Cauchy Multichannel Speech Enhancement with a Deep Speech Prior

Mathieu Fontaine* Aditya Arie Nugraha† Roland Badeau‡ Kazuyoshi Yoshii†§ Antoine Liutkus¶

*Université de Lorraine, CNRS, Inria, LORIA, Nancy, France †AIP, RIKEN, Tokyo, Japan
‡LTCI, Télécom ParisTech, Université Paris-Saclay, Paris, France ¶Inria, LIRMM, Montpellier, France
§Graduate School of Informatics, Kyoto University, Kyoto, Japan

***Abstract*—We propose a semi-supervised multichannel speech enhancement system based on a probabilistic model which assumes that both speech and noise follow the heavy-tailed multivariate complex Cauchy distribution. As we advocate, this allows handling strong and adverse noisy conditions. Consequently, the model is parameterized by the source magnitude spectrograms and the source spatial scatter matrices. To deal with the non-additivity of scatter matrices, our first contribution is to perform the enhancement on a projected space. Then, our second contribution is to combine a latent variable model for speech, which is trained by following the variational autoencoder framework, with a low-rank model for the noise source. At test time, an iterative inference algorithm is applied, which produces estimated parameters to use for separation. The speech latent variables are estimated first from the noisy speech and then updated by a gradient descent method, while a majorization-equalization strategy is used to update both the noise and the spatial parameters of both sources. Our experimental results show that the Cauchy model outperforms the state-of-art methods. The standard deviation scores also reveal that the proposed method is more robust against non-stationary noise.**

***Index Terms*—Multichannel speech enhancement, multivariate complex Cauchy distribution, variational autoencoder, nonnegative matrix factorization**

I. Introduction

Multichannel speech enhancement aims to extract speech from an observed multichannel noisy signal [1]. Usually, parametric models are used for the speech (target) signal, the additive noise, and the way both are captured by the microphones. The core feature of such models is to allow the reconstruction of the target signal from the noisy mixture, provided that the parameters are well estimated. The standard example is having sources parameterized by their spectrograms, and reconstructed through soft-masking (Wiener-like) strategies.

Deep neural networks (DNNs) have been increasingly used in this context [1]–[4]. Most approaches train *denoising* DNNs to estimate the model parameters, e.g., source spectrograms or the time-frequency mask of one or all of the sources. In this case, the training employs paired data consisting of noisy speech and clean speech. Although it has been shown that denoising DNNs could be robust to unseen environments [5], there is still a concern that they are not adaptive enough to unseen noises.

This issue has recently been addressed by several studies on semi-supervised speech enhancement [6]–[10]. The core idea of these studies is to formulate a probabilistic generative model in which both speech spectrogram and noise spectrogram are modeled by latent variable models. The speech spectrogram model is trained as the decoder in the variational autoencoder (VAE) framework [11], while the noise spectrogram is modeled by a *nonnegative matrix factorization* (NMF) [12] approach. The speech model is trained with clean speech only, thus independent from the actual noise that will be found in the observations at test time. Indeed, the core feature of this strategy is to let the noise parameters be estimated and adapted at test time, so that it is flexible and may achieve good denoising performance even in adversarial conditions not met at training time. In the case of multichannel enhancement [8], [9], the spatial parameters are also estimated at test time.

Most of these methods [6]–[9] tackle such a whole estimation problem under a Gaussian probabilistic setting. Although it is convenient because it leads to straightforward inference methods, it has the drawback of being sensitive to initialization and prone to be trapped in a local minimum [13].

As opposed to their Gaussian counterpart, heavy-tailed probabilistic models allow for outcomes that are far away from the expected values [14], [15]. From an inference perspective, this means that unlikely observations will not have a detrimental impact on the parameters, yielding *robust estimation*. Among them, non-Gaussian α-stable models are remarkable because they also satisfy the central limit theorem [16], which means that a sum of α-stable random vectors remains α-stable. This is an interesting feature in a context of speech enhancement where additive sources are combined to yield the observed mixture. Notwithstanding their attractive features, the main weakness of these distributions is the absence of a closed-form *probability density function* (PDF), except for $\alpha = 0.5$ (the Lévy distribution), $\alpha = 1$ (the Cauchy distribution), and $\alpha = 2$ (the Gaussian distribution).

An option is to express an α-stable random variable as conditionally Gaussian [16]. This may always be done in the scalar (single-channel) case and only in some cases for multichannel data. Put it simply, the trick is to write an α-stable random variable as a Gaussian variable with a covariance that is multiplied by a random *impulse variable*, distributed as a *positive* $\frac{\alpha}{2}$*-stable random variable*. This makes it possible to use the classical

This work was partly supported by the research programme KAMoulox (ANR-15-CE38-0003-01) funded by ANR, the French State agency for research, and JSPS KAKENHI No. 19H04137.

Gaussian methodology, provided that some specific method is found to estimate the impulse variables, notably some Markov chain Monte Carlo (MCMC) strategy [17]–[19]. However, the MCMC algorithm is often computationally demanding at test time, and [20] proposes an approximation to construct filters without requiring the estimation of impulse variables. Still, this strategy does not provide any convenient cost function to use for parameter estimation, which is inconvenient in our case.

In line with the present study, combining a VAE and general α-stable distributions has recently been proposed in [10]. It suffers from expensive MCMC schemes. A simplified approach undertaken in [21] is to focus on the Cauchy $\alpha = 1$ case, for which closed-form expressions of the likelihood are available. However, this study is limited to single-channel source separation based on low-rank source models.

In this paper, we go beyond related work in this respect and introduce a semi-supervised multichannel speech enhancement method that uses a VAE-based speech model and a low-rank noise model, that both parameterize Cauchy models for the sources. To the best of our knowledge, this is the first study that uses a computationally-tractable heavy-tailed model for multichannel *sources* unlike previous studies that use a heavy-tailed model for multichannel *mixtures* [22], [23]. Additionally, we show how deep latent variable models may be combined with more classical low-rank models in this setting.

II. Probabilistic Formulation

This section formulates a probabilistic model for the proposed Cauchy multichannel speech enhancement method.

A. Multivariate Complex Cauchy Distribution

Let $\mathbf{y}$ be a complex random vector of dimension K. Then, $\mathbf{y} \sim \mathcal{C}_{\mathbb{C}}(\mathbf{y}|\boldsymbol{\mu}, \mathbf{V})$ follows a *circularly-symmetric multivariate complex Cauchy distribution* iff. its probability density is

$$p_{\boldsymbol{\mu},\mathbf{V}}(\mathbf{y}) = A_{K,\mathbf{V}} \left(1 + (\mathbf{y}-\boldsymbol{\mu})^{\mathrm{H}} \mathbf{V}^{-1} (\mathbf{y}-\boldsymbol{\mu})\right)^{-K-\frac{1}{2}}, \quad (1)$$

where

$$A_{K,\mathbf{V}} = \prod_{k=1}^{K} \left(K - k + \frac{1}{2}\right) \pi^{-K} \det(\mathbf{V})^{-1} \quad (2)$$

and $.^{\mathrm{H}}$, $\mathbf{V}$ and $\boldsymbol{\mu}$ in (1) respectively stand for the Hermitian transposition, the positive definite *scatter matrix* and the *location parameter* [24], [25]. Similarly to the Gaussian distribution, a linear combination of Cauchy vectors remains a Cauchy vector. However, the tails of a Gaussian distribution are lighter than those of a Cauchy one (see Fig. 1).

B. Spatial Model

We work in the short time Fourier transform (STFT) domain. Let $f \in [1, F]$ and $t \in [1, T]$ be the frequency bin and time frame indexes, respectively. Following the literature, we assume that the observed mixture signal is a linear combination of the sources. Considering speech and noise as the sources, the STFT of a K-channel noisy speech $\mathbf{x} \in \mathbb{C}^{F \times T \times K}$ is expressed for each time-frequency (TF) bin ft as

$$\mathbf{x}_{ft} = \mathbf{x}^{\mathrm{s}}_{ft} + \mathbf{x}^{\mathrm{n}}_{ft}, \quad (3)$$

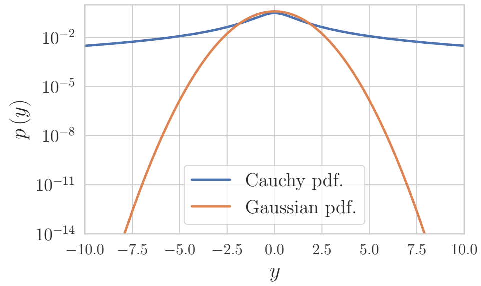


Fig. 1: Tails of unidimensional Cauchy and Gaussian PDF

where the speech $\mathbf{x}^{\mathrm{s}}_{ft} \in \mathbb{C}^K$ and the noise $\mathbf{x}^{\mathrm{n}}_{ft} \in \mathbb{C}^K$ are assumed to follow a Cauchy distribution:

$$\begin{cases} \mathbf{x}^{\mathrm{s}}_{ft} \sim \mathcal{C}_{\mathbb{C}}\left(\mathbf{x}^{\mathrm{s}}_{ft} \middle| \mathbf{0}, a^{\mathrm{s}}_{ft} \mathbf{R}^{\mathrm{s}}_{f}\right), \\ \mathbf{x}^{\mathrm{n}}_{ft} \sim \mathcal{C}_{\mathbb{C}}\left(\mathbf{x}^{\mathrm{n}}_{ft} \middle| \mathbf{0}, a^{\mathrm{n}}_{ft} \mathbf{R}^{\mathrm{n}}_{f}\right), \end{cases} \quad (4)$$

with $a^{\mathrm{s}}_{ft} \in \mathbb{R}_+$ and $a^{\mathrm{n}}_{ft} \in \mathbb{R}_+$ are the sources' magnitudes, while $\mathbf{R}^{\mathrm{s}}_{f}$ and $\mathbf{R}^{\mathrm{n}}_{f}$ are the $K \times K$ positive definite *spatial scatter matrices* of the sources.

C. Source Models

While the scatter matrices $\mathbf{R}^{\mathrm{s}}_{f}$ and $\mathbf{R}^{\mathrm{n}}_{f}$ are left unconstrained, specific models are picked for the speech and noise magnitudes.

First, the F-dimensional vector $\mathbf{a}^{\mathrm{s}}_{t}$ gathering the speech magnitudes a^{s}_{ft} for frame t is assumed to depend on low-dimensional *latent variables* written $\mathbf{z}_t \in \mathbb{R}^D$ with $D < F$. This mapping is given by a function called the *decoder* and written $\mathbf{a}^{\mathrm{s}}_{t} = \mu_\theta(\mathbf{z}_t)$, parameterized by θ.

Second, the noise magnitudes a^{n}_{ft} are modeled with a non-negative matrix factorization (NMF) [26] as follows:

$$a^{\mathrm{n}}_{ft} = \sum_{l=1}^{L} w_{fl} h_{lt} \text{ for } \forall f, t, \quad (5)$$

where L is the number of basis vectors.

At test time, the key idea becomes to use the observed mixture $\mathbf{x}$ to estimate the most likely latent vectors $\mathbf{z}_t$ and NMF parameters w and h as well as the scatter matrices. They are then used in conjunction to $\hat{\mathbf{a}}^{\mathrm{s}}_{t} = \mu_\theta(\mathbf{z}_t)$ to separate the signals with the technique presented in section II-D.

D. Projection-Based Wiener Filter

The Cauchy model above resembles its Gaussian counterpart [27]. Instead of *scatter* matrices, the Gaussian model has *covariance* matrices. The mixture covariance matrix is simply a linear combination of the source covariance matrices. In this case, given the model parameters, the multichannel Wiener filter can be used to extract the sources. Unfortunately, this linear combination between scatter matrices is usually not satisfied for the Cauchy model with $K > 1$.

We therefore propose to project the observation vectors $\mathbf{x}_{ft} \in \mathbb{C}^K$ to the complex plane $\mathbb{C}$. Let us consider M vectors $\mathbf{u}_1, \cdots, \mathbf{u}_M \in \mathbb{C}^K$ and let

$$x_{mft} = \mathbf{u}_m^{\mathrm{H}} \mathbf{x}_{ft} \text{ for } \forall m, f, t \quad (6)$$

be the m$^{\text{th}}$-projection of the observed signal $\mathbf{x}_{ft}$. As demonstrated in [28], the random variable x_{mft} is Cauchy distributed and the following posterior mean of the projected speech $\hat{x}^{\text{s}}_{mft}$ is tractable for all m, f, t:

$$\hat{x}^{\text{s}}_{mft} \triangleq \mathbb{E}\left[\mathbf{u}_m^{\text{H}} x^{\text{s}}_{ft} | x_{mft}, \mathbf{\Psi}\right] = \sqrt{\frac{v^{\text{s}}_{mft}}{v_{mft}}} x_{mft}, \tag{7}$$

where

$$\begin{cases} v^{\text{s}}_{mft} = a^{\text{s}}_{ft}\mathbf{u}_m^{\text{H}}\mathbf{R}^{\text{s}}_f\mathbf{u}_m, \\ v^{\text{n}}_{mft} = a^{\text{n}}_{ft}\mathbf{u}_m^{\text{H}}\mathbf{R}^{\text{n}}_f\mathbf{u}_m, \\ v_{mft} = \left(\sqrt{v^{\text{s}}_{mft}} + \sqrt{v^{\text{n}}_{mft}}\right)^2, \end{cases} \tag{8}$$

and $\mathbf{\Psi} \triangleq \left\{a^{\text{s}}_{ft}, a^{\text{n}}_{ft}, \mathbf{R}^{\text{s}}_f, \mathbf{R}^{\text{n}}_f\right\}$.

We then deduce an estimator $\hat{\mathbf{x}}^{\text{s}}_{ft}$ of $\mathbf{x}^{\text{s}}_{ft}$ by using $\mathbf{U}^\dagger$, which is the pseudo-inverse of $\mathbf{U} \triangleq [\mathbf{u}_1, \ldots, \mathbf{u}_M]^{\text{H}} \in \mathbb{C}^{M\times K}$:

$$\hat{\mathbf{x}}^{\text{s}}_{ft} = \mathbf{U}^\dagger \left[\hat{x}^{\text{s}}_{1ft}, \cdots, \hat{x}^{\text{s}}_{mft}\right]^{\text{T}}. \tag{9}$$

An estimator $\hat{\boldsymbol{x}}^{\text{n}}_{ft}$ of $\boldsymbol{x}^{\text{n}}_{ft}$ can be computed in a similar way. In this paper, in order to simplify the computation of (9), the projector $\mathbf{U}$ is chosen to be unitary so that $\mathbf{U}^\dagger = \mathbf{U}$.

III. Parameter estimation

This section explains how to estimate the parameters of the probabilistic model proposed in Section II.

A. Training Phase

To train the speech decoder model $\mu_\theta(\mathbf{z})$, we adopt the VAE framework [11]. For training, two DNNs are considered:

- *An encoder* that inputs $\mathbf{a}^{\text{s}}_t$ and outputs two D-dimensional vectors written $\boldsymbol{\mu}^q_\phi(\mathbf{a}^{\text{s}}_t)$ and $\boldsymbol{\sigma}^q_\phi(\mathbf{a}^{\text{s}}_t)$. Together, they define the distribution of the latent vectors: $q_\phi(\mathbf{z}_t|\mathbf{a}_t)$, defined as $\mathcal{N}(\mathbf{z}_t|\boldsymbol{\mu}^q_\phi(\mathbf{a}_t), \text{Diag}[\boldsymbol{\sigma}^q_\phi(\mathbf{a}_t)])$
- *A decoder* that outputs two F-dimensional vectors written $\mu_\theta(\mathbf{z}_t)$ and $\gamma_\theta(\mathbf{z}_t)$, that together describe the distribution of the speech magnitudes $\mathbf{a}^{\text{s}}_t$ given $\mathbf{z}$, written $p_\theta(\mathbf{a}^{\text{s}}_t|\mathbf{z}_t)$. We detail that distribution later.

In any case, the model parameters θ and ϕ are jointly optimized by minimizing the negative log-likelihood (NLL):

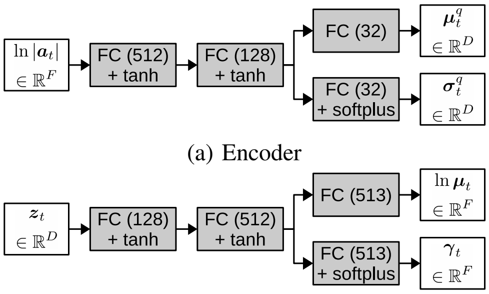
$$\begin{aligned} -\ln p_\theta(\mathbf{a}^{\text{s}}_t) &= -\ln \int_{\mathbf{z}_t} \frac{q_\phi(\mathbf{z}_t|\mathbf{a}^{\text{s}}_t)}{q_\phi(\mathbf{z}_t|\mathbf{a}^{\text{s}}_t)} p_\theta(\mathbf{a}^{\text{s}}_t, \mathbf{z}_t)\mathrm{d}\mathbf{z}_t \\ &\leq -\mathbb{E}_{q_\phi(\mathbf{z}_t|\mathbf{a}^{\text{s}}_t)}\left[\ln \frac{p_\theta(\mathbf{a}^{\text{s}}_t|\mathbf{z}_t)p_\theta(\mathbf{z}_t)}{q_\phi(\mathbf{z}_t|\mathbf{a}^{\text{s}}_t)}\right] \\ &= -\mathbb{E}_{q_\phi(\mathbf{z}_t|\mathbf{a}^{\text{s}}_t)}[\ln p_\theta(\mathbf{a}^{\text{s}}_t|\mathbf{z}_t)] + \text{KL}[q_\phi(\mathbf{z}_t|\mathbf{a}^{\text{s}}_t)\|p_\theta(\mathbf{z}_t)] \\ &\overset{\text{def}}{=} \mathcal{L}^{\text{mag}} + \mathcal{L}^{\text{reg}}, \end{aligned} \tag{10}$$

where $\text{KL}[q\|p]$ is the Kullback-Leibler (KL) divergence from p to q [29]. The reparameterization trick [11] is used to obtain $\mathbf{z}_t$ given the encoder outputs $\boldsymbol{\mu}^q_\phi(\mathbf{a}^{\text{s}}_t)$ and $\boldsymbol{\sigma}^q_\phi(\mathbf{a}^{\text{s}}_t)$.

For training, the observed magnitudes from a clean speech dataset, a^{s}_{ft}, are assumed to follow a real Cauchy distribution $\mathcal{C}_{\mathbb{R}}$ with location $[\mu_\theta(\mathbf{z}_t)]_f \in \mathbb{R}_+$ and scale $[\gamma_\theta(\mathbf{z}_t)]_f \in \mathbb{R}_+$:

$$p_\theta(a^{\text{s}}_{ft}|\mathbf{z}_t) = \mathcal{C}_{\mathbb{R}}\left(a^{\text{s}}_{ft}\middle| [\mu_\theta(\mathbf{z}_t)]_f, [\gamma_\theta(\mathbf{z}_t)]_f\right). \tag{11}$$

This complies with the $\alpha-$spectrogram assumption for $\alpha = 1$ [30]. Thus, the magnitude reconstruction loss $\mathcal{L}^{\text{mag}}$, serving as a cost function for training the parameters θ, may be picked as the negative log-likelihood (NLL):



(a) Encoder

(b) Decoder

Fig. 2: The encoder and decoder of the speech VAE. 'FC' represents a fully-connected layer whose size is shown inside the parentheses. The speech magnitude estimate is computed from the decoder output $\hat{\mathbf{a}}^{\text{s}}_t = \exp(\ln \boldsymbol{\mu}_t)$.

$$\mathcal{L}^{\text{mag}} \overset{c}{=} \frac{1}{T}\sum_{f,t=1}^{F,T}\left(\ln[\gamma_\theta(\mathbf{z}_t)]_f + \ln\left(1 + \frac{\left(a^{\text{s}}_{ft} - [\mu_\theta(\mathbf{z}_t)]_f\right)^2}{\gamma^2_{ft}}\right)\right). \tag{12}$$

Then, assuming a simple prior $p_\theta(\mathbf{z}_t) \sim \mathcal{N}(\mathbf{z}_t|\mathbf{0}, \mathbf{I})$, the regularization term $\mathcal{L}^{\text{reg}}$ [11] is computed as

$$\mathcal{L}^{\text{reg}} = \frac{1}{2T}\sum_{d,t=1}^{D,T}\left([\boldsymbol{\mu}^q_\phi(\mathbf{a}^{\text{s}}_t)]^2_d + [\boldsymbol{\sigma}^q_\phi(\mathbf{a}^{\text{s}}_t)]^2_d - \ln[\boldsymbol{\sigma}^q_\phi(\mathbf{a}^{\text{s}}_t)]^2_d - 1\right). \tag{13}$$

B. Test Phase

As advocated above, the projected mixtures x_{mft} are considered as isotropic complex random variables. They are thus parameterized through a scale parameter $\sqrt{v_{mft}}$ and the negative log-likelihood is given by

$$D(v) \overset{c}{=} \sum_{m,f,t=1}^{M,F,T} \frac{3}{2}\ln\left(v_{mft} + |x_{mft}|^2\right) - \frac{1}{2}\ln(v_{mft}), \tag{14}$$

where $\overset{c}{=}$ denotes equality up to a constant.

At test time, the latent variables $\mathbf{z}_t$ are initialized by sampling from $q_\phi(\mathbf{z}_t|\,|\mathbf{x}_t|)$, i.e., by applying the encoder to the average of the mixture magnitude spectrogram over channels. This in turn provides an initial estimate $\mu_\theta(\mathbf{z}_t)$ for the speech magnitude $\mathbf{a}^{\text{s}}_t$. Then, the latent variables $\mathbf{z}_t$ are updated by backpropagation with a gradient descent method to minimize the cost function (14). In this case, all parameters other than $\mathbf{z}_t$, including the decoder parameters, are kept fixed.

For estimating the noise parameters, we adopt a majorization-equalization (ME) strategy [26] as in [21]. Due to space constraints, we only provide here the multiplicative updates used for the parameters w and h as follows:

$$w_{fl} \leftarrow \frac{1}{3} w_{fl}\frac{\sum_{mt} h_{lt}\psi^{\text{n}}_{mf}}{\sum_{mt} h_{lt}\psi^{\text{n}}_{mf}\xi_{mft}}, \tag{15}$$

$$h_{lt} \leftarrow \frac{1}{3} h_{lt}\frac{\sum_{mf} h_{fl}\psi^{\text{n}}_{mf}}{\sum_{mf} w_{fl}\psi^{\text{n}}_{mf}\xi_{mft}}, \tag{16}$$

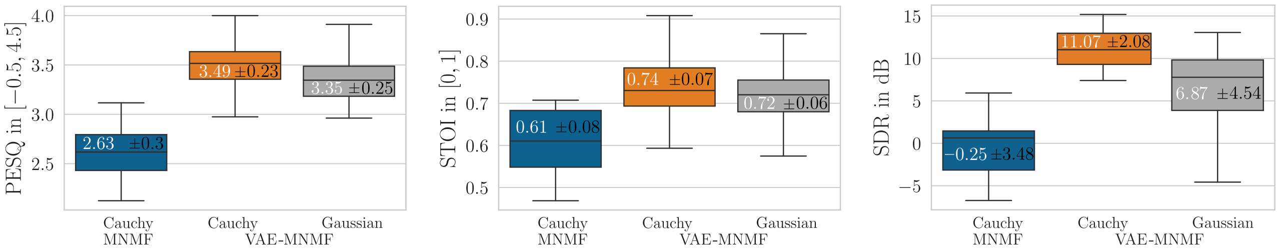


Fig. 3: Performance comparison in terms of PESQ (*left*), STOI (*middle*), and SDR (*right*). Higher is better. The mean and the standard deviation are respectively shown in white and black fonts.

$$\psi^{\rm n}_{mf} \triangleq \frac{\mathbf{u}_m^{\rm H}\mathbf{R}_f^{\rm n}\mathbf{u}_m}{\sqrt{v^{\rm n}_{mft}v_{mft}}}, \tag{17}$$

$$\xi_{mft} \triangleq 1 + \frac{|x_{mft}|^2}{v_{mft}}. \tag{18}$$

Similarly for noise parameter estimation, we define the estimator $\hat{v}^j_{ft} = a^j_{ft}\hat{\mathbf{R}}^j_f$ for $j \in \{\rm s, n\}$, where a^j_{ft} is provided by the source models and $\hat{\mathbf{R}}^j_f = \sum_{m'} r^j_{m',f}\mathbf{u}_{m'}\mathbf{u}^{\rm H}_{m'}$. It leads to

$$r^j_{m'f} \leftarrow \frac{1}{3} r^j_{m'f} \frac{\sum_{mt} a^j_{ft}\eta^j_{mm'ft}}{\sum_{mt}\eta^j_{mm'ft}\xi_{mft}}, \tag{19}$$

$$\eta^j_{mm'ft} \triangleq \frac{|\mathbf{u}^{\rm H}_{m'}\mathbf{u}_m|^2}{\sqrt{v^j_{mft}v_{mft}}}. \tag{20}$$

IV. Evaluation

In this section, we compare the performance of three different systems on a 5-channel speech enhancement task. Each of them includes at least one multichannel nonnegative matrix factorization (MNMF) spectrogram model [31]. The systems include: (1) *Cauchy VAE-MNMF*, that we propose above; (2) *Gaussian VAE-MNMF*, that is a system similar to ours, but based on a Gaussian model [9]; and (3) *Cauchy MNMF*, that is a semi-supervised multichannel Cauchy NMF, where the speech magnitude is also modeled with an NMF with basis vectors trained on clean speech beforehand. The Gaussian VAE-MNMF system is provided by the authors [9].

The performance is measured by the signal-to-distortion ratio (SDR) provided by the BSS-Eval toolbox [32], the Perceptual Evaluation of Speech Quality (PESQ) score [33], and the Short-Time Objective Intelligibility (STOI) score [34]. The SDR is computed on the enhanced 5-channel speech, while the PESQ and the STOI are computed on one of the channels.

A. Experimental Conditions

We consider the simulated training, development, and test sets of the CHiME-4 corpus [5]. All data are sampled at 16 kHz. We use 7138 single-channel clean speech signals of the training set for training the DNNs of the Cauchy VAE-MNMF and the speech basis vectors of the Cauchy MNMF. Moreover, we use 1640 single-channel clean speech signals from the development set as the validation set for the DNN training. The evaluation is then done on 10% of the full test set, i.e., 132 randomly selected 5-channel noisy utterances ($K = 5$). The STFT coefficients are extracted using a Hann window with a length of 1024 samples and an overlap of 75% ($F = 513$).

The encoder and the decoder of the speech VAE for the Cauchy VAE-MNMF is depicted in Fig. 2. These DNNs are trained by backpropagation with the Adam update rule whose learning rate is fixed to 10^{-3} [35]. The update is done for every minibatch of 8192 frames from 32 randomly selected utterances. The gradient is normalized with threshold fixed to 1 [36]. The weight normalization [37] is also employed. The training is started with a warm-up technique [38] for 100 epochs and stopped after 50 consecutive epochs that failed to obtain a better validation score. The latest model yielding the lowest error is kept. These DNNs are comparable in size to the ones used in the Gaussian VAE-MNMF.

For both VAE-MNMF methods, the latent variable dimension of the speech model is fixed to $D = 32$ and the number of bases of the noise model is fixed to $L = 32$. Similarly, for the Cauchy MNMF, the number of bases of both source models is fixed to $L = 32$. For the Cauchy MNMF and the Cauchy VAE-MNMF, the dimension of the projector $\mathbf{U}$ is fixed to $M = 8$. The NMF parameters are initialized randomly and the coefficients r^j_{mf} are initialized as 1. We consider 64 optimization iterations for the Cauchy MNMF and 50 for both VAE-MNMF methods.

B. Experimental Results

Fig. 3 shows that the Cauchy VAE-MNMF globally outperforms the Cauchy MNMF and the Gaussian VAE-MNMF. It provides an SDR improvement of 4.2 dB w.r.t. the Gaussian VAE-MNMF. We also observe that the standard deviation of the metrics is generally smaller for the Cauchy VAE-MNMF, suggesting that it has stronger robustness to noise.

As an illustration, we also displayed in Fig. 4 the log-magnitude spectrograms of estimated speech obtained with the Cauchy and the Gaussian VAE-MNMFs. We see that the one seems less robust against non-stationary noise.

V. Conclusion

We proposed a speech enhancement system combining a nonnegative matrix factorization (NMF) model for the noise and a variational autoencoder (VAE) for speech, that is trained and used under heavy-tailed probabilistic models. We derived the whole training and inference strategy and gave the details of the corresponding algorithms.

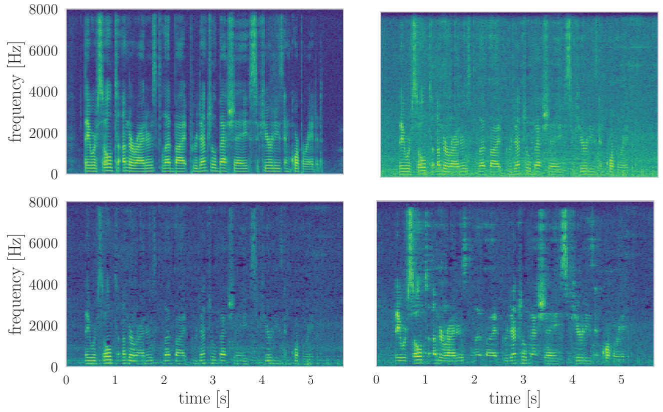


Fig. 4: Log-magnitude spectrograms of clean speech *(top-left)*, corrupted speech *(top-right)*, speech estimated with the Gaussian VAE-MNMF *(bottom-left)* and speech estimated with the Cauchy VAE-MNMF *(bottom-right)*. The utterance is `M05_447C020F_PED` from the test set `et05_ped_simu`.

In an evaluation on 5-channel mixtures from the CHiME-4 corpus, we found out that our proposed system achieves a significantly better performance than its Gaussian counterpart, yielding a 4 dB SDR improvement.

Acknowledgement

We thank S. Leglaive for providing the code of [9].

References

[1] E. Vincent, T. Virtanen, and S. Gannot, Eds., *Audio Source Separation and Speech Enhancement*. Wiley, 2018.

[2] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Trans. ASLP*, vol. 24, no. 9, pp. 1652–1664, 2016.

[3] J. Heymann, L. Drude, and R. Haeb-Umbach, "A generic neural acoustic beamforming architecture for robust multi-channel speech processing," *Computer Speech & Language*, vol. 46, pp. 374–385, 2017.

[4] T. N. Sainath, R. J. Weiss, K. W. Wilson, B. Li, A. Narayanan, E. Variani, M. Bacchiani, I. Shafran, A. Senior, K. Chin, A. Misra, and C. Kim, "Multichannel signal processing with deep neural networks for automatic speech recognition," *IEEE/ACM Trans. ASLP*, vol. 25, no. 5, pp. 965–979, May 2017.

[5] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech & Language*, vol. 46, pp. 535–557, 2017.

[6] Y. Bando, M. Mimura, K. Itoyama, K. Yoshii, and T. Kawahara, "Statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization," in *Proc. IEEE ICASSP*, 2018, pp. 716–720.

[7] S. Leglaive, L. Girin, and R. Horaud, "A variance modeling framework based on variational autoencoders for speech enhancement," in *Proc. IEEE MLSP*, 2018.

[8] K. Sekiguchi, Y. Bando, K. Yoshii, and T. Kawahara, "Bayesian multichannel speech enhancement with a deep speech prior," in *Proc. APSIPA*, 2018, pp. 1233–1239.

[9] S. Leglaive, L. Girin, and R. Horaud, "Semi-supervised multichannel speech enhancement with variational autoencoders and non-negative matrix factorization," in *Proc. IEEE ICASSP*, 2019.

[10] S. Leglaive, U. Şimşekli, A. Liutkus, L. Girin, and R. Horaud, "Speech enhancement with variational autoencoders and alpha-stable distributions," in *Proc. IEEE ICASSP*, 2019.

[11] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. ICLR*, 2014.

[12] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[13] C. Boutsidis and E. Gallopoulos, "SVD based initialization: A head start for nonnegative matrix factorization," *Pattern Recognition*, vol. 41, no. 4, pp. 1350–1362, 2008.

[14] R. A. Fisher, "Applications of "Student's" distribution," *Metron*, vol. 5, no. 3, pp. 90–104, 1925.

[15] R. Martin and C. Breithaupt, "Speech enhancement in the DFT domain using Laplacian speech priors," in *Proc. IWAENC*, vol. 3, 2003, pp. 87–90.

[16] G. Samoradnitsky and M. S. Taqqu, *Stable non-Gaussian random processes: stochastic models with infinite variance*. Chapman and Hall/CRC, 1994.

[17] S. Leglaive, U. Şimşekli, A. Liutkus, R. Badeau, and G. Richard, "Alpha-stable multichannel audio source separation," in *Proc. IEEE ICASSP*, 2017, pp. 576–580.

[18] M. Fontaine, F.-R. Stöter, A. Liutkus, U. Şimşekli, R. Serizel, and R. Badeau, "Multichannel audio modeling with elliptically stable tensor decomposition," in *Proc. LVA/ICA*, 2018, pp. 13–23.

[19] U. Şimşekli, H. Erdogan, S. Leglaive, A. Liutkus, R. Badeau, and G. Richard, "Alpha-stable low-rank plus residual decomposition for speech enhancement," in *Proc. IEEE ICASSP*, 2018.

[20] M. Fontaine, A. Liutkus, L. Girin, and R. Badeau, "Explaining the parameterized Wiener filter with alpha-stable processes," in *Proc. WASPAA*, 2017.

[21] A. Liutkus, D. Fitzgerald, and R. Badeau, "Cauchy nonnegative matrix factorization," in *Proc. WASPAA*, 2015, pp. 1–5.

[22] K. Kitamura, Y. Bando, K. Itoyama, and K. Yoshii, "Student's t multichannel nonnegative matrix factorization for blind source separation," in *Proc. IWAENC*, 2016, pp. 1–5.

[23] D. Kitamura, S. Mogami, Y. Mitsui, N. Takamune, H. Saruwatari, N. Ono, Y. Takahashi, and K. Kondo, "Generalized independent low-rank matrix analysis using heavy-tailed distributions for blind source separation," *EURASIP J. on Adv. in Signal Process.*, vol. 2018, no. 1, pp. 1–25, 2018.

[24] M. Lombardi and D. Veredas, "Indirect estimation of elliptical stable distributions," *Computational Statistics & Data Analysis*, vol. 53, no. 6, pp. 2309–2324, 2009.

[25] S. J. Press, "Multivariate stable distributions," *Journal of Multivariate Analysis*, vol. 2, no. 4, pp. 444–462, 1972.

[26] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the β-divergence," *Neural Computation*, vol. 23, no. 9, pp. 2421–2456, 2011.

[27] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. ASLP*, vol. 18, no. 7, pp. 1830–1840, 2010.

[28] D. Fitzgerald, A. Liutkus, and R. Badeau, "Projection-based demixing of spatial audio," *IEEE/ACM Trans. ASLP*, vol. 24, no. 9, pp. 1556–1568, 2016.

[29] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, 1951.

[30] A. Liutkus and R. Badeau, "Generalized Wiener filtering with fractional power spectrograms," in *Proc. IEEE ICASSP*, 2015, pp. 266–270.

[31] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. ASLP*, vol. 18, no. 3, pp. 550–563, 2009.

[32] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. P. Rosca, "First stereo audio source separation evaluation campaign: Data, algorithms and results," in *Proc. ICA*, 2007, pp. 552–559.

[33] ITU-T, "P.862 : Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," 2001.

[34] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. ASLP*, vol. 19, no. 7, pp. 2125–2136, 2011.

[35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.

[36] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *Proc. ICML*, 2013, pp. 1310–1318.

[37] T. Salimans and D. P. Kingma, "Weight normalization: A simple reparameterization to accelerate training of deep neural networks," in *Proc. NIPS*, 2016, pp. 901–909.

[38] C. K. Sønderby, T. Raiko, L. Maaløe, S. K. Sønderby, and O. Winther, "Ladder variational autoencoders," in *Proc. NIPS*, 2016, pp. 3738–3746.