LATENT VARIABLE APPROACH TO DIARIZATION OF AUDIO RECORDINGS USING AD-HOC RANDOMLY PLACED MOBILE DEVICES

Srikanth Raj Chetupalli, Anirban Bhowmick, and Thippur V. Sreenivas

Dept. of ECE, Indian Institute of Science, Bangalore, 560012.

ABSTRACT
Diarization of audio recordings from ad-hoc mobile devices using spatial information is considered in this paper. A two-channel synchronous recording is assumed for each mobile device, which is used to compute directional statistics separately at each device in a frame-wise manner. The recordings across the mobile devices are asynchronous, but a coarse synchronization is performed by aligning the signals using acoustic events, or real-time clock. Direction statistics computed for all the devices, are then modeled jointly using a Dirichlet mixture model, and the posterior probability over the mixture components is used to derive the diarization information. Experiments on real life recordings using mobile phones show a diarization error rate of less than 14%.

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

1. INTRODUCTION
Consider a meeting scenario with several participants carrying mobile devices with one or more microphones. Mobile devices placed on a table can be considered to form an ad-hoc network of microphones. Spatial diversity provided by the multi microphone signals can be used to improve the performance of speech applications such as enhancement, recognition, diarization etc. However, such a setup is characterized by asynchronous recording at different devices, although microphones on the same device can record synchronously. A similar scenario is encountered in wireless acoustic sensor networks, where each node in the network can record using multiple microphones in a synchronous manner, but the signals at different nodes are asynchronous. In this paper, we first address the diarization task, i.e., “who spoke when?” in an audio recording comprising of multiple sources (speakers), and signals recorded from an ad-hoc microphone array network.

Methods based on spectral features, spatial features or a combination of both are proposed for multi-channel diarization of audio recordings. In this paper, we consider the diarization of audio recordings using spatial features alone. Several solutions have been proposed utilizing spatial features, which use the time-difference-of-arrival (TDOA) features. However, the estimation of TDOA is sensitive to reverberation and noise. An alternate formulation based on a probabilistic spatial dictionary and Watson mixture modeling of directional features is proposed. A pre-trained (data-driven) or pre-computed (physics based) spatial dictionary is used, which limits the application of the method to a finite set of source positions and known microphone geometry. Synchronous recording is assumed for all the microphones in the network in the above approaches. For the ad-hoc microphone network considered in this paper, microphones across the different mobile devices are asynchronous, and hence network-wide computation of TDOA or the directional features is not possible. However, it is possible to compute the directional features independently at each device, which can be combined later using a stochastic formulation.

Spatial response function computed using steered response power with the phase transform (SRP-PHAT) filtering is used as the spatialization measure. Assuming known microphone geometry at each device, the SRP response function is computed for a set of directions and these measures can be normalized to form a stochastic measure. We can use this as a “directional statistic” feature and directional statistics of several devices can be combined using a latent variable mixture model. We use a Dirichlet mixture model for this purpose. The signals from different devices can be coarsely aligned using specific acoustic event such as a clap or a tap, or network time. Expectation-maximization is then used for maximum likelihood estimation of the latent variables and the diarization information is derived from the posterior mixture component probabilities. Experiments on real life meetings recorded using commercial off-the-shelf mobile phones show that diarization error rate (DER) of less than 14% is possible, even with coarse synchronization across the devices.

2. PROBLEM FORMULATION
Consider a meeting scenario with S number of sources and P number of mobile devices placed on a table as shown in . Let each device record a M-channel audio signal. Let denote the speech signal recorded at the mth microphone of the pth device. Given the recordings at all the devices , the goal in this paper is to perform diarization, i.e., to identify “who spoke when?” in the long conversation.
The mobile devices record the audio signal using their own independent clocks and no other external synchronization is used. However, the audio recordings across the different devices can be synchronized coarsely using the real-time log of network time or using acoustic events such as a tap or a clap. The synchronous recordings at a particular device provide us with beam steering computation of the source direction information (source direction statistics) independently at each mobile device. We consider the computation of directional statistics in a frame by frame manner. Each mobile device (randomly placed) forms its own SRP function and hence some what different directional statistics than other devices. We explore joint modeling of the directional statistics obtained at each of the devices using a latent variable mixture model. Even though the mobile devices are placed arbitrarily and the information about their own position and orientation is unknown (hence we cannot combine the individual directional statistics in a geometric formulation), we explore the power of stochastic modeling to derive the directional information. We note that in the proposed approach to diarization, the goal is not the exact position of the source, but to use the directional information to identify the source activity along the recorded time-line. We show that this is possible using a stochastic formulation of several mobile phone derived directional data.

3. STATISTICAL DETECTION

3.1. Directional statistics features

Let us consider steered response power (SRP) approach to compute the spatial features at each time-frame $n$ for each randomly placed mobile device separately. Let $x[n, k] = [x_1[n,k] \ldots x_M[n,k]]^T$ denote the multi-channel speech signal in the short time Fourier transform (STFT) domain for a microphone array. Let $a[\theta, k]$ denote the steering vector corresponding to a source at a spatial direction $\theta$ for the frequency bin $k$ with respect to a local coordinate system centered at the array. Assuming free field propagation and a compact array, we have

$$a[\theta, k] = \left[e^{\frac{i2\pi k M_1(\theta)}{N}} \ldots e^{\frac{i2\pi k M_M(\theta)}{N}}\right]^T,$$

where $\{\tau_2(\theta), \ldots, \tau_M(\theta)\}$ denote the TDOA values at the $M-1$ microphones with respect to the first microphone. SRP method can compute the spatial response function as,

$$s[n, \theta] = \sum_{k=1}^K |a[\theta, k]^H x_f[n, k]|^2,$$

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

The response function $s[n, \theta]$ is evaluated at $L$ discrete angular positions $\Theta = \{\theta_1, \ldots, \theta_L\}$ 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,

$$\tilde{s}[n, \theta_t] = \alpha \tilde{s}[n, \theta_t] + (1 - \alpha)s[n - 1, \theta_t].$$

Smoothed SRP function is then normalized to represent the estimated directional statistics which is then used as feature for the mixture density modeling.

Let $s[n] \triangleq \frac{1}{L} [\tilde{s}[n, \theta_1] \ldots \tilde{s}[n, \theta_L]]^T$, where $C = \sum_{t=1}^L \tilde{s}[n, \theta_t]$ is the normalization constant. Thus the vector $s[n]$ is a positive function and sums up to unity; hence, we can interpret the $s[n]$ as a PMF over the set $\Theta$ at each time-frame $n$.

In the present formulation, we compute the directional statistics independently for each mobile device, and obtain $P$ number of directional statistic features $\{s_p[n]\}$, one per mobile device, at each time frame. However, due to reverberation in the enclosure and other recording noise, $\{s_p[n]\}$ do have estimation errors and hence a further statistical formulation is required to effectively combine the information from several recording devices.

3.2. Latent variable joint modeling

![Fig. 2. Generative model of (a) the microphone observations and (b) the directional statistics.](image)

We model the set of distributions $\{s_p[n]\}$, $0 \leq n \leq N - 1$ jointly using a mixture model. A graphical model describing the generation of observations is shown in Fig. 2. The latent variable selection vector $z_n$ selects a directional position...
where standard Dirichlet distribution has the form, which results in the model, We assume the directional data to be independent across time, i.e., \( P(\{z_{ns}[n]\} | z_{ns} = 1, \Delta) = \prod_{p=1}^{P} P(s_p[n] | \delta_{sp}), \) (4)

where \( \Delta = \{\delta_{sp}, \forall s, p\} \) is the set of parameters of all the distributions. A Dirichlet distribution [9] is used to model the directional statistics, to suit the discrete nature of the directional data and to suit the EM derivation. Hence,

\[
P(s_p[n] | \delta_{sp}) = \frac{\Gamma \left( \sum_{l=1}^{L} \delta_{sp[l]} \right)}{\prod_{l=1}^{L} \Gamma(\delta_{sp[l]})} \prod_{l=1}^{L} s_p[n, l]^{\delta_{sp[l]}-1}.
\]

We assume the directional data to be independent across time, which results in the model,

\[
P(S, Z | \Delta, \pi) = \prod_{n=0}^{N-1} \prod_{s=1}^{S} \pi_s \prod_{p=1}^{P} D(s_p[n] | \delta_{sp})^{z_{ns}}, \quad (7)
\]

The parameters \( \Delta \) and \( \pi \) 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 \( P(Z | S, \Delta^{(i)}, \pi^{(i)}) \), and (ii) maximization of the expected joint likelihood objective \( Q(\Delta, \pi) = \mathbb{E}[\log P(S, Z | \Delta, \pi)] \).

It can be shown that, \( P(Z | S, \Delta^{(i)}, \pi^{(i)}) \) is an independent Bernoulli distribution with parameter,

\[
P \left( z_{ns} = 1 | \{s_p[n]\}, \Delta^{(i)}, \pi^{(i)} \right) = \frac{\pi_s^{(i)} \prod_{p=1}^{P} D(s_p[n] | \delta_{sp}^{(i)})}{\sum_{s=1}^{S} \pi_s^{(i)} \prod_{p=1}^{P} D(s_p[n] | \delta_{sp}^{(i)})},
\]

and \( \mathbb{E}(z_{ns}) \triangleq \gamma_{ns}^{(i+1)} = P(z_{ns} = 1 | \{s_p[n]\}, \Delta^{(i)}, \pi^{(i)}) \).

In the maximization step, the function \( Q(\Delta, \pi) \) is maximized:

\[
Q(\Delta, \pi) = \sum_{n=0}^{N-1} \sum_{s=1}^{S} \gamma_{ns}^{(i+1)} \log \pi_s + \sum_{n=0}^{N-1} \sum_{s=1}^{S} \gamma_{ns}^{(i)} \sum_{p=1}^{P} \log D(s_p[n] | \delta_{sp}). \quad (9)
\]

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

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

Maximization of (9) with respect to \( \delta_{sp} \) requires solving the problem:

\[
\delta_{sp}^{(i+1)} = \arg \max_{\delta_{sp}} \sum_{n=0}^{N-1} \gamma_{ns}^{(i+1)} \log D(s_p[n] | \delta_{sp}). \quad (11)
\]

Substituting for \( D(s_p[n] | \delta_{sp}) \), we get the optimization problem as,

\[
\delta_{sp}^{(i+1)} = \arg \max_{\delta_{sp}} \sum_{n=0}^{N-1} \gamma_{ns}^{(i+1)} \left[ \log \left( \sum_{l=1}^{L} \delta_{sp[l]} \right) - \sum_{l=1}^{L} \log \left( \delta_{sp[l]} \right) + \sum_{l=1}^{L} (\delta_{sp[l]} - 1) \log s_p[l] \right]. \quad (12)
\]

Gradient-descent based algorithm is used to solve for \( \{\delta_{sp}, \forall s, p\} \) as shown in [12].

### 3.3. Diarization

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

\[
\hat{s}[n] = \arg \max_{s} \gamma_{ns}^{*}.
\]

### 4. EXPERIMENTS AND RESULTS

Real-life meeting recordings are used for the evaluation of the proposed scheme. Three mobile phones are placed in an arbitrary orientation on a table in a reverberant enclosure (RT60 ≈ 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 STFT to 8 KHz. The sound from a tap on the table is used as the acoustic event to align the signals across the mobile devices. We consider five recordings with three participants in each recording for evaluating diarization. The participants are chosen from three male speakers and one female speaker. The duration of the recordings varied from 5 minutes to 10 minutes, and the recordings are annotated at the speaker level manually. The mobile phones and participants are placed randomly for all the five recordings.

STFT analysis is carried out using frames of size 64 ns with 50% overlap between successive frames. In the SRP-PHAT computation, the beam steering is performed with a
resolution of 4° (L = 46). Computation of the steering vector used in SRP-PHAT requires knowledge of the spacing between the microphones. For the commercial devices used in this experiment, since exact spacing is unknown, 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 directional statistics is chosen to be 0.9. EM algorithm for DMM estimation is initialized using the method suggested in [12], and the maximum number of iterations is chosen to be 100. The number of sources S is assumed to be known in this experiment. However, it is possible to estimate the number of sources, by using the histogram of the peak locations of directional statistics features. The diarization performance is measured using DER, and computed using the NIST speech recognition scoring toolkit\(^1\) with a collar interval of 0.25 s. The proposed algorithm assigns each frame to a single source, 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 shows an illustration of the directional statistic features computed at the three mobile devices along with the spectrogram of the speech recorded at one of the devices for one of the recordings (illustrations for all the recordings are available here\(^2\)). We see that the spatial features of the sources differ at the three devices, and the discriminability between two source positions is also different. However, there is one correspondence between the feature patterns across the devices. For example, in the first mobile recording, the directional statistics contain a clear peak only for one of the sources, 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. However, joint modeling along with the other devices helps in estimating the correct source regions. Source posterior \{γ_{ns}\} 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 the computation of directional statistics.

![Fig. 3](Image)

**Fig. 3.** (a) Spectrogram of a microphone signal, and (b,c,d) computed directional statistics \{s_p[n]\} for the three mobile devices, (e) estimated posterior speaker probability \{γ_{ns}, s = 1, 2, 3\} shown in respective color, and (f) ground truth source activity denoted by three colors.

Table 1. 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 |

Performance of the proposed algorithm on the five recorded conversations is shown in Tab. 1. The performance varies across the different recordings, due to the different source and microphone placements, and different amounts of overlap between the sources during the conversation. DER is found to be high for some conversations with more overlap. However, for all the recordings, the performance of the proposed algorithm is within 2% from the oracle performance.

## 5. CONCLUSIONS

Coarse synchronization of different mobile devices and joint modeling of directional statistics computed per device is found to be sufficient for identifying "who spoke when?" in audio recordings from randomly placed mobile devices. This is true despite the unknown variabilities such as the nature of the microphones, their orientation within different mobile devices and also the random placement of the mobile devices. 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.

\(^1\)https://www.nist.gov/itl/iad/mig/tools
\(^2\)http://www.ece.iisc.ernet.in/~sraj/mDiar.html
6. REFERENCES

[1] 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.

[2] X. Anguera, S. Bozonnnet, 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.

[3] M. Moattar and M. Homayounpour, “A review on speaker diarization systems and approaches,” *Speech Communication*, vol. 54, no. 10, pp. 1065 – 1103, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639312000696

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] C. M. Bishop, *Pattern recognition and machine learning*. Springer, 2006.

[10] 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.

[11] 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.

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