Speaker Embeddings to Improve Tracking of Intermittent and Moving Speakers

Taous Iatariene[12], Can Cui[2], Alexandre Guérin[1], Romain Serizel[2]
[1] *Orange Innovation*, Rennes, France
[2] *University de Lorraine, CNRS, Inria, Loria*, Nancy, France
taous.iatariene@orange.com, can.cui@inria.fr, alexandre.guerin@orange.com, romain.serizel@loria.fr

***Abstract*—Speaker tracking methods often rely on spatial observations to assign coherent track identities over time. This raises limits in scenarios with intermittent and moving speakers, i.e., speakers that may change position when they are inactive, thus leading to discontinuous spatial trajectories. This paper investigates the use of speaker embeddings, in a simple solution to this issue. We propose to perform identity reassignment post-tracking, using speaker embeddings. We leverage trajectory-related information provided by an initial tracking step and the multichannel audio signal. Beamforming is used to enhance the signal towards the speakers' positions in order to compute speaker embeddings. These are then used to assign new track identities based on an enrollment pool. We evaluate the performance of the proposed speaker embedding-based identity reassignment method on a dataset where speakers change position during inactivity periods. Results show that it consistently improves the identity assignment performance of neural and standard tracking systems. In particular, we study the impact of beamforming and input duration for embedding extraction.**

***Index Terms*—Speaker tracking, speaker embeddings, identity reassignment, beamforming, localization**

I. Introduction

Knowing the position of the speakers present in an acoustic scene is useful in many applications, such as speech enhancement [1] and automatic speech recognition [2]. For static speakers, this involves localization, which provides estimations of the speakers' Directions of Arrival (DoAs). In the presence of multiple and moving speakers, an additional step of tracking is necessary to link the DoAs through time and assign them coherent track identities.

Many studies employ a Bayesian filtering framework for tracking [3]–[7] where data association techniques are used to deal with multiple sources and track identity management [8]. Recently, deep learning methods were proposed, in which neural networks are trained to perform ordered localization, using permutation invariant loss functions [9]–[11].

Tracking intermittent and moving speakers, who may change position when silent, is a seldom explored problem [3]–[5]. Li *et al.* [5] proposed to track both active and silent speakers and assumed spatial continuity to update inactive speaker's trajectories. This raised limits when spatial continuity was not ensured, i.e, when speakers moved unpredictably while silent, which lead to discontinuous spatial trajectories [12].

Another solution for tracking intermittent and moving speakers would be to combine spatial and identity-related observations. Spectral signatures [6] and speaker's fundamental frequencies [7] have been investigated for speaker tracking, but never for tracking intermittent and moving speakers.

Speaker embeddings are identity-related observations obtained through recent advances in deep learning [13], which have gained popularity due to their superior ability to distinguish between speakers, compared to other speaker-related features [6], [7]. They are commonly used in speaker verification [14], [15] and diarization [16]. To the best of our knowledge, they have never been used to complement DoAs for speaker tracking.

This paper proposes to combine DoAs and speaker embeddings as identity-related observations, to correct the assignment errors caused by the discontinuous spatial trajectories of speakers changing position during inactivity periods. The provided solution performs identity reassignment over the output of a tracking system, given an enrollment pool of embeddings. To reduce the impact of noise and overlapping speakers for embedding extraction, beamforming is applied beforehand, that leverages the DoA information provided by the tracking system's spatial trajectories.

We evaluate the global performance of the proposed method for different enrollment pool sizes. We show that it allows for improvements in terms of identity assignment, proving that speakers embeddings can in fact be interesting identity-related observations for tracking intermittent and moving speakers. We study several factors impacting speaker embedding quality, namely, the impact of input signal length for extraction, as well as the impact of beamforming. We finally assess the robustness of the post-tracking method against several tracking systems, by studying the impact of spatial trajectory quality on the performance.

II. Proposed Method

A. Problem formulation

We consider a multichannel audio mixture $\mathbf{y}$ composed of J speakers in a reverberant environment:

$$\mathbf{y} = \sum_{j=1}^{J} \mathbf{x}_j + \mathbf{n} \tag{1}$$

where $\mathbf{n}$ is a diffuse noise. Each speaker, in addition to its clean speech $\mathbf{x}_j$, can be described by its spatial trajectory Θ_j, which is the set of DoAs associated over the entire duration of the mixture recording.

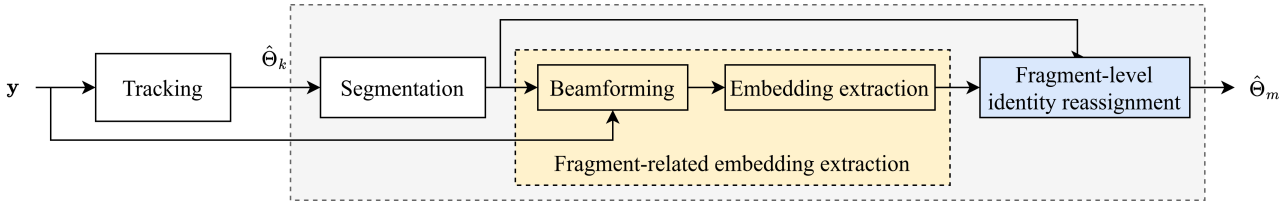

Fig. 1: Overview of the proposed speaker embedding-based reassignment method.

A tracking system can predict K trajectories $\hat{\Theta}_k$ with $k \in \{1, ..., K\}$. We define a fragment as a period of activity within a trajectory. Unlike a trajectory, which may include DoA discontinuities due to track inactivity, a fragment represents a continuous period of activity assigned by the tracker to a single identity, with no DoA interruption.

We consider scenarios with intermittent and moving speakers. Movements occurring during speaker activity can be managed with common tracking systems, for example by relying on data association techniques [8]. We therefore choose to discard them and focus on speakers whose movements occur only during inactivity periods. Such movements disrupt the spatial trajectory continuity, which the tracker relies on for identity assignment. This leads to potential errors like assigning multiple identities to the same speaker [5].

B. Fragment-level identity reassignment using speaker emebddings

To address the aforementioned issue, we propose a speaker embedding-based identity reassignment method, for which an overview is given in Fig. 1. Given a tracking system, our proposed method predicts M new trajectories $\hat{\Theta}_m$ with $m \in \{1, ..., M\}$, by computing fragment-related speaker embeddings, then performing fragment-level identity reassignment, i.e., assigning new identity to each trajectory fragment.

Each trajectory $\hat{\Theta}_k$ is first segmented into fragments. Robust fragment-related speaker embeddings are extracted with a model trained for speaker identification on a large scale dataset containing a high variety of speakers. Due to the presence of noise and overlapping speech induced by the presence of multiple speakers, beamforming is applied to the multichannel mixture $\mathbf{y}$ before embedding extraction. This spatial filtering technique can leverage the fragment DoA information and enhance the signal towards the fragment's direction, to extract more robust speaker embeddings.

To perform identity reassignment, a pool of enrollment embeddings is computed beforehand. The number of enrollments determines the maximum number of identities our system can predict, i.e. M. We consider two cases: the first case with $M = J$ enrollment embeddings corresponds to an enrollment at the beginning of a session; the second case $M > J$ assumes that the speakers present in the sessions are unknown beforehand but belong to a predefined pool.

Since one of our main focuses is to assess the interest of speaker embeddings as identity-related observation for tracking, we choose a naïve first-in-first-out method to perform fragment-level identity reassignment, in which fragments are reassigned by temporal order of appearance. The new fragment identity is attributed by selecting the identity of the enrollment embedding that has the highest cosine similarity score with the corresponding fragment-related speaker embedding. Since the same identity cannot be assigned to two overlapping fragments, enrollment embeddings of previously assigned and overlapping fragments are discarded.

III. Experimental Design

A. Datasets

Experiments are conducted on the LibriJump-2spk dataset [12], which is an evaluation dataset of 150 simulated acoustic scenes in the First Order Ambisonics (FOA) format, each containing $J = 2$ intermittent and moving speakers from the LibriSpeech [17] test-clean subset. While in LibriJump-2spk the speakers are spaced with angular distances of at least $60°$, we further add experiments with 150 other scenes containing closer speakers ($25 - 60°$). The SNR is fixed to 15 dB, and a 2-4 dB level difference is randomly set between the two speakers.

As detailed by Iatariene *et al.* [12], movement during silence is simulated by convolving each voice activity period from each speaker, with a unique Spatial Room Impulse Response (SRIR) from the same room but at a different location.

For localization and tracking, training datasets are constructed similarly to LibriJump acoustic scenes generation [12], except for the following differences : • Separate training subsets are used for SRIRs, noise samples, and speech utterances (Librispeech train-clean-100 and train-clean-360). • SNR varies randomly between 5 and 50 dB. • Speakers are static and do not move when silent. Movement is indeed unnecessary due to the models' short input sequence lengths [18], over which a speaker can be considered as static.

B. Tracking systems

We propose three tracking systems to assess the robustness of our fragment-level identity reassignment system, forming an exhaustive panel of current tracking approaches: 1) **GT**: Bayesian tracking based on ground truth DoAs. 2) **EST**: Bayesian tracking based on estimated DoAs. 3) **NN**: Neural tracking through the prediction of ordered DoAs.

The Bayesian tracker is a particle filter [19], in which the maximum number of identities predicted is fixed to the number of enrollment embeddings M used in the experiments. The probability of new source appearance is changed to adapt the nature of the observations (ground truth or estimated).

Estimated DoAs are obtained through a Convolutional Recurrent Neural Network (CRNN) [18]. The neural tracker is an adaptation of the source splitting approach [2] to the FOA multichannel format. It relies on Permutation Invariant Training (PIT) to predict ordered DoAs. The number of identities this neural tracker can predict is fixed by the number of output branches in the architecture (in our case, it is the number of speakers in the scenes, so 2).

C. Speaker Embedding Extraction

The state-of-the-art ECAPA-TDNN [14] model is used for speaker embedding extraction in our experiments. It takes as input MFCCs of a speech signal of variable length, to output a latent representation of the speaker, encoded in a vector of dimension 192. We use in particular a pre-trained and open-source version developed by speechbrain [20] which have been trained on the Voxceleb1 and Voxceleb2 datasets [21] for a speaker identification task[1].

Our method relies on fragment-related speaker embeddings, as well as enrollment embeddings for fragment-level identity reassignment. One robust enrollment embedding is extracted per speaker present in the considered scene, from 20 s wet speech signals, which are obtained similarly to what is described in III-A. We choose utterances that have not been used in the mixture computation, as well as an SRIR from another room. We run experiments with varying number of enrollment embeddings $\mathbf{M} = \mathbf{2}, \mathbf{10}, \mathbf{20}, \mathbf{30}$.

For the fragment-related embeddings, either the whole duration, or the fragment beginning is used for embedding extraction. We perform experiments using the fragment's beginning to go towards a low latency system, where track identity could be handled using speaker embeddings at each fragment appearance. We consider using the first **250, 500, 750, 1000, 1500 ms** of the fragments. We denote as *whole* the case where entire fragments lengths are used.

D. Beamforming

The fragment-related embedding quality can be impacted by the beamformer used [22]. We experiment using: 1) An ideal beamformer (**Ideal**)[2]. 2) A FOA Delay-and-Sum (**DS**) [23] beamformer. 3) A Minimum Variance Distortionless Response (**MVDR**) beamformer [24]. To estimate the noise spatial covariance matrix needed to compute the MVDR weights, we use the pretrained neural network of Bosca *et al.* [25], that estimates time-frequency masks, given FOA signal and DoA information.

E. Evaluation Metrics

We propose to use metrics measuring track identity assignment performance, namely the tracking association accuracy metric (**AssA**) [%] taken from the multi-object tracking community [26], and adapted to DoA tracking [12], as well as the LOCATA [27] Track Swap Rate (**TSR**) and Track Fragmentation Rate (**TFR**) $[s^{-1}]$ metrics. AssA evaluates the consistency of entire predicted trajectories against ground truths, in contrast with the LOCATA challenge metrics, which compares ground truths and predictions at the frame level [12].

For the spatial performance evaluation, we measure the Localization Error (**LE**) [°] as the mean angular distance between ground truth and predicted DoAs [28].

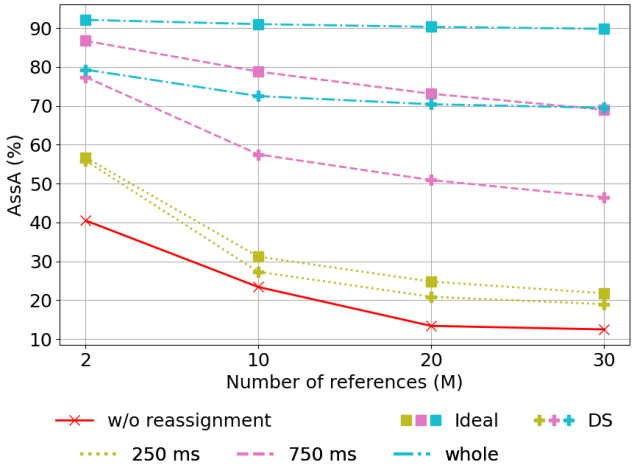

Fig. 2: Impact of the number of enrollments M on the reassignment system AssA score for three fragment input durations and two beamformers.

IV. Results

We start by assessing the utility of the proposed fragment-level embedding-based reassignment system to improve tracking of intermittent and moving speakers in IV-A. We study the factors impacting embedding extraction and thus the entire system (beamforming and fragment input duration) in IV-B. We finally analyze the impact of spatial trajectory quality by comparing in IV-C the performances of several trackers.

We run experiments with a $80 - 20\%$ bootstrap and mean metric scores are displayed. Standard deviation did not exceed 1% in all the experiments. Fig. 2 displays the impact of increasing the number of enrollments for the Ideal and DS beamformers[3], for three fragment input durations (250 ms, 750 ms and *whole*). Fig 3 displays the evolution of the AssA score when increasing the fragment input duration for the three beamformers (Ideal, DS, MVDR). The AssA score before reassignment is displayed in red. For both figures, the tracker GT is used on the $60 - 180°$ dataset.

A. Identity reassignment system global performances

From both figures, we can see that for any beamformer, input duration or number of enrollments, the proposed reassignment system improves performances over the baseline red scores before reassignment. This assesses the usefulness of speaker embeddings as identity-related observations to enhance tracking of intermittent and moving speakers.

[1]https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb

[2]In this paper, we use the term *ideal beamformer* to refer to the wet signals $\mathbf{x}_j$ before the mixture computation.

[3]We choose not to display the results using MVDR beamformer in Fig. 2, since they are similar to the ones using the DS beamformer.

TABLE I: Impact of speaker proximity on the AssA score for the three beamformers and three fragment input durations. Tracker: GT - Number of enrollments $M = J = 2$.

Beamformer	Dataset	Fragment duration [ms]		
		250	750	*whole*
Ideal	$25-60°$	55.8	86.9	93.0
	$60-180°$	56.7	86.7	92.1
DS	$25-60°$	52.1	71.4	67.6
	$60-180°$	**55.9**	**77.4**	**79.3**
MVDR	$25-60°$	52.9	72.2	72.1
	$60-180°$	**55.5**	**78.7**	**81.6**

As illustrated by Fig. 2, increasing the number of enrollments M has a negative impact on the AssA score before reassignment, going from 40.5% when $M = 2$ to 12.5% when $M = 30$. When given ideal conditions for fragment-level speaker embedding extraction , i.e, *whole* input durations and Ideal beamforming (blue curve with square marker), our proposed system becomes resilient to increasing M.

B. Impact of fragment-level speaker embedding quality

Speaker embedding quality depends on input signal duration for extraction [29], and on separation quality in the presence of overlapping speakers [22]. As illustrated by Fig 3, decreasing the input duration has a negative impact on the reassignment system performances. For the Ideal beamformer, it goes from 92.1% (*whole*) to 56.7% (250 ms).

As for beamforming, Fig. 3 shows that using the MVDR beamformer leads to better reassignment performances than using the DS. This emphasizes the need for robust separation for embedding extraction. Fig. 3 also shows that regardless of the beamformer, performances severely degrades on the shortest input durations. This shows the limit of the pre-trained speaker embedding model at extracting meaningful speaker representations from such short context.

Another factor impacting beamforming quality is the spatial placement of speakers. Table I displays AssA scores for the two datasets of distant ($60-180°$) and close ($25-60°$) speakers. For both beamformers, better performances are achieved with the distant speakers dataset. The DS beamformer is also more impacted by the speaker's spatial distance. Given the *whole* input duration, a drop of 11.7% can be measured for the DS beamformer when having closer speakers, while it is of 9.5% for the MVDR beamformer.

TABLE II: Spatial tracker performances without reassignment

Tracker	LE [°]	TSR [s^{-1}]	TFR [s^{-1}]	AssA [%]
GT	0.43	0.16	0.64	40.5
EST	**6.45**	**0.30**	0.93	36.7
NN	9.43	0.87	**0.87**	**38.1**

C. Impact of the tracking system

Previous results were obtained using the tracker GT to focus on factors impacting fragment-related speaker embedding extraction, given robust fragments of trajectories. However, spatial trajectories can have an impact on fragment quality and thus on the reassignment system. Table II and Table III display the performances of the three spatial trackers (GT, EST, NN)

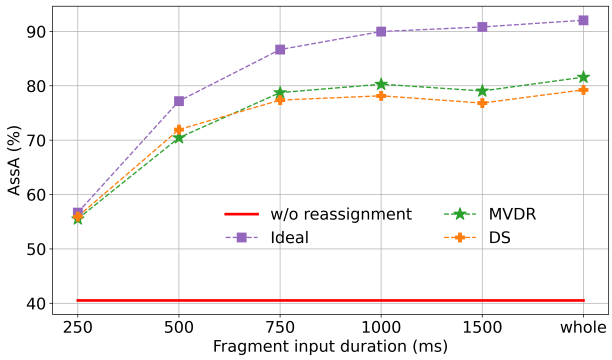

Fig. 3: Impact of the fragment input duration on the reassignment system AssA score for three beamformers - Number of enrollments $M = J = 2$

TABLE III: Spatial tracker performances (AssA score [%]) with reassignment for two beamformers and three input durations - Number of enrollments $M = J = 2$

Tracker	Beamformer	Fragment duration [ms]		
		250	750	*whole*
GT	Ideal	56.7	86.7	92.1
	DS	55.9	77.4	79.3
EST	Ideal	**53.1**	**79.0**	**82.1**
	DS	**50.3**	**71.2**	**74.4**
NN	Ideal	51.6	72.9	76.6
	DS	46.3	63.3	66.5

without and with reassignment (for 250 ms, 750 ms and *whole* input durations), respectively.

From Table II, we notice that EST is better than NN. We notice in particular that the TSR score is better for EST, and shows a rate of 0.87 swaps per second for NN. Since NN has been trained using Permutation Invariant Training (PIT), it probably suffers from the block permutation problem defined in speech separation community [30] and in speaker diarization community [31]. This leads NN to predict potentially permuted outputs from one inference run to another. LE score is lower for EST, while TFR score is similar for both EST and NN. The AssA score is 1.4% better for NN than it is for EST.

Table III finally illustrates that better spatial trajectories lead to better reassignment performances, since reassignment using the EST tracker performs better for both Ideal and DS beamformer. Given *whole* input durations, EST tracker is approx. 7.9% better with the DS beamformer and 5.5% better with the Ideal beamformer.

V. Conclusion

In this paper, we investigated the use of speaker embeddings as identity-related observation for tracking, and addressed the problem of tracking intermittent and moving speakers. We proposed a baseline solution that performs fragment-level identity reassignment in a post tracking step, given an enrollment pool. This method involves beamforming and speaker embedding extraction through a pretrained general model. Experimental results showed that such method improves significantly the identity assignment performance of several tracking systems, assessing the usefulness of speaker embeddings to comple-

ment DoAs for speaker tracking. However, performances were affected by beamforming, input duration for embedding extraction, and by the spatial trajectory quality. This illustrates the unsuitability of such general pretrained models to extract robust speaker embeddings, given short signals containing interfering speakers.

REFERENCES

[1] A. Xenaki, J. Bünsow Boldt, and M. Græsbøll Christensen, "Sound source localization and speech enhancement with sparse Bayesian learning beamforming," *The Journal of the Acoustical Society of America*, vol. 143, no. 6, pp. 3912–3921, Jun. 2018.

[2] A. S. Subramanian, C. Weng, S. Watanabe, M. Yu, and D. Yu, "Deep learning based multi-source localization with source splitting and its effectiveness in multi-talker speech recognition," *Comput. Speech Lang.*, vol. 75, p. 101 360, Sep. 2022.

[3] A. Quinlan, M. Kawamoto, Y. Matsusaka, H. Asoh, and F. Asano, "Tracking Intermittently Speaking Multiple Speakers Using a Particle Filter," en, *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2009, no. 1, pp. 1–11, Dec. 2009, Number: 1 Publisher: SpringerOpen.

[4] M. F. Fallon and S. J. Godsill, "Acoustic Source Localization and Tracking of a Time-Varying Number of Speakers," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1409–1415, May 2012, Conference Name: IEEE Transactions on Audio, Speech, and Language Processing.

[5] X. Li, Y. Ban, L. Girin, X. Alameda-Pineda, and R. Horaud, "Online Localization and Tracking of Multiple Moving Speakers in Reverberant Environments," en, *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 88–103, Mar. 2019.

[6] M. Crocco, S. Martelli, A. Trucco, A. Zunino, and V. Murino, "Audio Tracking in Noisy Environments by Acoustic Map and Spectral Signature," *IEEE Transactions on Cybernetics*, vol. 48, no. 5, pp. 1619–1632, May 2018.

[7] A. O. T. Hogg, C. Evers, and P. A. Naylor, "Multichannel Overlapping Speaker Segmentation Using Multiple Hypothesis Tracking Of Acoustic And Spatial Features," en, in *Proc. ICASSP*, Toronto, ON, Canada: IEEE, Jun. 2021, pp. 26–30.

[8] B.-n. Vo, M. Mallick, Y. Bar-shalom, S. Coraluppi, R. Osborne III, R. Mahler, and B.-t. Vo, "Multitarget Tracking," en, in *Wiley Encyclopedia of Electrical and Electronics Engineering*, John Wiley & Sons, Ltd, 2015, pp. 1–15.

[9] S. Adavanne, A. Politis, and T. Virtanen, "Differentiable Tracking-Based Training of Deep Learning Sound Source Localizers," in *Proc. WASPAA*, ISSN: 1947-1629, Oct. 2021, pp. 211–215.

[10] D. Diaz-Guerra, A. Politis, and T. Virtanen, "Position Tracking of a Varying Number of Sound Sources with Sliding Permutation Invariant Training," in *Proc. EUSIPCO*, Sep. 2023, pp. 251–255.

[11] E. Grinstein, C. M. Hicks, T. van Waterschoot, M. Brookes, and P. A. Naylor, "The Neural-SRP Method for Universal Robust Multi-Source Tracking," *IEEE Open Journal of Signal Processing*, vol. 5, pp. 19–28, 2024.

[12] T. Iatariene, A. Guérin, and R. Serizel, "Tracking of Intermittent and Moving Speakers : Dataset and Metrics," in *Proceedings of the 11th Convention of the European Acoustics Association Forum Acusticum 2025*, Malaga, Espagne, Spain, Jun. 2025.

[13] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-Vectors: Robust DNN Embeddings for Speaker Recognition," en, in *Proc. ICASSP*, Calgary, AB: IEEE, Apr. 2018, pp. 5329–5333.

[14] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," en, in *Proc. Interspeech*, Oct. 2020, pp. 3830–3834.

[15] M. Jakubec, R. Jarina, E. Lieskovska, and P. Kasak, "Deep speaker embeddings for Speaker Verification: Review and experimental comparison," en, *Engineering Applications of Artificial Intelligence*, vol. 127, p. 107 232, Jan. 2024.

[16] T. J. Park, N. Kanda, D. Dimitriadis, K. J. Han, S. Watanabe, and S. Narayanan, "A review of speaker diarization: Recent advances with deep learning," *Computer Speech & Language*, vol. 72, p. 101 317, Mar. 2022.

[17] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. ICASSP*, Apr. 2015, pp. 5206–5210.

[18] P.-A. Grumiaux, S. Kitic, L. Girin, and A. Guerin, "Improved feature extraction for CRNN-based multiple sound source localization," en, in *Proc. EUSIPCO*, Dublin, Ireland: IEEE, Aug. 2021, pp. 231–235.

[19] S. Kitić and A. Guérin, "TRAMP: Tracking by a Real-time AMbisonic-based Particle filter," in *IEEE-AASP Challenge on Acoustic Source Localization and Tracking - LOCATA*, Tokyo, Japan, Sep. 2018.

[20] M. Ravanelli, T. Parcollet, P. Plantinga, *et al.*, *SpeechBrain: A General-Purpose Speech Toolkit*, Jun. 2021.

[21] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A Large-Scale Speaker Identification Dataset," in *Proc. Interspeech*, 2017, pp. 2616–2620.

[22] S. Dowerah, R. Serizel, D. Jouvet, M. Mohammadamini, and D. Matrouf, *How to Leverage DNN-based speech enhancement for multi-channel speaker verification?* arXiv:2210.08834 [cs], Oct. 2022.

[23] M. Baque, "Analyse de scène sonore multi-capteurs : Un front-end temps-réel pour la manipulation de scène," fr, Ph.D. dissertation, Université du Maine, Jun. 2017.

[24] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969, Conference Name: Proceedings of the IEEE.

[25] A. Bosca, A. Guérin, L. Perotin, and S. Kitić, "Dilated U-net based approach for multichannel speech enhancement from First-Order Ambisonics recordings," in *2020 28th European Signal Processing Conference (EUSIPCO)*, ISSN: 2076-1465, Jan. 2021, pp. 216–220.

[26] J. Luiten, A. Osep, P. Dendorfer, P. Torr, A. Geiger, L. Leal-Taixe, and B. Leibe, "HOTA: A Higher Order Metric for Evaluating Multi-Object Tracking," *International Journal of Computer Vision*, vol. 129, no. 2, pp. 548–578, Feb. 2021.

[27] C. Evers, H. Loellmann, H. Mellmann, A. Schmidt, H. Barfuss, P. Naylor, and W. Kellermann, "The LOCATA Challenge: Acoustic Source Localization and Tracking," en, *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1620–1643, 2020.

[28] L. Perotin, R. Serizel, E. Vincent, and A. Guérin, "CRNN-Based Multiple DoA Estimation Using Acoustic Intensity Features for Ambisonics Recordings," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 22–33, Mar. 2019, Conference Name: IEEE Journal of Selected Topics in Signal Processing.

[29] C. Cui, I. A. Sheikh, M. Sadeghi, and E. Vincent, *Improving Speaker Assignment in Speaker-Attributed ASR for Real Meeting Applications*, arXiv:2403.06570 [cs], Sep. 2024.

[30] Z. Chen, T. Yoshioka, L. Lu, T. Zhou, Z. Meng, Y. Luo, J. Wu, X. Xiao, and J. Li, *Continuous speech separation: Dataset and analysis*, en, arXiv:2001.11482 [cs, eess], May 2020.

[31] K. Kinoshita, M. Delcroix, and N. Tawara, "Integrating End-to-End Neural and Clustering-Based Diarization: Getting the Best of Both Worlds," in *Proc. ICASSP*, ISSN: 2379-190X, Jun. 2021, pp. 7198–7202.