Switching Variational Auto-Encoders for Noise-Agnostic Audio-visual Speech Enhancement

Mostafa Sadeghi, Xavier Alameda-Pineda

SWITCHING VARIATIONAL AUTO-ENCODERS FOR NOISE-AGNOSTIC AUDIO-VISUAL SPEECH ENHANCEMENT

Mostafa Sadeghi[1] *and Xavier Alameda-Pineda,*[2] *IEEE Senior Member*

[1]Inria Nancy Grand-Est, [2]Inria Grenoble Rhône-Alpes & Univ. Grenoble Alpes, France

ABSTRACT

Recently, audio-visual speech enhancement has been tackled in the unsupervised settings based on variational auto-encoders (VAEs), where during training only clean data is used to train a generative model for speech, which at test time is combined with a noise model, e.g. nonnegative matrix factorization (NMF), whose parameters are learned without supervision. Consequently, the proposed model is agnostic to the noise type. When visual data are clean, audio-visual VAE-based architectures usually outperform the audio-only counterpart. The opposite happens when the visual data are corrupted by clutter, e.g. the speaker not facing the camera. In this paper, we propose to find the optimal combination of these two architectures through time. More precisely, we introduce the use of a latent sequential variable with Markovian dependencies to switch between different VAE architectures through time in an unsupervised manner: leading to switching variational auto-encoder (SwVAE). We propose a variational factorization to approximate the computationally intractable posterior distribution. We also derive the corresponding variational expectation-maximization algorithm to estimate the parameters of the model and enhance the speech signal. Our experiments demonstrate the promising performance of SwVAE.

***Index Terms*—** Audio-visual speech enhancement, robustness, variational auto-encoder, variational inference.

1. INTRODUCTION

Audio-visual speech enhancement (AVSE) refers to the task of removing background noise from a noisy speech with the help of visual information (lip movements) of the unknown speech [1, 2]. Several deep neural network (DNN)-based methods have been proposed for AVSE in the past. The majority of these methods are *supervised*, where the underlying idea is to learn a DNN that maps noisy speech and its associated visual data (video frames of mouth area) to clean speech [2–5]. To have a good generalization performance, a huge dataset with different noise types and various signal-to-noise ratio (SNR) levels is usually required.

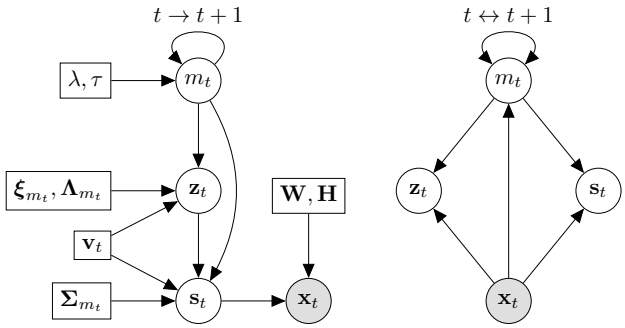
Xavier Alameda-Pineda acknowledges ANR JCJC ML3RI project (ANR-19-CE33-0008-01). This work has been partially supported by MIAI @ University Grenoble Alpes, (ANR-19-P3IA-0003)

Recently, some *unsupervised* AVSE methods have been proposed that do not need noise signals for training [6–8], meaning that their training is agnostic to the noise type. This approach builds upon the audio-only speech enhancement counterpart [9, 10] consisting of two main steps. First, modeling the probabilistic generative process of clean speech using VAEs [11]. Second, combining it with a noise model, e.g. NMF, to perform speech enhancement from noisy speech.

One critical issue with AVSE methods, shared with other AV-processing tasks such as speaker localisation and tracking [12, 13], is how to robustly handle noisy visual data at test time, e.g., when mouth area is heavily occluded or non-frontal. Exploiting such noisy visual data by an AVSE model trained on clean data may degrade the performance. In the supervised settings, this problem is usually addressed by proper data augmentation and efficient audio-visual fusion strategies during model training. For example, [14] proposes to combine speaker embedding with visual cues to achieve more robustness to occluded visual stream. Moreover, during training, some artificial occlusions are added to video frames. In the VAE-based unsupervised settings, a totally different perspective is pursued owning to its probabilistic nature. In this regard, a robust generative model has been proposed in [7] which is a mixture of trained audio-based (A-VAE) and audio-visual based (AV-VAE) model. As such, following a variational inference approach, for noisy visual data the A-VAE model is chosen, whereas for clean visual data the AV-VAE model is used, thus providing robustness.

In this paper, we build upon [7] and introduce a new model and associated robust AVSE algorithm, where a Markovian dependency is assumed to switch between different VAE-based generative models, and term them switching variational auto-encoder (SwVAE). Alternatively, the proposed model can be understood as a hidden Markov model (HMM) [15] with emission probabilities given by the decoder of several VAEs. Furthermore, we propose a variational factorization of the posterior distribution of the latent variables, enabling efficient inference and algorithm initialization. Experimental results demonstrate the superior performance of the proposed method compared to [7].

The rest of the paper is organized as follows. Section 2 introduces the proposed SwVAE. The inference and speech enhancement methodologies, and the relation of the present work to [7] are also detailed in this section. Section 3 presents and discusses the experiments.



(a) Graphical model (b) Variational approximation

Fig. 1: Graphical model (left) and proposed variational inference (right) assotiacted to switching variational autoencoders.

2. SWITCHING VARIATIONAL AUTOENCODERS

In this section, we present a generative model for short-time Fourier transform (STFT) time frames of clean speech consisting of audio-only and audio-visual VAE models plus a switching variable deciding which model to be used for each audio frame. The switching variable is modeled with an HMM. We also discuss how to structure the variance of the background noise via NMF. Then, a variational approximation is proposed to estimate the model parameters and infer the latent variables, including the clean speech signal, from the noisy mixture.

2.1. The generative model of SwVAE

We define $\mathbf{s}_t \in \mathbb{C}^F$ as the vector of clean speech STFT coefficients at time frame $t \in \{1, ..., T\}$. In the following, $\mathcal{N}_c$ and $\mathcal{N}$ stand for complex- and real-valued Gaussian distributions, respectively. The main methodological contribution of this paper is the use of a switching variable $m_t \in \{1, \dots, M\}$ modeled with a Markov chain in combination with a set of M non-linear generative models (i.e. VAE) to model clean speech. The full generative model describes the probabilistic relationship between the switching variable m_t, the clean speech $\mathbf{s}_t$, and the latent code $\mathbf{z}_t \in \mathbb{R}^L$, describing some hidden characteristics of $\mathbf{s}_t$, given the associated visual data representation $\mathbf{v}_t \in \mathbb{R}^V$. There are two possible, equivalent interpretations of this model. First, a hidden Markov model with emission probabilities given by the decoder of M VAEs. Second, a set of M VAEs switched by a selecting variable modeled with Markovian dependencies. More formally:

$$\begin{cases} p(m_1, \dots, m_T) \sim \mathcal{MC}(\lambda, \tau), \\ p(\mathbf{z}_t | m_t; \mathbf{v}_t) \sim \mathcal{N}\Big(\boldsymbol{\xi}_{m_t}(\mathbf{v}_t), \boldsymbol{\Lambda}_{m_t}(\mathbf{v}_t)\Big), \\ p(\mathbf{s}_t | \mathbf{z}_t, m_t; \mathbf{v}_t) \sim \mathcal{N}_c\Big(\mathbf{0}, \boldsymbol{\Sigma}_{m_t}(\mathbf{z}_t, \mathbf{v}_t)\Big), \end{cases} \tag{1}$$

where $\mathcal{MC}(\lambda, \tau)$ is short for a Markov chain with initial distribution λ and transition distribution τ, and $\boldsymbol{\xi}_{m_t}(.)$, $\boldsymbol{\Lambda}_{m_t}(.)$, and $\boldsymbol{\Sigma}_{m_t}(.,.)$ are non-linear transformations of their inputs indexed by $m_t \in \{1, \dots, M\}$ and realized as DNNs. For each generative model, the associated DNNs are trained by approximating the intractable posterior $p(\mathbf{z}_t | \mathbf{s}_t, m_t; \mathbf{v}_t)$ by another DNN-based parameterized Gaussian distribution called the encoder [6, 11]. So, there are M different distributions for the prior of $\mathbf{z}_t$ and for the likelihood of $\mathbf{s}_t$. Importantly, the switching variable m_t selects which one of the M models is used at each time step t, while ensuring temporal smoothing in the choice of this transformation. To complete the definition of the probabilistic model, we use an NMF structure for the additive noise [6, 9, 10]:

$$p(\mathbf{x}_t | \mathbf{s}_t) \sim \mathcal{N}_c\Big(\mathbf{s}_t, \mathrm{diag}\Big(\mathbf{W}\boldsymbol{h}_t\Big)\Big), \tag{2}$$

where $\mathbf{W} \in \mathbb{R}_+^{F \times K}$, $\mathbf{H} \in \mathbb{R}_+^{K \times T}$, and $\boldsymbol{h}_t$ denotes the t-th column of $\mathbf{H}$. The graphical representation of the full model is shown in Fig. 1 (a). The set of HMM and NMF parameters, i.e. $\{\lambda, \tau, \mathbf{W}, \mathbf{H}\}$ are then estimated following a variational inference method detailed in the next section, and represented in Fig. 1 (b). While for the generative model the dependencies are forward in time, at inference time, the latent code and spectrogram at any time t depend on the past and future noisy observations. It should be emphasized that the DNN parameters of (1), trained according to [6], are fixed.

2.2. Variational Inference

In the proposed formulation, the problem of speech enhancement is cast into the computation of the posterior probability $p(\mathbf{s}|\mathbf{x}, \mathbf{v})$, which is the marginal of the full posterior $p(\mathbf{s}, \mathbf{z}, \mathbf{m}|\mathbf{x}, \mathbf{v})$, where we define $\mathbf{x} = \{\mathbf{x}_t\}_{t=1}^T$ and analogously $\mathbf{s}, \mathbf{z}, \mathbf{m}, \mathbf{v}$. The full posterior being intractable, we propose the following variational factorization:

$$p(\mathbf{s}, \mathbf{z}, \mathbf{m}|\mathbf{x}, \mathbf{v}) \approx r^s(\mathbf{s}|\mathbf{m}) r^z(\mathbf{z}|\mathbf{m}) r^m(\mathbf{m}). \tag{3}$$

It is easy to see that r^s and r^z further factorize over time, meaning that: $r^s(\mathbf{s}|\mathbf{m}) = \prod_t r^s(\mathbf{s}_t|m_t)$ and analogously for $r^z(\mathbf{z}|\mathbf{m})$. Moreover, as a variational approximation, the posterior of the latent code $\mathbf{z}_t$ is assumed to follow a Gaussian distribution $r^z(\mathbf{z}_t|m_t) = \mathcal{N}(\mathbf{c}_{tm}, \boldsymbol{\Omega}_{tm})$, where the mean vector $\mathbf{c}_{tm}$ and the diagonal covariance matrix $\boldsymbol{\Omega}_{tm}$ are to be estimated along with r^s and r^m. To this end, we optimize the following lower-bound of the data log-likelihood $\log p(\mathbf{x}, \mathbf{v})$, as done in variational inference:

$$\mathbb{E}_{r^s r^z r^m}\left[\log \frac{p(\mathbf{x}, \mathbf{v}, \mathbf{s}, \mathbf{z}, \mathbf{m})}{r^s(\mathbf{s}|\mathbf{m}) r^z(\mathbf{z}|\mathbf{m}) r^m(\mathbf{m})}\right] \leq \log p(\mathbf{x}, \mathbf{v}). \tag{4}$$

2.2.1. E-s step

Optimizing (4) over r^s provides the following expression:

$$r^s(\mathbf{s}_t|m_t) \propto p(\mathbf{x}_t|\mathbf{s}_t) \cdot \exp\Big(\mathbb{E}_{r^z}\Big[\log p(\mathbf{s}_t|\mathbf{z}_t, m_t; \mathbf{v}_t)\Big]\Big).$$

Approximating the intractable expectation with a Monte-Carlo estimate, we obtain a Gaussian distribution: $r^s(\mathbf{s}_t|m_t) = \mathcal{N}_c(\boldsymbol{\eta}_t^{m_t}, \text{diag}[\boldsymbol{\nu}_t^{m_t}])$, where:

$$\eta_{ft}^{m_t} = \frac{\gamma_{ft}^{m_t}}{\gamma_{ft}^{m_t} + (\mathbf{WH})_{ft}} \cdot x_{ft},\ \nu_{ft}^{m_t} = \frac{\gamma_{ft}^{m_t} \cdot (\mathbf{WH})_{ft}}{\gamma_{ft}^{m_t} + (\mathbf{WH})_{ft}}, \quad (5)$$

$$\gamma_{ft}^{m_t} = \Big[\frac{1}{D}\sum_{d=1}^{D} \Sigma_{m_t,ff}^{-1}(\mathbf{z}_{m_t}^{(d)}, \mathbf{v}_t)\Big]^{-1}, \quad (6)$$

in which, $\Sigma_{m_t,ff}$ denotes the (f,f)-th entry of $\boldsymbol{\Sigma}_{m_t}$ (similarly for the rest of the variables), and $\{\mathbf{z}_{m_t}^{(d)}\}_{d=1}^D$ is a sequence sampled from $r^z(\mathbf{z}_t|m_t)$. The result in (5) must be interpreted as a Wiener filter, averaged over the latent variable $\mathbf{z}_t$ for a given VAE generative model m_t. The enhanced speech signal is the marginalisation over the switching variable at time t, and naturally writes:

$$\hat{\mathbf{s}}_t = \mathbb{E}_{r^m(m_t)}\Big[\mathbb{E}_{r^s(\mathbf{s}_t|m_t)}[\mathbf{s}_t]\Big] = \sum_{m_t} r^m(m_t)\boldsymbol{\eta}_t^{m_t}, \quad \forall t. \quad (7)$$

2.2.2. E-z step

After doing some derivations, the set of parameters of $r^z(\mathbf{z}_t|m_t)$ is estimated by solving:

$$\max_{\mathbf{c}_{tm},\boldsymbol{\Omega}_{tm}} \mathbb{E}_{r^m(m_t)}\Big[\mathbb{E}_{r^z(\mathbf{z}_t|m_t)}\Big[\mathbb{E}_{r^s(\mathbf{s}_t|m_t)}\Big[\log p(\mathbf{s}_t|\mathbf{z}_t, m_t; \mathbf{v}_t)\Big]\Big] - \text{KL}(r^z(\mathbf{z}_t|m_t)\|p(\mathbf{z}_t|m_t; \mathbf{v}_t))\Big]. \quad (8)$$

where, KL denotes the Kullback-Leibler divergence. In (8), the expectation over r^m and r^s can be evaluated in closed-form. This is also the case for the KL term as both the distributions are Gaussian. However, the expectation over r^z is intractable. Like in standard VAE, here we approximate this expectation with a single sample drawn from r^z. Furthermore, to be able to back-propagate through the posterior parameters, the reparametrization trick is utilized [11].

2.2.3. E-m step

For $r^m(\mathbf{m})$, we obtain:

$$r^m(\mathbf{m}) \propto p(\mathbf{m}) \cdot \prod_{t=1}^{T} \exp(-g_t(m_t)) \quad (9)$$

with:

$$g_t(m_t) = \mathbb{E}_{r^z}\Big[\text{KL}(r^s(\mathbf{s}_t|m_t)\|p(\mathbf{s}_t|\mathbf{z}_t, m_t; \mathbf{v}_t))\Big] - \mathbb{E}_{r^s}\Big[\log p(\mathbf{x}_t|\mathbf{s}_t)\Big] + \text{KL}(r^z(\mathbf{z}_t|m_t)\|p(\mathbf{z}_t|m_t; \mathbf{v}_t)) \quad (10)$$

Again, the KL terms and the expectation over r^s can be computed in closed-form. However, we approximate the expectation over r^z by a Monte-Carlo estimate. This allows us to compute (10). In order to compute the marginal variational posterior $r^m(m_t)$ required in the E-s and E-z steps, we realize that (9) has the same structure as standard HMM if we consider $\exp(-g_t(m_t))$ as the emission probability of the HMM. We therefore use the forward-backward algorithm [15] to efficiently compute $r^m(m_t)$.

Algorithm 1 SwVAE

1: **Input:** Trained A-VAE and AV-VAE models, noisy STFT frames $\{\mathbf{x}_t\}_{t=1}^T$, and visual embeddings $\{\mathbf{v}_t\}_{t=1}^T$.
2: **Initialize:**
- The latent codes $\{\mathbf{z}_{m_t}^{(d)}\}_{d=1}^D$ via the VAE encoders.
- The parameters of $r^s(\mathbf{s}|\mathbf{m})$ using (5).
- The posterior $r^m(\mathbf{m})$ uniformly.
- The parameters $\mathbf{W}$, $\mathbf{H}$, τ and λ (randomly).

3: **While** stop criterion not met **do**:
- **E-z step:** Using (8).
- **E-s step:** Using (5).
- **E-m step:** Compute $q_{mt} = \frac{\exp(-g_t(m_t))}{\sum_{m_t}\exp(-g_t(m_t))}$ using (10), and run the forward backward algorithm [15] to obtain the posterior probability $r^m(m_t)$ and the joint posterior probability $\zeta^m(m_{t-1}, m_t)$.
- **M step:** Update $\mathbf{W}, \mathbf{H}$ using (12) and (11), and λ, τ using the standard formulae with r^m and ζ^m [15].

4: **End while**
5: **Speech enhancement:** Using (7).

2.2.4. M step

After performing the E steps, the NMF parameters are updated by optimizing (4). The update formulas for $\mathbf{W}$ and $\mathbf{H}$ are then obtained by using standard multiplicative rules [16]:

$$\mathbf{H} \leftarrow \mathbf{H} \odot \frac{\mathbf{W}^\top\left(\mathbf{V} \odot (\mathbf{WH})^{\odot -2}\right)}{\mathbf{W}^\top(\mathbf{WH})^{\odot -1}}, \quad (11)$$

$$\mathbf{W} \leftarrow \mathbf{W} \odot \frac{\left(\mathbf{V} \odot (\mathbf{WH})^{\odot -2}\right)\mathbf{H}^\top}{(\mathbf{WH})^{\odot -1}\mathbf{H}^\top}, \quad (12)$$

where $\mathbf{V} = \Big[\sum_{m_t} r^m(m_t)(|x_{ft} - \eta_{ft}^{m_t}|^2 + \nu_{ft}^{m_t})\Big]_{(f,t)}$, and $\odot$ signifies entry-wise operation. The parameters of the HMM, i.e. λ and τ, are updated by the standard formulae using the joint posterior probabilities computed by the forward-backward algorithm in the E-m step. The complete inference and enhancement algorithm is summarized in Algorithm 1.

2.3. Novelty of SwVAE w.r.t. [7]

The closest work to ours is [7], which uses a mixture model, comprising an A-VAE and an AV-VAE, as the generative

Table 1: Average PESQ, SDR and STOI values of the enhanced speech signals. Here, "clean" and "noisy" refer to visual data.

Measure	PESQ					SDR (dB)					STOI				
SNR (dB)	-5	0	5	10	15	-5	0	5	10	15	-5	0	5	10	15
Input	1.44	1.67	2.04	2.30	2.72	-12.30	-7.30	-3.45	1.88	6.73	0.22	0.32	0.45	0.56	0.68
[7] - clean	**1.70**	1.92	2.29	2.48	2.66	**-3.51**	1.67	5.38	9.22	12.07	0.24	0.35	0.47	0.55	0.65
SwVAE - clean	1.67	**1.97**	**2.39**	**2.62**	**2.83**	-3.59	**2.00**	**6.24**	**10.73**	**14.12**	**0.25**	**0.36**	**0.51**	**0.61**	**0.72**
[7] - noisy	**1.66**	1.91	2.22	2.41	2.51	**-3.78**	1.50	5.18	8.72	10.88	0.23	0.34	0.45	0.53	0.63
SwVAE - noisy	1.65	**1.94**	**2.36**	**2.60**	**2.81**	-3.97	**1.84**	**6.14**	**10.51**	**14.06**	**0.24**	**0.35**	**0.50**	**0.59**	**0.67**

model of clean speech. Though sharing some similarities, there are several crucial differences between the two methods. First, here we assume a Markovian dependency on the switching variable that ensures smoothness over time. Second, in [7] the following variational factorization is proposed: $p(\mathbf{s}, \mathbf{z}, \mathbf{m}|\mathbf{x}) \approx r^s(\mathbf{s})r^z(\mathbf{z})r^m(\mathbf{m})$, where r^s and r^z are not conditioned on $\mathbf{m}$. This is in contrast to our proposed factorization given in (3), which provides a more effective approximation and a robust initialization for the latent codes $\mathbf{z}$, as required by the inference algorithm. More precisely, in the proposed framework, the parameters of $r^s(\mathbf{s}|\mathbf{m})$ are initialized using its respective set of latent codes $\mathbf{z}$, which themselves are initialized by the corresponding encoders (see Section 3), as opposed to [7] where a weighted combination of the latent codes (coming from different models) is used for initializing the parameters of $r^s(\mathbf{s})$. This might not be effective given that latent initialization is important in VAE-based AVSE [8]. Finally, the proposed posterior approximation $r^z(\mathbf{z}_t|m_t) = \mathcal{N}(\mathbf{c}_{tm}, \mathbf{\Omega}_{tm})$ makes sampling, needed by (6), more efficient than the method of [7] which relies on the computationally demanding Metropolis-Hastings algorithm [15].

3. EXPERIMENTS

Protocol We evaluate the performance of SwVAE and compare it with [7] using the same experimental protocol. We used two VAE models (A-VAE and AV-VAE)[1] from [6], trained on the NTCD-TIMIT dataset [17]. The test set includes 9 speakers, along with their corresponding lip region of interest, with different noise types: *Living Room (LR)*, *White*, *Cafe*, *Car*, *Babble*, and *Street*, and noise levels: $\{-5, 0, 5, 10, 15\}$ dB. From each speaker, we randomly selected 150 examples per noise level for evaluation.

The parameters for the algorithm of [7] where set as their proposed values. Both of the algorithms were run for 200 iterations, on the same test set. For optimizing (8), the Adam optimizer [18] was used with a learning rate of 0.05 for 10 iterations. Moreover, we used $D = 20$ samples to compute (6) and (10). The $\mathbf{c}_{tm}, \mathbf{\Omega}_{tm}$ parameters of r^z were, respectively, initialized with the means and variances at the output of the respective VAE encoders by giving $(\mathbf{x}_t, \mathbf{v}_t)$ as their inputs. The parameters of r^s are then initialized using (5) and (6).

The two AVSE algorithms were run on the test set with both clean visual data as well as artificially generated noisy versions, where about one third of the total video frames per test instance were occluded. Similarly to [7], the occlusions were simulated by random patches of standard Gaussian noise added to randomly selected sub-sequences of 20 consecutive video frames. We used three standard speech enhancement scores, i.e., signal-to-distortion ratio (SDR) [19], perceptual evaluation of speech quality (PESQ) [20], and short-time objective intelligibility (STOI) [21]. SDR is measured in decibels (dB), and PESQ and STOI values lie in the intervals $[-0.5, 4.5]$ and $[0, 1]$, respectively (the higher the better).

Results Table 1 summarizes the results, averaged over all the test samples, for the three performance measures, and clean as well as noisy visual data. From this table, we can see that in terms of PESQ and SDR, SwVAE outperforms [7], with the performance difference being more significant in high SNR values. In terms of the intelligibility measure, i.e., STOI, the proposed method exhibits much better performance than [7]. These observations are consistent for both clean and noisy visual data. Furthermore, the two algorithms show robustness to noisy visual data, which is especially noticeable in terms of STOI. However, for the algorithm of [7] the performance drop due to noisy visual data is higher than SwVAE. Supplementary materials are available online[2].

4. CONCLUSION

In this paper, we proposed a noise-agnostic audio-visual speech generative model based on a sequential mixture of trained A-VAE and AV-VAE models, combined with an NMF model for the noise variance. The switching variable allows us to seamlessly use either of the auto-encoders for speech enhancement, without requiring supervision. We detailed a variational expectation-maximization approach to estimate the parameters of the model as well as to enhance the noisy speech. The proposed algorithm, called switching VAE (SwVAE), exhibits promising performance when compared to the previous work [7] on robust AVSE. In the future, we would like to explore the use of Dynamical VAEs [22] for unsupervised AVSE.

[1] For A-VAE, the prior of $\mathbf{z}_t$ is a standard normal distribution, and $\mathbf{\Sigma}_{m_t}$ is a function of only $\mathbf{z}_t$; see (1).

[2] https://team.inria.fr/perception/research/swvae/

5. REFERENCES

[1] 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.

[2] D. Michelsanti, Z. H. Tan, S. X. Zhang, Y. Xu, M. Yu, D. Yu, and J. Jensen, "An overview of deep-learning-based audio-visual speech enhancement and separation," 2020, arXiv:2008.09586.

[3] Jen-Cheng Hou, Syu-Siang Wang, Ying-Hui Lai, Yu Tsao, Hsiu-Wen Chang, and Hsin-Min Wang, "Audio-visual speech enhancement using multimodal deep convolutional neural networks," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 117–128, 2018.

[4] 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.

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

[6] M. Sadeghi, S. Leglaive, X. Alameda-Pineda, L. Girin, and R. Horaud, "Audio-visual speech enhancement using conditional variational auto-encoders," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 28, pp. 1788 –1800, 2020.

[7] M. Sadeghi and X. Alameda-Pineda, "Robust unsupervised audio-visual speech enhancement using a mixture of variational autoencoders," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020.

[8] M. Sadeghi and X. Alameda-Pineda, "Mixture of inference networks for vae-based audio-visual speech enhancement," 2020, arXiv:1912.10647.

[9] 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.

[10] 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.

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

[12] Jan Cech, Ravi Mittal, Antoine Deleforge, Jordi Sanchez-Riera, Xavier Alameda-Pineda, and Radu Horaud, "Active-speaker detection and localization with microphones and cameras embedded into a robotic head," in *IEEE-RAS Humanoids*, 2013, pp. 203–210.

[13] Yutong Ban, Laurent Girin, Xavier Alameda-Pineda, and Radu Horaud, "Exploiting the complementarity of audio and visual data in multi-speaker tracking," in *IEEE ICCV Workshops*, 2017, pp. 446–454.

[14] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman, "My lips are concealed: Audio-visual speech enhancement through obstructions," in *INTERSPEECH*, 2019.

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

[16] 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.

[17] 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.

[18] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.

[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.

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

[22] Laurent Girin, Simon Leglaive, Xiaoyu Bie, Julien Diard, Thomas Hueber, and Xavier Alameda-Pineda, "Dynamical variational autoencoders: A comprehensive review," *arXiv preprint arXiv:2008.12595*, 2020.