Data-driven and Physical Model-based Designs of Probabilistic Spatial Dictionary for Online Meeting Diarization and Adaptive Beamforming

Nobutaka Ito Shoko Araki Tomohiro Nakatani
NTT Communication Science Laboratories, NTT Corporation
2-4, Hikaridai, Seika-cho, "Keihanna Science City" Kyoto 619-0237 Japan
Email: {ito.nobutaka, araki.shoko, nakatani.tomohiro}@lab.ntt.co.jp

***Abstract*—In this paper, we comparatively study alternative dictionary designs for recently proposed meeting diarization and adaptive beamforming based on a *probabilistic spatial dictionary*. This dictionary models the feature distribution for each possible direction of arrival (DOA) of speech signals and the feature distribution for background noise. The dictionary enables online DOA detection, which in turn enables online diarization. Here we describe *data-driven* and *physical model-based* designs of the dictionary. Experiments on a meeting dataset showed that a physical model-based dictionary gave a word error rate (WER) of 24.9 %, which is close to that for the best-performing data-driven dictionary (24.1 %). Therefore, the former has a significant advantage over the latter that it allows us to bypass the cumbersome measurement of training data without much degrading the performance of the automatic speech recognition (ASR).**

I. Introduction

Despite extensive research devoted to ASR, meeting ASR with distant microphones still remains a challenge. Meeting recordings with distant microphones usually contain not only desired speech but also other speakers' speech, background noise, and reverberation, all of which degrade the ASR performance significantly. Therefore, speech enhancement such as beamforming is necessary to suppress these unwanted components. Furthermore, in a meeting, multiple speakers speak at an arbitrary moment. Therefore, the estimation of speech intervals for each speaker (*i.e.*, diarization) is also significant. This paper deals with diarization and beamforming for meeting ASR in noisy, reverberant environments.

For robust ASR of a single (*i.e.*, not overlapped) speaker, mask-based adaptive beamforming [1]–[3] has recently turned out to be highly effective. This approach was employed in the best-performing system [2], [4] in CHiME-3 [5] and CHiME-4. The approach utilizes masks to learn an adaptive beamformer (*e.g.*, the minimum variance distortionless response (MVDR) beamformer [6]) from noisy observed signals. The masks indicate which signal (the speech signal or background noise in this case) dominates each time-frequency component of the observed signals.

The approach has also been extended to meetings involving multiple speakers [7]. Not only masks but also diarization information is utilized to learn a beamformer for each speaker from a meeting recording containing multiple speakers' speech. This is realized by learning a beamformer for each speaker using only the speech interval of that speaker indicated by the diarization information. For diarization, we proposed a probabilistic spatial dictionary, which models the feature distribution for each possible speaker DOA (see Fig. 1) and the feature distribution for background noise. The dictionary enables online DOA detection, which in turn enables online diarization.

Here we describe data-driven and physical model-based designs of the probabilistic spatial dictionary, and compare them in terms of ASR performance. The data-driven dictionary is pre-trained using measured training data, while the physical model-based dictionary is prepared using the array geometry only. We show that the latter has a significant advantage over the former that it allows us to bypass the cumbersome measurement without degrading the ASR performance much.

The rest of this paper is organized as follows. Section II briefly reviews diarization and adaptive beamforming based on the probabilistic spatial dictionary [7]. Section III describes the data-driven and the physical model-based dictionaries. Section IV compares these dictionaries in terms of the ASR performance, and finally Section V concludes this paper.

II. Review: Diarization and Adaptive Beamforming Based on Probabilistic Spatial Dictionary [7]

In this section, we briefly review meeting diarization and adaptive beamforming based on the probabilistic spatial dictionary [7], while the readers are referred to [7] for the details. Figure 2 shows the overall flow of this method.

A. Feature Extraction

In *feature extraction*, a source location feature vector $\boldsymbol{z}_{tf}$ is extracted from observed signals at each time-frequency point (t, f). Let us denote by $\boldsymbol{y}_{tf} = \begin{bmatrix} y_{tf}^{(1)} & \dots & y_{tf}^{(M)} \end{bmatrix}^{\mathsf{T}}$ the observed signals at all microphones in the short-time Fourier transform domain. Here, $y_{tf}^{(m)}$ denotes the observed signal at the mth microphone; M the number of microphones; $^{\mathsf{T}}$ the transposition. We here focus on the following feature vector for simplicity [8]:

$$\boldsymbol{z}_{tf} = \frac{\boldsymbol{y}_{tf}}{\|\boldsymbol{y}_{tf}\|}. \tag{1}$$

Here, $\|\cdot\|$ denotes the Euclidean norm.

ISBN 978-0-9928626-7-1 © EURASIP 2017

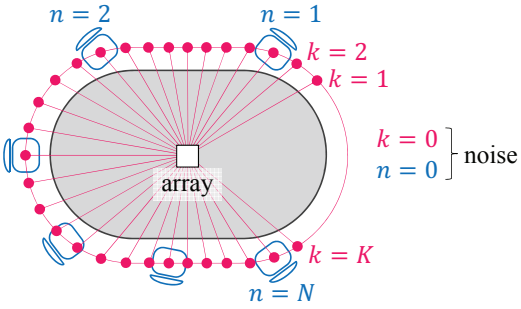

Fig. 1. The index k indicates potential speaker locations, and the index n seats, except for $k = 0$ and $n = 0$ indicating background noise.

B. Modeling Features Based on Probabilistic Spatial Dictionary

The probability distribution of $\boldsymbol{z}_{tf}$ for each of K potential speaker locations and background noise (see Fig. 1) is modeled by a complex Watson distribution

$$\mathcal{W}\left(\boldsymbol{z}_{tf}; \boldsymbol{a}_f^{(k)}, \kappa_f^{(k)}\right) \propto \exp\left(\kappa_f^{(k)} \left|\boldsymbol{a}_f^{(k)\mathsf{H}} \boldsymbol{z}_{tf}\right|^2\right) \tag{2}$$

proposed by Mardia *et al.* [9], where $^{\mathsf{H}}$ denotes the Hermitian transposition. The values $k = 1, 2, \ldots, K$ correspond to the potential speaker locations, and the value $k = 0$ to the background noise. $\boldsymbol{a}_f^{(k)}$ and $\kappa_f^{(k)}$ represent the centroid and the concentration of the distribution, and $\left\|\boldsymbol{a}_f^{(k)}\right\| = 1$. These parameters are given, and so are

$$\mathcal{W}\left(\boldsymbol{z}_{tf}; \boldsymbol{a}_f^{(k)}, \kappa_f^{(k)}\right), \ k = 0, 1, \ldots, K, \tag{3}$$

which we call a *probabilistic spatial dictionary*. See Section III regarding how to prepare $\boldsymbol{a}_f^{(k)}$ and $\kappa_f^{(k)}$.

The observed feature vector $\boldsymbol{z}_{tf}$ is modeled by a mixture of the distributions (3) as

$$p(\boldsymbol{z}_{tf}) = \sum_{k=0}^{K} \alpha_t^{(k)} \mathcal{W}\left(\boldsymbol{z}_{tf}; \boldsymbol{a}_f^{(k)}, \kappa_f^{(k)}\right), \tag{4}$$

where $\alpha_t^{(k)}$ is an *unknown, time-varying* mixture weight [10] satisfying $\sum_{k=0}^{K} \alpha_t^{(k)} = 1$. These mixture weights are estimated from the observed feature vector (see Section II-C) and utilized for diarization and adaptive beamforming (see Sections II-D–II-F).

Though Tran Vu *et al.* [11] have also employed a complex Watson mixture model, our modeling differs from theirs in the following aspects. First, $\boldsymbol{a}_f^{(k)}$ and $\kappa_f^{(k)}$ in (4) are given. Second, each mixture component in (4) corresponds to a potential speaker location instead of a speaker. Third, $\alpha_t^{(k)}$ in (4) depends on t, and can thus be utilized for diarization (see Section II-D).

C. Weight Estimation

In *weight estimation*, $\alpha_t^{(k)}$, $k = 0, 1, \ldots, K$, are estimated by using $\boldsymbol{z}_{tf}, f = 1, 2, \ldots, F$, and the dictionary (3). This is performed by the maximization of the likelihood function $\prod_{f=1}^{F} p(\boldsymbol{z}_{tf})$ based on, *e.g.*, gradient ascent. Here, F denotes the number of frequency bins up to the Nyquist frequency.

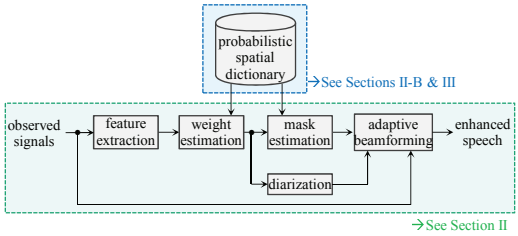

Fig. 2. Overall flow of meeting diarization and adaptive beamforming based on the probabilistic spatial dictionary.

D. Diarization

In *diarization*, binary variables $d_t^{(n)}$, $n = 1, 2, \ldots, N$, indicating whether there was voice activity from each seat are computed based on peak picking of the mixture weights. Here, n denotes the seat index, and N the number of seats (see Fig. 1). The values $d_t^{(n)} = 1$ and $d_t^{(n)} = 0$ indicate the presence and the absence of voice activity from the nth seat in the tth frame.

E. Mask Estimation

In *mask estimation*, masks $\mathcal{M}_{tf}^{(n)}$, $n = 0, 1, \ldots, N$, are estimated by using the mixture weights and the dictionary (3). The values $n = 1, 2, \ldots, N$ correspond to the seats, and the value $n = 0$ to the background noise.

F. Adaptive Beamforming

In *adaptive beamforming*, the steering vectors for the N seats are estimated based on the masks and the diarization result, MVDR beamformers are designed based on the steering vectors, and enhanced speech signals are obtained based on the MVDR beamformers.

III. Alternative Dictionary Designs

In this section, we present alternative designs of the probabilistic spatial dictionary, namely data-driven and physical model-based designs.

A. Data-driven Dictionary [7]

When a multichannel speech recording for each potential speaker location and a multichannel noise recording are available, the probabilistic spatial dictionary can be trained by using these data. We call such a dictionary a *data-driven* dictionary.

Let $\boldsymbol{y}_{tf}^{(k)}, k = 1, 2, \ldots, K$, be the multichannel speech recordings, and $\boldsymbol{y}_{tf}^{(0)}$ be the multichannel noise recording. Let $\boldsymbol{z}_{tf}^{(k)} = \frac{\boldsymbol{y}_{tf}^{(k)}}{\|\boldsymbol{y}_{tf}^{(k)}\|}$ be the feature vector (1) corresponding to $\boldsymbol{y}_{tf}^{(k)}$. The parameters $\boldsymbol{a}_f^{(k)}$ and $\kappa_f^{(k)}$ are estimated by maximizing the likelihood function $\prod_{t=1}^{T} \mathcal{W}\left(\boldsymbol{z}_{tf}^{(k)}; \boldsymbol{a}_f^{(k)}, \kappa_f^{(k)}\right)$, with T being the number of frames. This can be done for each k and f by the following algorithm, whose derivation is similar to [10] and omitted:

ISBN 978-0-9928626-7-1 © EURASIP 2017

1) Compute the empirical covariance matrix $\boldsymbol{R}_f^{(k)}$ of $\boldsymbol{z}_{tf}^{(k)}$ by $\boldsymbol{R}_f^{(k)} \leftarrow \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{z}_{tf}^{(k)} \boldsymbol{z}_{tf}^{(k)\mathsf{H}}$.
2) Let $\boldsymbol{a}_f^{(k)}$ be a principal eigenvector of $\boldsymbol{R}_f^{(k)}$ normalized so that $\left\|\boldsymbol{a}_f^{(k)}\right\| = 1$.
3) $\kappa_f^{(k)} \leftarrow \frac{M\lambda_f^{(k)}-1}{2\lambda_f^{(k)}\left(1-\lambda_f^{(k)}\right)}\left[1+\sqrt{1+\frac{4(M+1)\lambda_f^{(k)}\left(1-\lambda_f^{(k)}\right)}{M-1}}\right]$, where $\lambda_f^{(k)}$ denotes the principal eigenvalue of $\boldsymbol{R}_f^{(k)}$

B. Physical Model-based Dictionary

Alternatively, the probabilistic spatial dictionary can also be designed based on a physical model.

1) Physical Model-based Speech Dictionary: The speech dictionary (*i.e.*, $k = 1, 2, \ldots, K$) can be designed based on an anechoic propagation model, such as planewave and spherical wave models. Underlying assumptions are that reverberation is negligible, and that the array geometry is given. In the following, we focus on the planewave case.

If the speakers are in the far field of the array, a planewave propagation model can be employed [12]. In this case, $\boldsymbol{a}_f^{(k)}, k = 1, 2, \ldots, K$, can be computed by

$$\boldsymbol{a}_f^{(k)} = \frac{1}{\sqrt{M}} \begin{bmatrix} \exp\left(\sqrt{-1}\omega_f \delta^{(1,k)}\right) \\ \vdots \\ \exp\left(\sqrt{-1}\omega_f \delta^{(M,k)}\right) \end{bmatrix}. \tag{5}$$

Here, $\delta^{(m,k)}$ denotes the time difference of arrival (TDOA) between the mth microphone and the first microphone (reference) for the kth potential speaker location, $\omega_f = 2\pi f_\mathrm{s}(f-1)/F$ the angular frequency, and f_s the sampling frequency. The TDOA $\delta^{(m,k)}$ is computed by

$$\delta^{(m,k)} = \frac{\boldsymbol{e}^{(k)\mathsf{T}} \boldsymbol{r}^{(m)}}{c}, \tag{6}$$

where $\boldsymbol{e}^{(k)}$ denotes the given unit vector in the propagation direction of a speech signal from the kth potential speaker location, $\boldsymbol{r}^{(m)} \in \mathbb{R}^3$ the given Cartesian coordinates of the mth microphone with the first microphone being the origin, and c the sound velocity.

On the other hand, we determined $\kappa_f^{(k)}$ using heuristics in the experiments in Section IV. Since the feature vector (1) tends to have a smaller concentration (or, equivalently, a larger variance) in high frequencies, we determined $\kappa_f^{(k)}$ by the following inverse square law:

$$\kappa_f^{(k)} = \frac{A}{f^2}. \tag{7}$$

A is a dimensionless proportionality constant, which is to be tuned experimentally.

2) Physical Model-based Noise Dictionary [7]: The noise dictionary (*i.e.*, $k = 0$) can be designed based on an isotropic noise model: noise comes from every direction uniformly. Since the feature vector (1) has a small concentration (or, equivalently, a high variance) for such noise, we let

$$\kappa_f^{(0)} = 0, \tag{8}$$

for which the complex Watson distribution reduces to the uniform distribution on the hypersphere [9]. We let $\boldsymbol{a}_f^{(0)}$ be an arbitrary unit vector, since it does not affect the distribution.

Note that the above assumption of a uniform noise distribution is equivalent to that of spatially white noise. Although an isotropic noise is spatially colored in general, it becomes a spatially white noise asymptotically when the distances between microphones are much larger than the wavelength [13].

C. Pros and Cons of Two Dictionary Designs

Since the data-driven dictionary learns the characteristics of the environment, such as the acoustic transfer characteristics and the noise characteristics, a high performance is expected when the environments in training and testing are matched. On the other hand, collection of training data for each potential speaker location and noise is a cumbersome task. Contrarily, the physical model-based dictionary allows us to bypass the cumbersome data collection, while it may not be as effective as a matched data-driven dictionary because of model imperfection. The ASR performance for each design is evaluated in Section IV.

IV. Experimental Evaluation

A. Datasets

To train the data-driven dictionary, we employed a dataset composed of a multichannel speech recording for each potential speaker location, and a multichannel noise recording. The speech recordings were generated by convolving dry speech signals with measured room impulse responses, while the noise recording were made with babble noise played from ten loudspeakers outside the room.

We employed three meeting datasets recorded in different rooms:

- an office room next to an exhibition hall;
- a quiet office room;
- a sound-proof room.

The readers are referred to [7] for more details. Only the first dataset was used for evaluation. This dataset contains babble noise from outside the room and from standing audience with a signal-to-noise ratio of 3–15 dB, which simulates an exhibition situation. The reverberation time for this dataset was 500 ms. The number of speakers was four to six depending on sessions, and the array geometry was cubic with a side length of 4 cm. The other two datasets were used for ASR training only.

B. Front-end Configuration

We considered $K = 17$ potential speaker locations around a table (see Fig. 1). The parameter A in (7) were tuned on the development set of meeting data at $A = 12000$. The front-end processing was performed in an online manner.

ISBN 978-0-9928626-7-1 © EURASIP 2017

C. Back-end Configuration

We employed a DNN-HMM acoustic model [14] with seven hidden layers with 2048 units each. The input to the DNN consisted of 40 log-mel filterbank coefficients and their delta and acceleration with five left and five right context frames, which amount to a 1320-dimensional feature vector. The output of the DNN consisted of 4100 HMM states. The DNN was first trained on the clean training set of Corpus of Spontaneous Japanese (CSJ) [15], then retrained on the headset recordings in the training set of the three meeting datasets in Section IV-A, and finally retrained on all channels of the array recordings in the training set of the first meeting dataset in Section IV-A.

As for the language model, we employed a Kneser-Ney smoothed word trigram [16], trained on the CSJ, the training sets of the three meeting datasets in Section IV-A, and topic-related WWW data. The mixture weights were determined based on perplexity minimization on the development set of the meeting data.

We used manual annotation for VAD for ASR. Unlike the front-end processing, the back-end processing was performed in an utterance batch manner.

D. Results

Table I shows the WER on the evaluation set for the data-driven and physical-model based dictionaries. For reference, the WER was 18.8% for headset microphones, 40.5% for a distant microphone, and 53.3% for our conventional meeting speech enhancement method [17]. The data-driven design gave a lower WER for the speech dictionary, while the physical model-based design for the noise dictionary. The best-performing combination of the data-driven speech dictionary and the physical model-based noise dictionary achieved a WER reduction of $(0.405 - 0.241)/0.405 \times 100 = 40.5(\%)$ relative to the distant microphone. The physical model only case (the rightmost column in Table I) achieved a WER reduction of 38.5% relative to the distant microphone.

The lower WER for the data-driven speech dictionary is attributed to the facts that it models not only the direct path but also reflections on the walls, and that it was trained on the matched environment. It is interesting to note that the physical model-based design was more effective for the noise dictionary. This is probably because of environment mismatch between training and testing: the test set contains a variety of noise conditions, such as an open/close door and a high/low loudspeaker volume, while the training data for noise dictionary design was recorded in a specific condition with the door open. Interestingly, the physical model only case gave a WER very close to that for the best-performing design.

TABLE I
WER (%) FOR ALTERNATIVE DICTIONARY DESIGNS. 'D' STANDS FOR THE DATA-DRIVEN DICTIONARY, AND 'P' STANDS FOR THE PHYSICAL MODEL-BASED DICTIONARY.

speech dictionary	D	D	P	P
noise dictionary	D	P	D	P
WER	24.2	**24.1**	26.3	24.9

V. Conclusion

In this paper, we compared data-driven and physical model-based designs of the probabilistic spatial dictionary. The experiments have shown that the data-driven and the physical model-based designs gave a lower WER for the speech and the noise dictionaries, respectively. The experiments also showed that the design based solely on physical models gave a WER very close to that for the best-performing design. This greatly enhances the applicability of the proposed diarization and adaptive beamforming method, because we can bypass the cumbersome data collection with only little performance loss.

References

[1] M. Souden, S. Araki, K. Kinoshita, T. Nakatani, and H. Sawada, "A multichannel MMSE-based framework for speech source separation and noise reduction," *IEEE Trans. ASLP*, vol. 21, no. 9, pp. 1913–1928, Sept. 2013.

[2] T. Yoshioka, N. Ito, M. Delcroix, A. Ogawa, K. Kinoshita, M. Fujimoto, C. Yu, W.J. Fabian, M. Espi, T. Higuchi, S. Araki, and T. Nakatani, "The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices," in *Proc. ASRU*, Dec. 2015, pp. 436–443.

[3] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. ICASSP*, Mar. 2016.

[4] J. Du, Y.-H. Tu, L. Sun, F. Ma, H.-K. Wang, J. Pan, C. Liu, J.-D. Chen, and C.-H. Lee, "The USTC-iFlytek system for CHiME-4 challenge," in *Proc. CHiME2016*, 2016.

[5] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *Proc. ASRU*, Dec. 2015, pp. 504–511.

[6] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*, Springer, Berlin, Heidelberg, 2001.

[7] N. Ito, S. Araki, M. Delcroix, and T. Nakatani, "Probabilistic spatial dictionary based online adaptive beamforming for meeting recognition in noisy and reverberant environments," in *Proc. ICASSP*, Mar. 2017.

[8] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Trans. ASLP*, vol. 19, no. 3, pp. 516–527, Mar. 2011.

[9] K.V. Mardia and I.L. Dryden, "The complex Watson distribution and shape analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, no. 4, pp. 913–926, 1999.

[10] N. Ito, S. Araki, and T. Nakatani, "Permutation-free convolutive blind source separation via full-band clustering based on frequency-independent source presence priors," in *Proc. ICASSP*, May 2013, pp. 3238–3242.

[11] D.H. Tran Vu and R. Haeb-Umbach, "Blind speech separation employing directional statistics in an expectation maximization framework," in *Proc. ICASSP*, Mar. 2010, pp. 241–244.

[12] D.H. Johnson and D.E. Dudgeon, *Array Signal Processing: Concepts and Techniques*, Prentice Hall, Englewood Cliffs, NJ, 1993.

[13] N. Ito, H. Shimizu, N. Ono, and S. Sagayama, "Diffuse noise suppression using crystal-shaped microphone arrays," *IEEE Trans. ASLP*, vol. 19, no. 7, pp. 2101–2110, Sept. 2011.

[14] M. Delcroix, Y. Kubo, T. Nakatani, and A. Nakamura, "Is speech enhancement pre-processing still relevant when using deep neural networks for acoustic modeling?," in *Proc. Interspeech*, 2013, pp. 2992–2996.

[15] S. Furui, K. Maezawa, and H. Isahara, "A Japanese national project on spontaneous speech corpus and processing technology," in *Proc. ISCA ASR*, 2000.

[16] R. Kneser and H. Ney, "Improved backing-off for m-gram language modeling," in *Proc. ICASSP*, 1995, pp. 181–184.

[17] 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

ISBN 978-0-9928626-7-1 © EURASIP 2017

understanding using distant microphones and omni-directional camera,"
IEEE Trans. ASLP, vol. 20, no. 2, pp. 499–513, Feb. 2012.

ISBN 978-0-9928626-7-1