DOI: 10.61782/fa.2025.0526

LOCALIZE, FILTER, SEGMENT: TOWARDS MULTI-MICROPHONE SPEAKER SEGMENTATION

Théo Mariotte*[1]
[1] LIUM, Institut Claude Chappe, Le Mans Université, France

ABSTRACT

Speaker Diarization addresses the question *Who spoke and when?*—a crucial task in conversational AI. We propose **Lo**calize, **Fi**lter, **Seg**ment (LoFi-Seg), a novel framework for speaker segmentation in multi-microphone setups. LoFi-Seg consists of three modules: a direction-of-arrival (DOA) estimator, a spatial filter bank (beamforming), and a Voice Activity Detection (VAD) model. The framework processes multichannel audio by steering beamformers—one per speaker—toward predicted directions from the DOA module. The output of each beamformer is then passed through the VAD model to determine speaker activity. Combining explicit DoA estimation and VAD improves the model transparency, thus preserving physical interpretability during multichannel filtering. We validate LoFi-Seg through experiments on simulated multi-speaker, multi- microphone conversations, where speaker positions and acoustic conditions are controlled. The system is evaluated on speaker segmentation performance, with additional assessments of speaker localization performance drift across setups. This approach demonstrates how LoFi-Seg combines robust speaker segmentation with interpretable processing, making it a valuable tool for advancing speaker diarization in complex audio environments.

Keywords: *Speaker segmentation, Speaker localization, Beamforming, Interpretability, Data simulation*

1. INTRODUCTION

Deep learning is now commonly adopted for automatic speech processing, such as automatic speech recognition (ASR) [1], speaker recognition [2], or speaker diarization [3, 4]. Although these systems can reach impressively low error rates, most of the time this is at the cost of an increase in the number of trainable parameters [1], and a loss in model transparency [5].

This paper focuses on the speaker segmentation task [6] that consists of predicting the activity of each speaker in an audio segment. Speaker segments can later be used to perform speaker clustering in speaker diarization [7] or to perform speaker-wise ASR [8, 9]. Recently, several papers have shown that speaker segmentation can be performed by combining speech separation and Voice Activity Detection (VAD) [10, 11]. These systems offer several advantages. First, they can be very efficient regarding computational cost and can be deployed in the streaming scenario [11]. Second, they rely on the combination of speech separation and VAD models that can be trained in a short amount of time, requiring fewer resources than End-to-End approaches [4]. Third, applying separation before VAD improves the model transparency since we can access the separated sources before predicting speaker activities.

Although these systems offer strong segmentation performance, they are restricted to the single-microphone scenario, i.e., when the audio is encoded in a single channel. However, recording the acoustic scene with multiple microphones allows encoding spatial information that can be exploited to improve speech segmentation tasks [12–14].

Based on these previous works, we propose a new system inspired by the combination of source separation and VAD applied to the multi-microphone scenario. The system exploits spatial filtering to extract multiple representations of the multi-microphone input signal. Steering directions of these spatial filters – beamformers – are predicted by a neural network from the multi-microphone audio signal. Each spatially filtered signal is forwarded through a VAD model to predict the activity of each speaker. Since the model is trained on short time windows, e.g., 4s, the maximum number of overlapping speakers is considered as fixed [6]. Hence, the system includes one spatial filter for each possible active speaker. The model is later called **Lo**calize, **Fi**lter, **Seg**ment (LoFi-Seg). This paper explores several training schemes of the proposed model on simulated 2-speaker conversations. This can be seen as a proof-of-concept of the proposed architecture to perform speaker segmentation in the multichannel scenario. The contributions are the following:

- We propose a new data simulation algorithm to generate 2-speaker multi-microphone pseudo conversations ;

**Corresponding author*: *theo.mariotte@univ-lemans.fr.*

[1] https://jonathanbgn.com/2021/12/31/timeline-transformers-speech.html

- We show that combining steered beamformers with VAD can perform speaker segmentation on the proposed dataset ;
- We explore training schemes, such as pre-training the steering model with source localization objective or combining segmentation and localization losses ;
- We visualize the steering direction learned by the model to assess its transparency.

The code for model training, evaluation, and dataset generation is available at https://git-lium.univ-lemans.fr/speaker/speakseg.

2. RELATED WORKS

Speaker segmentation is a key task in speaker diarization [6] and appeared with the development of end-to-end speaker diarization models (EEND) [4]. Multiple works have been conducted to solve speaker segmentation in the single-channel scenario [4]. The Target Speaker VAD (TS-VAD) is also related since it predicts speaker activities from the audio signal with additional speaker information [15]. The proposed LoFi-Seg is inspired by the approach in [6] where speakers segments are predicted on short audio chunks (typically 5s).
Multi-microphone speaker diarization Several works have tackled speaker diarization in the multi-microphone scenario. The work from [16] is close to the proposed paper since it predicts speaker activity from spatial cues. Works that inspired the proposed method focused on multi-microphone Voice Activity Detection (VAD) and Overlapped Speech Detection (OSD) in the context of speaker diarization [12–14].
Neural source localization is related to the proposed method since we align beamformer steering direction on speakers' direction of arrival (DOA). Multiple methods and architectures have been proposed for source localization using deep learning [17]. The proposed architecture is inspired by [18, 19] and specifically [20]. Note that we do not intend to propose a new state-of-the-art source localization model in this paper.

3. SPEAKER SEGMENTATION AND LOCALIZATION

This section introduces the speaker segmentation (3.1) and localization (3.2) tasks. The formulation, training objectives, and evaluation procedures are described for each task.

3.1 Speaker segmentation

3.1.1 Problem formulation

Let $\mathbf{x} \in \mathbb{R}^{M\times N}$ be an audio segment with N samples and M channels recorded by a microphone array. Let $\mathbf{y} \in \mathbb{R}^{S\times T}$ be the speaker activity labels with S the number of possible speakers in $\mathbf{x}$ and T the number of time frames. Speaker segmentation consists of predicting $\mathbf{y}$ given the audio signal $\mathbf{x}$. This problem is commonly solved by defining a function $f_\theta : \mathbb{R}^{M\times N} \mapsto \mathbb{R}^{S\times T}$ with parameters θ. These parameters are optimized using gradient descent based on a loss function $\ell(\mathbf{y}, f_\theta(\mathbf{x}))$ measuring the prediction error between labels and model predictions. Since multiple speakers can be simultaneously active, speaker segmentation is a multi-label classification task. Hence, we consider Binary Cross Entropy (BCE) as a training objective as proposed in [6]. This loss is denoted as ℓ_{BCE}.

Similarly to source separation [21], speaker segmentation systems are subject to random permutations in the output [6]. To align the label y_s with its associated prediction $\hat{y}_s$, all the permutation pairs over $s \in \{1, S\}$ should be considered. The final permutation invariant training (PIT) objective is defined as follows:

$$\mathcal{L}_{BCE}\left(\mathbf{y}, f_\theta(\mathbf{x})\right) = \min_{\pi\in\Pi} \ell_{BCE}\left(\pi(\mathbf{y}), f_\theta(\mathbf{x})\right), \qquad (1)$$

where Π is the set of all possible permutations between labels and predictions.

3.1.2 Evaluation procedure and metrics

During inference, the audio file is chunked using a sliding window. The segmentation model f_θ predicts the likelihood of each speaker to be active at the frame level. The frame-wise likelihood is binarized by applying a threshold $\tau \in [0, 1]$ such that $\hat{y}_b = \{\hat{y} > \tau\}$. The predictions of consecutive sliding windows are concatenated to obtain the prediction over the entire file. We consider the Oracle speaker assignation scenario. The stitching between windows is obtained using (1). The predictions are converted to the RTTM file format before computing the metrics using `pyannote.metrics` [7].

Speaker diarization is usually evaluated in terms of Diarization Error Rate (DER) [4]. This metric combines segmentation errors and speaker confusion. In this work, we focus on the segmentation task. The performance of the models is reported as the Missed Detection (Miss.) and False Alarm (FA) rates. The sum of both metrics is also reported and denoted as Segmentation Error Rate (SER). This is equivalent to a DER with a speaker confusion set to 0.

3.2 Speaker localization

3.2.1 Regression vs. Classification

Acoustic source localization consists of predicting the position of an emitting acoustic source from a multi-microphone signal recorded by a microphone array. In this context, several speakers can be active in the audio segment. Thus, we consider the multiple source localization problem. In this paper, we restrict the localization problem to an azimuthal direction-of-arrival (DOA) prediction problem [17].

The multi-source localization problem can be formulated as a classification [17] or a regression [19, 20] problem. Since it remains difficult to identify which formulation offers the best performance [19], we consider the formulation proposed in [20]

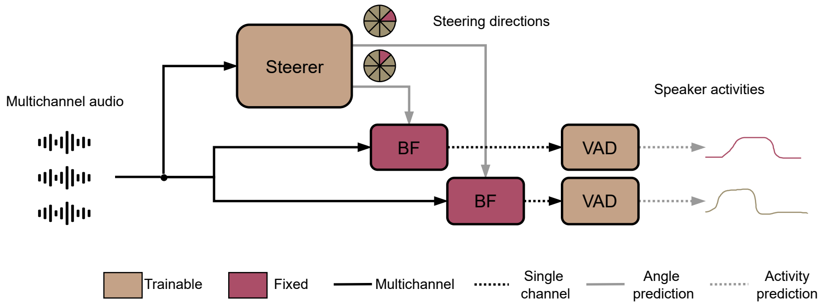

Figure 1. Overview of the system. The multichannel signal is forwarded to the steerer to predict beamformer steering directions. The signal is then filtered by all the beamformers steered in different directions. VAD is applied on each beamforming output to predict speaker activities.

where the model is trained to predict the likelihood of each DOA as a regression problem.

3.2.2 *Problem formulation*

Let $\varphi_s \in [-\pi, \pi]$ be the ground-truth DOA of the s-th speaker in radians. The φ_s DOA is encoded into a likelihood function with support $l = [1, L]$. Following [20], the target likelihood vector $\{o_l\}$ is defined as follows:

$$o_l = \begin{cases} \max_{s=1}^{S} \left\{ e^{-d(\varphi_s, \varphi_l)^2/\sigma^2} \right\} & \text{if } S > 0 \\ 0 & \text{otherwise} \end{cases}, \qquad (2)$$

where σ controls the width of the Gaussian-like functions and $d(\cdot, \cdot)$ denotes the angular distance between the support DOA φ_l and the ground-truth φ_s. An example of DOA encoding with equation (2) is in Figure 2.

The localization model $g_\phi : \mathbb{R}^{M \times N} \mapsto \mathbb{R}^L$ is optimized to predict the appropriate $\hat{o}$ function from the audio signal. The Mean Squared Error (MSE) is used as a loss function to optimize the parameters ϕ of the model:

$$\mathcal{L}_{MSE} = \frac{1}{L} \sum_l \|o_l - \hat{o}_l\|_2^2. \qquad (3)$$

3.2.3 *Evaluation procedure and metrics*

Once the g_ϕ model is trained to predict the best likelihood function $\hat{o}$ according to the cost (3), it is evaluated on the evaluation set. The DOA can be predicted (i.e. decoded) from the $\hat{o}$ prediction. DOA predictions can be inferred from $\hat{o}$ by finding peaks above a threshold $\xi \in [0, 1]$: $\hat{\varphi} = \{\varphi_l \mid \hat{o}_l > \xi\}$.

In the proposed experiments, the maximum number of active speakers is supposed to be known for a given segment. Hence, if the number of predicted DOAs in $\hat{\varphi}$ is greater than the number of possible speakers S, we restrict the DOA predictions to the ones with the highest likelihood.

The quality of the model predictions is evaluated by Mean Absolute Error (MAE) on the DOA, denoted MAE (°). The MAE is computed considering the shortest path on the unit circle.

In some situations, the model can miss or overpredict sources,i.e., when the likelihood is too low, or too high. The detection performance is reported in terms of FA and Miss scores. Note that the MAE is computed only on the detected sources, measuring the prediction quality of the sources that the model can predict.

4. LoFi-Seg SYSTEM OVERVIEW

LoFi-Seg is composed of three main parts described in this section. The multichannel audio is processed by a set of spatial filters implemented as a Superdirective beamformer (4.1). A steering model predicts the steering direction of each filter (4.2). Finally, speaker activities are predicted by a VAD model after each spatial filter (4.3). An abstract view of the model is proposed in Figure 1.

4.1 Super-directive Beamforming

Super-directive beamforming is a commonly used algorithm for spatial filtering [22]. It can be seen as a simplified version of the so-called Minimum Variance Distortionless Response (MVDR) beamforming [23]. The narrowband weights of such a filter can be expressed as follows:

$$\mathbf{w}^H(f, \varphi_s) = \frac{\mathbf{v}^H(f, \varphi_s)\mathbf{\Sigma}^{-1}(f)}{\mathbf{v}^H(f, \varphi_s)(\mathbf{\Sigma}(f) + \lambda_{reg}\mathbf{I})^{-1}\mathbf{v}(f, \varphi_s)}, \qquad (4)$$

where f is the frequency, $\mathbf{v}(f, \varphi_s) \in \mathbb{C}^{M \times 1}$ a steering vector and $\mathbf{\Sigma}(f) \in \mathbb{R}^{M \times M}$ the noise covariance matrix. λ_{reg} is a regularization parameter to prevent singularities in matrix inversion [24]. Under isotropic noise assumption, an element $\mathbf{\Sigma}_{m,n}(f)$ of the covariance matrix can be expressed as [22]: $\mathbf{\Sigma}_{m,n}(f) = \text{sinc}(2\pi f d_{m,n}/c)$, where $d_{m,n}$ is the distance between m-th and n-th microphones.

When considering a uniform circular array (UCA), the m-th element of the steering vector v_m, oriented towards the φ_s angular direction, can be expressed as [25] :

$$v_m(f, \varphi_s) = \exp\left(j2\pi f r c^{-1} \cos(\varphi_s - \psi_m)\right), \tag{5}$$

with ψ_m the angle of the m-th microphone, c the speed of sound and r the radius of the UCA.

Let $\mathbf{X} \in \mathbb{C}^{M \times K \times T}$ be the STFT of the multi-microphone input signal $\mathbf{x}$. The output of filter steered in the φ_s direction at frequency f, is obtained following

$$\mathbf{Y}(t, f, \varphi_s) = \mathbf{w}^H(f, \varphi_s)\mathbf{X}(f, t). \tag{6}$$

The broadband output $\mathbf{Y}(\varphi_s) \in \mathbb{C}^{K \times T}$ is the STFT of a single-channel signal steered towards the φ_s direction, obtained by applying (6) for each considered frequency. In this work, one beamformed signal is obtained for each possible speaker. A steering model predicts the angular direction of each filter $\hat{\varphi}_s$. Its design is described in the next section.

4.2 Steering model

The steering model predicts the speakers' DOA from the STFT of the multi-microphone signal $\mathbf{X}$. The steering architecture is designed to require a few trainable parameters to limit the complexity of the LoFi-Seg model. The architecture is inspired by [18] and combines a convolutional neural network (CNN) to extract features from the STFT, followed by sequence modeling layers. BLSTM layers, commonly used for sequence modeling, are replaced with a Temporal Convolutional Network (TCN) [26]. The TCN has shown impressive sequence modeling performance [12, 13, 27] with a limited number of trainable parameters.

Specifically, the model is composed of a 2D convolutional encoder that maps the multichannel STFT $\mathbf{X} \in \mathbb{R}^{2M \times K \times T}$ to a new representation $\mathbf{X}_e \in \mathbb{R}^{D \times K \times T}$, where $D > 2M$ is the number of channels in the encoded representation [28]. Note here that $2M$ holds for the concatenation of the real and imaginary parts of the STFT. The $\mathbf{X}_e$ representation is processed by a CNN composed of 3 Conv-blocks that reduce the frequency dimension before processing the sequence with the TCN model. The conv-blocks are structured with a 2D convolutional layer, followed by Batch Normalization (BN), ReLU activation, and a max pooling operation. The number of channels is doubled after each conv-block. We use depthwise separable convolutions [29] to reduce the number of trainable parameters. The CNN outputs a representation of shape $\mathbf{X}_c \in \mathbb{R}^{D' \times T}$, where D' is the final number of channels.

Finally, the $\mathbf{X}_c$ representation is processed by the TCN. The architecture is similar to [12]. After the encoding layer, the sequence is processed by N_B TCN blocks of depth N_L. The output of the model is of shape $\hat{o} \in \mathbb{R}^L$, with L the number of DOA in the likelihood function predicted by the model (See Section 3.2).

4.3 VAD model

The VAD model follows the architecture we used in [13]. The output of the s-th beamformer $\mathbf{Y}(\varphi_s)$, represented in the STFT domain, is transformed into a Mel spectrogram with F filters. These features of shape $\mathbb{R}^{F \times T}$ are forwarded through a TCN model predicting the output $\hat{y} \in \mathbb{R}^{S \times T}$.

5. SIMULATED CONVERSATION DATASET

In this section, we present the proposed methodology to simulate conversational data to train and evaluate LoFi-Seg.

5.1 Motivations

Traditional benchmark datasets for speaker diarization mainly consider the single-channel scenario, e.g., in the context of broadcast news or meetings. Few datasets propose multi-microphone recordings in meetings like AMI [30] or AISHELL-4 [31]. Although widely adopted, these datasets lack annotations such as speaker localization. However, speaker position can be very informative in the context of speaker diarization. Since our system relies on a steering model trained with a speaker localization objective, it requires the ground-truth direction of arrival (DOA) labels. Thus, we consider data simulation to cover this need.

5.2 Conversation dynamics generation

Data simulation for speaker diarization has been considered for EEND training [4]. Originally, simulations combined speech samples to create overlapping speech segments as done for speech separation. However, this strategy provides limited generalization capacities to real data. To tackle this issue, conversation simulation has been proposed by modeling real conversation statistics and using them to perform simulations [32]. The proposed strategy lies between these approaches. Instead of modeling real conversation dynamics, we simplify the conversation generation process with few parameters.

The first step of data simulation is to generate the conversation dynamics. This consists of defining each speaker's starting and ending time in the conversation.

The segment generation is controlled by a few parameters: $\mathcal{N}_p(\mu_p, \sigma_p)$ represents the pause statistics, $\mathcal{N}_s(\mu_s, \sigma_s)$ represents the segment statistics, p_{ov} is the probability of a segment to overlap with the previous one, and r_{ov} is the average overlap ratio when overlap is occurring. The speaker index, starting and ending time are randomly sampled according to these parameters to create pseudo-conversations of a target duration T_c. Once the segments are defined, one has to sample segments from a source dataset to generate the audio signal aligned with the conversation dynamics. This step is described in the next section.

5.3 Conversation generation and spatialization

In this work, we use Librispeech [33], a dataset of clean read speech, to create the conversations. We use the `train-100` split to generate the training set, and `dev-clean` to generate the test set. For each conversation, we ensure selecting different speakers and that those have enough audio segments to create a conversation. Then, we randomly sample the utterances from the original data without replacement and crop them to fit the target duration of the current segment in the conversation.

Once the clean audio signal is generated for each speaker, we simulate a room using the `Pyroomacoustics` toolkit [34] with random source positions around the microphone array. The reverberation time of the room–in seconds–is also randomly sampled following $T_{60} \sim \mathcal{U}(0.5, 1)$.

5.4 Simulated dataset

The dataset used in the following experiments is split into two subsets: `train-100` for training and `dev-clean` for evaluation. The source audio signals are sampled from the eponymous subsets of Librispeech to prevent overlapping speakers between the training and evaluation sets. Both subsets contain simulated pseudo-conversations in random rooms with random source positions, recorded by a uniform circular microphone array with a radius $r = 0.1$m and $M = 8$ microphones. The training set contains 2985 audio files, each being a short conversation between 60s and 180s. It represents about 175 hours of simulated conversations, with a 50% probability of overlapping speech between speakers and an average overlap ratio of 70%. The evaluation set is about 10 hours of conversation and follows a similar setup.

During training, dynamic mixing [35] is applied between speaker segments with a relative SNR sampled following $SNR \sim \mathcal{U}(-5, 5)$dB. 10% of the training set is kept for validation purposes. Dynamic mixing is disabled during validation and evaluation. The evaluation conversations are generated without relative SNR between speakers.

6. EXPERIMENTAL RESULTS

6.1 Speaker segmentation in various training configurations

In this section, we train LoFi-Seg considering different training schemes. The training is conducted on the `train-100` simulated data, and evaluated on the`dev-clean` evaluation set. This section explores the impact of each scenario on the performance. We encourage the reader to refer to the code and model configurations for more details about the training configurations.

6.1.1 Experimental setup

Five configurations are considered for the following experiments. The two first configurations consist of training LoFi-Seg from scratch: *(i)* only the segmentation loss (1) is used. *(ii)* The model is trained with a combination of the segmentation (1) and localization (3) losses: $\mathcal{L}_{BCE} + \mathcal{L}_{MSE}$. Hence, localization and segmentation are jointly optimized.

In the 3 other configurations, the steering model g_ϕ parameters are initialized from a model trained with a source localization objective. The training procedure of such a model is detailed in Section 6.1.2. Three configurations are considered with initialized ϕ parameters: *(iii)* the steering model g_ϕ is frozen during segmentation optimization with $\mathcal{L}_{BCE}$; *(iv)* LoFi-Seg is optimized to minimize $\mathcal{L}_{BCE}$; *(v)* LoFi-Seg is optimized to minimize $\mathcal{L}_{BCE} + \mathcal{L}_{MSE}$.

6.1.2 Pre-training the speaker localization model

Training configuration We explore several configurations where the steering model g_ϕ is pre-trained to predict speakers' angular directions. The model is trained on 4-second segments randomly sampled from the `train-100` set and grouped into batches of 64 examples. DOA labels are encoded similarly to [20] with equation (2) by setting $\sigma = 10$ and $L = 360$. Models are trained to minimize the MSE on DOA likelihood functions (3). 10% of the training set is kept for validation purposes. Model parameters are optimized using the ADAM optimizer [36] with a learning rate of 0.001. The learning rate is halved if no improvement in the validation loss is seen during 10 epochs. We select the model reaching the lowest MSE on the validation set.

Model architecture The localization model predicts DOA likelihood from the STFT of the multichannel signal extracted on 512-sample windows with a 160-sample shift. The convolutional encoder has 32 hidden channels with a 2D kernel of 1 × 1 to keep the frequency and time dimensions. The CNN has the same structure as [18] by doubling the number of channels after each convolutional layer. These layers have a kernel size of 3×3 with a stride of 1. Each max-pooling operation downsamples the frequency axis by 8 while keeping the temporal axis the same. The TCN model has 3 TCN blocks composed of 3 convolution layers with exponentially increasing dilation. The hidden dimension is set to 128. Finally, the model has only 240k trainable parameters.

6.1.3 Training the LoFi-Seg speaker segmentation model

Training configuration All models are trained with the same data distribution. The optimization objective depends on the considered configuration (6.1.1). Parameters are optimized with the ADAM optimizer with an initial learning rate of 0.001. The learning rate is halved if no improvement is seen in the validation loss after 15 epochs. The batch size is set to 64. We select the model reaching the best validation loss for evaluation.

Model architecture For each setup, the architecture is the same. STFT is computed on 512-sample widows with a hop size of 160 samples. The mel-spectrogram is computed with $F = 128$ filters. The beamformer weights are calculated with a regularization term set to $\lambda_{reg} = 10^{-4}$. The VAD model contains 3 repeating TCN blocks, composed of 3 TCN layers with expanding dilation. The TCN bottleneck comprises 128 channels; the hidden channel number is 512. The model has 2 VAD model instances since we are processing 2-speaker conversations. In the end, the model has 2.7M trainable parameters.

Table 1. Segmentation performance of the LoFi-Seg model in various training scenarios. The results of each configuration are compared with the single-channel segmentation baseline.

# Param. (M)	Init. g_ϕ	Frozen g_ϕ	$\mathcal{L}_{MSE}$	SER (%) ↓	Miss (%) ↓	FA (%) ↓
2.46	✓	✓	✗	26.9 ± 6.8	21.2	5.9
2.70	✓	✗	✗	26.9 ± 6.8	21.2	5.9
2.70	✓	✗	✓	26.0 ± 6.2	20.0	6.1
2.70	✗	✗	✗	25.8 ± 6.3	19.6	6.4
2.70	✗	✗	✓	25.8 ± 6.2	20.1	6.0

6.1.4 Results

Speaker segmentation is evaluated on the `dev-clean` subset. Speaker activity predictions are obtained on 4s sliding widows with no overlap and metrics are computed following the procedure described in 3.1.2. Table 1 presents the results for each model configuration.

The two models trained from scratch reached the best segmentation performance with 25.8% with or without MSE loss. Adding a pre-trained localization model tends to slightly degrade the performance. In fact, including a pre-trained g_ϕ model leads to 26.9% SER whether it is frozen or not. However, adding the MSE loss with this setup seems to improve the segmentation with a 26% SER. The Wilcoxon signed-rank test was performed on the SER between the worst-performing model (initializing g_ϕ and freezing it) and the others to evaluate the statistical significance. All the tests obtain a p-value $p > 0.1$, indicating the SER difference between models is not statistically significant.

Globally, the models can perform speaker segmentation on the simulated dataset. However, the SER remains high with the best model reaching 25.8%, mostly because of a high missed detection rate. This is probably due to silences in Librispeech utterances used to create the simulated conversations, which are not labeled. The model predicts a low likelihood for a given speaker when there is silence, while the label indicates that this speaker is active. This highlights necessary improvements in the labeling strategy of the conversation simulation algorithm.

6.2 Drift of speaker localization across setups

While the segmentation performance does not seem to strongly rely on the training configuration, the model's speaker localization capacities are not guaranteed to be kept. The model could find optimal parameters without predicting the source positions. This would drastically reduce the physical interpretability of LoFi-Seg. This section explores the drift in speaker localization performance across configurations.

6.2.1 Evaluation of speaker localization

The localization performance is evaluated on 1000 4s segments randomly sampled from the `dev-clean` evaluation set. For each model, the threshold for DOA prediction is set to $\xi = 0.6$. We report the MAE in degrees along with FA and Miss detection metrics, as described in the protocol defined in section 3.2.3. The performance of the g_ϕ model for each configuration is presented in Table 2.

6.2.2 Results

First, table 2 shows that the models trained from scratch, i.e., without a pre-trained localization model, cannot perform source localization (e.g., MAE of 66.3° without the MSE loss). The model can find a set of parameters that solves the segmentation task without predicting the accurate source DOA. Thus, the model loses transparency in this situation. The detection metrics are low because the steering model activates almost all the DOAs in the output. By listening to beamforming outputs, it seems that the audio signals are not strongly discriminative. That would not facilitate the system to predict the source DOA and require further investigations.

Then, when adding a pre-trained localization model, the localization performance remains the same whether the weights are frozen or not. The model does not modify the ϕ parameters, as observed in the segmentation experiments.

Finally, when the initialized model is trained with the additional MSE loss, the localization performance is improved. In this set-up, the model can better segment the speakers (table 1) while providing an accurate localization of the sources with an average MAE of 13.4°. Note that the detection performance is still mitigated, with 39.7% of missed detection. However, the MAE is better on the sources the model was able to detect.

Figure 2 illustrates the improvement obtained with the MSE loss during training. It presents the DOA prediction for a given segment of the 3 initialized models. It shows how the model trained with BCE and MSE (`ckpt+MSE`) improves the localization w.r.t. the two other models.

7. CONCLUSIONS AND DISCUSSIONS

This paper proposes **Lo**calize, **Fi**lter, **Seg**ment (LoFi-Seg), a new model for multi-microphone speaker segmentation. This model consists of a localization module that predicts the speaker's direc-

Table 2. Drift of speaker localization prediction with the LoFi-Seg model across setups. The number of sources is supposed to be known.

Init. g_ϕ	Frozen g_ϕ	$\mathcal{L}_{MSE}$	MAE (°) (↓)	FA (%) (↓)	Miss (%) (↓)
✓	✓	✗	21.4	2.3	39.9
✓	✗	✗	21.4	2.3	39.9
✓	✗	✓	**13.4**	1.9	39.7
✗	✗	✗	66.3	4.1	0.0
✗	✗	✓	66.8	4.1	0.0

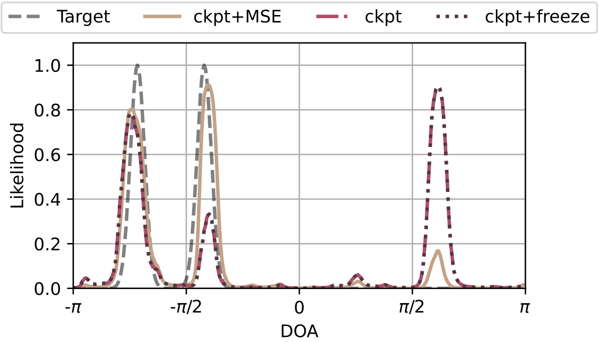

Figure 2. Example of DOA prediction with the g_ϕ model for various training setups.

tions of arrival (DOA), a set of spatial filters, and a Voice Activity Detection (VAD) model. The DOA is used to steer the spatial filters in the speaker directions to obtain a filtered single-channel signal, before applying VAD. The LoFi-Seg aims at conserving physical interpretability in the steering step of the model, to better understand the model behavior. The paper also introduces a new simulation strategy, combining pseudo-conversation generation with acoustic simulation to create multi-microphone conversation datasets.

The proposed model can perform speaker segmentation on the generated dataset. While the performance remains mitigated, some configurations allow to keep the speaker localization information while accurately predicting speaker segments. The best configuration in terms of performance and transparency is to initialize the localization model with weights optimized with a speaker localization objective and to train LoFi-Seg with joint segmentation and localization losses. This improves the localization quality and the segmentation performance, providing a transparent and efficient model.

In general, the current model offers limited segmentation performance and should be compared to other concurrent models (e.g. Pyannote [6]). Furthermore, by listening to the beamforming output, the audio quality seems very low, and the difference between different directions is hard to perceive. Some work is required in that direction to improve the quality of the beamforming step. The current version of the model requires one VAD model for each beamforming output to perform well (which is probably related to the poor quality of beamforming outputs). The number of parameters highly depends on the number of speakers to segment. Future work will focus on preventing this limitation to scale the system to more speakers. Finally, the segmentation performance highlighted some limitations in the label generation from the simulated conversation, paving the way for improvements in the simulation algorithm.

8. ACKNOWLEDGMENTS

I want to thank Pr. Slim Essid for the discussions regarding data simulation and Pr. Marie Tahon and Pr. Anthony Larcher for the discussions on the experimental study.

9. REFERENCES

[1] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, “Speech recognition using deep neural networks: A systematic review,” *IEEE access*, vol. 7, pp. 19143–19165, 2019.

[2] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, “X-vectors: Robust dnn embeddings for speaker recognition,” in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5329–5333, IEEE, 2018.

[3] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, “Speaker diarization: A review of recent research,” *IEEE Transactions on audio, speech, and language processing*, vol. 20, no. 2, pp. 356–370, 2012.

[4] 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. 101317, 2022.

[5] A. Akman and B. W. Schuller, “Audio explainable artificial intelligence: A review,” *Intelligent Computing*, vol. 2, p. 0074, 2024.

[6] H. Bredin and A. Laurent, “End-to-end speaker segmentation for overlap-aware resegmentation,” *arXiv preprint arXiv:2104.04045*, 2021.

[7] H. Bredin, “pyannote. audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe,” in *24th INTERSPEECH Conference (INTERSPEECH 2023)*, pp. 1983–1987, ISCA, 2023.

[8] C. Boeddeker, A. S. Subramanian, G. Wichern, R. Haeb-Umbach, and J. Le Roux, "Ts-sep: Joint diarization and separation conditioned on estimated speaker embeddings," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1185–1197, 2024.

[9] J. Kalda, C. Pagés, R. Marxer, T. Alumäe, and H. Bredin, "Pixit: Joint training of speaker diarization and speech separation from real-world multi-speaker recordings," in *The Speaker and Language Recognition Workshop (Odyssey 2024)*, pp. 115–122, 2024.

[10] G. Morrone, S. Cornell, L. Serafini, E. Zovato, A. Brutti, and S. Squartini, "End-to-End Integration of Speech Separation and Voice Activity Detection for Low-Latency Diarization of Telephone Conversations," *Speech Communication*, vol. 161, p. 103081, June 2024. arXiv:2303.12002 [eess].

[11] E. Gruttadauria, M. Fontaine, and S. Essid, "Online speaker diarization of meetings guided by speech separation," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 11356–11360, IEEE, 2024.

[12] S. Cornell, M. Omologo, S. Squartini, and E. Vincent, "Overlapped speech detection and speaker counting using distant microphone arrays," *Computer Speech & Language*, vol. 72, p. 101306, 2022.

[13] T. Mariotte, A. Larcher, S. Montrésor, and J.-H. Thomas, "Channel-combination algorithms for robust distant voice activity and overlapped speech detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[14] T. Mariotte, A. Larcher, S. Montrésor, and J.-H. Thomas, "Asobo: Attentive beamformer selection for distant speaker diarization in meetings," in *Interspeech 2024*, pp. 1620–1624, 2024.

[15] I. Medennikov, M. Korenevsky, T. Prisyach, Y. Khokhlov, M. Korenevskaya, I. Sorokin, T. Timofeeva, A. Mitrofanov, A. Andrusenko, I. Podluzhny, A. Laptev, and A. Romanenko, "Target-speaker voice activity detection: A novel approach for multi-speaker diarization in a dinner party scenario," in *Interspeech 2020*, pp. 274–278, 2020.

[16] N. Zheng, N. Li, J. Yu, C. Weng, D. Su, X. Liu, and H. Meng, "Multi-channel speaker diarization using spatial features for meetings," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7337–7341, IEEE, 2022.

[17] P.-A. Grumiaux, S. Kitić, L. Girin, and A. Guérin, "A survey of sound source localization with deep learning methods," *The Journal of the Acoustical Society of America*, vol. 152, pp. 107–151, July 2022.

[18] L. Perotin, R. Serizel, E. Vincent, and A. Guérin, "Crnn-based joint azimuth and elevation localization with the ambisonics intensity vector," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 241–245, 2018.

[19] L. Perotin, A. Defossez, E. Vincent, R. Serizel, and A. Guerin, "Regression Versus Classification for Neural Network Based Audio Source Localization," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, (New Paltz, NY, USA), pp. 343–347, IEEE, Oct. 2019.

[20] W. He, P. Motlicek, and J.-M. Odobez, "Deep Neural Networks for Multiple Speaker Detection and Localization," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 74–79, May 2018. arXiv:1711.11565 [cs].

[21] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 241–245, IEEE, 2017.

[22] M. Wölfel and J. McDonough, *Distant speech recognition*. John Wiley & Sons, 2009.

[23] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 260–276, 2010.

[24] G. Huang, J. Benesty, and J. Chen, "Subspace superdirective beamforming with uniform circular microphone arrays," in *2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 1–5, 2016.

[25] J. Benesty, J. Chen, and I. Cohen, *Design of Circular Differential Microphone Arrays*, vol. 12 of *Springer Topics in Signal Processing*. Cham: Springer International Publishing, 2015.

[26] S. Bai, J. Z. Kolter, and V. Koltun, "An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling," *arXiv:1803.01271 [cs]*, 2018.

[27] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[28] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, and S. Watanabe, "Tf-gridnet: Integrating full-and sub-band modeling for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 3221–3236, 2023.

[29] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1800–1807, 2017.

[30] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, *et al.*, "The ami meeting corpus: A pre-announcement," in *Machine Learning for Multimodal Interaction: Second International Workshop, MLMI 2005, Edinburgh, UK, July 11-13, 2005, Revised Selected Papers 2*, pp. 28–39, Springer, 2006.

[31] Y. Fu, L. Cheng, S. Lv, Y. Jv, Y. Kong, Z. Chen, Y. Hu, L. Xie, J. Wu, H. Bu, X. Xu, J. Du, and J. Chen, "AISHELL-4: An Open Source Dataset for Speech Enhancement, Separation, Recognition and Speaker Diarization in Conference Scenario," in *Proc. Interspeech 2021*, pp. 3665–3669, 2021.

[32] F. Landini, A. Lozano-Diez, M. Diez, and L. Burget, "From simulated mixtures to simulated conversations as training data for end-to-end neural diarization," in *Interspeech 2022*, pp. 5095–5099, 2022.

[33] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5206–5210, IEEE, 2015.

[34] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 351–355, IEEE, 2018.

[35] N. Zeghidour and D. Grangier, "Wavesplit: End-to-end speech separation by speaker clustering," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2840–2849, 2021.

[36] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.