Robust Statistical Processing of TDOA Estimates for Distant Speaker Diarization

Pablo Peso Parada*, Dushyant Sharma*, Toon van Waterschoot† and Patrick A. Naylor‡*
*Nuance Communication Inc, Burlington, USA. Email: dushyant.sharma@nuance.com
† ESAT-STADIUS, KU Leuven, Belgium. Email: Toon.vanwaterschoot@esat.kuleuven.be
‡ Imperial College London, UK. Email: p.naylor@imperial.ac.uk

***Abstract*—Speaker diarization systems aim to segment an audio signal into homogeneous sections with only one active speaker and answer the question "who spoke when?" We present a novel approach to speaker diarization exploiting spatial information through robust statistical modeling of Time Difference of Arrival (TDOA) estimates obtained using pairs of microphones. The TDOAs are modeled with Gaussian Mixture Models (GMM) trained in a robust manner with the expectation-conditional maximization algorithm and minorization-maximization approach. In situations of multiple microphone deployment, our method allows for the selection of the best microphone pair as part of the modeling and supports ad-hoc microphone placement. Such information can be useful for subsequent speech processing algorithms. We show that our method, which uses only spatial information, achieves up to 36.1% relative reduction in speaker error time compared to an open source toolkit using TDOA features and tested on the NIST RT05 multiparty meeting database.**

I. Introduction

Speaker diarization systems have gained much importance over the past five years in overcoming key challenges faced by automatic meeting transcription systems. These systems aim at segmenting the audio signal into homogeneous sections with only one active speaker and answer the question "who spoke when?". Speaker diarization provides important information for many speech processing applications and can be used to improve the performance of Automatic Speech Recognition (ASR) systems by allowing effective speaker acoustic model adaptation. In this paper the term diarization refers to the process of identifying fragments of audio which correspond to the same speaker regardless of the speaker's identity and we concentrate on the meeting scenario.

Although spatial information can be estimated from single-channel recordings for diarization [1], current state-of-the-art algorithms can only utilize spatial information when multi-microphone recordings are available. In the multi-microphone case this information is usually in one of two forms: (a) TDOA [2] which represents the time delay of the same signal in two different microphones, or (b) based on the steered response power method [3] that seeks the location where a beamformer created with all microphones provides the maximum power output. In single-microphone scenarios this feature is infeasible to compute and therefore common speech features like Mel-Frequency Cepstral Coefficients (MFCC) or Perceptual Linear Prediction (PLP) are typically used to perform diarization.

State-of-the-art diarization approaches [4] fall into two main categories: bottom-up and top-down. The former is initialized for the entire audio input with many clusters (typically more than the expected number of speakers), where a cluster refers to a collection of features corresponding to temporal segments of the speech signal, which are then merged successively until a stopping criteria is reached. The latter starts with only one cluster and adds new clusters until a stop criteria is achieved. The aim of this clustering is to group all the features of one speaker in one cluster. Feature extraction, cluster initialization, split/merging procedure or stop criterion are important issues in speaker diarization systems for which various solutions have been proposed in the literature [4][5]. Single-channel speaker diarization algorithms generally discriminate different speakers using speech dependent features such as MFCC or PLP coefficients [6] commonly extracted from data captured by close talking microphones [7]. In recent years, Log Mel-filter banks are employed in DNN-based systems [8] or i-vector features that are widely used in speaker recognition [9]. When multi-channel signals are available, TDOA estimates are frequently used to perform diarization commonly determined using the Generalized Cross Correlation with Phase Transform (GCC-PHAT) [10]. In [2], a framework to combine these TDOAs with MFCCs is proposed based on information theory. In [11] the diarization is performed using the TDOAs obtained from all possible combinations of microphones. An unsupervised discriminant analysis method with a Linear Discriminant Analysis (LDA)-like formulation, without the need of speaker labels, is then applied to these TDOAs to transform the input feature space into a new feature space. These new features are then used to diarize using a standard agglomerate clustering approach. The diarization system in [12] is based on estimates of the phoneme, vowel and consonant classes, which are extracted from a phoneme recognizer. Speaker change points and speaker clusters are calculated using the Bayesian Information Criterion (BIC) [13]. This criterion is computed from Gaussian models fitted to MFCC features computed from two successive speech segments, always using different models for each segment and for each phoneme class. A real-time meeting analyzer is presented in [14]. Several blocks of the full system are presented (e.g. dereverberation [15], source separation, speech recognition) along with speaker diarization which is based on clustering the Direction of arrival (DOA). Speaker diarization decisions are extracted by averaging the

ISBN 978-0-9928626-7-1 © EURASIP 2017

per frame diarization decisions over the word length. A front-end for speaker diarization based on beamforming is presented in [16].

The method proposed in this paper is based on the standard clustering of the TDOAs using a Gaussian Mixture Model (GMM). However, in order to provide robustness against noise, a mixture to explicitly model the feature vectors that do not form part of any cluster is added to the GMM, in addition to the speaker mixtures, whose parameters are learned using linear constraints on the means and variances of the mixtures. Furthermore, the speaker index is found by maximizing the *a posteriori* probability of each mixture given the TDOA estimate and the decisions are smoothed using a Hidden Markov Model (HMM). To the best of the authors' knowledge, it is the first time these linear constraints are applied for speaker diarization purposes.

The remainder of this paper is organized as follows. In Section II we present the proposed method and the evaluation in Section III and conclusions in Section IV.

II. Multi-channel diarization based on robust TDOA modeling

Figure 1 outlines the main components of the proposed method. Each of the main blocks represented in this diagram is described in the following sections.

A. TDOA computation

The TDOA is a common feature extracted in multi-microphone speech acquisition and represents the difference in the arrival times when a signal originating from a point source is recorded by microphones at two different positions. This feature is extracted per frame and a TDOA stream is created by concatenating in chronological order all these TDOA features computed for a given recording. The total number of TDOA streams J that are possible to compute from an N_{mic} microphone setup is given by, $J = 0.5 \cdot N_{mic} \cdot (N_{mic} - 1)$, and each comprises a total of N_{TDOA} samples. The TDOA, τ_l^j, for frame l and stream j is commonly computed as the maximum of the inverse Fourier transform of the GCC-PHAT [10], which computes the normalized cross-correlation between two signals in the frequency-domain. A frame size of 500 ms with a 87.5% overlap between consecutive frames was used in this paper (determined empirically on a development database).

B. Speaker Modeling

A GMM $\boldsymbol{\theta}$ can be parametrized by the *a priori* vector ($\boldsymbol{\lambda}$), the mean vector ($\boldsymbol{\mu}$) and the covariance matrix ($\boldsymbol{\sigma}$). The parameters of the individual mixtures for a given stream j are represented by $\boldsymbol{\theta}_i^j$ = (λ_i^j, μ_i^j, σ_i^j). An important aspect of our approach is that a maximum of N_{spk} + 1 mixtures are considered, i.e. $\boldsymbol{\theta}_i^j$ = ($\boldsymbol{\theta}_B^j$, $\boldsymbol{\theta}_1^j$, $\cdots$, $\boldsymbol{\theta}_{N_{spk}}^j$), N_{spk} mixtures to model the speakers' TDOAs and an additional mixture $\boldsymbol{\theta}_B^j$ to model the noisy estimates. The Maximum Likelihood Estimate (MLE) [17] of the model parameters given the data (i.e. TDOA stream) can be used to obtain $\boldsymbol{\theta}^j$ as $\underset{\boldsymbol{\theta}^j}{\arg\max} \log P(\boldsymbol{\tau}^j|\boldsymbol{\theta}^j)$, where $\boldsymbol{\tau}^j$ = (τ_1^j, $\tau_2^j \cdots$, $\tau_{N_{TDOA}}^j$). In common applications, $\boldsymbol{\tau}^j$ can be inaccurate due to noise, overlapping speakers, non-speech acoustic events and/or reverberation. Thus, $\boldsymbol{\theta}^j$ needs to be estimated robustly to these spurious TDOA estimates. In order to robustly estimate these model parameters $\boldsymbol{\theta}^j$, linear constraints are applied on the mean and the standard deviation in the Expectation-Maximization (EM) algorithm. These constraints are described in the following subsections.

1) Constraints on the mean: Linear constraints on the distribution means are determined *a priori* and defined such that the mean of the noise model, μ_B, is independent of the speakers' means. Additionally, the speakers' means are also constrained to be separated by at least a minimum separation to avoid them being determined indefeasibly close to each other. Thus, the constrained means are computed as follows

$$\boldsymbol{\mu} = \boldsymbol{\mathcal{M}}\boldsymbol{\beta} + \boldsymbol{C}, \tag{1}$$

where the former constraint, i.e. the noise model mean is independent of the speakers' means, is achieved with the matrix $\boldsymbol{\mathcal{M}}$ whereas the latter constraint, i.e. minimum separation between speakers' means, is defined using the matrix $\boldsymbol{\mathcal{C}}$ and the remaining unknown term $\boldsymbol{\beta}$ is computed by maximizing the likelihood of the model parameters given the TDOAs using Expectation-Conditional Maximization [18].

The expression (1) can be rewritten as,

$$\begin{bmatrix} \mu_B \\ \mu_1 \\ \mu_2 \\ \ldots \\ \mu_{N_{spk}} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ \ldots & \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ C_2 \\ \ldots \\ C_{N_{spk}} \end{bmatrix}, \tag{2}$$

where $C_{N_{spk}} = \tau_{maxN_{spk}} - \tau_{max1}$,

$$\tau_{max1} = \arg\max_{\tau} \left\{ p(\tau) \mid \frac{dp(\tau)}{d\tau} = 0 \right\}, \tag{3}$$

and

$$p(\tau) = \frac{1}{N_{TDOA}} \sum_{l=1}^{N_{TDOA}} \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-\frac{\|\tau - \tau_l^j\|^2}{2\sigma^2}}. \tag{4}$$

The unknown elements in $\boldsymbol{C}$ are computed following the same procedure but replacing N_{spk} by the speaker model index, where the $\tau_{maxN_{spk}}$ term is computed following (3) with the additional constraint of $\tau \neq \left\{\tau_{max1}, \tau_{max2}, \ldots, \tau_{max(Nspk-1)}\right\}$. The standard deviation σ of the Gaussian kernel in (4) is computed using Silverman's rule of thumb [19]. In order to provide robustness to the estimation of $p(\tau)$, negative and positive extreme values are removed from $\boldsymbol{\tau}^j$. This is carried out by removing the tail of the estimated density such that the limits are greater than 5% of the maximum peak of the density. Density kernels are used instead of histograms to estimate the probability density because this approach does not depend on the bin width [20] and the peaks are therefore more accurately estimated.

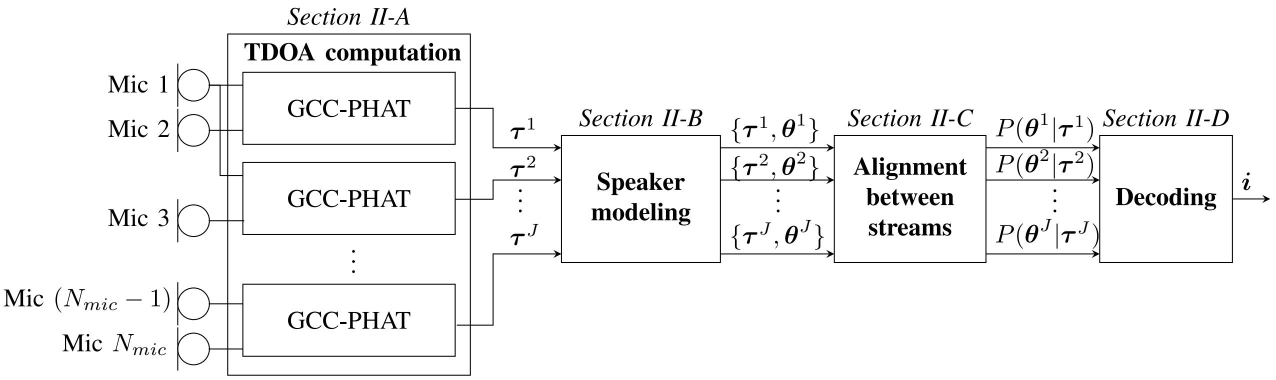
Section II-A
TDOA computation
Mic 1
Mic 2
Mic 3
Mic $(N_{mic}-1)$
Mic N_{mic}
GCC-PHAT
GCC-PHAT
GCC-PHAT
$\boldsymbol{\tau}^1$
$\boldsymbol{\tau}^2$
$\boldsymbol{\tau}^J$
Section II-B
Speaker modeling
$\{\boldsymbol{\tau}^1, \boldsymbol{\theta}^1\}$
$\{\boldsymbol{\tau}^2, \boldsymbol{\theta}^2\}$
$\{\boldsymbol{\tau}^J, \boldsymbol{\theta}^J\}$
Section II-C
Alignment between streams
$P(\boldsymbol{\theta}^1|\boldsymbol{\tau}^1)$
$P(\boldsymbol{\theta}^2|\boldsymbol{\tau}^2)$
$P(\boldsymbol{\theta}^J|\boldsymbol{\tau}^J)$
Section II-D
Decoding
$\boldsymbol{i}$

Fig. 1. Block diagram of the proposed method. The description of each module is in the section indicated on top of the blocks.

2) Constraints on the standard deviation: Linear constraints on the standard deviation are also fixed *a priori* and defined such that the deviation of the noise model is wider than the deviations of the speakers' models and additionally, head movements of all speakers are assumed to be similar, therefore the standard deviation of the speakers' models is the same. Hence, the linear constraints on the standard deviation are given by (5),

$$\boldsymbol{\iota} = \boldsymbol{\mathcal{G}}\boldsymbol{\Upsilon}, \tag{5}$$

where the elements of $\boldsymbol{\iota}$ represent the inverse of the standard deviations, $\boldsymbol{\mathcal{G}}$ comprises the defined constraints and $\boldsymbol{\Upsilon}$ is estimated by maximizing the likelihood of the parameters given the input data, solved by the Minorization-Maximization algorithm [18].

Equation (5) can we reformulated as follows,

$$\begin{bmatrix} 1/\sigma_B \\ 1/\sigma_1 \\ \cdots \\ 1/\sigma_{N_{spk}} \end{bmatrix} = \begin{bmatrix} \iota_B \\ \iota_1 \\ \cdots \\ \iota_{N_{spk}} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ \cdots & \cdots \\ 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} \Upsilon_1 \\ \Upsilon_2 \end{bmatrix}. \tag{6}$$

Additionally, variance upper and lower bounds are applied to avoid unlikely values. These variances are set to 1.25 ms and 0.03125 ms respectively, which are found experimentally.

C. Alignment between streams

An alignment is needed to ensure that the N_{spk} speaker indexes represent the same speaker across the different J streams. To explain this point let's assume $N_{spk} = 2$, then the alignment between streams verifies whether speaker model $\boldsymbol{\theta}_1^1$ represents the TDOAs of the speaker index that is modeled with $\boldsymbol{\theta}_1^j$ or the speaker index that is modeled with $\boldsymbol{\theta}_2^j$ for $j = \{2, \cdots, J\}$, where the superscript j indicates the stream evaluated. This verification is carried out by finding the vector $\hat{\boldsymbol{d}}$ such that,

$$\arg\max_{\hat{d} \in \boldsymbol{d}} \sum_{n=1}^{N_{TDOA}} s(d^1(n), \hat{d}(n)), \tag{7}$$

where, $s(x, y) = 1$ if $x = y$ and 0 otherwise. The term $\boldsymbol{d}$ is defined as a set of candidate vectors $\boldsymbol{d}^j$ and $\tilde{\boldsymbol{d}}^j$ where the latter vector is the permutation of the former, as, $\tilde{\boldsymbol{d}} = \boldsymbol{d}^j \pmod 2 + 1$, and the individual decision vectors are defined as $\boldsymbol{d}^j = \arg\max_{\boldsymbol{i}} P(\boldsymbol{\theta}_i^j|\boldsymbol{\tau}^j)$ where $\boldsymbol{\tau}^j$ represents the TDOAs computed from stream j. Finally, the terms $d^1(n)$ and $\hat{d}(n)$ are the nth elements on the vector $\boldsymbol{d}^1$, comprising decisions extracted from the first stream, and $\hat{\boldsymbol{d}}$ respectively. The magnitudes $P(\boldsymbol{\theta}_1^j|\boldsymbol{\tau}^j)$ and $P(\boldsymbol{\theta}_2^j|\boldsymbol{\tau}^j)$ are swapped if $\boldsymbol{d} = \tilde{\boldsymbol{d}}^j$. This approach can be applied to any $N_{spk} > 2$ by forming $\boldsymbol{d}$ such that it comprises $N_{spk}!$ vectors with all possible decision permutations. In this case, same decisions within each vector permute to the same values. This alignment has a complexity of $O(N_{spk}!)$, consequently the execution time rapidly increases when more speakers are considered. In order to reduce this complexity, a stochastic search is performed using a Genetic Algorithm (GA)[21] when $N_{spk} \geqslant 7$. In this case, the chromosomes encode the speaker index permutations and the fitness function is derived from (7) and the crossover and mutation probabilities are set empirically to 0.9 and 0.05 respectively.

D. Decoding

The aim of the decoding block is to find, for each frame l, the speaker index i that maximizes the posterior probability of the speaker model $\boldsymbol{\theta}_i^j$ given the TDOA sample τ_l^j as $\arg\max_i P(\boldsymbol{\theta}_i^j|\tau_l^j)$, where,

$$P(\boldsymbol{\theta}_i^j|\tau_l^j) = \frac{P(\tau_l^j|\boldsymbol{\theta}_i^j) \cdot P(\boldsymbol{\theta}_i^j)}{\sum_{e=1}^{N_{spk}} P(\tau_l^j|\boldsymbol{\theta}_e^j) \cdot P(\boldsymbol{\theta}_e^j)}. \tag{8}$$

The denominator of (8) is independent of i and hence it can be omitted from the maximization, thus the final Maximum A Posteriori (MAP) expression is $\arg\max_i P(\tau_l^j|\boldsymbol{\theta}_i^j) \cdot P(\boldsymbol{\theta}_i^j)$.

1) Stream Selection: We now consider multiple microphones in a pair-wise setup such as may be relevant for estimating TDOA streams, one stream per pair of microphones. *A priori*, the pair of microphones that is closer to the speaker is likely to be the best pair but the position of speakers and microphones is assumed unknown in general and noise sources can degrade the TDOAs computed in those pairs of microphones that are close to a noise source. The stream

ISBN 978-0-9928626-7-1 © EURASIP 2017

selection aims at choosing the TDOA stream from the best microphone pair to diarize, i.e the model that provides the lowest Diarization Error Rate (DER), using the commonly applied metric in model selection [22] of the Bayesian Information Criterion (BIC). This criterion is shown in (9) which is used to find the optimal pair of microphones j as follows

$$\text{BIC}(\boldsymbol{\theta}^j, \boldsymbol{\tau}^j) = -2 \log \mathcal{L}(\boldsymbol{\theta}^j | \boldsymbol{\tau}^j) + N_{fp} \cdot \log(N_{TDOA}), \quad (9)$$

where $\mathcal{L}(\boldsymbol{\theta}^j | \boldsymbol{\tau}^j)$ is the likelihood of the model $\boldsymbol{\theta}^j$ given the data $\boldsymbol{\tau}^j$ of the TDOA stream j and N_{fp} is the number of free parameters to be estimated in $\boldsymbol{\theta}$. Since N_{fp} and N_{TDOA} are the same for all J streams, expression (9) is equivalent to the Maximum Likelihood criterion.

2) Stream Combination: Alternatively, rather than selecting only one TDOA stream to perform MAP speaker labeling, the MAP can be performed over the average of all J streams as follows

$$\arg \max_i \frac{1}{J} \sum_{j=1}^{J} P(\boldsymbol{\theta}_i^j | \tau_l^j), \text{where } i = \{1, \cdots, N_{spk}\}. \quad (10)$$

3) HMM: A Hidden Markov Model (HMM) is implemented in order to include prior models for utterance duration and thereby potentially avoid very unlikely short utterances from one speaker [23]. Each state of the Hidden Markov Models (HMM) represents one speaker index and the transition probabilities a_{qr} and observation probabilities b_q are computed as follows for two speakers,

$$a_{12} = a_{21}, \; a_{11} = 1 - a_{12}, \; a_{22} = 1 - a_{21}, \quad (11)$$

$$b_1(\tau_l^j) = P(\boldsymbol{\theta}_1^j | \tau_l^j), \; b_2(\tau_l^j) = P(\boldsymbol{\theta}_2^j | \tau_l^j). \quad (12)$$

The a_{21} term in (11) is computed as the ratio of TDOA frame increment over the average speaker duration. Assuming an approximate average speaker duration of 2.5 s [4] and the TDOA frame increment of 62.5 ms, then $a_{21} = 0.025$. This ratio is derived from the fact that the number of steps in the same state is geometrically distributed [24] and its expected value is $1/(1 - a_{qq})$ for $q \in \{1, 2, \cdots, N_{spk}\}$. Therefore $1/(1-a_{qq})$ is set to be the average speaker duration in frames. For $N_{spk} > 2$, all the states are still interconnected and the $1/(1 - a_{qq})$ is still computed as the average speaker duration in frames, however $a_{qr} = (1 - a_{qq})/(N_{spk} - 1)$. Finally, the speaker label estimate at frame l can be extracted by applying the Viterbi algorithm [25].

III. Evaluation

A. Experimental Setup

The distant multi-microphones partition of the conference room meetings corpora from NIST RT-05 [26] is used for evaluation of the proposed method, as they provide real scenarios with highly interactive discussions between multiple speakers. The results are obtained by setting the maximum number of speakers to 10 for the proposed and baseline methods. Thus both systems can be compared in the same test conditions. The baseline algorithm for comparison in this paper is DiarTK [2]. This open source toolkit was given a multi-dimension feature vector comprising TDOA streams from all microphone pairs (TDOAs were computed with 500 ms frames and 62.5 ms frame increment as for the proposed method). Since DiarTk also requires a VAD input, the ground truth VAD labels from RT05 are provided (our method was not given this information). The evaluation was restricted to speech active regions and thus the speaker error rates were the metric used.

Meeting	N_{spk}	N_{mic}	Stream Selection	Stream Combination
AMI1	4	8	54.1	85.6
AMI2	4	8	-6.0	31.3
CMU1	4	3	75.2	77.1
CMU2	4	3	77.4	38.0
ICSI1	7	6	84.6	70.8
ICSI2	9	6	50.1	49.9
NIST1	10	7	-54.3	-56.9
NIST2	4	7	0.0	31.2
VT1	5	2	8.3	8.3
VT2	5	2	25.9	25.9
Mean RRSE(%)			31.5	36.1

TABLE I

RRSE (%) RESULTS ON THE NIST RT05 DATABASE FOR THE PROPOSED SYSTEM. THE N_{spk} COLUMN HIGHLIGHTS THE NUMBER OF SPEAKERS AND N_{mic} THE NUMBER OF MICROPHONES IN THE MEETINGS.

B. Results

In the following we present the Relative Reduction in Speaker Error (RRSE) as a metric for comparing the baseline with the various modes of our proposed method. This metric is computed as

$$\text{RRSE} = \frac{SE_{baseline} - SE_{proposed}}{SE_{baseline}} \cdot 100 \quad (13)$$

where $SE_{baseline}$ and $SE_{proposed}$ are the speaker error achieved with the baseline and proposed method respectively.

The detailed results are presented in Table I. The stream selection and combination modes of the proposed method outperform the baseline algorithm on the RT05 method on average. Overall, the stream combination approach gives the highest RRSE (and correspondingly the lowest speaker errors). The mean RRSE obtained without using the HMM is 29.2% and 25.3% for stream selection and combination modes respectively which indicate the suitability of these models for diarization purposes.

The proposed method performs poorly on the NIST1 meeting, which has 10 active speakers, one of whom is a remote participant, joining the meeting through a speaker. More importantly, only 22.2% of the evaluated segment of this meeting contains speech (on average this is 93.78% across all remaining 9 meetings), therefore the proposed method tries to model the remaining 77.8% of the meeting with only one mixture $\boldsymbol{\theta}_B^j$ which is likely to have a negative impact on the speakers' models due to the large amount of data without speech. In contrast, DiarTK has a large advantage in terms of prior knowledge used for modeling since it is provided with the ground truth speech active segment information, avoiding thus this problem. Also DiarTK estimates the number of speakers internally through an agglomerate process and thus tries to iteratively reduce the number of speakers until the optimal is reached whereas our proposed method builds 10 Gaussian models for each TDOA stream. By setting the correct number of speakers in the latter method, the speaker error is further reduced to RRSE = 51.9% with the stream combination approach.

In general, Table I suggests that the performance of the different methods is independent of the number of speakers. The error of AMI2 which comprises 8 speakers is relatively high while the error of AMI1 which comprises 8 speaker is relatively low.

Lastly, our method is not very sensitive to errors resulting from overestimating the number of speakers when the speaker activity is well distributed. The performance for the proposed method is the same for the VT meetings as there is only 1 TDOA stream available.

IV. Conclusions

In this paper we presented a novel speaker diarization method that uses spatial features in the form of TDOAs extracted using for example the GCC-PHAT algorithm and modeled to be robust to noise and reverberation by applying linear constraint on the variances and means of these GMMs models' parameters. The evaluation of the proposed method was carried out on the distant multi-microphone condition of the NIST RT05 database and showed that our method outperforms DiarTk by 36.1% relative reduction in speaker error, using only spatial features and by setting the number of speakers to maximum value. Further improvements can be gained when the number of speakers is known *a priori*. Although this paper focused on TDOA-based features only, additional improvements in performance are expected by combining additional speech features such as MFCCs with the proposed method. In relation to the diarization output, a confidence measure associated with each decision can be derived from (10) by computing the averaged probability that maximizes this expression for a given frame l. In addition to the diarization output, our method can be used to select the best microphone pair, which can provide valuable side information for a number of speech signal processing algorithms.

References

[1] Mathieu Hu, P. Peso Parada, D. Sharma, S. Doclo, T. van Waterschoot, M. Brookes, and P. A. Naylor, "Single-channel speaker diarization based on spatial features," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, October 2015, pp. 1–5.

[2] D. Vijayasenan, F. Valente, and H. Bourlard, "An information theoretic combination of MFCC and TDOA features for speaker diarization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 2, pp. 431–438, February 2011.

[3] D. Korchagin, "Audio spatio-temporal fingerprints for cloudless real-time hands-free diarization on mobile devices," in *Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, May 2011, pp. 25–30.

[4] X. Anguera Miro, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: a review of recent research," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 2, pp. 356–370, 2012.

[5] T. Stafylakis and V. Katsouros, "A review of recent advances in speaker diarization with bayesian methods," in *Speech and Language Technologies*, I. Ipsic, Ed., chapter 11, pp. 217–240. INTECH Open Access Publisher, Rijeka, 2011.

[6] R. Sinha, S. E. Tranter, M. J. F. Gales, and P. C. Woodland, "The Cambridge University March 2005 speaker diarisation system," in *Proc. European Conf. on Speech Communication and Technology*, Lisbon, Portugal, September 2005, pp. 2437–2440.

[7] S.E. Tranter and D.A. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, pp. 1557–1565, 2006.

[8] R. Milner, O. Saz, S. Deena, M. Doulaty, R. W. M. Ng, and T. Hain, "The 2015 Sheffield system for longitudinal diarisation of broadcast media," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, December 2015, pp. 632–638.

[9] G. Sell and D. Garcia-Romero, "Speaker diarization with PLDA i-vector scoring and unsupervised calibration," in *IEEE Spoken Language Technology Workshop (SLT)*, December 2014, pp. 413–417.

[10] Charles H Knapp and G Clifford Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[11] N.W.D. Evans, C. Fredouille, and J.-F. Bonastre, "Speaker diarization using unsupervised discriminant analysis of inter-channel delay features," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, April 2009, pp. 4061–4064.

[12] T. Oku, S. Sato, A. Kobayashi, S. Homma, and T. Imai, "Low-latency speaker diarization based on bayesian information criterion with multiple phoneme classes," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, March 2012, pp. 4189–4192.

[13] Gideon Schwarz et al., "Estimating the dimension of a model," *The annals of statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[14] T. Hori, S. Araki, T. Yoshioka, M. Fujimoto, S. Watanabe, T. Oba, A. Ogawa, K. Otsuka, D. Mikami, K. Kinoshita, T. Nakatani, A. Nakamura, and J. Yamato, "Low-latency real-time meeting recognition and understanding using distant microphones and omni-directional camera," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 2, pp. 499–513, February 2012.

[15] P. A. Naylor and N. D. Gaubitch, Eds., *Speech Dereverberation*, Springer, London, 2010.

[16] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 7, pp. 2011–2022, September 2007.

[17] G. McLachlan and T. Krishnan, *The EM algorithm and extensions*, vol. 382, John Wiley & Sons, New York, 2007.

[18] Didier Chauveau and David R. Hunter, "ECM and MM algorithms for normal mixtures with constrained parameters," working paper or preprint, August 2013.

[19] Bernard W Silverman, *Density estimation for statistics and data analysis*, vol. 26, CRC press, 1986.

[20] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer-Verlag New York, Inc., 2006.

[21] Luca Scrucca, "GA: A package for genetic algorithms in R," *Journal of Statistical Software*, vol. 53, no. 4, pp. 1–37, 2013.

[22] Kenneth P. Burnham and David R. Anderson, *Model selection and multimodel inference: a practical information-theoretic approach*, Springer Science & Business Media, 2002.

[23] C.D. Mitchell and L.H. Jamieson, "Modeling duration in a hidden markov model with the exponential family," in *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on*, April 1993, vol. 2, pp. 331–334.

[24] C. R. Shelton and G. Ciardo, "Tutorial on structured continuous-time markov processes.," *Journal of Artificial Intelligence Research*, vol. 51, pp. 725–778, 2014.

[25] Andrew Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE transactions on Information Theory*, vol. 13, no. 2, pp. 260–269, 1967.

[26] J. G. Fiscus, N. Radde, J. S. Garofolo, A. Le, J. Ajot, and C. Laprun, "The rich transcription 2005 spring meeting recognition evaluation," in *Machine Learning for Multimodal Interaction*, pp. 369–389. Springer, 2005.

ISBN 978-0-9928626-7-1 © EURASIP 2017