2019 27th European Signal Processing Conference (EUSIPCO)

Ad-hoc mobile array based audio segmentation using latent variable stochastic model

Srikanth Raj Chetupalli, Anirban Bhowmick, and Thippur V. Sreenivas
Department of Electrical Communication Engineering, Indian Institute of Science, Bengaluru, 560012.

***Abstract*—Segmentation/diarization of audio recordings using a network of ad-hoc mobile arrays and the spatial information gathered is a part of acoustic scene analysis. In this ad-hoc mobile array network, we assume fine (sample level) synchronization of the signals only at each mobile node and a gross synchronization (frame level) across different nodes is sufficient. We compute spatial features at each node in a distributed manner without the overhead of signal data aggregation between mobile devices. The spatial features are then modeled jointly using a Dirichlet mixture model, and the posterior probabilities of the mixture components are used to derive the segmentation information. Experiments on real life recordings in a reverberant room using a network of randomly placed mobile phones has shown a diarization error rate of less than 14% even with overlapped talkers.**

***Index Terms*—Diarization, Dirichlet distribution, steered response power, acoustic sensor network, mobile devices.**

I. Introduction

Ad-hoc network of microphone arrays [1] is an interesting technology for speech and audio applications due to the improved spatial coverage and the path diversity, which can be used to improve the performance of applications such as speech enhancement, recognition, segmentation/diarization etc. However, such a setup is characterized by asynchronous recording at different sensor nodes, although microphones at the same node can record synchronously. Random placement of sensor nodes is also a challenge for aggregating their individual signals, without the actual geometry of their placement. In this paper, we consider one specific task of speech segmentation in a meeting scenario; i.e., "who spoke when?" in a multi-channel audio recording of multiple speakers.

Methods based on spectral features, spatial features or a combination of both have been proposed for multi-channel signal diarization of audio recordings [2]–[4]. However, a single array of microphones is the commonly used approach. Audio/speaker diarization using an ad-hoc microphone network has not been attempted. In this paper, we consider the segmentation/diarization of audio recordings using the spatial features alone, but it can be augmented using spectral features to obtain further audio intelligence.

Several solutions have been proposed utilizing spatial features, such as the time-difference-of-arrival (TDOA) features [5]–[8]. In this, the estimation of TDOA is sensitive to room reverberation and other acoustic interferences. To overcome this a formulation based on a pre-trained spatial dictionary and Watson mixture modeling of directional features is proposed in [9]. However, all these methods require synchronous signal data from distributed microphones. Instead for the ad-hoc microphone network considered in this paper, microphones across the different nodes are asynchronous, the geometry of the placement of nodes is random/unknown, and hence network-wide computation of TDOA or beam-forming is not feasible. Instead, we compute the directional features independently at each device, and then combine them using a stochastic formulation.

Spatial response function computed using steered response power with the phase transform (SRP-PHAT) filtering [10] is used as the spatialization measure. Assuming known microphone geometry at each ad-hoc node, the SRP response function is computed for a set of directions and these measures are normalized to result in a stochastic representation. We use the stochastic representation as the spatial feature and features of several ad-hoc devices are combined using a latent variable mixture model. We use a Dirichlet mixture model [11] after the signals from different devices are aligned coarsely using a specific acoustic event such as a clap or a tap, or network time. Expectation-maximization [12] approach is used for maximum likelihood estimation of the latent variables and the segmentation/diarization information is derived from the posterior mixture component probabilities. Experiments on real life meeting speech recorded using commercial off-the-shelf randomly placed mobile phones has shown diarization error rate (DER) of less than 14%.

II. Problem Formulation

Consider a general audio recording scenario with S number of sources and P number of microphone array nodes. Let M_p- be the number of microphones at the p^{th} node, and let $x_{m,p}[t]$ denote the audio signal recorded at the m^{th} microphone of the p^{th} node. Given the recordings at all the devices $\{x_{m,p}[t], \forall m \in [1\ M_p]; \forall p \in [1\ P]\}$, the goal of this paper is to perform segmentation/diarization of the recorded signal, i.e., to identify "who spoke when?" in the long conversation recording. We assume the sources to be fixed (non moving) and a single source at a given spatial region, which is true in most of the meeting recording scenarios. In such a scenario, the spatial information alone can provide the source (speaker) activity along the recording time-line.

The audio signal is recorded at each node using the local clock without any other external synchronization. However, for further processing, the estimated features from different nodes can be synchronized coarsely. The synchronous recording at a particular node provides for beam steering to compute the source direction information. We consider computation

978-9-0827-9703-9/19/$31.00 ©2019 IEEE

of source spatial information statistics in a frame by frame manner, independently at each mobile node, using SRP-PHAT method. We then formulate a joint modeling of the directional statistics obtained at each of the nodes using a latent variable mixture model. Since the sensor nodes are placed arbitrarily and the information about their own position and orientation are unknown, we cannot combine the individual spatial features through a geometric formulation. Hence, we resort to stochastic modeling to derive the segmentation information. We note that in the proposed approach, the goal is not the exact position of the source, but to use the directional information to recognize the presence of source activity along the recording time-line. We show that this is possible using a stochastic formulation of directional data derived from several ad-hoc microphone arrays.

III. Statistical Detection

A. Spatial features

We consider steered response power (SRP) approach to compute the spatial features at each time-frame n for each of the nodes separately. We omit the index of the node p in the following discussion for brevity. Let $\mathbf{x}[n,k] = [x_1[n,k] \ldots x_M[n,k]]^T$ denote the multi-channel signal in the short time Fourier transform (STFT) domain for the microphone array of a single node, where n, k denote the discrete time and frequency indices respectively. Let $\mathbf{a}[\theta,k]$ denote the steering vector corresponding to a source at a spatial direction θ for the frequency bin k with respect to a local coordinate system centered at the array. Assuming free field sound propagation and a compact array, we have

$$\mathbf{a}[\theta,k] = \left[1 \; e^{\left(-\frac{j2\pi k\tau_{21}(\theta)}{K}\right)} \ldots e^{\left(-\frac{j2\pi k\tau_{M1}(\theta)}{K}\right)}\right]^T, \quad (1)$$

where K is the size of the discrete Fourier transform used for the STFT computation, and $\{\tau_{21}(\theta), \ldots, \tau_{M1}(\theta)\}$ denote the TDOA values at the $M-1$ microphones with respect to the first (reference) microphone. In the SRP method [10], the spatial response function is computed as,

$$y[n,\theta] = \sum_{k=1}^{K} \left|\mathbf{a}[\theta,k]^H \mathbf{x}_f[n,k]\right|^2, \quad (2)$$

where $\mathbf{x}_f[n,k] = \frac{\mathbf{x}[n,k]}{|\mathbf{x}[n,k]|}$ is the signal phase vector obtained after PHAT filtering.

We evaluate the response function $y[n,\theta]$ at L discrete angular positions $\mathbf{\Theta} = \{\theta_1, \ldots, \theta_L\}$ with respect to the array. Since the source can be assumed to be relatively stationary compared to STFT/SRP computation, we smooth the discrete SRP function across time using recursive averaging,

$$\tilde{y}[n,\theta_l] = \alpha\tilde{y}[n,\theta_l] + (1-\alpha)y[n-1,\theta_l], \; l \in \; [1,L]. \quad (3)$$

This is to minimize the time variation of SRP function due to the diffuse reverberation component. Smoothed SRP function is then normalized to represent the estimated source direction statistic which is used as a feature for the mixture density modeling.

Let $\mathbf{y}[n,\mathbf{\Theta}] \triangleq \frac{1}{C}\left[\tilde{y}[n,\theta_1] \ldots \tilde{y}[n,\theta_L]\right]^T$, where $C = \sum_{l=1}^{L} \tilde{y}[n,\theta_l]$ is the normalization constant. Thus the vector $\mathbf{y}[n,\mathbf{\Theta}]$ is a positive function and sums up to unity over θ, to be viewed as a probability measure.

In the present formulation, we compute the spatial features independently at each node, and obtain P number of features $\{\mathbf{y}_p[n,\mathbf{\Theta}_p]\}$, one vector per node, at each time frame n. Due to reverberation in the enclosure and other recording noise, $\mathbf{y}_p[n,\mathbf{\Theta}_p], \forall \; n,p$ does have estimation errors and hence a further statistical formulation is required to effectively combine the information from several nodes.

B. Latent variable source modeling

We model the spatial features computed at each of the microphone arrays of the P nodes $\{\mathbf{y}_p[n,\mathbf{\Theta}_p], \; 1 \le p \le P, \; 0 \le n \le N-1\}$ jointly using a stochastic mixture model. The generative model of the observations can be stated as follows: the latent variable selection vector $\mathbf{z}_n$ (S dimensional binary vector) selects a source (an audio source or a speaker) from a set of S sources based on a Bernoulli distribution with parameter $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_S]^T$, $\mathbb{P}(\mathbf{z}_n; \boldsymbol{\pi}) = \prod_{s=1}^{S} \pi_s^{z_{ns}}$, and the signal from the selected source results in the spatial feature observations $\{\mathbf{y}_p[n,\mathbf{\Theta}_p]\}$ at all the P nodes. Spatial feature (SRP function) computation at a node depends on the placement/orientation of the microphones at the node and the relative distance of the node with respect to the source, and it is independent of the other nodes. Hence, $\{\mathbf{y}_p[n,\mathbf{\Theta}_p]\}$ computed at the P nodes can be combined together as a joint probability of,

$$\mathbb{P}(\{\mathbf{y}_p[n,\mathbf{\Theta}_p]\}|z_{ns}=1,\mathbf{\Delta}) = \prod_{p=1}^{P} \mathbb{P}(\mathbf{y}_p[n,\mathbf{\Theta}_p]|\boldsymbol{\delta}_{sp}), \quad (4)$$

where $\mathbf{\Delta} = \{\boldsymbol{\delta}_{sp}, \forall s,p\}$ is the set of parameters of the mixture component densities of the S sources, at each of the P nodes.

Since the spatial feature $\mathbf{y}_p[n]$ represents a probability mass function (PMF). We propose to model them using a Dirichlet distribution [11] to suit the discrete nature of the sources.

$$\begin{aligned}\mathbb{P}\left(\mathbf{y}_p[n,\mathbf{\Theta}_p]\middle|\boldsymbol{\delta}_{sp}\right) &= \mathcal{D}(\mathbf{y}_p[n,\mathbf{\Theta}_p];\boldsymbol{\delta}_{sp}), \quad (5)\\ &= \frac{\Gamma\left(\sum_{l=1}^{L}\delta_{sp}[l]\right)}{\prod_{l=1}^{L}\Gamma\left(\delta_{sp}[l]\right)} \prod_{l=1}^{L} \left(\mathbf{y}_p[n,\theta_{pl}]\right)^{\delta_{sp}[l]-1},\end{aligned}$$

where $\Gamma(.)$ denotes the Gamma function. It may be noted that δ_{sp} parameters provide for the discrete form of the PMF at each node, although θ_l are different at different nodes. Also, assuming the directional data to be independent across time, we get the overall model as:

$$\mathbb{P}(\mathbf{Y}|\mathbf{Z},\mathbf{\Delta}) = \prod_{n=0}^{N-1}\prod_{s=1}^{S}\left[\prod_{p=1}^{P}\mathcal{D}(\mathbf{y}_p[n,\mathbf{\Theta}_p];\boldsymbol{\delta}_{sp})\right]^{z_{ns}}. \quad (6)$$

The independence assumption across time may not be always true, since, (i) the spatial features are computed after smoothing (eqn. (3)), and (ii) the sources do not change their position arbitrarily. However, this dependence is not considered in the present formulation, but can be introduced with a first order Markov dependence of the latent variables.

The formulation is equivalent to mixture modeling of the spatial features separately at each node, with the latent selection variable $\mathbf{z}_n$ shared (common) across all the nodes. Accordingly, we show in the next section that the parameter estimation of the component densities is independent for each node, while the computation of posterior over the latent variable requires central aggregation of frame likelihoods.

C. Parameter estimation

The parameters $\mathbf{\Delta}$ and $\boldsymbol{\pi}$ are estimated by maximizing the total likelihood function using the expectation-maximization (EM) algorithm. At iteration-i, the EM algorithm involves computation of (i) the posterior distribution $\mathbb{P}\left(\mathbf{Z}|\mathbf{Y},\mathbf{\Delta}^{(i)},\boldsymbol{\pi}^{(i)}\right)$, and (ii) maximization of the expected joint likelihood $Q(\mathbf{\Delta},\boldsymbol{\pi})=\mathbb{E}\{\log\mathbb{P}(\mathbf{Y},\mathbf{Z}|\mathbf{\Delta},\boldsymbol{\pi})\}$.

It can be shown that, $\mathbb{P}\left(\mathbf{Z}|\mathbf{Y},\mathbf{\Delta}^{(i)},\boldsymbol{\pi}^{(i)}\right)$ is an independent Bernoulli distribution with parameter,

$$\mathbb{P}\left(z_{ns}=1|\{\mathbf{y}_p[n,\mathbf{\Theta}_p]\},\mathbf{\Delta}^{(i)},\boldsymbol{\pi}^{(i)}\right)= \frac{\pi_s^{(i)}\prod\limits_{p=1}^{P}\mathcal{D}(\mathbf{y}_p[n,\mathbf{\Theta}_p];\boldsymbol{\delta}_{sp}^{(i)})}{\sum\limits_{s=1}^{S}\pi_s^{(i)}\prod\limits_{p=1}^{P}\mathcal{D}(\mathbf{y}_p[n,\mathbf{\Theta}_p];\boldsymbol{\delta}_{sp}^{(i)})} \tag{7}$$

and $\mathbb{E}\{z_{ns}\}\triangleq\gamma_{ns}^{(i+1)}=\mathbb{P}(z_{ns}=1\big|\{\mathbf{y}_p[n,\mathbf{\Theta}_p]\},\mathbf{\Delta}^{(i)},\boldsymbol{\pi}^{(i)})$.

In the maximization step, the function $Q(\mathbf{\Delta},\boldsymbol{\pi})$ is maximized:

$$Q(\mathbf{\Delta},\boldsymbol{\pi})=\mathbb{E}\{\log\mathbb{P}(\mathbf{Y}|\mathbf{Z},\mathbf{\Delta})+\mathbb{E}\{\log\mathbb{P}(\mathbf{Z}|\boldsymbol{\pi})\} \tag{8}$$

Substituting for the component densities and simplifying, we get,

$$\begin{aligned}Q(\mathbf{\Delta},\boldsymbol{\pi})=&\sum_{n=0}^{N-1}\sum_{s=1}^{S}\gamma_{ns}^{(i+1)}\log\pi_s+\\ &\sum_{n=0}^{N-1}\sum_{s=1}^{S}\gamma_{ns}^{(i+1)}\sum_{p=1}^{P}\log\mathcal{D}(\mathbf{y}_p[n,\mathbf{\Theta}_p];\boldsymbol{\delta}_{sp}).\end{aligned} \tag{9}$$

Maximization of eqn. (9) with respect to π_s subject to the constraint $\sum\limits_{s=1}^{S}\pi_s=1$ results in the estimate,

$$\pi_s^{(i+1)}=\frac{N_s}{N},\quad\text{where}\quad N_s=\sum_{n=0}^{N-1}\gamma_{ns}^{(i+1)}. \tag{10}$$

Maximization of (9) with respect to $\boldsymbol{\delta}_{sp}$ requires solving the problem:

$$\boldsymbol{\delta}_{sp}^{(i+1)}=\underset{\boldsymbol{\delta}_{sp}}{\arg\max}\sum_{n=0}^{N-1}\gamma_{ns}^{(i+1)}\log\mathcal{D}(\mathbf{y}_p[n,\mathbf{\Theta}_p];\boldsymbol{\delta}_{sp}). \tag{11}$$

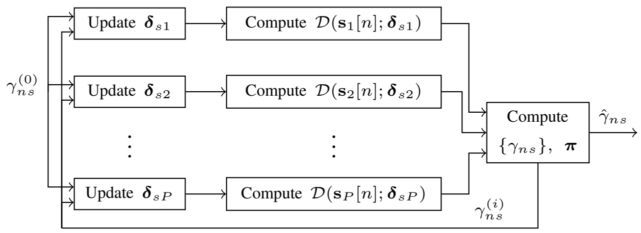
$\gamma_{ns}^{(0)}$
Update $\boldsymbol{\delta}_{s1}$
Update $\boldsymbol{\delta}_{s2}$
Update $\boldsymbol{\delta}_{sP}$
Compute $\mathcal{D}(\mathbf{s}_1[n];\boldsymbol{\delta}_{s1})$
Compute $\mathcal{D}(\mathbf{s}_2[n];\boldsymbol{\delta}_{s2})$
Compute $\mathcal{D}(\mathbf{s}_P[n];\boldsymbol{\delta}_{sP})$
Compute $\{\gamma_{ns}\}$, $\boldsymbol{\pi}$
$\hat{\gamma}_{ns}$
$\gamma_{ns}^{(i)}$

Fig. 1. Block diagram of the EM estimation algorithm

Substituting for $\mathcal{D}(\mathbf{s}_p[n];\boldsymbol{\delta}_{sp})$ using (5), we get the optimization problem as,

$$\begin{aligned}\boldsymbol{\delta}_{sp}^{(i+1)}=\underset{\boldsymbol{\delta}_{sp}}{\arg\max}\sum_{n=0}^{N-1}\gamma_{ns}^{(i+1)}\Bigg[\log\Gamma\left(\sum_{l=1}^{L}\delta_{sp}[l]\right)-\\ \sum_{l=1}^{L}\log\Gamma(\delta_{sp}[l])+\sum_{l=1}^{L}(\delta_{sp}[l]-1)\log\mathbf{y}_p[n,\theta_{pl}]\Bigg].\end{aligned} \tag{12}$$

Gradient-descent algorithm is used to solve for $\{\boldsymbol{\delta}_{sp}\}$ [13].

The parameters $\boldsymbol{\delta}_{sp},\forall s,p$ are estimated independently for each node p which allows for distributed computation, but the estimation of $\{\gamma_{ns}\}$ and $\boldsymbol{\pi}$ require the likelihood computed at all the nodes (eqn. (7)). A block-diagram description of the algorithm is shown in Fig. 1. In each iteration, the parameters $\boldsymbol{\delta}_{sp}$ are updated using (12) independently at each node using the posterior estimates γ_{ns} from the previous iteration. The updated $\boldsymbol{\delta}_{sp}$ are then used to compute the likelihood $\mathcal{D}(\mathbf{y}_p[n,\mathbf{\Theta}_p];\boldsymbol{\delta}_{sp})$ at each node, which is then shared to a common computing node, which updates the posterior γ_{ns}.

D. Segmentation/Diarization

At the convergence of the EM algorithm, the posterior parameter, γ_{ns}^{*} denotes the probability of s^{th} source being active at n^{th} time frame. The segmentation information is obtained as the source label s at each time frame n using the max-rule over s,

$$\hat{s}[n]=\underset{s}{\arg\max}\ \gamma_{ns}^{*}. \tag{13}$$

IV. Experiments and Results

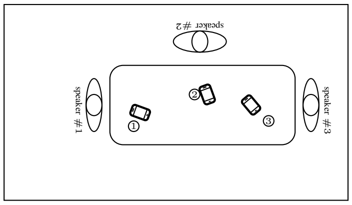

Fig. 2. Meeting recording scenario

Real-life meeting recordings are used for the evaluation of the proposed scheme. Three mobile phones (from three different manufacturers, running android OS) are placed in an arbitrary orientation on a table of dimensions $1.23\ m\times 0.77\ m\times 0.77m$ in a reverberant enclosure (RT60 $\approx$ 650 ms). Each mobile is configured to record stereo signals at $F_s=48\ KHz$. The recorded signals are down-sampled to

16 KHz to confine the STFT range to 8 KHz. The sound from a tap on the table is used as the acoustic event to align the STFTs across the mobile devices. We consider five recordings with three participants (seated on three sides of the table as shown in Fig. 2) in each recording. The three speakers are chosen among three male speakers and one female speaker. The duration of the recordings varied from $5-10$ minutes, and the recordings are annotated manually for the speaker-ID. The mobile phones and participants are placed freely for all the five recordings, without any specific orientation.

STFT analysis is carried out using a frame size of 64 ms with 50% overlap between successive frames. In the SRP-PHAT computation, the beam steering is performed with a resolution of 4^o ($L = 46$). The steering vector of SRP-PHAT requires knowledge of the spacing between the microphones. For the commercial mobile devices used here, we do not know the exact mic spacing, hence we choose a maximum spacing of 0.16 m. This will affect only the local angle θ_l and does not alter the probability measures. The parameter α used for obtaining smooth spatial features is chosen to be 0.9. EM algorithm for DMM estimation is initialized using the method suggested in [13], and the maximum number of iterations is limited to 100. The number of sources S is assumed to be known in this experiment. However, it is possible to estimate it by using the histogram of the peak locations of the spatial feature. The performance is measured using the diarization error rate (DER), and computed using the NIST speech recognition scoring toolkit [14], with a collar interval of 0.25 s. The proposed algorithm assigns each frame to a single source, and an estimate of the oracle performance is obtained using ground truth labels where we assigned the label of previous frame to frames with speaker overlap.

Fig. 3 illustrates the spatial features computed at the three mobile devices and the spectrogram of the speech recorded at one of the devices for one of the recordings (illustrations for all the recordings are available online[1]). The spatial features of the sources differ at the three devices, and the discriminability between source positions is also different for the three mobiles. We can see that m_2 and m_3 show clearer directional features than m_1. This is likely due to the differences in the placement and sensitivity of the microphones on the mobile devices. However, there is one-one correspondence between the feature patterns across the devices. For example, in the first mobile recording, the spatial features contain a clear peak only for one of the sources (green), and the energy is less directional for the other two sources. This may be due to the directionality and placement of the microphones in the mobile device. The joint modeling at all the devices does help in estimating the correct source regions. Source posterior $\{\gamma_{ns}\forall n\}$ is shown in Fig. 3(e). We see that estimated speaker activity closely matches the ground truth shown in Fig. 3(f). We note that, silence regions and also segments with overlapped speakers are assigned to the previous segmented speaker. This is because of the smoothing step in feature computation.

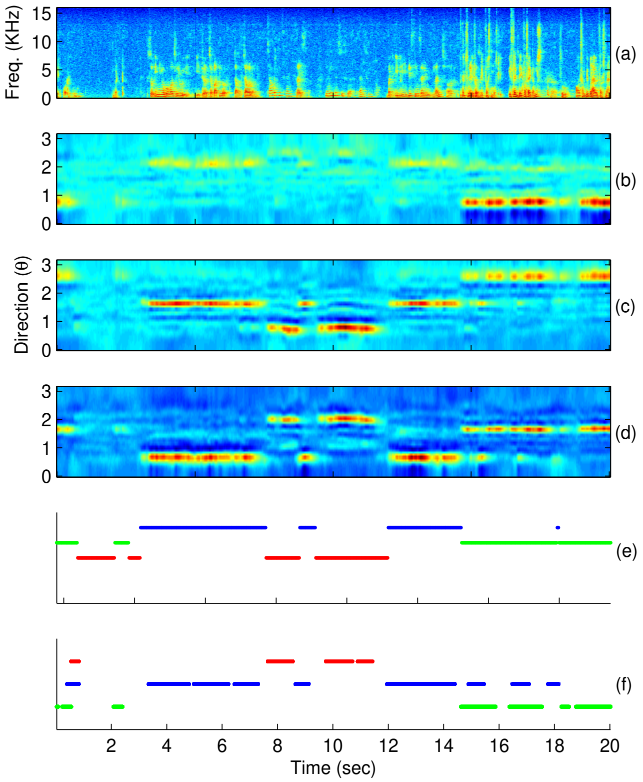

Fig. 3. (a) Spectrogram of a microphone signal. (b,c,d) computed spatial features $\{\mathbf{s}_p[n]\}$ for the three mobile devices ($m_1 - m_3$), (e) estimated source activity, and (f) ground truth source activity (s_1 = red, s_2 = blue, s_3 = green) shown in respective color.

TABLE I
DER PERFORMANCE (%) FOR FIVE RECORDINGS $R1 - R5$

ID	R1	R2	R3	R4	R5	Avg.
Proposed	13.1	12.5	20.9	14.0	6.5	13.4
Oracle	11.3	10.8	20.5	13.7	5.6	12.4

Overall performance of the new scheme for all the five recorded conversations is shown in Tab. I. The performance varies across the different recordings, due to the different sources and the different microphone placements; also there will be different amounts of overlap between the sources during the conversation. The DER is found to be high for some conversations that have higher overlap. However, for all the recordings, the performance of the proposed algorithm is with in 2% from the oracle performance.

V. Conclusions

Dirichlet mixture modeling of spatial features computed per node and a shared latent space is found to be good for identifying "who spoke when?" in audio recordings from a network of ad-hoc mobile microphone arrays. This is true despite the unknown variabilities such as the nature of microphones, their orientation within different nodes, unknown/random placement of the nodes and asynchronous recording. Presently a single source is assigned for each time-frame, but the method can be extended to predict multiple source activity, which can further improve the diarization performance.

[1] http://www.ece.iisc.ernet.in/∼sraj/mDiar.html

VI. Acknowledgements

This research is partially supported by the TCS research fellowship program.

References

[1] A. Bertrand, "Applications and trends in wireless acoustic sensor networks: A signal processing perspective," in *IEEE Symposium on Communications and Vehicular Technology in the Benelux (SCVT)*, Nov 2011, pp. 1–6.

[2] S. E. Tranter and D. A. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 5, pp. 1557–1565, Sept 2006.

[3] X. Anguera, 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, Feb 2012.

[4] M. Moattar and M. Homayounpour, "A review on speaker diarization systems and approaches," *Speech Communication*, vol. 54, no. 10, pp. 1065 – 1103, 2012.

[5] 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, Sept 2007.

[6] N. W. D. Evans, C. Fredouille, and J. Bonastre, "Speaker diarization using unsupervised discriminant analysis of inter-channel delay features," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, April 2009, pp. 4061–4064.

[7] P. P. Parada, D. Sharma, T. van Waterschoot, and P. A. Naylor, "Robust statistical processing of TDOA estimates for distant speaker diarization," in *Proc. European Signal Process. Conf. (EUSIPCO)*, Aug 2017, pp. 86–90.

[8] K. Nakamura and T. Mizumoto, "Blind spatial sound source clustering and activity detection using uncalibrated microphone array," in *2017 25th European Signal Processing Conference (EUSIPCO)*, Aug 2017, pp. 2438–2442.

[9] N. Ito, S. Araki, and T. Nakatani, "Data-driven and physical model-based designs of probabilistic spatial dictionary for online meeting diarization and adaptive beamforming," in *Proc. European Signal Process. Conf. (EUSIPCO)*, Aug 2017, pp. 1165–1169.

[10] J. H. DiBiase, *A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays*. Brown University, Providence RI, USA, May 2000.

[11] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[12] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.

[13] T. Minka, "Estimating a dirichlet distribution," https://tminka.github.io/papers/dirichlet/minka-dirichlet.pdf, 2012, [Online; accessed 28-October-2018].

[14] NIST speech recognition scoring toolkit. [Online]. Available: https://www.nist.gov/itl/iad/mig/tools