Robust Unsupervised Audio-visual Speech Enhancement Using a Mixture of Variational Autoencoders

Mostafa Sadeghi, Xavier Alameda-Pineda

ROBUST UNSUPERVISED AUDIO-VISUAL SPEECH ENHANCEMENT USING A MIXTURE OF VARIATIONAL AUTOENCODERS

Mostafa Sadeghi and Xavier Alameda-Pineda, IEEE Senior Member

Inria Grenoble Rhône-Alpes, France

ABSTRACT

Recently, an audio-visual speech generative model based on variational autoencoder (VAE) has been proposed, which is combined with a nonnegative matrix factorization (NMF) model for noise variance to perform unsupervised speech enhancement. When visual data is clean, speech enhancement with audio-visual VAE shows a better performance than with audio-only VAE, which is trained on audio-only data. However, audio-visual VAE is not robust against noisy visual data, e.g., when for some video frames, speaker face is not frontal or lips region is occluded. In this paper, we propose a robust unsupervised audio-visual speech enhancement method based on a per-frame VAE mixture model. This mixture model consists of a trained audio-only VAE and a trained audio-visual VAE. The motivation is to skip noisy visual frames by switching to the audio-only VAE model. We present a variational expectation-maximization method to estimate the parameters of the model. Experiments show the promising performance of the proposed method.

***Index Terms*—** Robust audio-visual speech enhancement, generative models, variational auto-encoder, mixture mode, variational expectation-maximization

1. INTRODUCTION

Speech enhancement – or how to estimate clean speech from a noisy signal – has attracted a lot of attention, both for single- and multi-channel audio recordings [1–4]. Recently, generative models have been utilized for speech enhancement [5–10]. Specifically, some works proposed to use variational autoencoder (VAE) to model speech spectrogram, and then perform speech enhancement by considering an NMF noise variance model [5,6]. This is done in an unsupervised way. A common characteristic of all these methods, which we refer to as audio VAE (A-VAE), is the use of audio recordings only.

Audio-visual speech enhancement methods incorporate also the visual information (video frames) associated with the noisy speech, aiming to improve the quality of the enhanced speech signal [11–13]. Using the video modality is well-motivated by the fact that lips movements provide information about what is being uttered. As an audio-visual extension of VAE-based methods of [5, 6], an audio-visual VAE (AV-VAE) model has recently been proposed in [14], training a VAE model conditioned on visual features, e.g., lips region of interest (ROI). For speech enhancement, AV-VAE has been shown to outperform A-VAE specially in high noise levels [14].

A critical problem with audio-visual methods is "noisy visual data", e.g., when speakers do not face the camera or the lips are occluded. Specifically, the AV-VAE based speech enhancement method presented in [14] uses clean visual data to train the speech spectrogram prior. Therefore, it expects clean visual data as well in the test (enhancement) phase for unseen data. Otherwise, its performance could degrade below audio-only methods, as we will see later in this paper.

The present work aims to provide a solution to the above-mentioned problem. That is, to make AV-VAE speech enhancement robust against noisy visual data. To achieve this goal, we propose a VAE mixture model consisting of a trained A-VAE and a trained AV-VAE model. As said before, AV-VAE yields poor results in the presence of noisy visual data. However, the proposed mechanism would skip visual data whenever it is not reliable, and uses the A-VAE model instead. Importantly, the choice between A-VAE and AV-VAE is unsupervised and must be done for every frame at test time. We present a variational inference framework to tackle all these issues. Experimental results on clean as well as noisy visual frames are provided, demonstrating the effectiveness of our method.

The rest of this paper is organized as follows. Section 2 briefly reviews the audio-only and audio-visual VAE-based speech spectrogram modeling methods. Section 3 presents the proposed mixture generative model. Sections 4 details the inference and speech reconstruction steps. Finally, experimental results are provided in Section 5.

2. VAE-BASED SPEECH SPECTRA MODELING

2.1. Audio-only VAE

In this section, we briefly review the VAE generative speech model that was first proposed in [5]. Let s_{fn} denote the

Xavier Alameda-Pineda acknowledges ANR and the IDEX for funding the ML3RI project. This work has been partially supported by MIAI @ Grenoble Alpes, (ANR-19-P3IA-0003)

complex-valued speech short-time Fourier transform (STFT) coefficient at frequency index $f \in \{0, ..., F-1\}$ and at frame index $n \in \{0, ..., N-1\}$. At each time frequency (TF) bin, we have the following probabilistic generative model, which will be referred to as A-VAE:

$$s_{fn}|\mathbf{z}_n \sim \mathcal{N}_c(0, \sigma_f^a(\mathbf{z}_n)), \quad (1)$$

$$\mathbf{z}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad (2)$$

where $\mathbf{z}_n \in \mathbb{R}^L$, with $L \ll F$, is a latent random variable describing a speech generative process, $\mathcal{N}(\mathbf{0}, \mathbf{I})$ is a zero-mean multivariate Gaussian distribution with identity covariance matrix, and $\mathcal{N}_c(0, \sigma)$ is a univariate complex proper Gaussian distribution with zero mean and variance σ. Let $\mathbf{s}_n \in \mathbb{C}^F$ be the vector whose components are the speech STFT coefficients at frame n. The set of non-linear functions $\{\sigma_f^a : \mathbb{R}^L \mapsto \mathbb{R}_+\}_{f=0}^{F-1}$ are modeled as neural networks sharing the input $\mathbf{z}_n \in \mathbb{R}^L$. These parameters are estimated using variational inference by defining another neural network, called encoder (inference) network, which approximates the intractable posterior of $\mathbf{z}_n$ given $\mathbf{s}_n$ [5, 6, 15].

2.2. Audio-visual VAE

In the AV-VAE framework proposed in [14], the following latent space model is considered, independently for all $l \in \{0, ..., L-1\}$ and all TF bins (f, n):

$$s_{fn}|\mathbf{z}_n, \mathbf{v}_n \sim \mathcal{N}_c(0, \sigma_f^{av}(\mathbf{z}_n, \mathbf{v}_n)), \quad (3)$$

$$z_{ln}|\mathbf{v}_n \sim \mathcal{N}\left(\bar{\mu}_l(\mathbf{v}_n), \bar{\sigma}_l(\mathbf{v}_n)\right), \quad (4)$$

where $\mathbf{v}_n \in \mathbb{R}^M$ is an embedding for the image of the speaker lips at frame n, and the non-linear functions $\{\sigma_f^{av} : \mathbb{R}^L \times \mathbb{R}^M \mapsto \mathbb{R}_+\}_{f=0}^{F-1}$ are modeled as a neural network taking $\mathbf{z}_n$ and $\mathbf{v}_n$ as input. Furthermore, the non-linear functions $\{\bar{\mu}_l : \mathbb{R}^M \mapsto \mathbb{R}\}_{l=0}^{L-1}$ and $\{\bar{\sigma}_l : \mathbb{R}^M \mapsto \mathbb{R}_+\}_{l=0}^{L-1}$, yielding $\mathbf{z}_n$'s prior, are collectively modeled with a neural network which takes $\mathbf{v}_n$ as input. In a way similar to A-VAE, an encoder network, approximating the intractable posterior of $\mathbf{z}_n$ given $\mathbf{s}_n$ and $\mathbf{v}_n$, is defined and trained jointly with the decoder (3) and prior (4) using a clean set of speech spectrogram frames and the corresponding clean visual data, i.e., images of lips region [14].

Now that both A-VAE and AV-VAE are trained for their specific tasks, in the following, we present the VAE mixture model to automatically select one of the two at inference time.

3. VAE MIXTURE MODEL

To make the speech enhancement robust to noisy visual data, we propose an automatic selection mechanism between A-VAE and AV-VAE. Ideally, such a mechanism would allow to select the best-suited method at each frame: when visual information is clean, AV-VAE, and otherwise, A-VAE. We formalise this with a mixture model, named VAE mixture model (VAE-MM):

$$p(s_{fn}|\mathbf{z}_n, \mathbf{v}_n, \alpha_n) = \left[\mathcal{N}_c(0, \sigma_f^a(\mathbf{z}_n))\right]^{\alpha_n} \times \left[\mathcal{N}_c\left(0, \sigma_f^{av}(\mathbf{z}_n, \mathbf{v}_n)\right)\right]^{1-\alpha_n}, \quad (5)$$

$$p(\mathbf{z}_n|\mathbf{v}_n, \alpha_n) = \left[\mathcal{N}(\mathbf{0}, \mathbf{I})\right]^{\alpha_n} \times \left[\prod_l \mathcal{N}\left(\bar{\mu}_l(\mathbf{v}_n), \bar{\sigma}_l(\mathbf{v}_n)\right)\right]^{1-\alpha_n} \quad (6)$$

$$p(\alpha_n) = \pi^{\alpha_n} \times (1-\pi)^{1-\alpha_n}, \quad (7)$$

where, $\alpha_n \in \{0, 1\}$ is a latent variable specifying the component of the mixture model that is used by the n-th frame, for both s_{fn} and $\mathbf{z}_n$. The prior distribution of α_n is chosen as a Bernoulli distribution with parameter π.

4. VAE-MM INFERENCE & LEARNING

The observed noisy microphone signal writes:

$$x_{fn} = s_{fn} + b_{fn}, \quad (8)$$

for all TF bins (f, n). Similarly as done in the previous works [5–10], we use an unsupervised NMF-based Gaussian noise model that assumes independence across TF bins:

$$b_{fn} \sim \mathcal{N}_c\left(0, (\mathbf{W}_b\mathbf{H}_b)_{fn}\right), \quad (9)$$

where $\mathbf{W}_b \in \mathbb{R}^{F \times K}$ is a nonnegative matrix of spectral power patterns and $\mathbf{H}_b \in \mathbb{R}^{K \times N}$ is a nonnegative matrix of temporal activations, with K such that $K(F+N) \ll FN$.

The set of parameters to be estimated is defined as $\Theta = \{\mathbf{W}_b, \mathbf{H}_b, \pi\}$. We use a variational expectation-maximization (VEM) approach [16] to estimate these parameters. To do so, the intractable posterior $p(\mathbf{s}_n, \mathbf{z}_n, \alpha_n|\mathbf{x}_n)$ is approximated by a variational distribution factorizing as follows:

$$r(\mathbf{s}_n, \mathbf{z}_n, \alpha_n) = r(\mathbf{s}_n)\, r(\mathbf{z}_n)\, r(\alpha_n). \quad (10)$$

The variational factors in (10) are then estimated by minimizing the Kullback-Leibler divergence between (10) and the true posterior. The final update formulas for the variational distributions are given below and are used in an alternating optimization strategy. The details of the derivations are provided in a supporting document which is available online.[1]

4.1. E-$\mathbf{s}_n$ step

The variational distribution of $\mathbf{s}_n$ factorizes over f. For each component, we obtain the following:

$$r(s_{fn}) = \mathcal{N}_c(m_{fn}, \nu_{fn}), \quad (11)$$

[1] https://team.inria.fr/perception/research/vae-mm-se/

where

$$\begin{cases} m_{fn} &= \frac{\gamma_{fn}}{\gamma_{fn}+(\mathbf{W}_b\mathbf{H}_b)_{fn}} \cdot x_{fn} \\ \nu_{fn} &= \frac{\gamma_{fn}\cdot(\mathbf{W}_b\mathbf{H}_b)_{fn}}{\gamma_{fn}+(\mathbf{W}_b\mathbf{H}_b)_{fn}} \end{cases}, \tag{12}$$

$$\gamma_{fn}^{-1} = \sum_{\alpha_n\in\{0,1\}} r(\alpha_n)\cdot\eta_{fn}^{\alpha_n}, \tag{13}$$

$$\eta_{fn}^{\alpha_n} = \mathbb{E}_{r(\mathbf{z}_n)}\left[\frac{1}{\sigma_f^{\alpha_n}(\mathbf{z}_n,\mathbf{v}_n)}\right] \approx \frac{1}{D}\sum_{d=1}^{D}\frac{1}{\sigma_f^{\alpha_n}(\mathbf{z}_n^{(d)},\mathbf{v}_n)}, \tag{14}$$

and $\left\{\mathbf{z}_n^{(d)}\right\}_{d=1}^{D}$ is a sequence sampled from $r(\mathbf{z}_n)$. Moreover, we have defined $\sigma_f^{\alpha_n}(\mathbf{z}_n,\mathbf{v}_n)$ as follows:

$$\sigma_f^{\alpha_n}(\mathbf{z}_n,\mathbf{v}_n) = \begin{cases} \sigma_f^{a}(\mathbf{z}_n) & \alpha_n = 1 \\ \sigma_f^{av}(\mathbf{z}_n,\mathbf{v}_n) & \alpha_n = 0 \end{cases}. \tag{15}$$

4.2. E-$\mathbf{z}_n$ step

For $r(\mathbf{z}_n)$ we obtain the following result:

$$r(\mathbf{z}_n) \propto \exp\Big(\sum_{\alpha_n\in\{0,1\}} r(\alpha_n)\cdot\Big[\log p(\mathbf{z}_n|\mathbf{v}_n,\alpha_n)+ \sum_f -\log\left(\sigma_f^{\alpha_n}(\mathbf{z}_n,\mathbf{v}_n)\right) - \frac{|m_{fn}|^2+\nu_{fn}}{\sigma_f^{\alpha_n}(\mathbf{z}_n,\mathbf{v}_n)}\Big]\Big). \tag{16}$$

The above distribution cannot be computed in closed-from. Nevertheless, we can draw samples from it using the Metropolis-Hastings (MH) algorithm [16].

Let $\tilde{r}(\mathbf{z}_n)$ denote the right-hand side of (16). At the iteration m of the MH algorithm, given a current sample $\mathbf{z}_n^{(m-1)}$, to obtain the next one, i.e., $\mathbf{z}_n^{(m)}$, we first draw a candidate sample from a proposal Gaussian distribution centered around $\mathbf{z}_n^{(m-1)}$ and with $\epsilon\mathbf{I}$ as the covariance matrix. The candidate sample, denoted $\mathbf{z}_n^{(T)}$, is accepted by the probability $p = \min\left(1, \frac{\tilde{r}(\mathbf{z}_n^{(T)})}{\tilde{r}(\mathbf{z}_n^{(m-1)})}\right)$. If the sample is accepted, we set $\mathbf{z}_n^{(m)} = \mathbf{z}_n^{(T)}$. Otherwise, $\mathbf{z}_n^{(m)} = \mathbf{z}_n^{(m-1)}$. The above procedure is repeated until the required number of samples is achieved. Furthermore, the initial samples, obtained during the so-called burn-in period, are discarded.

4.3. E-α_n step

To update the variational distribution of α_n, we can write:

$$r(\alpha_n) \propto \exp\Big(\mathbb{E}_{r(\mathbf{s}_n)\cdot r(\mathbf{z}_n)}\Big[\log p(\mathbf{s}_n|\mathbf{z}_n,\mathbf{v}_n,\alpha_n)+ \log p(\mathbf{z}_n|\mathbf{v}_n,\alpha_n)+\log p(\alpha_n)\Big]\Big) \tag{17}$$

which is a Bernoulli distribution with

$$\pi_n = g\Big(\mathbb{E}_{r(\mathbf{s}_n)\cdot r(\mathbf{z}_n)}\Big[\log\frac{p(\mathbf{s}_n,\mathbf{z}_n|\mathbf{v}_n,\alpha_n=1)}{p(\mathbf{s}_n,\mathbf{z}_n|\mathbf{v}_n,\alpha_n=0)}\Big]+\log\frac{\pi}{1-\pi}\Big) \tag{18}$$

as the parameter, where $g(.)$ denotes the sigmoid function defined as $g(x) = 1/(1+\exp(-x))$. The above expression is further simplified to the following:

$$\pi_n \approx g\Big(\frac{1}{D}\sum_{d=1}^{D}\sum_f \log\frac{\sigma_f^{av}(\mathbf{z}_n^{(d)},\mathbf{v}_n)}{\sigma_f^{a}(\mathbf{z}_n^{(d)})}+ \Big(1/\sigma_f^{av}(\mathbf{z}_n^{(d)},\mathbf{v}_n) - 1/\sigma_f^{a}(\mathbf{z}_n^{(d)})\Big)\cdot\Big(|m_{fn}|^2+\nu_{fn}\Big)+ \sum_l \log\bar{\sigma}_l(\mathbf{v}_n) + \frac{(z_{ln}^{(d)}-\bar{\mu}_l(\mathbf{v}_n))^2}{2\bar{\sigma}_l(\mathbf{v}_n)} - \frac{(z_{ln}^{(d)})^2}{2} + \log\frac{\pi}{1-\pi}\Big) \tag{19}$$

where the expectation with respect to $r(\mathbf{z}_n)$ has been approximated by a Monte-Carlo average.

4.4. M step

The parameters of the mixture model, that is, $\{\mathbf{W}_b,\mathbf{H}_b,\pi\}$ are estimated by optimizing the expected data log-likelihood, which takes the following form:

$$Q(\mathbf{W}_b,\mathbf{H}_b,\pi) = \sum_{(f,n)} -\frac{|x_{fn}-m_{fn}|^2+\nu_{fn}}{(\mathbf{W}_b\mathbf{H}_b)_{fn}} - \log\left(\mathbf{W}_b\mathbf{H}_b\right)_{fn} + \pi_n\cdot\log\pi + (1-\pi_n)\cdot\log(1-\pi). \tag{20}$$

The update formulas for $\mathbf{W}_b$ and $\mathbf{H}_b$ are then obtained by using standard multiplicative update rules [17]:

$$\mathbf{H}_b \leftarrow \mathbf{H}_b \odot \frac{\mathbf{W}_b^\top\left(\mathbf{V}\odot(\mathbf{W}_b\mathbf{H}_b)^{\odot-2}\right)}{\mathbf{W}_b^\top(\mathbf{W}_b\mathbf{H}_b)^{\odot-1}}, \tag{21}$$

$$\mathbf{W}_b \leftarrow \mathbf{W}_b \odot \frac{\left(\mathbf{V}\odot(\mathbf{W}_b\mathbf{H}_b)^{\odot-2}\right)\mathbf{H}_b^\top}{(\mathbf{W}_b\mathbf{H}_b)^{\odot-1}\mathbf{H}_b^\top}, \tag{22}$$

where

$$\mathbf{V} = \Big[|x_{fn}-m_{fn}|^2+\nu_{fn}\Big]_{(f,n)}. \tag{23}$$

The prior probability of α_n is also updated as follows:

$$\pi = \frac{1}{N}\sum_{n=1}^{N}\pi_n. \tag{24}$$

4.5. Speech enhancement

After the convergence of the VEM, the speech STFT frames are estimated as their posterior means. That is, $\forall(f,n)$:

$$\hat{s}_{fn} = \mathbb{E}_{r(s_{fn})}[s_{fn}] = \frac{\gamma_{fn}}{\gamma_{fn}+(\mathbf{W}_b\mathbf{H}_b)_{fn}}\cdot x_{fn} \tag{25}$$

where all the involved parameters are set to their optimal values obtained in the VEM framework.

The complete speech enhancement algorithm is summarized in Algorithm 1. Note that by setting $\forall n:\ \pi_n = \pi = 1$ (respectively, $\forall n:\ \pi_n = \pi = 0$), this algorithm reduces to an A-VAE (respectively, AV-VAE) based speech enhancement method.

Algorithm 1 VAE-MM for speech enhancement

1: **Inputs:**

- Learned A-VAE and AV-VAE models [14]
- Noisy microphone STFT frames $\mathbf{x} = \{\mathbf{x}_n\}_{n=0}^{N-1}$
- Visual embeddings $\mathbf{v} = \{\mathbf{v}_n\}_{n=0}^{N-1}$

2: **Initialization:**

- Initialize the NMF noise parameters $\mathbf{H}_b$ and $\mathbf{W}_b$ with random nonnegative values, and set $\pi = 0.5$.
- Initialize the latent codes $\mathbf{z}^a = \{\mathbf{z}_n^a\}_{n=0}^{N-1}$ (A-VAE) and $\mathbf{z}^{av} = \{\mathbf{z}_n^{av}\}_{n=0}^{N-1}$ (AV-VAE) using the corresponding learned encoder networks with $\mathbf{x}$ and $\mathbf{v}$. Then, set $\mathbf{z} = \{\pi \cdot \mathbf{z}_n^a + (1-\pi) \cdot \mathbf{z}_n^{av}\}_{n=0}^{N-1}$.

3: **while** stop criterion not met **do**:

- **E-z step:** Sample from (16) by the MH algorithm.
- **E-s step:** Update $\mathbf{s}_n$'s posterior by (12).
- **E-α step:** Update α_n's posterior by (19).
- **M-step:** Update the parameters using (21) – (24).

4: **end while**

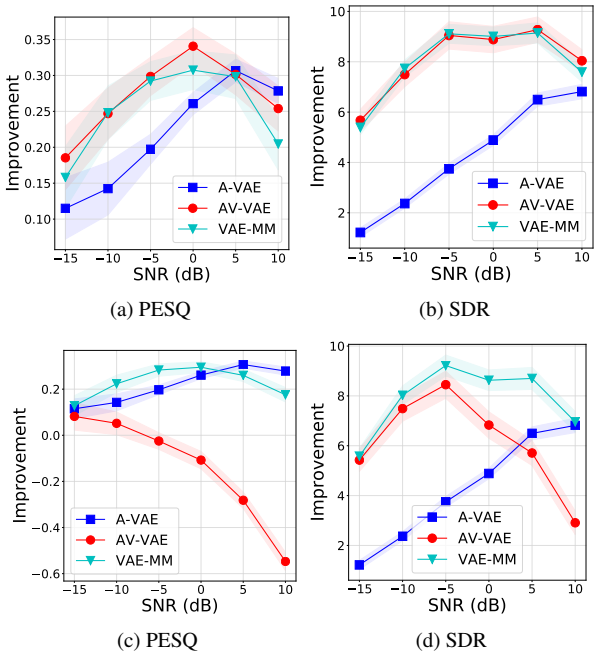
5: **Speech enhancement:** using (25).

5. EXPERIMENTS

Dataset and models. We considered two trained VAE models: an A-VAE and an AV-VAE, from [14], which have been trained on the NTCD-TIMIT dataset [18]. The test set includes about 1 hour noisy speech, along with their corresponding lips ROIs, with six different noise types, including *Living Room (LR)*, *White*, *Cafe*, *Car*, *Babble*, and *Street*, with noise levels: $\{-15, -10, -5, 0, 5, 10\}$ dB.

Parameters settings. For the MH algorithm, similarly to [6, 14], we run 40 iterations using $\epsilon = 0.01$ for the proposal distribution. The first 30 samples were discarded. The VEM steps were performed for 200 iterations. We initialize $m_{fn} = x_{fn}$ and $\nu_{fn} = 0, \forall(f, n)$. The latent codes, $\mathbf{z} = \{\mathbf{z}_n\}_{n=0}^{N-1}$, were initialized as described in Algorithm 1.

Experimental protocol. In the following, we compare the performance of A-VAE, AV-VAE, and the proposed VAE-MM. First, we run the three methods on clean visual data. In the second experiment, we randomly corrupt about $1/3$ of the total lips images per test instance. The occluded images are created by randomly selecting sub-sequences of 20 consecutive video frames and adding to the associated lips images random patches of standard Gaussian noise. We used two standard speech enhancement scores, i.e., the signal-to-distortion ratio (SDR) [19] and the perceptual evaluation of speech quality (PESQ) [20] scores. SDR is measured in decibels (dB) while PESQ values lie in the interval $[-0.5, 4.5]$ (higher the better). For computing SDR, the mir_eval Python library was used.[2] For each measure, we report the difference between the output value, i.e., evaluated on the enhanced speech signal, and the input value, i.e., evaluated on the noisy mixture.



(a) PESQ (b) SDR

(c) PESQ (d) SDR

Fig. 1: Speech enhancement performance for clean (top) and noisy (bottom) visual data.

Results. Figure 1 summarizes the results. First, as can be seen, AV-VAE shows a much better performance than A-VAE when visual data is clean. Second, in the case of the clean visual data, VAE-MM and AV-VAE show similar performances in terms of SDR. In the PESQ measure, we see some small drops in the performance of VAE-MM compared to that of AV-VAE. For noisy visual data, AV-VAE's performance drops significantly. However, VAE-MM seems to have successfully skipped the occluded lips images by switching to A-VAE. Its performance is still better than that of A-VAE, as some of the video frames contain clean and usable visual data.

6. CONCLUSION

In this paper, we proposed an audio-visual speech generative model based on a VAE mixture consisting of a trained A-VAE and a trained AV-VAE. Combined with an NMF model for noise variance [5, 6, 14], the goal was to make audio-visual VAE speech enhancement robust against noisy visual frames by switching to A-VAE in an unsupervised way. We presented a variational expectation-maximization approach to estimate the parameters of the model as well as the clean speech. The promising performance of the proposed method was demonstrated through some experiments.

[2] https://github.com/craffel/mir_eval

7. REFERENCES

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

[2] W. DeLiang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[3] P. C. Loizou, *Speech enhancement: theory and practice*, CRC press, 2007.

[4] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2015.

[5] 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 International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2018, pp. 716–720.

[6] S. Leglaive, L. Girin, and R. Horaud, "A variance modeling framework based on variational autoencoders for speech enhancement," in *Proc. IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2018, pp. 1–6.

[7] K. Sekiguchi, Y. Bando, K. Yoshii, and T. Kawahara, "Bayesian multichannel speech enhancement with a deep speech prior," in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2018, pp. 1233–1239.

[8] S. Leglaive, L. Girin, and R. Horaud, "Semi-supervised multichannel speech enhancement with variational autoencoders and non-negative matrix factorization," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019, pp. 101–105.

[9] S. Leglaive, U. Şimşekli, A. Liutkus, L. Girin, and R. Horaud, "Speech enhancement with variational autoencoders and alpha-stable distributions," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019, pp. 541–545.

[10] M. Pariente, A. Deleforge, and E. Vincent, "A statistically principled and computationally efficient approach to speech enhancement using variational autoencoders," in *Proc. Conference of the International Speech Communication Association (INTERSPEECH)*, 2019.

[11] L. Girin, J.-L. Schwartz, and G. Feng, "Audio-visual enhancement of speech in noise," *The Journal of the Acoustical Society of America*, vol. 109, no. 6, pp. 3007–3020, 2001.

[12] T. Afouras, J. S. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement," in *Proc. Conference of the International Speech Communication Association (INTERSPEECH)*, 2018, pp. 3244–3248.

[13] A. Gabbay, A. Shamir, and S. Peleg, "Visual speech enhancement," in *Proc. Conference of the International Speech Communication Association (INTERSPEECH)*, 2018, pp. 1170–1174.

[14] M. Sadeghi, S. Leglaive, X. Alameda-Pineda, L. Girin, and R. Horaud, "Audio-visual speech enhancement using conditional variational auto-encoder," *arXiv preprint arxiv.org/abs/1908.02590*, 2019.

[15] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *International Conference on Learning Representations (ICLR)*, 2014.

[16] C. Bishop, *Pattern Recognition and Machine Learning*, Springer-Verlag Berlin, Heidelberg, 2006.

[17] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural computation*, vol. 21, no. 3, pp. 793–830, 2009.

[18] A.-H. Abdelaziz, "NTCD-TIMIT: A new database and baseline for noise-robust audio-visual speech recognition," in *Proc. Conference of the International Speech Communication Association (INTERSPEECH)*, 2017, pp. 3752–3756.

[19] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[20] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2001, pp. 749–752.