The Impact of Removing Head Movements on Audio-visual Speech Enhancement

Zhiqi Kang, Mostafa Sadeghi, Radu Horaud, Xavier Alameda-Pineda, Jacob Donley, Anurag Kumar

THE IMPACT OF REMOVING HEAD MOVEMENTS ON AUDIO-VISUAL SPEECH ENHANCEMENT

Zhiqi Kang[1]*, Mostafa Sadeghi*[2]*, Radu Horaud*[1]*, Xavier Alameda-Pineda*[1]*, Jacob Donley*[3] *and Anurag Kumar*[3]

[1]Inria Grenoble Rhône-Alpes & Univ. Grenoble Alpes, France, [2]Inria Nancy Grand-Est, France
[3]Reality Labs Research, Redmond WA, USA

ABSTRACT

This paper investigates the impact of head movements on audio-visual speech enhancement (AVSE). Although being a common conversational feature, head movements have been ignored by past and recent studies: they challenge today's learning-based methods as they often degrade the performance of models that are trained on clean, frontal, and steady face images. To alleviate this problem, we propose to use robust face frontalization (RFF) in combination with an AVSE method based on a variational auto-encoder (VAE) model. We briefly describe the basic ingredients of the proposed pipeline and we perform experiments with a recently released audio-visual dataset. In the light of these experiments, and based on three standard metrics, namely STOI, PESQ and SI-SDR, we conclude that RFF improves the performance of AVSE by a considerable margin.[1]

Index Terms— Audio-visual speech enhancement, variational auto-encoder, face frontalization.

1. INTRODUCTION

It has long been established that speech communication is multimodal. In particular, vision provides an alternative representation of some of the information that is present in the audio, with the advantage that it is not affected by acoustic noise. This has led to speech recognition and speech enhancement systems that combine audio and visual features to achieve robust performance in noise [1]. While there is a long history of research in audio-visual speech enhancement (AVSE), the most performant methods are based on deep neural networks (DNNs). Most DNN-based methods for AVSE are *supervised* [1]–[4]: they learn a direct mapping from the noisy speech and clean visual inputs onto the clean speech output. This mode of doing requires large-scale datasets with a high variety of noise types and signal-to-noise ratio (SNR) levels, which suffers from poor generalization to unseen noise environments. Alternatively, *unsupervised* AVSE methods do not necessitate noise signals for training, e.g. [5], [6]. They use variational auto-encoders (VAEs) [7] to model the generative process of audio-visual speech using multimodal datasets of clean audio and visual speech. Then, this model is combined with an audio noise model, e.g. nonnegative matrix factorization (NMF), to perform speech enhancement from noisy audio and visual speech. Thanks to the independence of the noise type, the unsupervised methods exhibit better generalization than their supervised counterpart.

Nevertheless, to date, both supervised and unsupervised AVSE methods assume clean visual information, namely they use lip regions that are cropped from clean, frontal and steady face images. In practice, the visual data are corrupted by various sources of perturbation, such as partial occlusion with an object and by head movements. The performance of existing AVSE systems rapidly degrades in the presence of corrupted visual information. Recently, there have been a few attempts to incorporate knowledge about the quality of the visual data at hand and to ignore the visual input whenever it cannot be properly exploited [8]–[10]. Here, we investigate AVSE models that put audio and visual information on an equal footing and that can deal with both noisy audio signals and noisy lip movements – a largely uninvestigated topic. We propose to remove rigid head movements using a face frontalization method that preserves non-rigid facial deformations.

The rest of the paper is organized as follows. Section 2 reviews the VAE-based AVSE model [5]. Section 3 sumarizes the face frontalization method of [11] . Section 4 describes the experiments and discusses the results.

2. AUDIO-VISUAL VARIATIONAL AUTOENCODER

In this section, we briefly summarize the conditional VAE (CVAE) based AVSE model of [5], denoted by AV-CVAE. The whole framework consists of two steps: training and testing (inference). In the first step, a prior distribution for clean speech is learned from the concatenation of a clean audio signal with an embedding of associated lip images. The second step infers clean speech from the noisy-speech and lip-embedding inputs: the learned prior distribution is combined

This work has been partially supported by the H2020 SPRING project #871245 and by the Multidisciplinary Institute of Artificial Intelligence (MIAI) ANR-19-P3IA-0003.

[1]https://team.inria.fr/robotlearn/head-movements-avse/

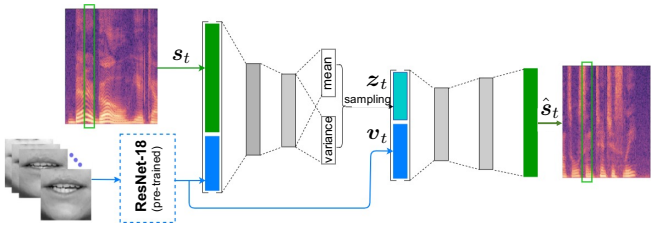

Fig. 1. AV-CVAE model with a pre-trained ResNet-18 network as visual feature extractor.

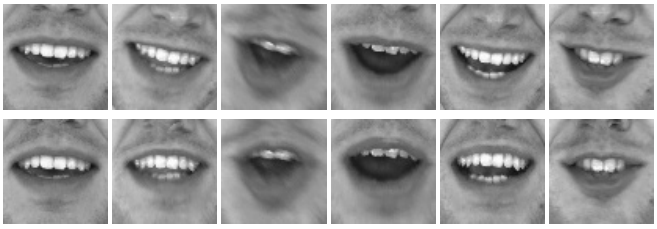

Fig. 2. An example of the effect of face frontalization on the lip regions cropped from the images of a head-moving speaker. Top: lip regions cropped from the original images (with head motions). Bottom: lip regions cropped from the frontalized images (head motions removed).

with a noise model, whose parameters together with the clean speech ones are estimated following a variational expectation-maximization (VEM) procedure.

Given a dataset of complex-valued short-time Fourier transform (STFT) frames of a clean-speech signal, denoted $\boldsymbol{s}_t \in \mathbb{C}^F$, and the corresponding lip embedding obtained from a lip bounding-box cropped from the image of a speaker face, denoted $\boldsymbol{v}_t \in \mathbb{R}^M$, a latent-variable generative model is trained using the VAE framework. This involves defining a parametric distribution for the likelihood $p_\Theta(\boldsymbol{s}_t|\boldsymbol{z}_t, \boldsymbol{v}_t)$, and a parametric prior distribution for the latent code $\boldsymbol{z}_t \in \mathbb{R}^L$, $L \ll F$, $p_\Gamma(\boldsymbol{z}_t|\boldsymbol{v}_t)$. These distributions are implemented by some deep neural network architectures, whose parameters, $\{\Theta, \Gamma\}$, are learned following an amortized variational inference [7], where an encoder network is introduced to approximate the intractable posterior distribution of the latent codes. Fig. 1 illustrates the AV-CVAE architecture. The main difference between this architecture and the one proposed in [5], [6] is the presence of a ResNet backbone from a pretrained model specialized for lip reading [12].

With the parametric prior distribution for clean speech being learned, one considers an observation model as $\boldsymbol{o}_t = \boldsymbol{s}_t + \boldsymbol{b}_t$, in which $\boldsymbol{o}_t \in \mathbb{C}^F$ and $\boldsymbol{b}_t \in \mathbb{C}^F$ denote observed speech and noise, respectively. Considering an NMF-based model for noise, and combining with the speech model, the set of NMF parameters are then learned by a variational inference procedure. Once learned, the clean speech estimate $\hat{\boldsymbol{s}}_t$ is obtained via a probabilistic Wiener filtering. More details can be found in [5].

3. ROBUST FACE FRONTALIZATION

We briefly describe a method for removing head movements with the goal to provide frontal and steady lip regions to the AV-CVAE model outlined above, e.g. Figure 2. The core idea of the robust face frontalization (RFF) method that we recently proposed [11], is to estimate the pose (scale σ, 3D rotation $\mathbf{R}$ and 3D translation $\boldsymbol{t}$) and the 3D deformable shape $\boldsymbol{s}$ of an arbitrarily-viewed input face, and to warp it onto a frontally viewed synthesized face. The main feature of the method is to perform pose and shape-deformation estimation sequentially. This way of doing enables a better estimation of the pose parameters in the presence of facial expressions, e.g. lip movements.

The pose is estimated by aligning two 3D points sets: a set of *observed* facial landmarks extracted from the input face, $\{\boldsymbol{X}_j\}_{j=1}^J \subset \mathbb{R}^3$, and a set of *model* landmarks associated with a neutral and frontal view of a mean face, $\{\boldsymbol{Z}_j\}_{j=1}^J \subset \mathbb{R}^3$. Because the landmark locations of the input face are inherently affected by detection errors as well as by *non-rigid facial deformations*, it is suitable to use a robust rigid-parameter estimation technique. For this purpose, we assume that the errors between the model and frontalized landmarks, i.e. $\boldsymbol{e}_j = \boldsymbol{Z}_j - \sigma\mathbf{R}\boldsymbol{X}_j - \boldsymbol{t}$, are samples of a random variable drawn from the Student t-distribution – a heavy tailed pdf that is able to deal with both Gaussian (inliers) and non-Gaussian (outliers) noise in the data, by assigning a weight random variable to each observed landmark [13]. The corresponding EM algorithm alternates between the estimation of (i) the weight posteriors, (ii) the pdf parameters and (iii) the rigid parameters. At convergence, EM assigns high weight posteriors to landmark pairs that are linked by a rigid transformation and low weights to landmark pairs that are affected by detection errors or by non-rigid facial deformations.

Next, the exact shape of the face is estimated by fitting a 3D morphable model (3DMM) to the frontalized 3D landmarks denoted $\{\boldsymbol{Y}_j\}_{j=1}^J$, with $\boldsymbol{Y}_j = \sigma^*\mathbf{R}^*\boldsymbol{X}_j + \boldsymbol{t}^*$, where * indicates the optimal parameter values previously computed. We use a linear deformation model which consists of a 3D mesh whose vertices are parameterized by a low-dimensional embedding $\boldsymbol{s}$. Once the face embedding is estimated, a frontal dense depth map of the face is built, which in turn enables the input face to be warped onto the frontalized one. The final step is to crop a lip region from the frontalized face.

The face alignment method of [14] (being one of the best performing 3D face alignment methods to date), is used to extract 3D landmarks. The Student EM algorithm is described in detail in [11]. All the computations inside EM are in closed

form, with the exception of the rotation. The latter is parameterized by a unit quaternion which allows a compact representation of the rotation and the use of a sequential least-squares programming (SLSQP) solver in combination with a root-finding software package [15]. Without loss of generality, we use a linear deformation model, namely the publicly available Basel Shape Model [16]. This provides registered face scans of 200 identities: 200 frontal scans with a neutral expression as well as 200 expressive scans. Each scan is described with $N = 53490$ vertices. Among these vertices, $J = 68$ are used for pose and shape estimation, $\{\boldsymbol{Z}_j\}_{j=1}^{J}$. The low-dimensional embedding is set to $K = 200$.

4. EXPERIMENTS

All the experiments reported below use the MEAD dataset [20] which contains short videos of talking faces with large-scale facial expressions. For all 46 publicly available participants, there are recordings of eight different emotions at three different intensity levels and seven camera viewpoints. Many participants have natural head motions, which challenges state-of-the-art AVSE. Among all videos, we select the videos of all emotion categories taken at the frontal view and at the level 3 (the highest) of emotion intensity. These high-intensity emotions are associated with large head movements and exaggerated lip motions, thus allowing to assess the effect of head movements. In total, there are around 5 hours of videos for training, 0.7 hours for validation and 0.7 hours for testing.

We process the input videos with three different frontalization methods to compare their effectiveness of removing head movements and hence of improving the quality of the speech output. We consider the following three methods. The first one [11], denoted RFF and outlined in Section 3, directly estimates a rigid motion and hence it preserves the facial expressions. The second method combines 3D-to-2D shape-to-image fitting with a style-transfer GAN model [18], denoted ST-GAN. The third one learns a non-linear image-to-image dual-attention GAN model [19], denoted DA-GAN. We also consider the case of directly using the raw input without any form of face frontalization, denoted WithHM (with head movements). In all these four cases we crop the lip region, which yields 67×67 images, which are then converted to gray scale and normalized to facilitate the downstream processing.

We consider three speech enhancement models. The Audio-only VAE (A-VAE), [5], [17] has an encoder and a decoder composed of fully-connected (fc) layers. The extracted audio feature vector is of size $F = 513$ whereas the latent space is of size $L = 32$. One audio-visual VAE model shares a similar encoder-decoder architecture as A-VAE, with the additional fully-connected layers to encode the visual information [5]. We denote this model as AV-CVAE. Furthermore,

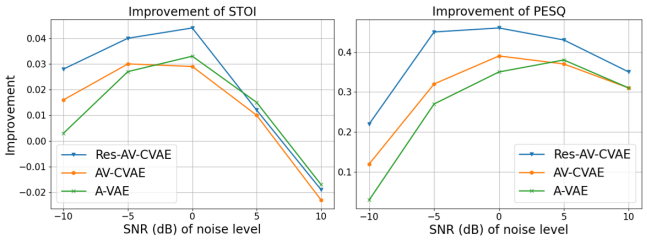

Fig. 3. Performance comparison of A-VAE, AV-CVAE and Res-AV-CVAE based on STOI (left) and PESQ (right).

we propose to use the ResNet backbone from a pretrained model specialized for lip reading [12] (shown in a dashed box in Figure 1) for visual feature extraction. The backbone follows the standard design of ResNet-18 [21] except for the first convolutional layer, which is replaced by a 3D convolutional layer to incorporate temporal information from neighbouring frames. This variant is denoted as Res-AV-CVAE. In practice, the dimension of the visual embedding is $M = 128$.

All the VAE models are trained in an end-to-end manner. The A-VAE model is trained on pure audio data. The pretrained model AV-CVAE [5] is fine-tuned on the MEAD dataset, whereas Res-AV-CVAE models are trained from scratch. Note that the ResNet backbone is frozen without requiring the gradients. It is hence a static feature extractor. We set $5e^{-5}$ as the learning rate for the fine-tuning model and $1e^{-4}$ for the training from scratch. The optimizer is Adam and the batch size is of 128. We also applied early stopping with a patience of 10 epochs. Note that we trained and tested one model with one specific lip preprocessing method at a time. At test time, noise from the DEMAND [22] dataset is combined with the clean speech to construct the audio input. For each noise type there are five noise levels: -10 dB, -5 dB, 0 dB, 5 dB and 10 dB. Three standard speech enhancement metrics are used for quantitative evaluation: scale-invariant signal-to-distortion ratio (SI-SDR) [23], short-time objective intelligibility (STOI) [24] and perceptual evaluation of speech quality (PESQ) [25]. SI-SDR is measured in decibels (dB), while STOI and PESQ values are in the range $[0, 1]$ and $[-0.5, 4.5]$, respectively (the higher the better).

We start with evaluating the impact of different frontalization methods on AVSE performance, i.e. Table 1, where the average scores for different levels of noise (SNR) are presented. Selecting the RFF, the best-performing method in the table, as an example, we remark that the difference between with and without frontalization is significant. This confirms that the head motions interfere the capture of visual speech patterns. In other words, separating the rigid head movements from the non-rigid lip deformations allows the model to learn a better clean speech model. Moreover, the comparison between A-VAE and Res-AV-CVAE with RFF further validates the contribution of the visual modality.

Table 1. Average STOI, PESQ, SI-SDR values.

Measure	STOI $[0, 1] \uparrow$					PESQ $[-0.5, 4.5] \uparrow$					SI-SDR (dB) $\uparrow$				
SNR (dB)	-10	-5	0	5	10	-10	-5	0	5	10	-10	-5	0	5	10
Noisy audio input	0.40	0.53	0.66	0.78	**0.86**	0.90	1.24	1.67	2.05	2.42	-15.92	-10.62	-5.44	-0.40	4.60
A-VAE [17]	0.41	0.56	0.70	**0.79**	0.85	0.93	1.51	2.02	2.43	2.73	-7.01	-0.29	5.08	**9.41**	**12.74**
AV-CVAE [5]	0.42	0.57	0.69	**0.79**	0.84	1.02	1.56	2.06	2.42	2.73	-6.96	-0.04	5.01	9.06	12.25
Res-AV-CVAE-WithHM	0.41	0.55	0.67	0.77	0.83	1.02	1.53	1.99	2.35	2.70	-7.84	-0.60	4.68	8.81	12.30
Res-AV-CVAE-DA-ST-GAN [18]	0.40	0.55	0.68	0.78	0.84	1.01	1.54	2.01	2.39	2.72	-7.92	-1.14	4.13	9.27	11.77
Res-AV-CVAE-DA-GAN [19]	0.39	0.55	0.66	0.68	0.72	0.76	1.42	1.87	1.66	1.96	-9.08	-0.45	3.88	4.55	5.23
Res-AV-CVAE-RFF [11]	**0.43**	**0.58**	**0.71**	**0.79**	0.85	**1.12**	**1.69**	**2.13**	**2.48**	**2.77**	**-6.30**	**0.10**	**5.24**	9.30	12.60

The choice of the face frontalization method is important. While RFF shows significant improvements, ST-GAN yields a minor difference compared to the presence of head movements. Indeed, GAN-based image generation models have no theoretical guarantee for preserving the lip shape – they add some form of visual noise, which neutralizes the gain of frontalization. This explanation is also supported by the results of DA-GAN: its performance is falling far behind the counterparts. As the results of ST-GAN are conditioned on the transformation-based process, the model possesses a prior knowledge about the frontalized face. Moreover, the direct mapping from an arbitrary viewpoint to a frontal view of DA-GAN introduces even more dramatic modifications in the lip shape. Thus, the model has more difficulties to learn the correct speech patterns from lip movements.

We then compare the performance of different VAE architectures in Figure 3, where the improvement of scores are shown as a function of different levels of noise (SNR). More precisely, the improvement refers to the difference between the score obtained by using the raw noisy speech and that obtained by using the enhanced speech. First, it is remarkable to see that Res-AV-CVAE significantly outperforms AV-CVAE, showing the gain of using a more powerful feature extractor. Second, we observe that with a noise level at between -5 dB and 0 dB, the Res-AV-CVAE model reaches an optimal stage (a peak in the curve) for fusing the audio-visual data. That is, with the noise level going higher (smaller SNR), the audio would be too corrupted to be enhanced. In contrast, with the noise level going lower (higher SNR), the importance of the visual data is decreasing and the already clean speech becomes harder to be enhanced. While the superior performance of the Res-AV-CVAE models is more significant at high noise levels, it is quite remarkable to observe that Res-AV-CVAE-RFF performs almost equally well as A-VAE for low noise levels. These experiments confirm the complementary roles of the visual and audio modalities for the task of speech enhancement.

To give an insight on the impact of removing head movements, Figure 4 shows the horizontal and vertical displacements of a landmark located on the upper lip. Both the vertical and horizontal trajectories of this lip landmark are strongly affected by head motions. In the light of this experiment, one may interpret the process of separating rigid head movements and non-rigid lip movements, as a way of extracting clean visual-speech information from the raw videos.

Fig. 4. Left: landmark location on the upper lip. Middle: horizontal landmark motion. Right: vertical landmark motion. Blue plots: Lip motions after frontalization (using RFF). Orange plots: lip motions in the presence of head movements (without frontalization).

5. CONCLUSION

In this paper, we investigated the effect of head movements on the task of audio-visual speech. We showed that the combination of face frontalization for removing head movements in combination with a ResNet backbone considerably improves the performance of state-of-the-art VAE AVSE, based on several speech enhancement metrics. We compared a recently proposed RFF method [11] with two state-of-the-art frontalization methods, both based on GANs. We showed that the built-in expression-preserving property of [11] yields much better results than methods based on a GAN architecture – the latter cannot guarantee that the frontalization process preserves the lip movements of the input. In the future, we foresee supplementary experiments with large head movements and more extreme head poses. In addition, combining face frontalization with the recently proposed switching-VAE AVSE model [10] is an interesting topic as well.

REFERENCES

[1] D. Michelsanti *et al.*, "An overview of deep-learning-based audio-visual speech enhancement and separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.

[2] J.-C. Hou *et al.*, "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.

[3] T. Afouras, J. S. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement," in *INTERSPEECH*, 2018, pp. 3244–3248.

[4] A. Gabbay, A. Shamir, and S. Peleg, "Visual speech enhancement," in *INTERSPEECH*, 2018, pp. 1170–1174.

[5] M. Sadeghi *et al.*, "Audio-visual speech enhancement using conditional variational auto-encoders," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1788–1800, 2020.

[6] M. Sadeghi and X. Alameda-Pineda, "Mixture of inference networks for VAE-based audio-visual speech enhancement," *IEEE Transactions on Signal Processing*, vol. 69, pp. 1899–1909, 2021.

[7] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *ICLR*, 2014.

[8] T. Afouras, J. S. Chung, and A. Zisserman, "My lips are concealed: Audio-visual speech enhancement through obstructions," in *INTERSPEECH*, 2019.

[9] M. Sadeghi and X. Alameda-Pineda, "Robust unsupervised audio-visual speech enhancement using a mixture of variational autoencoders," in *ICASSP*, 2020.

[10] M. Sadeghi and X. Alameda-Pineda, "Switching variational auto-encoders for noise-agnostic audio-visual speech enhancement," in *ICASSP*, 2021.

[11] Z. Kang, R. Horaud, and M. Sadeghi, "Robust face frontalization for visual speech recognition," in *ICCV Workshops*, 2021.

[12] B. Martinez, P. Ma, S. Petridis, and M. Pantic, "Lipreading using temporal convolutional networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 6319–6323.

[13] F. Forbes and D. Wraith, "A new family of multivariate heavy-tailed distributions with variable marginal amounts of tailweight: Application to robust clustering," *Statistics and Computing*, vol. 24, no. 6, pp. 971–984, 2014.

[14] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3d facial landmarks)," in *ICCV*, 2017.

[15] D. Kraft, "A software package for sequential quadratic programming," DLR German Aerospace Center – Institute for Flight Mechanics, Koln, Germany, Tech. Rep. DFVLR-FB 88-28, 1988.

[16] P. Paysan *et al.*, "A 3D face model for pose and illumination invariant face recognition," in *IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2009, pp. 296–301.

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

[18] H. Zhou *et al.*, "Rotate-and-render: Unsupervised photorealistic face rotation from single-view images," in *CVPR*, 2020.

[19] Y. Yin, S. Jiang, J. P. Robinson, and Y. Fu, "Dual-attention GAN for large-pose face frontalization," in *IEEE International Conference on Automatic Face and Gesture Recognition*, 2020, pp. 24–31.

[20] K. Wang *et al.*, "MEAD: A large-scale audio-visual dataset for emotional talking-face generation," in *ECCV*, Springer, 2020.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[22] J. Thiemann, N. Ito, and E. Vincent, "Demand: A collection of multi-channel recordings of acoustic noise in diverse environments," in *Proc. Meetings Acoust*, 2013, pp. 1–6.

[23] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?" In *ICASSP*, 2019.

[24] 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.

[25] 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 *ICASSP*, 2001.