

Neural Multichannel Distant Speaker Diarization With Heavy-tailed Source Separation Model

Sicheng Mao* Baihan Li*[†] Mathieu Fontaine* Anthony Larcher[‡] Roland Badeau*

* LTCI, Telecom Paris, Institut Polytechnique Paris, France

[†] SJTU Paris Elite Institute of Technology, Shanghai Jiao Tong University, China

[‡] LIUM, Le Mans University, France

Abstract—Distant speaker diarization remains challenging due to difficult acoustic environments, varying numbers of speakers and overlapping speech. Model-driven methods are proposed to exploit the speech source features in multi-channel recordings that help diarization. This paper generalizes a neural model that jointly learns to perform blind source separation and diarization over speech mixtures (neural FCASA) with heavy-tailed models. The popular Gaussian distribution has been applied for variance modeling in the original source separation model, which we replace with two families of heavy-tailed models (Leptokurtic Generalized Gaussian distribution and Student’s t distribution) to better capture the heavy-tailedness in speech signals. Thanks to the Gaussian scale mixture model, we are able to unify the proposed method and the original one under the same form of learning objective. Our experiments show consistent large improvements in Diarization Error Rate (DER) and Jaccard Error Rate (JER) compared to the baseline on various corpora.

Index Terms—heavy-tailed distribution, distant speaker diarization, blind source separation, meeting recording

I. INTRODUCTION

Speaker diarization (SD) addresses the question of “who spoke when” by logging speaker-specific salient events in audio data. Distant speaker diarization remains challenging due to difficult acoustic environments such as meeting scenarios [1–3], where speech signals are often recorded by a distant microphone array and contain more noise and reverberation, with a varying number of speakers and overlapping speech.

To enhance the robustness of a distant diarization system, data-driven paradigm performs data augmentation by noise, reverberation and synthesized overlapping speech and trains powerful speaker feature extractors [4] within pipeline systems [5] or end-to-end neural diarization (EEND) systems [6, 7]. This paradigm recently drives the application of Large Language Models (LLMs) [8, 9] that demonstrate the scaling laws on the diarization task.

On the other hand, model-driven methods still hold certain interest by incorporating traditional signal processing knowledge, especially for multi-channel distant speech recordings. The direction-of-arrival (DOA), the inter-channel time delay (ITD) and other distant speech specific features captured by multi-channel recordings can be used to characterize the speech sources and hence contribute to multi-channel diarization tasks. [10, 11] incorporate beamforming models and extract the steering vector of a signal to enhance the speech features. The multi-channel EEND [12] adapts the Transformer encoder to extract spatio-temporal information and the neural FCASA [13]

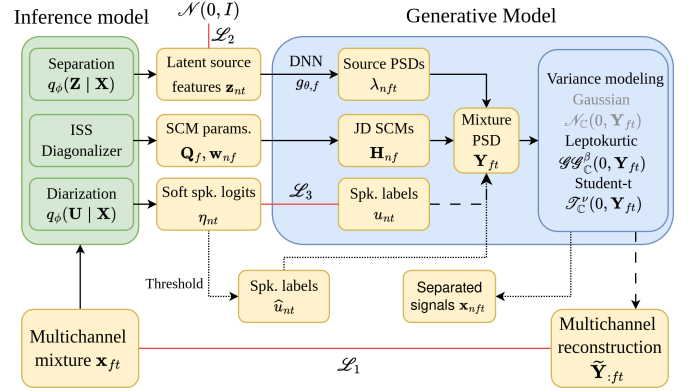


Fig. 1: The overview of our model, training and inference. Dashed: training pipeline; Dotted: inference pipeline; Solid: both; Red: loss terms. The replaced Gaussian model is faded in gray.

incorporates a source separation model that performs joint speaker separation and diarization.

While the neural FCASA focuses more on enhancing the source separation performance with diarization information, from a multi-task learning perspective, it is also beneficial to the diarization system with an incorporated separation model, because, intuitively, a well-separated speech mixture would reduce the speaker diarization to a simpler voice activity detection task. Therefore, we hypothesize that a better separation model would benefit the diarization performance more. It is known that heavy-tailed separation models [14] are more robust to a wide range of difficult scenarios (e.g. noisy conditions, dynamic range, reverberation etc.). Hence, in this work, we replace the Gaussian model in neural FCASA by heavy-tailed distributions from two typical families: the Leptokurtic Generalized Gaussian distribution and the Student’s t distribution. We then train models and conduct a thorough evaluation of diarization performance over various corpora including AMI [15], AliMeeting [16] and CHiME-6 [2] across different setups and parameter settings. Our results show that this method outperforms the baseline by a large margin on these corpora.

The main contribution of this work is to propose a neural speaker diarization method with a heavy-tailed separation model as an improvement and generalization of the neural FCASA.

II. BLIND SOURCE SEPARATION AND DIARIZATION IN THE NEURAL FCASA

This section describes the formulation of the neural FCASA, which we further extend with heavy-tailed separation models.

A. Generative model of multichannel mixture signal

Let us denote $\mathbf{x}_{ft} \in \mathbb{C}^M$ as an M -channel mixture signal, which is a sum of N image signals $\mathbf{x}_{nft}$ generated from a source signal $s_{nft} \in \mathbb{C}$, represented in the Short Time Fourier Transform (STFT) domain. The neural FCASA has been proposed to jointly train the separation and diarization by introducing a source activity temporal mask $u_{nt} \in \{0, 1\}$ to the typical blind source separation model [17, 18]:

$$\mathbf{x}_{ft} = \sum_{n=1}^N u_{nt} \mathbf{x}_{nft} = \sum_{n=1}^N u_{nt} \mathbf{a}_{nf} s_{nft}, \quad (1)$$

where $\mathbf{a}_{nf} \in \mathbb{C}^M$ is the steering vector for source n , and $t = 1, \dots, T$ and $f = 1, \dots, F$ represent the time and frequency indices, respectively.

Following the typical variance modeling [19], each s_{nft} is assumed to independently follow a circularly-symmetric complex Gaussian distribution with its power spectrum density (PSD) $\lambda_{nft} \in \mathbb{R}_+$ as the variance of the Gaussian model:

$$s_{nft} \sim \mathcal{N}_{\mathbb{C}}(0, \lambda_{nft}). \quad (2)$$

By marginalizing s_{nft} from Eqs. (1) and (2), the mixture signal model is obtained as:

$$\mathbf{x}_{ft} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \mathbf{Y}_{ft}), \quad (3)$$

where $\mathbf{Y}_{ft} \triangleq \sum_{n=1}^N u_{nt} \lambda_{nft} \mathbf{H}_{nf}$ and $\mathbf{H}_{nf} = \mathbf{a}_{nf} \mathbf{a}_{nf}^H \in \mathbb{S}_+^{M \times M}$ is called the spatial covariance matrix (SCM) [17, 20] of source n at frequency f that characterizes the signal propagation under a certain geometric configuration.

To facilitate the inference, SCMs are further assumed to be jointly diagonalizable (JD) by a common projection matrix $\mathbf{Q}_f \in \mathbb{C}^{M \times M}$ across all the sources [21, 22]:

$$\mathbf{H}_{nf} = \mathbf{Q}_f^{-1} \text{diag}(\mathbf{w}_{nf}) \mathbf{Q}_f^{-H}, \quad (4)$$

where $\mathbf{w}_{nf} \in \mathbb{R}_+^M$ are the diagonal coefficients for source n . Both $\mathbf{Q} \triangleq \{\mathbf{Q}_f\}_{f=1}^F$ and $\mathbf{W} \triangleq \{\mathbf{w}_{nf}\}_{n,f=1}^{N,F}$ can be efficiently estimated by the iterative source steering (ISS) algorithm [21, 23].

Finally, the deep spectral modeling is applied to the source PSDs λ_{nft} . It introduces D -dimensional latent source features $\mathbf{z}_{nt} \in \mathbb{R}^D$ to generate the PSD λ_{nft} with a deep neural network (DNN) $g_{\theta,f} : \mathbb{R}^D \rightarrow \mathbb{R}_+$ as follows:

$$\lambda_{nft} = g_{\theta,f}(\mathbf{z}_{nt}), \quad (5)$$

where θ represents the model parameters of $g_{\theta,f}$ and the latent features $\mathbf{z}_{nt}$ are supposed to be spectral characteristics (e.g., pitches and envelopes) of the n -th source. Such a neural generative model can be trained as the decoder of a variational autoencoder (VAE) [24–27] for isolated signals by assuming $\mathbf{z}_{nt} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

B. Inference model

Given the observed mixture signals $\mathbf{X} \triangleq \{\mathbf{x}_{ft}\}_{f,t=1}^{F,T}$ as input, an inference model h_{ϕ} is designed to estimate its latent source features $\mathbf{Z} \triangleq \{\mathbf{z}_{nt}\}_{n,t=1}^{N,T}$, speaker activity masks $\mathbf{U} \triangleq \{u_{nt}\}_{n,t=1}^{N,T}$, projection matrices $\mathbf{Q}$ and the diagonal coefficients $\mathbf{W}$. To this end, h_{ϕ} is constructed in a hybrid manner, which contains a neural encoder modeling the posterior distribution $q_{\phi}(\mathbf{Z}|\mathbf{X})$, $q_{\phi}(\mathbf{U}|\mathbf{X})$ of $\mathbf{Z}$ and $\mathbf{U}$ and an ISS diagonalizer [28] computing $\mathbf{Q}_f$ and $\mathbf{w}_{nf}$

$$\{\mathbf{Q}, \mathbf{W}, q_{\phi}(\mathbf{Z}|\mathbf{X}), q_{\phi}(\mathbf{U}|\mathbf{X})\} \leftarrow h_{\phi}(\mathbf{X}). \quad (6)$$

where the posterior variational distribution $q_{\phi}(\mathbf{Z}|\mathbf{X})$ is assumed to follow Gaussian distribution parametrized by mean $\mu_{\phi,ntd} \in \mathbb{R}$ and variance $\sigma_{\phi,ntd}^2 \in \mathbb{R}_+$ and $q_{\phi}(\mathbf{U}|\mathbf{X})$ to Bernoulli distribution parametrized by the logits $\eta_{\phi,nt} \in [0, 1]$.

Once all the parameters are inferred from Eq. (6), the separation is done by estimating isolated image signals $\mathbf{x}_{nft}$ using a multichannel Wiener filter [18, 21] $\mathbf{x}_{nft} = u_{nt} \mathbf{Y}_{nft} \mathbf{Y}_{ft}^{-1} \mathbf{x}_{ft}$ and the diarization is done by thresholding the logits $\eta_{\phi,nt}$ to 0 or 1.

III. HEAVY-TAILED NEURAL BLIND SOURCE SEPARATION AND DIARIZATION

We propose the heavy-tailed neural FCASA as an extension to the neural FCASA in Eq. (2) to better estimate the PSD of the isolated signals by heavy-tailed models in order to improve the diarization performance.

A. Heavy-tailed modeling for power spectrum density

A typical variance modeling in source separation assumes the Gaussianity of the signal's time-frequency representation. However, heavy-tailed models show better robustness against impulsive noise or uncommon scenario [14]. Moreover, the real meeting recordings have a wider dynamic range of the overlap ratio [2] comparing to full-overlap signals in source separation task, hence such signals have a stronger heavy-tailed property.

A heavy-tailed extension can be simply formulated by adding a random positive impulse variable r_{ft} to the Gaussian variance, leading to a (circularly-symmetric) Gaussian Scale Mixture (GSM) model:

$$s_{nft}|r_{ft} \sim \mathcal{N}_{\mathbb{C}}(0, r_{ft} \lambda_{nft}) \quad (7)$$

where we can assume different prior distribution of r_{ft} . By marginalizing r_{ft} , we can get a specific heavy-tailed model for the mixture signal $\mathbf{x}_{ft}$. Particularly in this paper, we are interested in two common heavy-tailed model families, the Leptokurtic Generalized Gaussian distribution and the Student's t distribution [14] (Fig. 2):

$$\begin{aligned} \mathbf{x}_{ft} &\sim \mathcal{GG}_{\mathbb{C}}^{\beta}(0, \mathbf{Y}_{ft}) \quad (\text{Leptokurtic}), \\ \mathbf{x}_{ft} &\sim \mathcal{T}_{\mathbb{C}}^{\nu}(0, \mathbf{Y}_{ft}) \quad (\text{Student's t}), \end{aligned} \quad (8)$$

where $\beta \in (0, 2)$, $\nu \in \mathbb{R}^+$ are shape parameters for Leptokurtic Generalized Gaussian and Student's t distribution respectively. Note that the GSM reduces to a Gaussian model hence to the vanilla neural FCASA when r_{ft} is deterministic.

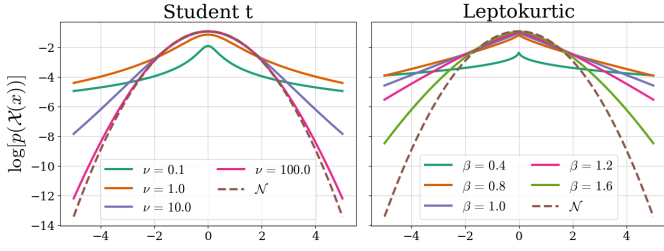


Fig. 2: Standard univariate heavy-tailed models

Left: Student's t distributions with degrees of freedom $\nu \in \mathbb{R}^+$. Right: Leptokurtic Generalized Gaussian distributions with shape parameters $\beta \in (0, 2)$. They tend to Gaussian when $\nu \rightarrow +\infty$ and $\beta \rightarrow 2$, respectively

B. Training and inference

We train the model parameters in Eq. (8) by maximizing its log-likelihood (LL) $\log p_\theta(\mathbf{X}, \mathbf{U})$ of the mixture observations $\mathbf{X}$ and the speaker label $\mathbf{U}$, which is similar to [13, 18]. We use the amortized variational inference to optimize the evidence lower bound (ELBO) of the LL [27, 29]:

$$\begin{aligned} \log p_\theta(\mathbf{X}, \mathbf{U}) &\geq \underbrace{\mathbb{E}_{q_\phi(\mathbf{Z}|\mathbf{X}, \mathbf{U})}[\log p_\theta(\mathbf{X} | \mathbf{U}, \mathbf{Z})]}_{\triangleq \mathcal{L}_1} \\ &\quad \underbrace{-\text{KL}[q_\phi(\mathbf{Z} | \mathbf{X}, \mathbf{U}) || p_\theta(\mathbf{Z})]}_{\triangleq \mathcal{L}_2} + \underbrace{\log p(\mathbf{U})}_{\triangleq \mathcal{L}_3}, \end{aligned} \quad (9)$$

where KL stands for the Kullback-Leiber (KL) divergence. For simplicity, we approximate the variational family $q_\phi(\mathbf{Z}|\mathbf{X}, \mathbf{U})$ by the inference model $q_\phi(\mathbf{Z}|\mathbf{X})$.

For the first term, we use the one-point Monte-Carlo estimation with one sample $\mathbf{Z}$ sampled from $q(\mathbf{Z}|\mathbf{X})$ and $\mathbf{U}$ directly from the ground-truth label as teacher-forcing. According to [14], the conditional likelihood is:

$$\mathcal{L}_1 = \sum_{f,t=1}^{F,T} \tilde{\mathcal{L}}_1 + T \sum_{f=1}^F \log |\mathbf{Q}_f \mathbf{Q}_f^H|, \quad (10)$$

where $\tilde{\mathcal{L}}_1$ is (up to an additive constant)

$$-\left(\sum_{m=1}^M \frac{|\tilde{x}_{ftm}|^2}{\tilde{y}_{ftm}} \right)^{\frac{\beta}{2}} - \sum_{m=1}^M \log \tilde{y}_{ftm}, \quad (11)$$

for Leptokurtic Generalized Gaussian distribution and

$$-\left(\frac{\nu}{2} + M \right) \log \left(1 + \frac{2}{\nu} \sum_{m=1}^M \frac{|\tilde{x}_{ftm}|^2}{\tilde{y}_{ftm}} \right) - \sum_{m=1}^M \log \tilde{y}_{ftm}, \quad (12)$$

for Student's t distribution, where $\tilde{x}_{ftm}, \tilde{y}_{nftm}$ are the m -th entry of $\tilde{\mathbf{x}}_{ft} \triangleq \mathbf{Q}_f \mathbf{x}_{ft}, \tilde{\mathbf{y}}_{nft} \triangleq u_{nt} g_{\theta,f}(\mathbf{z}_{nt}) \mathbf{w}_{nf}$ and $\tilde{y}_{ftm} \triangleq \sum_{n=1}^N \tilde{y}_{nftm}$.

The second term is a KL-divergence between two Gaussians and can be computed in closed form [27]. For the third term, as we don't assume a prior model $p(\mathbf{U})$ for the speaker activity mask $\mathbf{U}$ [13, 30], instead, we directly maximize the log-posterior of surrogate model, $\log q_\phi(\mathbf{U}|\mathbf{X})$, which corresponds

to minimize the KL-divergence between the empirical posterior distribution, $p_{data}(\mathbf{U}|\mathbf{X})$ and the surrogate posterior, $q_\phi(\mathbf{U}|\mathbf{X})$, which is equivalent to minimizing the binary cross entropy (BCE) between the ground truth label, u_{nt} , and the logit $\eta_{\phi,nt}$, i.e.

$$\mathcal{L}_3 \approx \log q_\phi(\mathbf{U}|\mathbf{X}) = - \sum_{n,t=1}^{N,T} \text{BCE}[u_{nt} | \eta_{\phi,nt}]. \quad (13)$$

Note that this shares the same insights as the extended objective function in the semi-supervised VAE [29], where we want to train the inference model for $\mathbf{U}$ directly from the labelled data as well.

Meanwhile, the permutation invariant training (PIT) is applied to solve the speaker permutation ambiguity [30]. Combining them altogether, we deduce the final form of our maximization objective:

$$\mathcal{L} = \frac{1}{TF} (\mathcal{L}_1 + \delta \mathcal{L}_2) + \gamma \frac{1}{TN} \mathcal{L}_3, \quad (14)$$

where δ and γ are scaling hyperparameters. It turns out that the heavy-tailed neural FCASA can fully inherit the neural FCASA training and inference procedure by just replacing the original objective to Eq. (14), leading to a straightforward extension.

IV. EXPERIMENTS

To fairly compare with the neural FCASA as our baseline model, we follow the same experimental settings as in [13]¹.

A. Dataset description

We evaluate the proposed method on three public corpora that provide a multi-channel far-field subset:

- 1) AMI corpus [15]. This dataset contains 100 hours of 16 kHz English meeting recordings with 3 to 5 participants. Audio was captured using an 8-mic circular array with a radius of 10 cm placed on the table. Our experiments adopt the official split: training (80.7h) for training, development (9.7h) for validation and evaluation (9.1h) for evaluation.
- 2) AliMeeting corpus [16]. This dataset contains 118.75 hours of 16 kHz Mandarin meeting recordings with 2 to 4 participants. Audio was captured using an 8-mic circular array with a radius of 5 cm placed on the table. Our experiments adopt the official split: training (104.75h) for training, evaluation (4h) for validation and test (10h) for evaluation.² Besides, we use the aligned RTTM transcription from [31] as the ground truth labels instead of the original one.
- 3) CHiME-6 corpus [32] contains 50 hours of 16 kHz English dinner party conversation recordings with 4 participants. Audio was captured using 6 distant Microsoft Kinect devices. Each device has a linear array of 4 sample-synchronized microphones. The participants

¹Codes available at <https://github.com/alephi/neural-fcasa>.

²Note that the dataset split naming convention of AliMeeting is different from the one of AMI or CHiME-6.

are moving around different rooms (kitchen, dining room and living room) in the house. Our experiments adopt the official split: `training` (40.5h) for training, `development` (4.5h) for validation and `evaluation` (5h) for evaluation. For a fair comparison with AMI and AliMeeting, we only use the recordings from device 1 and 2 and merge them as 8 channel recordings.

B. Configurations

The network architecture of the heavy-tailed neural FCASA follows that of the vanilla neural FCASA and we refer the reader to [13] for more details. All the signals are dereverberated in advance using weighted prediction error (WPE) [33]. The spectrograms are generated via STFT with a window size of 512 and a hop size of 160. The source number $N = 6$ containing 5 speaker channels and a noise channel. The latent dimension $D = 10$ for the noise channel and $D = 64$ for the speaker channels to avoid channel modeling ambiguity [34]. Speaker activations u_{nt} are derived from the oracle diarization labels, while the noise channel is always set to active to prevent confusion with silence. δ, γ in Eq. (14) are both set to 1.0. During training, the audio recordings are segmented into 20s clips, from which 10s continuous crops are randomly sampled and fed into the training pipeline with a batch size of 128. The model is trained for 200 epochs using the AdamW optimizer with a learning rate of 1e-4 and a weight decay of 1e-5. For evaluation, the audio recordings are segmented into 10s clips and fed into the inference pipeline. The speaker activity predictions $\eta_{\phi, nt}$ are smoothed by a median filter of 11 frames for stability and then cast to 0 or 1 using a threshold of 0.5.

We use Pyannote [35] to report the diarization error rate (DER) and its three components: miss, false alarm (FA) and speaker confusion (Conf.), as well as the Jaccard Error Rate (JER).³ To evaluate the diarization performance under different standards, we adopt four different evaluation setups. Three of them follow the convention proposed in [5], where Forgiving applies a 0.25s collar without overlap, Fair applies a 0.25s collar within overlap and Full applies no collar within overlap. In addition, we adopt the Overlap setup, which applies no collar but evaluates only on the overlap, since the overlapping part of speech is a major challenge for diarization systems.

We first train all our models on the AMI corpus training set. We have trained five models in the Leptokurtic Generalized Gaussian family and four in the Student's t family, with shape parameters $\beta \in \{0.4, 0.8, 1.0, 1.2, 1.6\}$ and $\nu \in \{0.1, 1, 10, 100\}$ respectively. We also retrained a baseline Gaussian model proposed in the neural FCASA. Note that the Gaussian model can be regarded as Leptokurtic Generalized Gaussian with $\beta = 2$ or Student's t with $\nu \rightarrow \infty$. For each model setting, we track the diarization loss $\mathcal{L}_3$ on the validation set and find the best checkpoints and evaluate all the metrics on the evaluation set.

³Note that we do not evaluate the objective source separation metrics since all of these corpora consist of real conversations without isolated ground truth.

Once found the optimal parameter settings on the AMI corpus, we launch the model training and evaluate their performances on AliMeeting corpus and CHiME-6 corpus to verify the generalization ability of our method. We also perform cross-tests for models trained on one corpus and test them on another corpus.

TABLE I: Diarization metrics of the trained models on AMI evaluation set. All metrics are reported in percentage with the best score in **bold** and the second-best score underlined. We also list the single channel diarization baseline Pyannote 3.1 [35] as a reference. Note that the evaluation of our method is performed chunk-wisely (10 seconds), which is not directly comparable to the Pyannote baselines.

Model	Shape param.	Setup	Miss (↓)	FA (↓)	Conf. (↓)	DER (↓)	JER (↓)
Gaussian (baseline)	-	Forgiving	5.59	8.16	0.73	14.48	13.50
		Fair	8.61	6.33	0.79	15.73	23.16
		Full	10.77	6.81	1.16	18.73	27.65
		Overlap	19.40	3.80	0.90	24.11	26.53
Leptokurtic	0.4	Forgiving	7.98	6.23	0.86	15.07	15.81
		Fair	11.52	4.97	0.92	17.41	27.33
		Full	13.85	5.42	1.28	20.54	31.52
		Overlap	24.07	3.05	0.90	28.02	31.35
	0.8	Forgiving	6.14	7.84	0.83	14.81	13.98
		Fair	9.35	6.29	0.96	16.61	24.36
		Full	11.63	6.82	1.30	19.75	28.70
		Overlap	20.79	4.20	0.96	25.95	28.10
	1.0	Forgiving	5.45	7.36	0.79	13.60	13.01
		Fair	8.84	5.76	0.80	15.41	22.66
		Full	11.08	6.26	1.13	18.47	27.38
		Overlap	20.13	3.52	0.80	24.45	27.02
	1.2	Forgiving	5.13	9.12	0.86	15.11	13.90
		Fair	8.30	6.80	0.95	16.06	23.59
		Full	10.38	7.58	1.34	19.30	28.17
		Overlap	18.98	4.11	1.03	24.12	26.78
	1.6	Forgiving	6.04	6.65	0.71	13.39	13.32
		Fair	9.55	5.08	0.75	15.39	23.33
		Full	11.82	5.65	1.06	18.52	27.92
		Overlap	21.09	3.10	0.74	24.93	27.82
Student's t	0.1	Forgiving	5.04	5.59	0.61	11.24	10.78
		Fair	<u>7.49</u>	4.42	0.62	12.53	18.38
		Full	<u>9.23</u>	5.86	0.92	16.01	23.45
		Overlap	<u>16.18</u>	3.18	0.67	<u>20.03</u>	<u>21.99</u>
	1	Forgiving	4.30	6.44	0.67	11.41	11.04
		Fair	6.92	<u>5.02</u>	<u>0.69</u>	<u>12.62</u>	18.80
		Full	8.55	6.52	0.98	16.05	23.82
		Overlap	15.57	<u>3.44</u>	<u>0.70</u>	19.71	21.92
	10	Forgiving	5.45	7.58	0.78	13.80	12.17
		Fair	8.55	6.17	0.80	15.52	20.98
		Full	10.93	5.99	1.07	17.99	25.95
		Overlap	19.76	3.71	0.74	24.21	26.01
	100	Forgiving	5.57	6.78	0.72	13.08	12.48
		Fair	8.93	5.39	0.80	15.12	22.09
		Full	11.25	5.96	1.09	18.30	26.82
		Overlap	20.12	3.55	0.78	24.44	26.72
Pyannote 3.1 IDM		Full	9.5	3.6	5.7	18.8	-
Pyannote 3.1 SDM		Full	11.2	3.8	7.5	22.4	-

C. Results and discussions

The evaluation results on AMI corpus are illustrated in Table I. We first observe that for each model, the error rates become larger when the evaluation setup becomes stricter on overlaps. We then observe some heavy-tailed models have consistent improvements against the baseline in error rates across different setups. For example, we have better results from Leptokurtic Generalized Gaussian models with $\beta = 1.0, 1.6$, whereas Student's t models outperform the baseline by a large

margin (2-3% in absolute and 10-15% in relative) with $\nu = \{0.1, 1\}$. This validates the effectiveness of the proposed heavy-tailed neural FCASA method and also demonstrates the choice of the shape parameters substantially affects model performance. Note that for both heavy-tailed families, a smaller parameter implies a heavier tail. In particular, when $\nu = 1$, the Student's t corresponds to the Cauchy distribution, which preserves the law under linear combination, making it naturally compatible with the linear model assumption in Eq. (1). Larger shape values ($\nu = 100, \beta = 1.6$) converge towards Gaussian model leading to worse performance. Also, there are high potentials that we could find a better performing shape parameter when $\beta \in [1.0, 1.6]$ or $\nu \in [0.1, 1]$.

TABLE II: Diarization metrics of the trained models on AliMeeting test set. All metrics are reported in percentage with the best score in bold and the second-best score underlined. We also list the single channel diarization baseline Pyannote 3.1 [35] as a reference. Note that the evaluation of our method is performed chunk-wisely (10 seconds), which is not directly comparable to the Pyannote baseline.

Model	Shape param.	Setup	Miss (↓)	FA (↓)	Conf. (↓)	DER (↓)	JER (↓)
Gaussian (baseline)	-	Forgiving	2.52	2.53	0.37	5.42	7.04
		Fair	3.46	1.90	0.40	5.75	10.29
		Full	7.25	<u>4.52</u>	0.70	12.47	19.89
		Overlap	14.98	2.00	0.47	17.46	20.25
Student's t	0.1	Forgiving	2.28	2.35	0.22	4.85	<u>5.97</u>
		Fair	2.93	1.76	0.22	4.91	8.46
		Full	6.38	4.60	0.49	<u>11.47</u>	18.11
		Overlap	13.42	2.01	0.35	15.78	18.28
	1	Forgiving	1.87	2.16	0.28	4.32	5.88
		Fair	2.62	1.59	0.28	4.49	8.47
		Full	6.41	4.09	0.52	11.03	18.18
		Overlap	<u>14.17</u>	1.76	0.33	<u>16.26</u>	<u>18.88</u>
Pyannote 3.1 channel 1		Full	10.0	4.4	10.0	24.4	-

TABLE III: Diarization metrics of the trained models on CHiME-6 evaluation set. All metrics are reported in percentage with the best score in bold and the second-best score underlined. We also list the single channel diarization official baseline from CHiME-6 [2] as a reference. Note that the evaluation of our method is performed chunk-wisely (10 seconds), which is not directly comparable to the CHiME-6 baseline.

Model	Shape param.	Setup	Miss (↓)	FA (↓)	Conf. (↓)	DER (↓)	JER (↓)
Gaussian (baseline)	-	Forgiving	10.73	61.51	6.96	79.20	42.08
		Fair	14.72	38.46	7.46	60.65	49.27
		Full	17.48	35.61	8.64	61.74	52.89
		Overlap	29.84	6.74	6.56	43.14	51.68
Student's t	0.1	Forgiving	9.16	30.47	5.75	<u>45.37</u>	32.40
		Fair	14.83	19.30	6.22	40.35	41.67
		Full	17.85	29.01	7.23	54.10	46.61
		Overlap	27.15	<u>10.95</u>	6.14	<u>44.25</u>	46.60
	1	Forgiving	15.89	19.20	6.58	41.67	38.41
		Fair	20.52	12.54	6.80	39.86	46.97
		Full	23.72	15.54	7.57	46.84	51.44
		Overlap	32.61	3.66	6.08	42.35	51.89
CHiME-6 official baseline Single channel		Full	-	-	-	62.0	71.4

The performances of models trained respectively on AliMeeting and CHiME-6 training sets with optimal parameter settings found on AMI corpus (i.e. Student's t distribution with $\nu \in \{0.1, 1\}$) are reported in Table II and Table III. We observe consistent improvements against the baseline. For AliMeeting corpus, we observe 1%-2% in absolute (10%-20% in relative)

TABLE IV: Diarization metrics of the cross-tests. All metrics are reported percentage. We only select the DER of the Full setup for simplicity.

Model	Shape param.	test train	AMI	AliMeeting	CHiME-6
Gaussian (baseline)	-	AMI	18.73	41.99	80.53
		AliMeeting	44.11	12.47	70.71
		CHiME-6	47.06	38.17	61.74
Student's t	0.1	AMI	16.01	39.35	64.25
		AliMeeting	43.28	11.47	74.41
		CHiME-6	75.45	43.15	54.10
	1	AMI	16.05	38.48	80.96
		AliMeeting	43.33	11.03	65.52
		CHiME-6	40.21	35.94	46.84

improvements compared to the baseline. For CHiME-6 corpus, we even observe up to 20% in absolute (35% in relative) improvements compared to the baseline, due to the difficulty of CHiME-6 corpus and the relatively poor performance of the baseline. Therefore, this verifies the generalization ability of our proposed method and especially justifies the necessity of the heavy-tailed models for difficult acoustic environments.

The cross-tests results are reported in Table IV. We observe that for train-test pairs such as AMI-AliMeeting and AliMeeting-AMI, there are minor but consistent improvements over the baseline. However, the performances are still far worse than the in-domain results. This can be justified by the similarity of the meeting scenario and the microphone array configuration between AMI and AliMeeting and the major mismatch is the working language (English vs. Mandarin). On the other hand, cross-tests involving CHiME-6 show less consistency, we think this is due to the large gap between the scenario (meeting vs. party, steady vs. moving), the microphone array configuration (circular vs. linear) and language.

D. Limitations

We notice that for some parameter configurations of Lep-tokurtic Generalized Gaussian distributions, there is a diarization performance degradation compared to the baseline, e.g. for $\beta \in \{0.4, 0.8, 1.2\}$. This suggests the potential sensitivity of our method to hyperparameter tuning. We will leave the investigation and improvements as future work such as direct estimation of those parameters.

V. CONCLUSION AND FUTURE WORK

We propose a heavy-tailed extension to the neural FCASA system which jointly performs speaker separation and diarization. The proposed method can be formulated as a Gaussian scale mixture model with a random impulse variable to the variance, resulting in a fairly simple objective modification of the original baseline. Through extensive experiments, we find various parameter settings that outperform the baseline by a large margin. In future work, we would like to explore the extension to more heavy-tailed model families and their performance on speaker separation. Alternative directions include finding a better shape parameter tuning strategy, incorporating the speaker activity model into the heavy-tailed

modeling [36] and incorporating the geometrical knowledge of microphone array [37].

VI. ACKNOWLEDGEMENT

We acknowledge the use of Claude (<https://claude.ai/>) to polish this paper. We pasted the draft of the abstract, the first three paragraphs of Section I, Section IV-A and Section IV-B with the instruction “Please rephrase the text in an academic manner”. The outputs were then modified further to better represent our own writing.

We thank Dr. Yoshiaki Bando for insightful discussions. This project is funded by ANR Project SAROUMANE (ANR-22-CE23-0011) and granted access to the HPC resources of IDRIS under the allocation 20191014876 attribution made by GENCI.

VII. REFERENCE

- [1] J. M. Pardo et al., “Speaker diarization for multiple distant microphone meetings: Mixing acoustic features and inter-channel time differences,” in *Proc. Interspeech*, 2006, paper 1337-Thu1A10.5–.
- [2] S. Watanabe et al., “CHiME-6 Challenge: Tackling Multi-speaker Speech Recognition for Unsegmented Recordings,” in *Proc. CHiME*, 2020, pp. 1–7.
- [3] S. Cornell et al., “The CHiME-8 DASR challenge for generalizable and array agnostic distant automatic speech recognition and diarization,” in *8th International Workshop on Speech Processing in Everyday Environments*, 2024, pp. 1–6.
- [4] D. Snyder et al., “Deep Neural Network Embeddings for Text-Independent Speaker Verification,” in *Proc. Interspeech*, ISCA, 2017, pp. 999–1003.
- [5] F. Landini et al., “Bayesian HMM clustering of x-vector sequences (VBx) in speaker diarization: Theory, implementation and analysis on standard tasks,” *Comput. Speech Lang.*, vol. 71, p. 101254, 2022.
- [6] Y. Fujita et al., “End-to-End Neural Speaker Diarization with Permutation-free Objectives,” in *Interspeech*, 2019, pp. 4300–4304.
- [7] S. Horiguchi et al., “Encoder-Decoder Based Attractors for End-to-End Neural Diarization,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 1493–1507, 2022.
- [8] Q. Wang et al., “DiarizationLM: Speaker diarization post-processing with large language models,” in *Proc. Interspeech*, 2024, pp. 3754–3758.
- [9] H. Yin et al., *SpeakerLM: End-to-End Versatile Speaker Diarization and Recognition with Multimodal Large Language Models*, 2025. arXiv: 2508.06372 [cs.LG].
- [10] X. Anguera et al., “Acoustic beamforming for speaker diarization of meetings,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 7, pp. 2011–2022, 2007.
- [11] T. Mariotte et al., “ASoBO: Attentive beamformer selection for distant speaker diarization in meetings,” in *Proc. Interspeech*, 2024, pp. 1620–1624.
- [12] S. Horiguchi et al., “Multi-Channel End-to-End Neural Diarization with Distributed Microphones,” in *Proc. ICASSP*, 2022, pp. 7332–7336.
- [13] Y. Bando et al., “Neural Blind Source Separation and Diarization for Distant Speech Recognition,” in *Proc. Interspeech*, 2024, pp. 722–726.
- [14] M. Fontaine et al., “Generalized Fast Multichannel Nonnegative Matrix Factorization Based on Gaussian Scale Mixtures for Blind Source Separation,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 1734–1748, 2022.
- [15] J. Carletta et al., “The AMI meeting corpus: A pre-announcement,” in *Proceedings of the Second International Conference on Machine Learning for Multimodal Interaction*, ser. MLMI’05, Berlin, Heidelberg: Springer-Verlag, 2005, pp. 28–39.
- [16] F. Yu et al., “M2MeT: The ICASSP 2022 multi-channel multi-party meeting transcription challenge,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, 2022, pp. 6167–6171. arXiv: 2110.07393 [eess.AS].
- [17] N. Q. K. Duong et al., “Under-Determined Reverberant Audio Source Separation Using a Full-Rank Spatial Covariance Model,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1830–1840, 2010.
- [18] Y. Bando et al., “Neural Fast Full-Rank Spatial Covariance Analysis for Blind Source Separation,” in *Proc. EUSIPCO*, IEEE, 2023, pp. 51–55.
- [19] E. Vincent, “Probabilistic modeling paradigms for audio source separation,” in *Machine Audition: Principles, Algorithms and Systems*, 2011.
- [20] H. Sawada et al., “Multichannel Extensions of Non-Negative Matrix Factorization With Complex-Valued Data,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 5, pp. 971–982, 2013.
- [21] K. Sekiguchi et al., “Fast Multichannel Nonnegative Matrix Factorization With Directivity-Aware Jointly-Diagonalizable Spatial Covariance Matrices for Blind Source Separation,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2610–2625, 2020.
- [22] N. Ito and T. Nakatani, “FastMNMF: Joint Diagonalization Based Accelerated Algorithms for Multichannel Nonnegative Matrix Factorization,” *Proc. ICASSP*, pp. 371–375, 2019.
- [23] R. Scheibler and N. Ono, “Fast and Stable Blind Source Separation with Rank-1 Updates,” in *Proc. ICASSP*, 2020, pp. 236–240.
- [24] S. Leglaive et al., “A Recurrent Variational Autoencoder for Speech Enhancement,” *Proc. ICASSP*, pp. 371–375, 2020.
- [25] Y. Bando et al., “Neural Full-Rank Spatial Covariance Analysis for Blind Source Separation,” *IEEE Signal Process. Lett.*, vol. 28, pp. 1670–1674, 2021.
- [26] L. Li et al., “FastMVAE2: On improving and accelerating the fast variational autoencoder-based source separation algorithm for determined mixtures,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 96–110, 2022.
- [27] D. P. Kingma and M. Welling, “Auto-Encoding Variational Bayes,” in *Proc. ICLR*, 2014.
- [28] R. Scheibler and M. Togami, “Surrogate Source Model Learning for Determined Source Separation,” in *Proc. ICASSP*, 2021, pp. 176–180.
- [29] D. P. Kingma et al., “Semi-supervised learning with deep generative models,” in *NIPS*, vol. 27, 2014.
- [30] S. Maiti et al., “EEND-SS: Joint End-to-End Neural Speaker Diarization and Speech Separation for Flexible Number of Speakers,” in *Proc. SLT*, 2023, pp. 480–487.
- [31] S. Horiguchi et al., “Can we really repurpose multi-speaker asr corpus for speaker diarization?” In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2025.
- [32] S. Watanabe et al., “CHiME-6 challenge: Tackling multi-speaker speech recognition for unsegmented recordings,” in *6th International Workshop on Speech Processing in Everyday Environments (CHiME 2020)*, 2020, pp. 1–7.
- [33] T. Yoshioka and T. Nakatani, “Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 10, pp. 2707–2720, 2012.

- [34] Y. Bando et al., “Weakly-Supervised Neural Full-Rank Spatial Covariance Analysis for a Front-End System of Distant Speech Recognition,” in *Proc. Interspeech*, 2022, 3824–3828.
- [35] H. Bredin, “pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe,” in *Proc. Interspeech*, 2023.
- [36] S. Mao et al., *Neural multichannel distant speaker diarization and source separation with beta speaker activity prior*, 2026. arXiv: 2608.28661 [cs.SD].
- [37] Y. Sumura et al., “Joint audio source localization and separation with distributed microphone arrays based on spatially-regularized multichannel NMF,” in *2024 18th International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2024, pp. 145–149.