Bayesian time-domain multiple sound source localization for a stochastic machine

1st Raphael Frisch
Univ. Grenoble Alpes, LIG, France
38000 Grenoble, France
raphael.frisch@univ-grenoble-alpes.fr

2nd Marvin Faix
Univ. Grenoble Alpes, LIG, France
38000 Grenoble, France
marvin.faix@inria.fr

3rd Jacques Droulez
ISIR: CNRS/Sorbonne Univ.
75005 Paris, France
jacques.droulez@isir.upmc.fr

4th Laurent Girin
Univ. Grenoble Alpes, Grenoble-INP, GIPSA-Lab,
38000 Grenoble, France
laurent.girin@gipsa-lab.grenoble-inp.fr

5th Emmanuel Mazer
Univ. Grenoble Alpes, LIG, France
38000 Grenoble, France
emmanuel.mazer@inria.fr

***Abstract*—We propose a time-domain multiple sound source localization (SSL) method based on Bayesian inference. This method is specifically designed to run on the stochastic machines (SM) that we are currently developing to perform efficient low-level sensor signal processing with ultra-low power consumption. The proposed SSL method is divided into two main parts. First, a probabilistic model is run on 50 very short time frames (3.75ms each) of multichannel recorded signals. Second, the results obtained on the different frames are fused to obtain a final localization map. Using the system in a supervised way allows to extract estimated source locations by selecting as many maxima as there are sources in the room. We explain how this method is implemented on a SM. Experiments are presented to illustrate the performance and robustness of the resulting system.**

***Index Terms*—Multiple sound source localization, time-domain processing, Bayesian stochastic machine, specific hardware.**

I. Introduction

The long-term goal of the research project this study is part of is the development of efficient computational architectures dedicated to Bayesian inference, called stochastic machines (SMs) (or Bayesian machines), and their application to practical low-level signal processing problems such as sound source localization (SSL) and sound source separation (3S) in very-low-consumption embedded devices. These computational architectures must take advantage of new nano-devices to solve probabilistic inference problems, while consuming as little energy as possible. Within the former Bambi European project,[1] we have designed several prototypes of such machines, e.g. [4], [6], leading to a first generation of stochastic machines. In the ongoing MicroBayes project,[2] applications to realistic problems such as SSL and 3S are tackled.

Mono-source localization with a stochastic machine applied to signals pre-processed in the time-frequency (TF) domain was presented in [8] and deeply analyzed in [7]. However, one keypoint of stochastic machines is that they should avoid as much as possible signal pre-processing and focus directly on the Bayesian inference process. In the method presented in [8], the pre-processing part including the Fourier transform and the feature computation takes up to 35% of the final circuit. Therefore, it is important to avoid this pre-processing. This is because i) their architecture is not necessarily well suited for the pre-processing part, and ii) we want to limit energy consumption. For those reasons, we present in this paper a Bayesian multiple source (in the present case, two speakers) localization method that works in the time domain and that is directly applicable to the stochastic machines introduced above.

A number of time-domain SSL methods exist in the literature, e.g. methods based on signals time difference of arrival (TDOA) [3]. They are actually part of fundamental microphone array processing techniques [1], [2]. However, they do not perform well when several sources overlap. Some researchers have used very short analysis windows, e.g. a 10 ms-window in [11]. Most state-of-the-art multiple sound source localization (and separation) methods work in the TF domain, generally obtained by applying the short-time Fourier transform (STFT) on the microphone signals, see e.g. among many others [5], [12]–[14], [16]. Those methods exploit the so-called audio signal sparsity in the TF domain, i.e. each TF bin is assumed to be dominated by one single dominant sound source [15]. They are thus more efficient than time-domain methods, where sparsity is more difficult to exploit (we have to detect time slots where only one source is active).

As explained in more details in Section II, the method proposed in the present paper exploits a very specific form of time-domain sparsity in the case of a two-speaker mixture, by evaluating Bayesian evidence values on very short time-frames. Integration of localization results over several such frames leads to multiple speaker localization. Since the evidence values are directly computed from time-domain sensor signals, this method is directly applicable to stochastic machines, as targeted in our project. To our knowledge, this is the first time-domain multi-source SSL method working with very-short time frames and designed to be implemented in

[1] https://www.bambi-fet.eu/

[2] https://persyval-lab.org/en/sites/content/microbayes

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

stochastic machines.

II. Multi-source localization method

In this section, we first present the probabilistic model used to perform the localization for a single (very) short time frame. Then, we present the fusion of localization results over several frames for the multi-source localization. Finally, the implementation in the Bayesian stochastic machine is explained.

A. Probabilistic model for single-frame single-source SSL

Let $\boldsymbol{X}$ be the $I \times N$ matrix concatenating the microphone signals, with I being the number of microphones and N being the number of signal (time) samples in the considered frame. $x_{i,n}$ denotes the n-th sample of signal (in the current frame) recorded at microphone i, with $i \in \{1, \ldots, I\}$ and $n \in \{1, \ldots, N\}$. The signals are recorded at 8-kHz sampling frequency and 8-bit quantization. $\boldsymbol{S}$ represents a set of candidate positions of the sound sources in the room, placed on a 2D regular grid, and s is an index representing a candidate source position on that grid.

The probabilistic model used for the sound source localization for a single time frame is classically based on the TDOA and amplitude attenuation between the signals recorded by several microphones placed in the room, which locations are assumed to be known. In this work, an anechoic free-field source-to-microphone propagation model is assumed. One microphone is chosen as a reference for TDOA and relative attenuation of the other microphones. Without loss of generality, let this reference microphone be microphone 1. Given the source location s, and the locations of microphone 1 (reference) and microphone i, one can compute the "ideal" signal TDOA in between the two microphones expressed in number of samples $\tau_i(s) = (d(s,i) - d(s,1))F_s/c$, where $d(s,i)$ is the distance between source and microphone i, c is the sound celerity, and F_s is the sampling frequency, as well as the attenuation factor $a(s,i)$ according to the model in Equation (1) in [9]. The following probabilistic formulation allows us to take into account model approximations and the various sources of noise, including recording noise and ambient noise. For a given source position s and a given signal at the reference microphone (denoted by $\boldsymbol{x}_1$), the distribution of the signal at microphone i is given by:

$$P(x_{i,n}|s, \boldsymbol{x}_1) = \mathcal{N}\Big(x_{i,n}; \frac{a(s,i)}{a(s,1)} x_{1,n-\tau_i(s)}, \sigma^2\Big), \qquad (1)$$

where $\mathcal{N}(\mu, \sigma^2)$ denotes the Gaussian distribution with mean μ and variance σ^2. The latter is set to an arbitrary fixed value in the above model. The Gaussian distribution is used since it is a good general model for modeling uncertainty and it is easily implementable in hardware.

Then, we assume that given the source location and the reference microphone signal, the signals at all other microphones are independent. We also assume that the signal at the reference microphone $\boldsymbol{x}_1$ is independent of source location (which is reasonable if we consider that the relative spatial information across different microphone signals is correlated to source location, but the speech content of each microphone signal taken individually is not). We thus have:

$$P(\boldsymbol{X}, s) = P(s)P(\boldsymbol{x}_1) \prod_{i=2}^{I} \prod_{n=1}^{N} P(x_{i,n}|s, \boldsymbol{x}_1). \qquad (2)$$

$P(s)$ is assumed to be a uniform distribution over the candidate source locations. $P(\boldsymbol{x}_1)$ is an unknown distribution. However, it is not needed for the inference. Using Bayes' theorem, the posterior distribution of source location s given the microphone signals $\boldsymbol{X}$ is given by:

$$P(s|\boldsymbol{X}) \propto \prod_{i=2}^{I} \prod_{n=1}^{N} P(x_{i,n}|s, \boldsymbol{x}_1), \qquad (3)$$

where each conditional likelihood $P(x_{i,n}|s, \boldsymbol{x}_1)$, aka *evidence* in this context, is given by (1). Source location s is estimated on a very short time-frame basis by finding the maximum value of the product in the r.h.s. of (3).

A keypoint here is that our stochastic machine computes the posterior probability values (3) for all candidate values of s on the grid $\boldsymbol{S}$, in parallel and in a very efficient/rapid manner, as explained in the upcoming sections. Indeed, the inference is basically done by multiplication of the different evidences, and our SM is precisely dedicated to perform matrix multiplications on probabilistic variables in a very efficient way, using AND-gates [7], [8].

Another keypoint is that, as briefly stated in the introduction and as confirmed by our experiments, this approach to SSL works for a reasonable number of very short frames even if the microphone signals are composed of overlapping speech signals produced by two speakers speaking simultaneously from two different locations. Indeed, for a reasonable amount of such frames, the speech signal energy from one speaker has much larger energy than the signal from the other speaker, even if the speakers are speaking simultaneously. This is because a speech signal is a centered signal with fluctuating energy. For example, in a vowel, some successive samples have high energy, corresponding to a vocal fold pulse, and some successive samples have low energy, corresponding to the end of the vocal tract response to the pulse (before the next pulse reinjects energy). Therefore, at many occasions a few successive speech samples with high energy produced by one speaker correspond to a few successive samples with low energy produced by the other speaker, and vice versa. For very short time-frame containing such portions of speech signals, the proposed localization method will work well.

In practice, we have to find a trade-off between limiting the number of samples (for the above assumption to remain valid) and ensuring robust calculation of the posterior (3). A short window of $N = 30$ samples, i.e. 3.75 ms at 8-kHz sampling rate, was selected in our experiments. This indeed corresponds to less than one period of voiced speech with fundamental frequency within 100-200 Hz. Note that this comes in contrast with the usual short-term frames used in the STFT-based SSL methods and in speech/audio analysis in general (typically within 20-30 ms).

Of course, frame-wise localization will not work well on frames where the signals from the two speakers are both of either low or high energy. However, the fusion process described in the next section deals with this problem.

B. Fusion of frame-wise results for multiple-source SSL

The next step of the proposed SSL method is the fusion of the information provided by the probabilistic model on different very short time-frames, in order to provide multi-speaker localization. Let F be the number of frames used to perform this process. Let $\boldsymbol{X}_f$ denote the matrix of microphone signals at frame f.

First, as stated in the previous subsection, a frame-wise source location estimate $\hat{s}_f$ is computed for each frame f, as the candidate location which has the maximum posterior probability in that frame:

$$\hat{s}_f = \underset{s \in \boldsymbol{S}}{argmax} P(s|\boldsymbol{X}_f). \tag{4}$$

Then a global distribution map P_{glob} is created which counts for each candidate position $s \in S$, how many times it was selected as the most probable position $\hat{s}_f$ over all F frames. Typically, this counting process is done for a "reasonably large" number of frames, $F = 50$ in our experiments. Moreover, we spaced the frames with an interval of 70 samples (8.75 ms) in order to maximize the chance to capture different configurations of speech signal mixtures from the two speakers (see previous subsection). A bloc of microphone signal used to perform multi-source localization thus represents 0.625 s, which is quite reasonable for such task.

As a result of choosing a sufficiently large number of very short time-frames, the obtained global distribution P_{glob} is in general very peaky and the locations of the two sources are clearly visible on the map (examples are provided in Section III-B). Final multi-source SSL then simply consists in selecting the two predominant peaks in the map. Note that, by doing so, we adopt a supervised mode for SSL, since we assume a priori that there are two sources in the scene and we thus select two peaks. The alternative way is the unsupervised mode where a threshold is set for peak selection, see e.g. [12]. This latter approach has the advantage to automatically provide an estimate of the number of sources, which is generally unknown in practice, but it has the drawback of being sensible to the threshold setting. In the present study, because two peaks corresponding to the two speakers are generally predominant in P_{glob}, the boundary between the supervised and unsupervised modes gets thin.

After each multi-source localization on a bloc of F frames, the map P_{glob} is reset to zero, and a new counting is processed on the next bloc.

C. Implementation on the stochastic machine

In this section, the stochastic machine used to perform the inference of the probabilistic model is presented. Since the architecture is based on stochastic computing, the evidences which are multiplied in the inference equation (3) are represented as stochastic bit streams. Stochastic computing allows to easily compute the multiplication between two probabilities represented as stochastic bit streams using an AND-gate. Then, the calculation of (3) is done by a cascade of AND-gates.

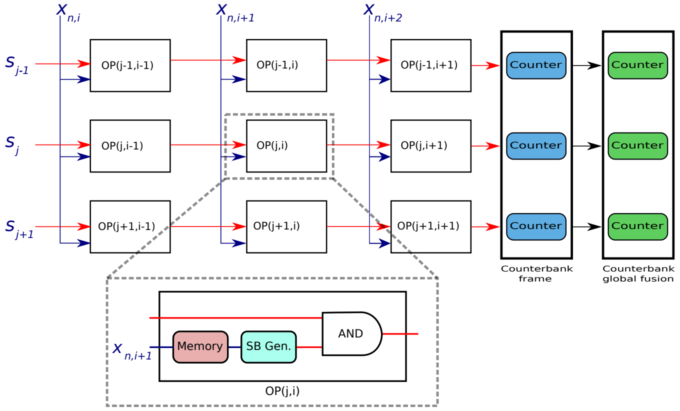


Fig. 1: Simplified representation of the Bayesian machine architecture used to compute the inference equation for SSL.

For this aim, the SM is structured as a $S \times (N \times (I-1))$ matrix, where $S = \text{card}(\boldsymbol{S})$ is the number of source location candidates on the grid. Fig. 1 shows a schematic representation of the SM architecture for 3 lines and 3 columns (for all 3 microphones i, $i+1$ and $i+2$ at a given time sample n). Each line of the machine computes (3) for a specific value of source location s. In each column one microphone $i \in [2, I]$ at a given time $n \in [1, N]$ is compared to microphone 1 which leads to the evaluation of $P(x_{i,n}|\boldsymbol{s}, \boldsymbol{x}_1)$. An AND-gate is present in each OP-block, which represents the multiplication in stochastic computing. In our experiments we used $I = 4$ microphones. The machine had thus $N \times (I-1) = 30 \times (4-1) = 90$ columns and $1,024$ lines considering a grid of 32×32 possible source locations. At the end of each line, counters (in blue) count the number of "1"s in the output stochastic bit stream. At the end of each frame, the line with the maximum counting activates the corresponding counter in the further global fusion counterbank (in green) which implements the fusion process explained in Section II-B, and frame-wise (blue) counters are reset to zero.

Due to stochastic computing, the present SM architecture is not suited for most pre-processing routines. For example, working in the time-frequency domain using the STFT requires a considerable amount of pre-processing. We thus focused on developing a localization technique in the time domain. Pre-processing is kept as light as possible to reduce power consumption.

In this study, the experiments were run on our stochastic machine simulator which emulates the actual SM electronic circuit. Currently, the evidences are pre-computed according to (1) on a conventional computer before being stored in memory blocks for the SM. In the near future, this pre-computation will also be implemented in hardware by adding a module responsible for the initialization of the machine.

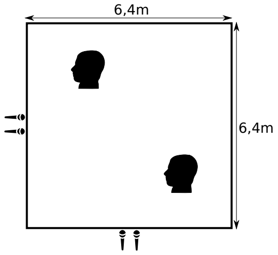


Fig. 2: Simulated room setup.

Note that preliminary power consumption measurements have been made for a similar Bayesian machine architecture with 100 columns and $4,096$ lines. A dynamic consumption around 6 mW was obtained, which can be considered as very low-power for such application as SSL.

III. Experiments

In this section, we present the experiments we conducted to evaluate the proposed SSL method.

A. Setup

Simulations have been processed with a $6.4\,\mathrm{m} \times 6.4\,\mathrm{m}$ room discretized in $20\,\mathrm{cm} \times 20\,\mathrm{cm}$ tiles, leading to a grid of $32 \times 32 = 1{,}024$ candidate 2D source positions (x, y) for s. As shown in Fig. 2, $I = 4$ microphones have been placed in the room. Each pair was located in the middle of two adjacent walls. The inter-microphone distance was $20\,\mathrm{cm}$. For our first experiments, the sources have been placed in the middle of the grid cells located at $(8, 24)$ and $(24, 8)$, as illustrated in Fig. 2.

Utterances from the TIMIT database [10], resampled at $8\,\mathrm{kHz}$, have been used as speech material. To generate the multichannel mixture signal recorded at the microphone array, a simple spatial sound simulator has been used which implements the *spherical wave model* as defined in Eq. (1) in [9].

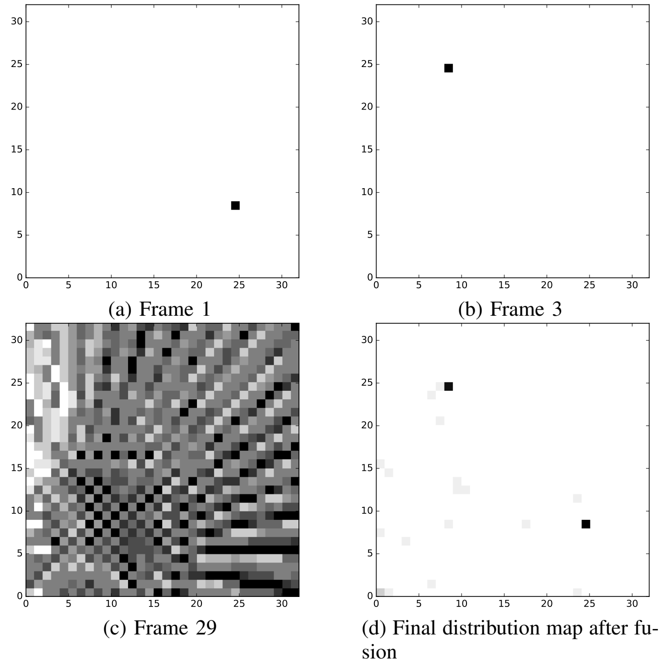
Let us remind that in our experiments, we performed the "individual" frame-wise localization on $F = 50$ frames of $N = 30$ time samples each (i.e. $3.75\,\mathrm{ms}$) and the multi-source localization with frame-wise results fusion is performed on successive blocs of $0.625\,\mathrm{s}$.

As previously mentioned, the design of our localization method was driven by the low-power stochastic architecture that computes the inference of our probabilistic model. In these experiments, after the evidences have been pre-computed as explained in Section II, the inference was run on our stochastic machine simulator.

B. A detailed example of results

Let us first present the localization results obtained on single very short time-frames, before showing the final distribution obtained by the fusion process explained in Section II-B.

Fig. 3 shows the posterior distribution $P(s|\boldsymbol{X}_f)$ obtained for 4 different frames, as a function of candidate 2D source location on the 32×32 room grid. The goal is to detect the two sources which are located at $(8, 24)$ and $(24, 8)$. The darker a cell is, the higher the probability for this cell (in other words, 1 is black and 0 is white). For the purpose of illustrating well the behavior of the proposed mehod, we selected 4 frames with quite different, though representative, type of results. Indeed, as one can see, frame 1 and frame 3 each locate perfectly one of the two sources. However, in some frames, the localization does not provide a clear maximum position, as shown in frame 29 and frame 46. However, most frames provide results similar to frame 1 and frame 3.



(a) Frame 1

(b) Frame 3

(c) Frame 29

(d) Final distribution map after fusion

Fig. 3: (a) to (c): Posterior distribution map obtained for 3 very short time-frames of a given 50-frame bloc. (d) Final distribution map after fusion over 50 frames. The two black squares correspond to the actual positions of the two sources.

When fusing the information from 50 frames following the strategy explained in Section II-B, the resulting map shows two clear maxima located at the actual source positions, as seen in Fig. 3(d). The map has two black dots which correspond to the two source locations. Some cells are in light grey (representing a low probability) which shows that these cells obtained the maximum of the posterior distribution $P(s|\boldsymbol{X}_f)$ for some frame(s). The proposed localization method is robust in the sense that the contrast between the detected/actual sources locations (black dots) and the other cells with non-zero probability (light grey) is strong enough: Quantitatively, for this example, the source located at $(24, 8)$ is detected by the frame-wise posterior maximum in 16 frames, and the source at $(24, 8)$ is detected in 15 frames. All other cells (in light grey) are counted at most 3 times as frame-wise maximum. This means that the remaining randomly-positioned 19 frame-wise maxima do not impact on the final result. This shows the importance to find a good setting for the frame size N and for the number of frames F. In summary, thanks to the fusion process over the F frames, a robust blind source localization method is obtained.

As briefly mentioned in Section II-B, because of this high contrast between the two main peaks and the other non-zero values in the global localization map, the system could automatically determine the number of sources by counting the number of peaks over a certain threshold. However, this was not systematically tested in the present experiments.

C. Average localization performance

To analyze the performance of the proposed localization method more quantitatively, experiments with 2 sources were conducted for various source locations. In total, 12 different setups have been evaluated with different randomly chosen source positions. The true source position has been compared to the estimated sources positions.

Results show that out of the 12 setups, in 4 cases, both sources were located at the right position in the grid. Moreover, in 6 out of the 12 cases, one source was estimated at the true position and the other one was estimated in an adjacent cell in the grid. Finally, in 2 cases, both sources were positioned in a neighbouring cell next to their respective true position. In practice, considering that one cell is 20 cm wide, the maximum error done by the system is about 28cm (in case of a diagonal neighbor cell). Analyzing the error distance over the 24 estimated positions (2×12), an overall average error distance of 0.12 m between the estimated position and the true position is obtained.

IV. Conclusion & Future work

In this work, a multiple sound source localization method has been presented. Compared to conventional approaches in SSL, the principal novelties are the following: i) The method works directly in the time domain (as opposed to most state-of-the-art SSL methods working in the TF domain) and the probabilistic model applies directly to the microphone waveform samples; ii) The frame-wise detection is obtained by evaluation of posterior probabilities on a very short time frame (3.75 ms); Interestingly, time-domain processing makes the method reminiscent of pioneering works on single-source SSL with a microphone array based on TDOA, while the probabilistic approach and the choice of very short time frames make the method capable of detecting several (at least two) speakers. Using so few samples (30 time samples of the signal in our experiments) is something new; To our knowledge, this has never been proposed in the SSL literature; iii) Because of i) and ii), the method is directly implementable in a stochastic machine based on Bayesian evidence calculation and integration with binary stochastic streams, which was the primary goal of the proposed study. Note that fusion of the results obtained on several frames is a classic in speech/audio processing in general, and it has been shown to provide excellent results in the present setting, with a very simple fusion process.

Experiments have been conducted to demonstrate the robustness of the localization method, which is based on the peakiness of the global localization map. Importantly, some preliminary power consumption simulations show encouraging results in the mW range and make the proposed technique a promising low power signal processing method for SSL.

Acknowledgment

This work has been partially supported by the LabEx PERSYVAL-Lab (ANR-11-LABX-0025-01).

References

[1] J. Benesty, J. Chen, and Y. Huang, *Microphone array signal processing*. Springer Science & Business Media, 2008.

[2] M. Brandstein and D. Ward, *Microphone arrays: signal processing techniques and applications*. Springer Science & Business Media, 2013.

[3] J. Chen, J. Benesty, and Y. Huang, "Time delay estimation in room acoustic environments: an overview," *EURASIP Journal on Advances in Signal Processing*, vol. 2006, no. 1, 2006.

[4] A. Coninx, P. Bessière, E. Mazer, J. Droulez, R. Laurent, M. Aslam, and J. Lobo, "Bayesian sensor fusion with fast and low power stochastic circuits," in *IEEE International Conference on Rebooting Computing (ICRC)*, 2016.

[5] Y. Dorfan and S. Gannot, "Tree-based recursive expectation-maximization algorithm for localization of acoustic sources," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 10, pp. 1692–1703, 2015.

[6] M. Faix, E. Mazer, R. Laurent, M. Abdallah, R. Le Hy, and J. Lobo, "Cognitive computation: a bayesian machine case study," in *IEEE International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC)*, 2015, pp. 67–75.

[7] R. Frisch, M. Faix, E. Mazer, L. Fesquet, and A. Lux, "A cognitive stochastic machine based on Bayesian inference: A behavioral analysis," in *IEEE International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC)*, 2018, pp. 124–131.

[8] R. Frisch, R. Laurent, M. Faix, L. Girin, L. Fesquet, A. Lux, J. Droulez, P. Bessière, and E. Mazer, "A Bayesian stochastic machine for sound source localization," in *IEEE International Conference on Rebooting Computing (ICRC)*, Nov 2017, pp. 1–8.

[9] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, 2017.

[10] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT acoustic phonetic continuous speech corpus," in *Linguistic data consortium*, 1993.

[11] H. Kayser and J. Anemller, "A discriminative learning approach to probabilistic acoustic source localization," in *2014 14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, Sep. 2014, pp. 99–103.

[12] X. Li, L. Girin, R. Horaud, and S. Gannot, "Multiple-speaker localization based on direct-path features and likelihood maximization with spatial sparsity regularization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1007–2012, 2017.

[13] M. Mandel, R. Weiss, and D. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 382–394, 2010.

[14] D. Pavlidi, A. Griffin, M. Puigt, and A. Mouchtaris, "Real-time multiple sound source localization and counting using a circular microphone array," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2193–2206, 2013.

[15] S. Rickard and O. Yilmaz, "On the approximate W-disjoint orthogonality of speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, 2002, pp. 529–532.

[16] J. Woodruff and D. Wang, "Binaural localization of multiple sources in reverberant and noisy environments," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 5, pp. 1503–1512, 2012.