm{y} \mid \theta_s) = \sum_{i=1}^{F} \alpha_i \, \mathcal{N}(\mathrm{y} \mid \mu_i, \mathrm{C}_i) \tag{3}$$

where α_i, $i = 1, \ldots, F$ are the mixing weights, and $\mathcal{N}(\mathrm{y}|\mu_i, \mathrm{C}_i)$ represents a Gaussian distribution density with mean μ_i and covariance matrix C_i.

$\theta = \{\alpha_1, \mu_1, \mathrm{C}_1, \ldots, \alpha_F, \mu_F, \mathrm{C}_F\}$, (the index s is omitted for the sake of notation simplicity) is the set of unknown parameters to be estimated that specify the Gaussian mixture.

An estimate of θ, with training data $\mathcal{W}$ can be obtained by the *maximum likelihood* (ML)

$$\hat{\theta}_{\mathrm{ML}} = \underset{\theta}{\operatorname{argmax}} \{ \log p(\mathcal{W} \mid \theta) \} \tag{4}$$

however as (4) is difficult to be solved analytically since (4) contains the log of a sum, the usual choice for solving ML estimate of the mixture parameters is the expectation maximization (EM) algorithm.

The EM algorithm, which has been adopted in this work, is based on the interpretation of $\mathcal{W}$ as incomplete data and the set $\mathcal{H}$ as the missing part of the complete data $\mathcal{X} = \{\mathcal{W}, \mathcal{H}\}$. The complete data log-likelihood, i.e. the log-likelihood of $\mathcal{X}$ as though $\mathcal{H}$ was observed, is

$$\log\left[p\left(\mathcal{W}, \mathcal{H}|\theta\right)\right] = \sum_{\ell=1}^{L}\sum_{i=1}^{F} \mathrm{h}_i^{(\ell)} \log\left[\alpha_i \mathcal{N}(\mathrm{y}^{(\ell)}|\mu_i, \mathrm{C}_i)\right] \ . \tag{5}$$

In general the EM algorithm computes a sequence of parameter estimates $\{\hat{\theta}(p), p = 0, 1, \ldots\}$ by iteratively performing two steps:

- *Expectation step:* compute the expected value of the complete log-likelihood, given the training set $\mathcal{W}$ and the current parameter estimate $\hat{\theta}(p)$. The result is the so-called *auxiliary function*

$$Q\left(\theta|\hat{\theta}(p)\right) = E\left\{\log\left[p\left(\mathcal{W}, \mathcal{H}|\theta\right)\right]|\mathcal{W}, \hat{\theta}(p)\right\} \ . \tag{6}$$

- *Maximization step:* update the parameter estimate

$$\hat{\theta}(p+1) = \underset{\theta}{\operatorname{argmax}}\left\{Q\left(\theta|\hat{\theta}(p)\right)\right\} \tag{7}$$

by maximizing the Q-function.

Usually for 8 kHz (16 kHz) bandwidth speech, the vector y has a dimension $N = 128$ (256). So that a too large amount of training data would be necessary to estimate the pdf $p(\mathrm{y}|\theta)$ and, in any case, with such a dimension the estimation problem is impractical.

The usual choice is to solve this problem is to reduce the vector y to a vector k_M of lower dimension by a linear transform H such that

$$\mathrm{k}_M = \mathrm{H}\,\mathrm{y} \ , \tag{8}$$

where y is a $N \times 1$ vector, k_M an $M \times 1$ vector, H an $M \times N$ matrix, and $M \ll N$. The vector k_M represents the so-called feature-vector belonging to an appropriate M-dimension subspace [6, 12].

Principal component analysis (PCA) [7] has proven to be an excellent technique for dimensionality reduction in many application areas including data compression, image analysis, visualization to mention just a few. The main property of PCA [5], is that for a set of observed N-dimensional data vectors $y[n]$, $n = 0, \ldots, N-1$, M *principal axes* ϕ_j, $j = 1, \ldots, M$, can be derived such that they are orthonormal axes onto which the retained *variance* under projection is maximal.

The PCA of y is derived from the Karhunen-Love transform (KLT), defined by the couple of equations

$$\mathrm{y} = \boldsymbol{\Phi}\ \mathrm{k}\ , \tag{9}$$

$$\mathrm{k} = \boldsymbol{\Phi}^T \mathrm{y}\ , \tag{10}$$

where $\boldsymbol{\Phi} = \left[\phi_1, \ldots, \phi_N\right]$ is an $N \times M$ matrix and $\mathrm{k} = \left[k_1, \ldots, k_N\right]^T$ is the transformed random vector.

The M principal axes are identified as those corresponding to the M maximal eigenvalues $\lambda_j, j = 1, \ldots M$ of $\mathrm{R}_{\mathrm{yy}}\phi_j = \lambda_j\phi_j\ , j = 1, \ldots, N$, where R_{yy} is the autocorrelation function. Thus $\boldsymbol{\Phi}$ decomposes as $\boldsymbol{\Phi} = [\boldsymbol{\Phi}_M\ ,\ \boldsymbol{\Phi}_\eta]$, and (9) can be rewritten as:

$$\mathrm{y} = \boldsymbol{\Phi}\ \mathrm{k} = \boldsymbol{\Phi}_M\ \mathrm{k}_M + \boldsymbol{\Phi}_\eta\ \mathrm{k}_\eta = \mathrm{x}_M + \eta_{\mathrm{y}}\ , \tag{11}$$

being $\Phi_M = \left[\phi_1, \ldots, \phi_M\right]$ an $N \times M$ matrix, k_M an $M \times 1$ vector. In a similar way (10) becomes:

$$\begin{bmatrix} \mathrm{k}_M \\ \mathrm{k}_\eta \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Phi}_M^T \\ \boldsymbol{\Phi}_\eta^T \end{bmatrix}\ \mathrm{y}\ . \tag{12}$$

In (11) the term

$$\mathrm{x}_M = \boldsymbol{\Phi}_M\ \mathrm{k}_M\ , \tag{13}$$

represents the truncated expansion, and it is equivalent to the approximations

$$\mathrm{y} \approx \mathrm{x}_M, \qquad \mathrm{k} \approx \mathrm{k}_T = \begin{pmatrix} \mathrm{k}_M \\ 0 \end{pmatrix}\ , \tag{14}$$

Thus, as k_M is given by $\mathrm{k}_M = \boldsymbol{\Phi}_M^T\ \mathrm{y}$, comparing with (8) yields $\mathrm{H} = \boldsymbol{\Phi}_M^T$.

On the basis of previous results a Bayesian classification scheme which is consistent with PCA can be derived.

Given a group of S speakers, let us define the pdfs $p_s(\mathrm{k}_T) = p(\mathrm{k}_T \mid \theta_s)$, $s = 1, 2, \ldots, S$, where k_T is the truncation of k. Consequently the pdf $p_s(\mathrm{k}_T) = p_s(\mathrm{k}_M)\ \delta(\mathrm{k}_\eta)$, represents an approximation of the pdf in (1). Thus (2) becomes:

$$\hat{\mathrm{s}}(\mathrm{y}) = \underset{1 \le s \le S}{\operatorname{argmax}} \left\{p_s(\mathrm{k}_M)\ \delta(\mathrm{k}_\eta)\right\} = \underset{1 \le s \le S}{\operatorname{argmax}} \left\{p_s(\mathrm{k}_M)\right\}\ . \tag{15}$$

As you can see comparing (15) with (2), the dimensionality of classification problem is reduced from N to M, with $M < N$.

2.2 Multi Frame Classification

The accuracy of speaker identification can be considerably improved using a sequence of frames instead of a single frame alone. To this end let us refer to a sequence of V frames defined as $Y = \{\mathrm{y}^{(1)}, \ldots, \mathrm{y}^{(V)}\}$, where $\mathrm{y}^{(v)}$ represents the vth frame. Using (15) we can determine the class each frame $\mathrm{y}^{(v)}$ belongs to. Thus the S sets $\mathcal{Z}_s = \left\{\mathrm{y}^{(v)} \mid \mathrm{y}^{(v)} \text{ belongs to class } s\right\}, s = 1, \ldots, S$, are univocally determined.

Given Y, we define the score for each class s as

$$r_s(Y) = \sum_{\mathrm{y}^{(v)} \in \mathcal{Z}_s} p(\mathrm{y}^{(v)}) \tag{16}$$

where $p(\mathrm{y}^{(v)})$ represents the probability achieved by the frame $\mathrm{y}^{(v)}$.

Finally the multi-frame speaker identification is based on:

$$\hat{s}(Y) = \underset{1 \le s \le S}{\operatorname{argmax}} \left\{ r_s(Y) \right\}. \tag{17}$$

3 Experimental Results

Experiments are conducted using two different corpora, *(i)* a data base called DBT that was specifically designed to subject the algorithm to a severe test, where a large percentage of overlapping speech and different consistency of framing material is considered, *(ii)* the well known AMI Meeting Corpus, as it represents a widely accepted test for the evaluation of speaker diarization system.

3.1 Features Extractor

Figure 1 shows the block diagram of the proposed front-end employed for feature extraction. At the input of the processing chain a voice activity detection block drops all non speech segments from the input audio records, exploiting the energy

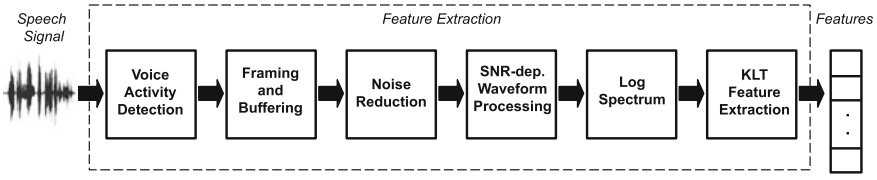

Fig. 1 The proposed front-end for feature extraction

acceleration associated with voice onset. The signal is then divided into overlapping frames of 25 ms (200 samples), with a frame shift of 10 ms (80 samples). Hence buffering is required for storing overlapping regions among frames. Besides, before computing the DKLT features, each frame is cleaned up by a noise reduction block based on the Wiener filter. Further enhancements are then performed by a SNR-dependent waveform processing phase, that weights the input noise-reduced frame according to the positions of its smoothed instant energy contour maxima. It is worth noting that noise reduction introduces an overall latency of 30 ms (3 frames) due to its algorithm requiring internal buffering.

3.2 Experiments on Data Base DBT

A first set of experiments was carried out on a large database, called DBT, which was formed by collecting several audio recordings of five different speakers, two females (A, B) and three males (C, D, E) as reported in Table 1. All recordings extracted are mono, 8 kilosamples per second, 16 bit. The consistency of DBT database in terms of number of frames used for each speaker is reported in Table 1. In order to test several different models, the databases DB1, DB2, and DB3, with different percentage consistency of training subsets, have been derived.

A meeting scenario has beeen simulated by interleaving 45 audio segments extracted from database *liber liber* (http://www.liberliber.it/) to achieve a 20 min audio track. In the conversation the 5 speakers alternate each other with short turn durations. More specifically two audio tracks have been derived: in the former the audio segments follow one another without overlap, in the latter an overlapping of 20 % is taken into account to test the robustness of the algorithm to the overlapping phenomenon.

The widely adopted metric for diarization performance measurement is the Diarization Error Rate (DER). It has been introduced by the NIST in 2000 within the Speaker Recognition evaluation [9] for their speaker segmentation task [4]. The

Table 1 Recordings used for the creation of the identification corpus

Database			DBT	80 % DBT	50 % DBT	20 % DBT
Speaker	Gender	Duration (s)	Model 1	Model 2	Model 3	Model 4
A	F	761	58903	47122	29451	11780
B	F	2593	195591	156472	97795	39118
C	M	251	18867	1509	9431	3773
D	M	838	63713	50970	31856	12742
E	M	1162	91253	73002	45626	18250
Total		5605	428327	342659	214161	85663

Source liber liber (http://www.liberliber.it/). The material was used for training purposes. The consistency of the databases used for modeling is shown in terms of number of frames

DER is defined as the ratio of incorrectly detected speaker time to total speaker time. The metric is computed by mapping the system output speaker segment sets to reference speaker segment sets so as to minimize the total error. By defining the following errors:

- *Speaker assignment errors* (E_{spkr}): percentage of scored time that a speaker ID is assigned to the wrong speaker.
- *Missed detections* (E_{miss}): percentage of scored time that a hypothesized non-speech segment corresponds to a reference speaker segment.
- *False alarm detections* (E_{fa}): percentage of scored time that a hypothesized speaker is labelled as a non-speech in the reference.

the final DER is given by

$$DER = E_{spkr} + E_{miss} + E_{fa} \tag{18}$$

Figures 2 and 3 report for speech without and with overlapping respectively, the value of the DER as a function of the number of frames used in the identification algorithm and for the four models previously defined. As you can see the DER drastically decreses as the sequence length of frame increases, while only minor differences are due to the various models. It is worth to notice that due to robustness of the algorithm, the performance are not appreciably degraded in a meeting conversation with overlapping.

Tables 2 and 3 report the values of the three parameters (E_{spkr}, E_{miss}, E_{fa}) that contribute to the *DER*.

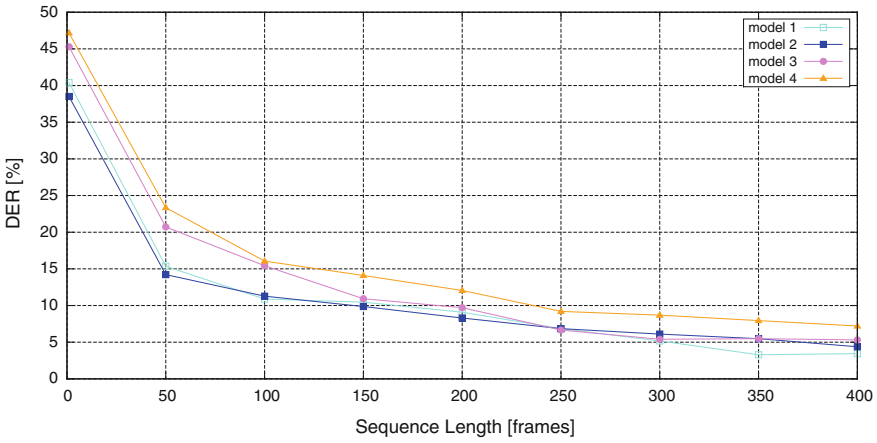

Fig. 2 DER in a meeting conversation without overlapping speech as a function of the sequence length, for different training models

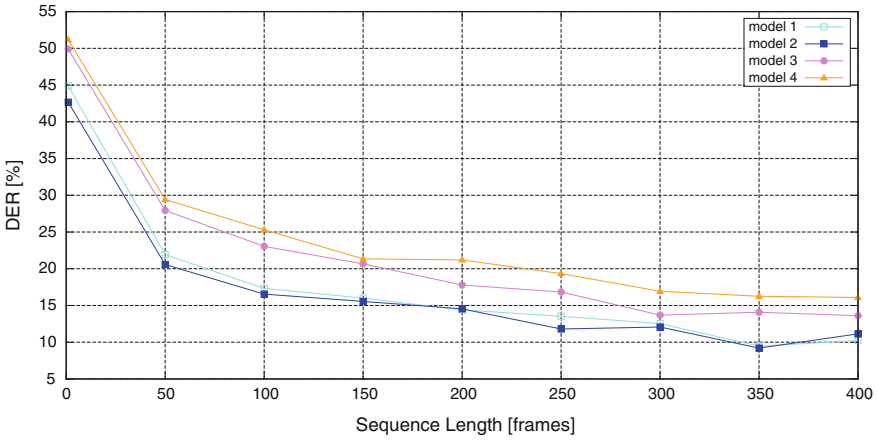

Fig. 3 DER in a meeting conversation with 20 % of overlapping speech as a function of the sequence length, for different training models

3.3 Experiments on AMI Meeting Corpus

A second set of experiments evaluating the speaker identification system were performed using meeting audio data from the *AMI Meeting Corpus* (http://www.idiap.ch/dataset/ami/). AMI is a large, multi-site and multi-disciplinary project with the aim of developing meeting browsing technologies that improve work group effectiveness. As part of the development process, the project is collecting a corpus of 100 h of meetings using instrumentation that yields high quality, synchronized multi-modal recording, with, for technical reasons, a focus on groups of four people [2].

Experiments are conducted with a subset of 20 meetings of the AMI Corpus, belonging to the IDIAP subset ('IS' meetings) of the corpus. This subset comprises 38 meetings, each involving four participants engaged in a scenario-based meeting ranging in duration from 13 to 40 min. The meetings contain approximately 18 % overlapping speech. The AMI meetings are a convenient choice since the 20 meetings are split into five different sessions, each one containing four meetings with the same four participants. One meetings is randomly chosen to train the four speaker models. The classification with a window length variable from 1 to 4 s is used on the rest of the data to perform diarization. The amount of speech used (per speaker) to train the models are shown in Table 4. We also train an additional 60 s room-specific non-speech model. Surprisingly, with only 60 s of speech per speaker the system is able to obtain good performance.

Table 2 Speaker identification performance for different training models as function of the sequence length

Frames	E_{spkr} (%)	E_{miss} (%)	E_{fa} (%)	DER (%)
Model 1				
400	2.19	1.25	0.00	3.44
350	1.64	1.09	0.54	3.29
300	2.82	2.11	0.23	5.17
250	2.15	4.11	0.58	6.85
200	3.13	5.01	0.94	9.09
150	3.52	5.05	1.88	10.46
100	4.31	3.91	2.66	10.89
50	7.48	2.93	4.93	15.36
1	25.18	4.31	10.90	40.40
Model 2				
400	2.19	1.88	0.31	4.38
350	2.19	2.74	0.54	5.48
300	2.35	2.82	0.94	6.11
250	1.95	4.31	0.58	6.85
200	2.35	4.85	1.09	8.30
150	2.82	4.58	2.46	9.87
100	4.15	3.36	3.76	11.28
50	6.26	2.82	5.13	14.22
1	23.15	4.10	11.25	38.51
Model 3				
400	3.13	1.56	0.62	5.32
350	3.01	1.92	0.54	5.48
300	3.05	1.64	0.70	5.40
250	2.54	3.52	0.58	6.66
200	4.23	3.60	1.88	9.71
150	4.34	3.40	3.17	10.93
100	6.11	3.44	5.87	15.43
50	10.30	2.50	7.9	20.72
1	27.76	3.89	13.63	45.30
Model 4				
400	1.88	4.70	0.62	7.21
350	3.29	4.38	0.27	7.95
300	3.52	3.99	1.17	8.69
250	3.91	3.91	1.37	9.20
200	5.01	4.85	2.19	12.06
150	4.58	5.64	3.87	14.10
100	6.58	4.31	5.17	16.06
50	11.48	3.60	8.26	23.35
1	28.74	4.67	13.76	47.18

Table 3 Speaker identification performance for different training models as function of the sequence length

Frames	E_{spkr} (%)	E_{miss} (%)	E_{fa} (%)	DER (%)
Model 1				
400	8.66	1.23	0.30	10.21
350	8.39	1.08	0.00	9.47
300	9.98	2.08	0.46	12.53
250	9.86	3.09	0.58	13.54
200	8.97	4.02	1.39	14.39
150	10.56	3.36	2.08	16.01
100	10.60	3.17	3.55	17.33
50	13.69	2.39	5.84	21.93
1	29.18	4.06	11.64	44.89
Model 2				
400	8.66	2.16	0.30	11.14
350	7.85	1.35	0.00	9.20
300	7.89	3.71	0.46	12.07
250	7.93	3.28	0.58	11.80
200	8.82	4.79	0.92	14.54
150	9.05	4.52	1.97	15.55
100	10.44	2.94	3.17	16.55
50	12.76	2.66	5.14	20.58
1	27.26	4.03	11.38	42.68
Model 3				
400	12.38	0.61	0.61	13.61
350	12.18	0.81	1.08	14.08
300	10.67	2.08	0.92	13.69
250	13.54	2.12	1.16	16.83
200	12.69	2.32	2.78	17.79
150	14.39	2.20	4.06	20.66
100	14.54	1.70	6.80	23.05
50	16.67	1.66	9.59	27.93
1	31.57	3.37	14.96	49.91
Model 4				
400	11.45	3.71	0.92	16.09
350	11.91	3.25	1.08	16.25
300	12.07	4.17	0.69	16.94
250	14.12	3.86	1.35	19.34
200	13.92	4.95	2.32	21.20
150	14.16	3.59	3.59	21.35
100	14.77	4.02	6.50	25.30
50	17.79	2.47	9.16	29.44
1	32.29	4.51	14.37	51.18

The meeting data contain overlapping speech

Table 4 DER for IDIAP AMI Corpus using small training model and short sequences of speech frame

Frames	Model		
	30 (s)	60 (s)	90 (s)
	DER (%)		
400	12.26	11.90	12.98
300	14.87	12.71	14.33
200	18.93	15.87	16.59
100	28.76	25.96	24.61

4 Conclusion

The paper describes a speaker identification scheme that is able to identify a speaker in a meeting, that is when speech from one speaker can abruptly change to, or can be overlapped with, speech from another speaker. Although these problems are common in speaker diarization, in such a case the output is limited to labeling speaker region with number or letters, without detecting the speaker's identity.

Experiments conducted on two distinct database have shown the robustness of the approach in a meeting scenario.

References

1. Biagetti, G., Crippa, P., Curzi, A., Orcioni, S., Turchetti, C.: Speaker identification with short sequences of speech frames. In: Proceedings of the International Conference on Pattern Recognition Applications and Methods, pp. 178–185 (2015)
2. Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., et al.: The AMI meeting corpus: a pre-announcement. Springer, Berlin (2005)
3. Friedland, G., Vinyals, O.: Live speaker identification in conversations. In: Proceedings of the 16th ACM International Conference on Multimedia, pp. 1017–1018. ACM (2008)
4. Galibert, O.: Methodologies for the evaluation of speaker diarization and automatic speech recognition in the presence of overlapping speech. In: Proceedings of INTERSPEECH, pp. 1131–1134 (2013)
5. Hotelling, H.: Analysis of a complex of statistical variables into principal components. J. Educ. Psychol. **24**(6), 417–441 (1933)
6. Jiang, X.: Linear subspace learning-based dimensionality reduction. IEEE Sig. Process. Mag. **28**(2), 16–26 (2011)
7. Jolliffe, I.T.: Principal Component Analysis. Springer Series in Statistics. Springer, New York (1986)
8. Luque, J., Hernando, J.: Robust speaker identification for meetings: UPC CLEAR'07 meeting room evaluation system. In: Multimodal Technologies for Perception of Humans, pp. 266–275. Springer (2008)
9. NIST: 2000 speaker recognition evaluation—evaluation plan. (2000). http://www.itl.nist.gov/iad/mig/tests/spk/2000/spk-2000-plan-v1.0.htm

10. Reynolds, D.A.: Experimental evaluation of features for robust speaker identification. IEEE Trans. Speech Audio Process. **2**(4), 639–643 (1994)
11. Reynolds, D.A., Rose, R.C.: Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Trans. Speech Audio Process. **3**(1), 72–83 (1995)
12. Singh-Miller, N., Collins, M., Hazen, T.J.: Dimensionality reduction for speech recognition using neighborhood components analysis. In: Proceedings of INTERSPEECH, pp. 1158–1161 (2007)
13. Togneri, R., Pullella, D.: An overview of speaker identification: accuracy and robustness issues. IEEE Circ. Syst. Mag. **11**(2), 23–61 (2011)
14. Yella, S.H., Bourlard, H.: Overlapping speech detection using long-term conversational features for speaker diarization in meeting room conversations. IEEE/ACM Trans. Audio Speech Lang. Process. **22**(12), 1688–1700 (2014)