IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

SPEECH ENHANCEMENT USING DEEP LEARNING AVE-NET

[1]MANNEM ANUSHA, [2]KAKA JHANSI RANI,

[1]M.Tech, [2]Assistant professor,

[1]Electronics and Communication Engineering,

[1]University collage of engineering-JNTU (k), Kakinada, India

Abstract: Most approaches to speech enhancement focus solely on audio features in order to create filters or transfer functions that convert noisy speech signals to clean ones. Many speech-related approaches have integrated visual data with audio data to achieve more effective speech processing performance. We propose audio-visual VAE variants for single-channel and speaker-independent speech enhancement in this project. We propose a conditional VAE (CVAE) in which the generative process of audio speech is conditioned on visual information from the lip region. Experiments are carried out using the recently released NTCD-TIMIT dataset and the GRID corpus. The results show that the proposed audio-visual CVAE effectively fuses audio and visual information, and it outperforms the audio-only VAE model in terms of speech enhancement performance, especially when the speech signal is heavily corrupted by noise.

Index Terms—Audio-visual speech enhancement, deep generative models, variational auto-encoders, nonnegative matrix factorization, Monte Carlo expectation-maximization.

1. INTRODUCTION:

The problem of speech enhancement (SE) is to estimate clean-speech signals from noisy single-channel or multiple-channel audio recordings. There is a long history of audio speech enhancement (ASE) methods and associated algorithms, software, and systems, such as [1]-[3]. In this paper, we address the issue of audio-visual speech enhancement (AVSE) by utilizing the benefits of visual speech information available with video recordings of lip movements in addition to audio. The rationale behind AVSE is that, unlike audio information, visual information (lip movements) is not corrupted by acoustic perturbations, and thus visual information can aid in speech enhancement, particularly in the presence of audio signals with low signal-to-noise ratios (SNRs).Although it has been demonstrated that combining visual and audio information is beneficial for various speech perception tasks, such as [4–6], AVSE has received far less attention than ASE. The origins of AVSE methods can be traced back to [7] and subsequent work, such as [8]-[13]. Not surprisingly, AVSE has recently been addressed in the context of deep neural networks (DNNs), and a number of interesting architectures and high-performance algorithms have been developed, for example, [14]-[18]. In this paper, we propose using variational auto-encoders to combine single-channel audio and single-camera visual information for speech enhancement (VAEs). This could be considered multimodal extension of [19]-[24] VAE-based methods that, to our knowledge, yield state-of-the-art ASE performance in an unsupervised learning setting We propose using conditional variational auto-encoders (CVAEs) to incorporate visual observations into the VAE speech enhancement framework [25]. We proceed in three steps, as in [20].

First, using synchronized clean audio-speech and visual-speech data, the parameters of the audio-visual CVAE (AV-CVAE) architecture are learned. As a result, an audio-visual speech prior model is produced. The training is completely unsupervised, as no speech signals mixed with various types of noise signals are required. This contrasts with supervised DNN methods, which must be trained in the presence of multiple noise types and noise levels to ensure generalization and good performance, as shown in [14–16], [26]. Second, the learned speech prior is used in conjunction with a mixture model and a nonnegative matrix factorization (NMF) noise variance model to infer both the gain and the NMF parameters, which model the time-varying loudness of the speech signal. Third, using the speech prior (VAE parameters) as well as the inferred gain and noise variance, the clean speech is reconstructed. The latter can be thought of as a probabilistic Wiener filter. Using the NTCDTIMIT dataset [27] and the GRID corpus [28], the learned VAE architecture and its variants, the gain- and noise-parameter inference algorithms, and the proposed speech reconstruction method are thoroughly tested and compared with a state-of-the-art method.

2. LITERATURE SURVEY:

Speech enhancement has been a highly researched topic for decades, and a comprehensive state of the art is beyond the scope of this paper. We review the literature on single-channel SE briefly before discussing the most significant work in AVSE.

Classical methods employ spectral subtraction [29] and Wiener filtering [30] in the short-time Fourier transform (STFT) domain to estimate noise and/or speech power spectral density (PSD). Another popular method family is the short-term spectral amplitude estimator [31], which was originally based on a local complex-valued Gaussian model of the speech STFT coefficients and was later extended to other density models [32], [33], and a log-spectral amplitude estimator [34], [35]. NMF, for example, [37]-[39], is a popular technique for modelling the PSD of speech signals [36]. SE has recently been addressed in the context of DNNs [40].

Supervised methods learn mappings between noisy-speech spectrograms and clean-speech spectrograms, which are then used to reconstruct a speech waveform [41-43]. Alternatively, the noisy input is mapped onto a time frequency (TF) mask before being applied to the input to remove noise while preserving as much speech information as possible [26], [44]-[46]. In order for these supervised learning methods to generalize well and produce cutting-edge results, the training data must be highly variable in terms of speakers and, more importantly, noise types and noise levels [42], [44]; in practice, this leads to time-consuming learning processes. Alternatively, generative (or unsupervised) DNNs do not use any kind of noise information for training, and as a result, they have excellent generalization capabilities. VAEs [47] provide an intriguing generative formulation. When combined with NMF, VAE-based methods produce cutting-edge SE performance [19]-[24] in an unsupervised learning setting. Speaker-dependent VAEs have also been used for speaker-dependent multi-microphone speech separation [48], [49], and de-reverberation [50].

The use of visual cues to supplement audio when it is noisy, ambiguous, or incomplete has been extensively researched in psychophysics [4]-[6]. Speech production necessitates simultaneous air circulation through the vocal tract as well as tongue and lip movements, implying that speech perception is multimodal. Several computational models, such as [9], [12], have been proposed to exploit the correlation between audio and visual information for speech perception. [8] Proposed a multi-layer perceptron architecture for mapping noisy speech linear prediction features concatenated with visual features onto clean speech linear prediction features. Wiener filters were then developed for denoising. Later, phoneme-specific Gaussian mixture regression and filter bank audio features were used to extend audio-visual Wiener filtering [51]. Other AVSE methods make use of noise-free visual information [10, 11], or twin hidden Markov models (HMMs) [13].

DNNs underpin cutting-edge supervised AVSE methods. The idea behind [14, 16] is to use visual information to predict a TF soft mask in the STFT domain, which is then applied to the audio input to remove noise. [16] Trains a video-to-speech architecture for each speaker in the dataset, resulting in a speaker-dependent AVSE method. [14]'s architecture is made up of a magnitude sub-network that accepts both visual and audio data as inputs and a phase sub-network that only accepts audio data as inputs. Both sub-networks are trained using clean speech as the ground truth. The magnitude sub-network then predicts a binary mask, which is applied to both the magnitude and phase spectrograms of the input signal, resulting in a filtered speech spectrogram. [17] And [15] have very similar architectures in that they are made up of two sub-networks, one for processing noisy speech and one for processing visual speech. The two encodings are then concatenated and processed to produce an improved speech spectrogram. The primary distinction between [17] and [15] is that the former predicts both enhanced visual and audio speech, whereas the latter only predicts audio speech. The concept of obtaining a binary mask for distinguishing speech from unknown noise was used in [18]: a hybrid DNN model integrates a stacked long short-term memory (LSTM) and convolutional LSTM for audio-visual (AV) mask estimation.

Generalization to unseen data is a critical issue in the supervised deep learning methods just mentioned. The two most serious issues are noise and speaker variability. To ensure generalization, training these methods requires noisy mixtures with a large number of noise types and speakers. In comparison, the proposed method is completely unsupervised: it is trained using VAEs and requires only clean audio and visual speech. A Monte Carlo expectation maximization (MCEM) algorithm is used to estimate the gain and noise variance during testing [52]. The learned parameters are then used to reconstruct clean speech from audio and visual inputs. The latter can be thought of as a probabilistic Wiener filter. This differs from the vast majority of supervised DNN-based AVSE methods, which predict a TF mask to be applied to the noisy input. Empirical validation using a widely used publicly available dataset and standard SE scores shows that our method outperforms the ASE method [20] as well as the state-of-the-art supervised AVSE method [15].

3. Proposed Method:

Speech processing typically entails a basic representation of a speech signal in a digital domain, which necessitates limiting the signal's band width, sampling it at a specific corresponding rate, and storing each sample with sufficient resolution. However, our primary focus in the field of speech processing is communication. Speech can be represented as a signal with message content or information. Speech signals are generated by our vocal cords and transmitted from speaker to listener via pressure waves propagated through air, resulting in continuous or analogue signals. This project combines audio and visual information to improve the quality of speech even further. The goal of enhancement is to improve intelligibility. The most important field of speech enhancement is the enhancement of speech that has been degraded by noise, also known as noise reduction. Mobile phones, VoIP, teleconferencing systems, speech recognition, and hearing aids are examples of applications.

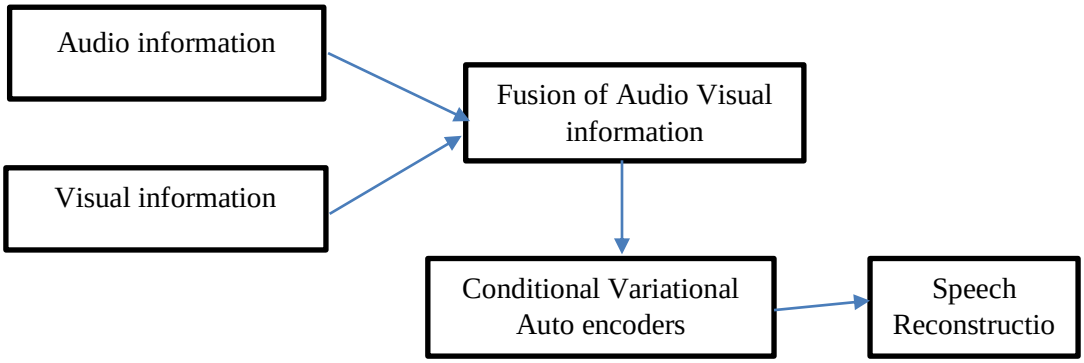

Figure 1 Block Diagram of proposed model

A.Input Dataset:

We used the NTCD-TIMIT dataset, which contains AV recordings of 56 Irish-accented English speakers saying 98 different sentences. So there are 5698 = 5488 videos with a length of about 5 seconds. The visual data consists of 30 FPS lip ROI videos. Each ROI frame is 6767 pixels in size. The speech signal is sampled at a rate of 16 kHz.

B.Variational Auto-encoder:

A variational auto-encoder, like a standard auto-encoder, is an architecture that includes both an encoder and a decoder and is trained to minimize the reconstruction error between the encoded-decoded data and the initial data. To introduce some regularization of the latent space, we modify the encoding-decoding process slightly: instead of encoding an input as a single point, we encode it as a distribution over the latent space.

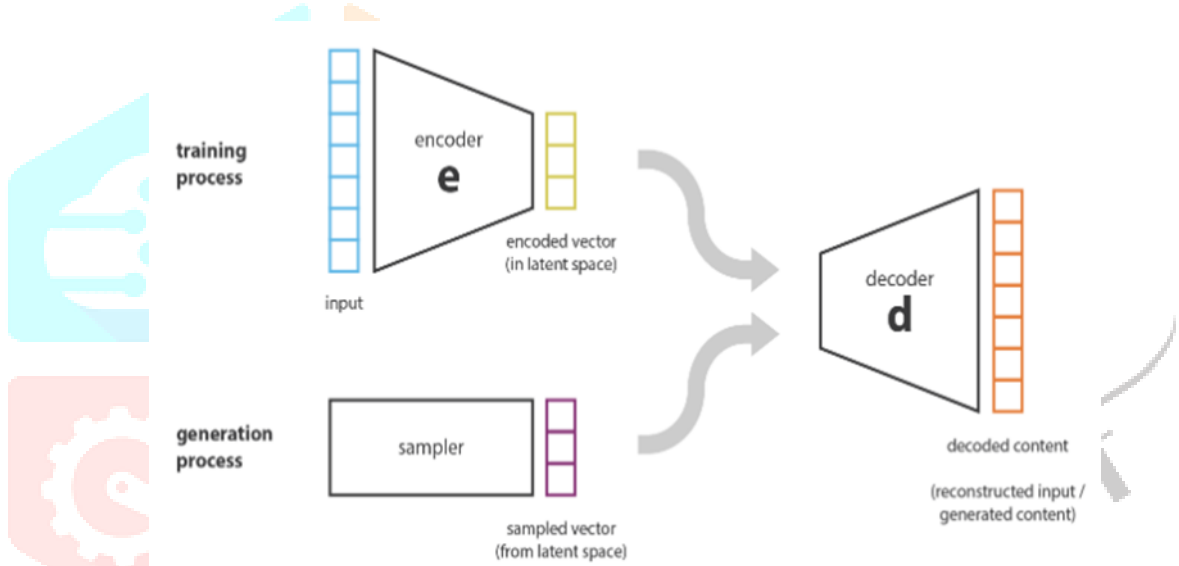

Figure 2: Block diagram of Auto Variational Encoder

C. Short Time Fourier Transform:

The Short-Time Fourier Transform (STFT) (or short-term Fourier transform) is a versatile audio signal processing tool. It defines a particularly useful class of time-frequency distributions for any signal, which specify complex amplitude versus time and frequency. We are mostly interested in fine-tuning the STFT parameters for the following applications:

a. Approximating the time-frequency analysis performed by the ear for spectral display purposes.

b. Model parameter measurement in a short-time spectrum.

D. Audio Variational Auto Encoder:

The deep generative speech model, as well as its parameter estimation procedure using VAEs, was first proposed. Let s fn represent the complex-valued speech STFT coefficient at frequency indexes f 0,....,F-1 and frame index 'n'. We have the following model at each Time Frequency bin, which will be referred to as audio VAE (A-VAE):

$$S\_fn\ |Z_n \sim N_c\left(0, \sigma_f(Z_n)\right) \quad \text{..... [1]}$$

$$Z\_n \sim N(0, I)$$

N(0,I) is a zero mean multivariate Gaussian distribution with an identity covariance matrix, and _c (0,) is a unilabiate complex proper Gaussian distribution with zero mean and variance. The parameters of these neural networks are denoted collectively as. This variance can be interpreted as a model for the short-term PSD of the speech signal. VAEs have the important property of providing an efficient method of learning the parameters of such generative models using ideas from variational inference. The parameters in the VAE framework are estimated by maximizing a lower bound of the log-likelihood, ln p(s;), known as the evidence lower bound (ELBO), which is defined as:

$$L(S, \theta, \varphi) = E\_(q(Z, S, \varphi))\ [lnp(S \mid Z; \theta] - D\_KL\ (q(Z|S; \varphi||p(Z)) \text{..... [2]}$$

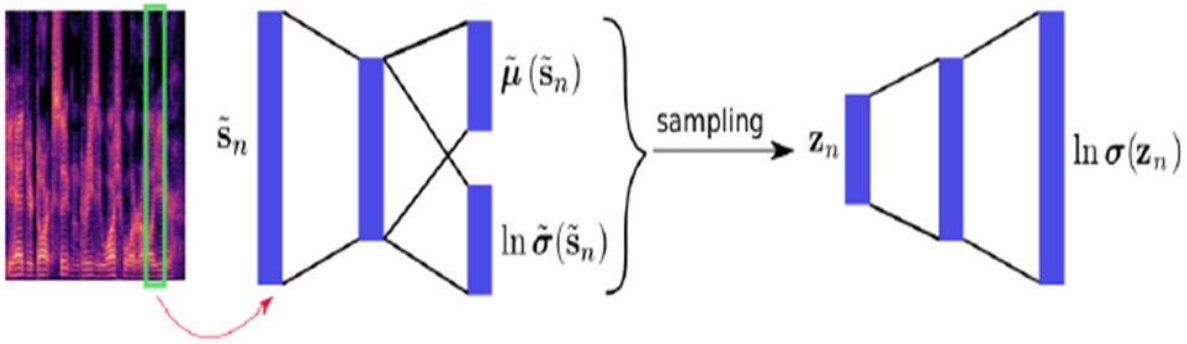

Figure 3: A-VAE Network

E. Visual Variational Auto Encoder (V-VAE):

In this paper, two VAE network variants for learning the speech prior from visual data are introduced: base visual VAE (V-VAE) and augmented V-VAE. Standard computer vision algorithms are used to extract a fixed-sized bounding-box from the image. Optionally, as part of a network specifically trained for the task of supervised audio-visual speech recognition, an additional pre-trained front-end network (dashed box) composed of a 3D convolution layer followed by a RESNET with 50 layers can be used. This second option is known as augmented V-VAE. The V-VAE and A-VAE in this project share the same CNN decoder architecture.

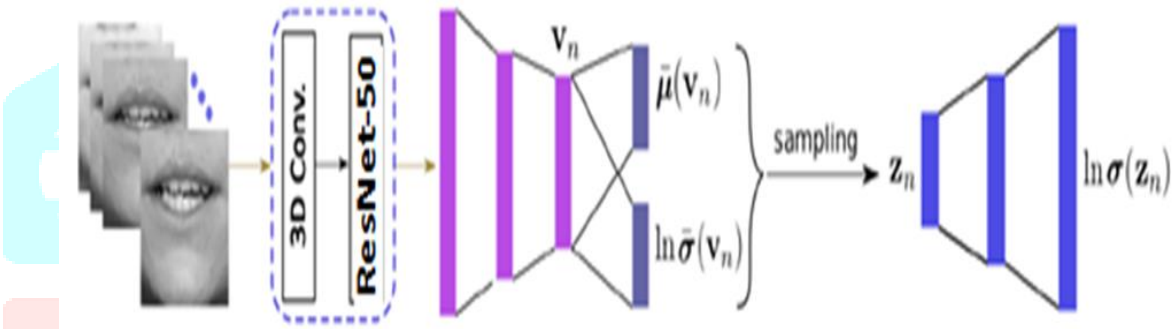

Figure .4 Two V-VAE Network Variant

F. Audio-Visual Variational Auto Encoders:

An audio-visual VAE model, that is, a model that combines audio and visual speech, is created. The reasoning behind this multimodal approach is that while audio data is frequently corrupted by noise, visual data is not. Without limiting the generality, it will be assumed that audio and visual data are synchronized, i.e. that each audio frame is accompanied by a video frame. We consider the CVAE framework to learn structured output representations in order to combine the above A-VAE and V-VAE formulations. During training, a CVAE is given data as well as associated class labels, allowing the network to learn a structured data distribution. At test, the trained network is given a class label and asked to generate samples from that class.

The observed visual speech is used to condition the clean audio speech, which is only available during training. However, because visual information is available both during training and testing, it serves as a deterministic prior on the desired clean audio speech. The variable z n is sampled from the approximate posterior modelled by the encoder and passed to the decoder during AV-CVAE training. During testing, however, only the decoder and prior networks are used, with the encoder being discarded.

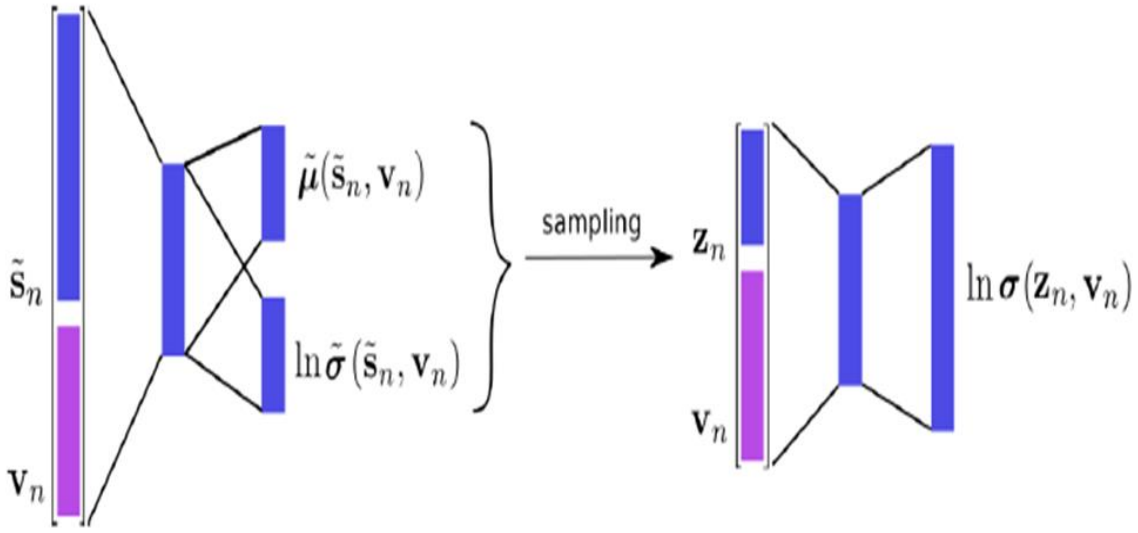

Figure 5. Proposed AC-CVAE Model

G.AV-CVAE For Speech Enhancement:

This section describes the proposed AV-CVAE speech model-based speech enhancement algorithm. It closely resembles the algorithm proposed for audio-only speech enhancement with VAE. The unsupervised noise model is presented first, followed by the mixture model and the proposed algorithm for estimating the noise model's parameters. Finally, the procedure for clean-speech inference is described. Through this section, v={v_n }_(n=0)^(N-1), s={s_n }_(n=0)^(N-1), z={z_n }_(n=0)^(N-1) denote the test sets of visual features, clean speech STFT features and latent vectors, respectively. These variables are linked to an N-frame noisy-speech test sequence.

H. Speech Reconstruction:

The speech is reconstructed without the noise using a scaled version of the STFT coefficients. The last step is to calculate these coefficients based on their posterior mean. This estimation is a "probabilistic" version of Wiener filtering, with the filter averaging over the posterior distribution of the latent variables.

4. Results and Discussion:

The database is taken from the NTCD-TIMIT dataset as well as the GRID corpus. We are using 3parameters PESQ,SDR,STOI.Here are 6 different noise types such as LR noise, street noise, car noise, white noise, cafe noise and babble noise are used. The clean speech file is corrupted with above mentioned noises at different SNR levels, i.e., -15dB, -10dB, -5dB, 0dB, 5dB, 10 dB and 15 dB, to build a multi-condition training set of pairs of noisy and clean speech signals. All these noise types are taken for the training of DNN.

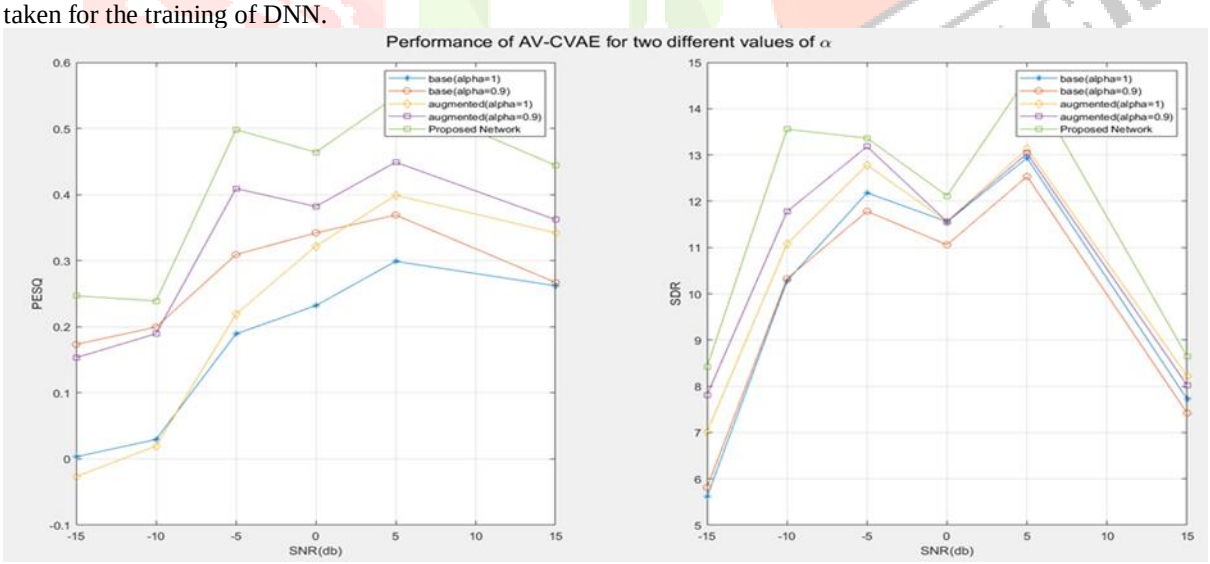

Figure 6. Performance of AV-CVAE for two different values of alpha

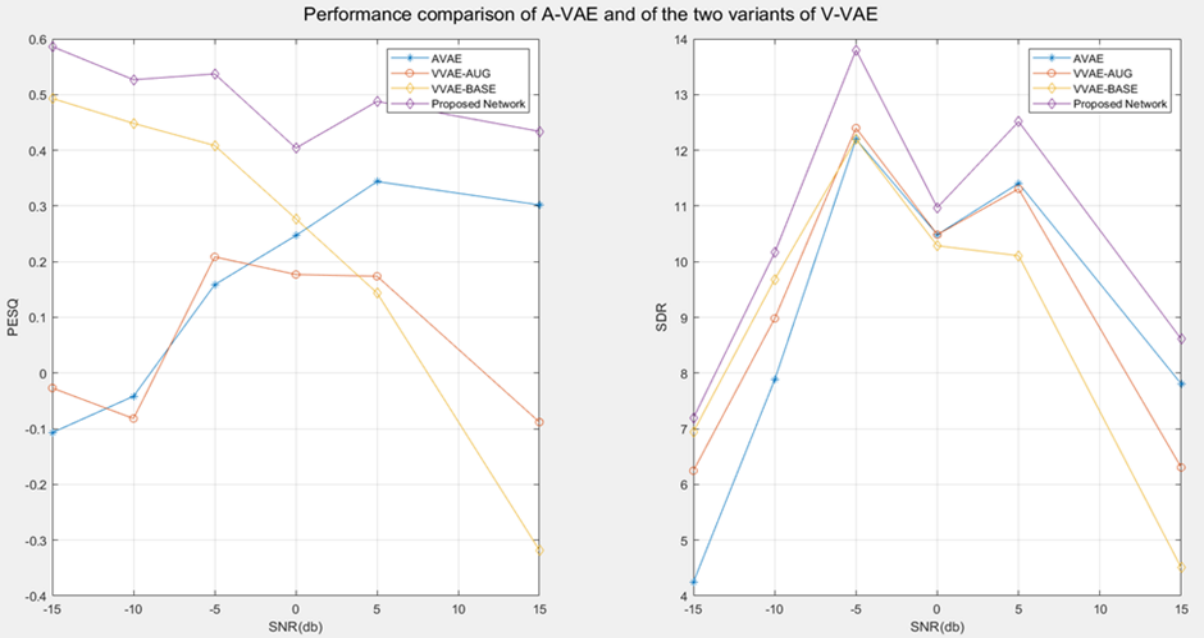

Figure 7. Performance of V-VAE and the two variants of A-VAE.

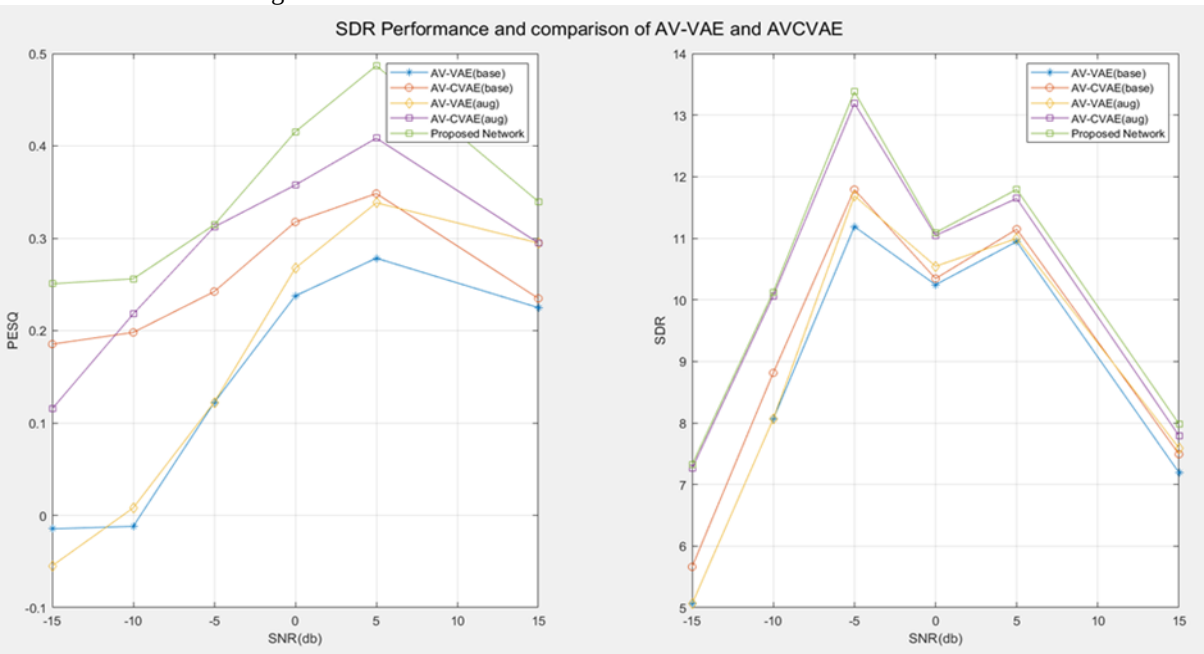

Figure 8. SDR Performance comparison of AV-VAE and AC-CVAE

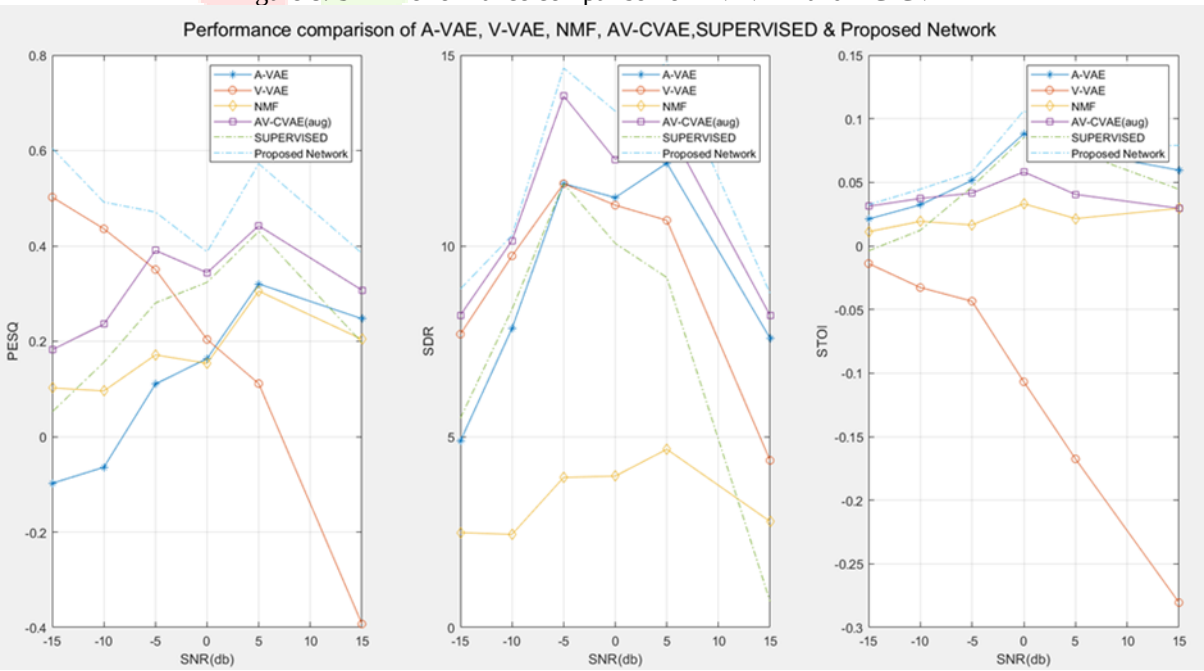

Figure 9. Performance comparison of A-VAE, V-VAE, Supervised, and proposed network.

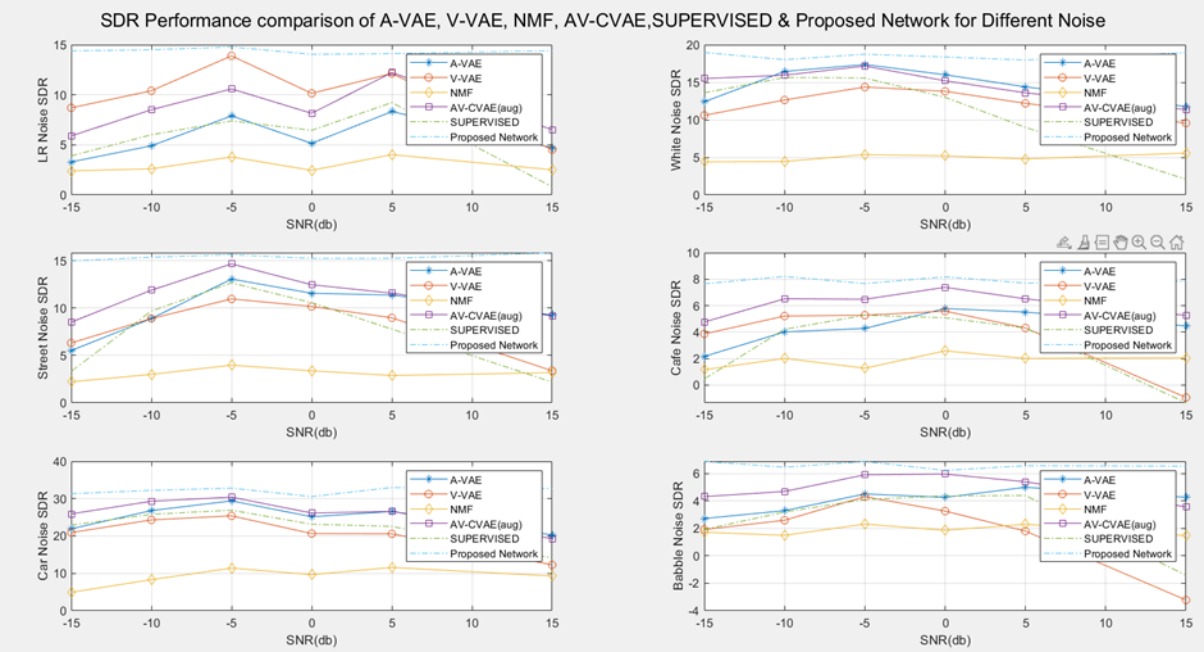

Figure 10. SDR Performance comparison of A-VAE, V-VAE, Supervised, and proposed network.

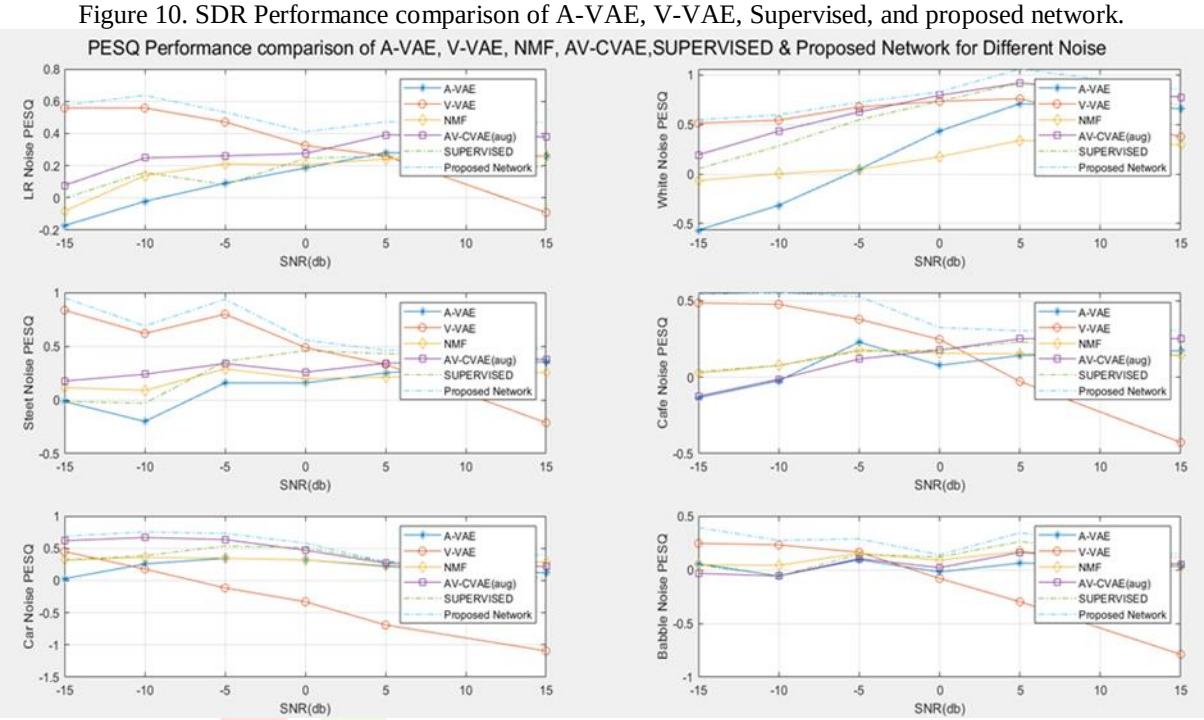

Figure 11. PSEQ Performance comparison of A-VAE, V-VAE, Supervised, and proposed network.

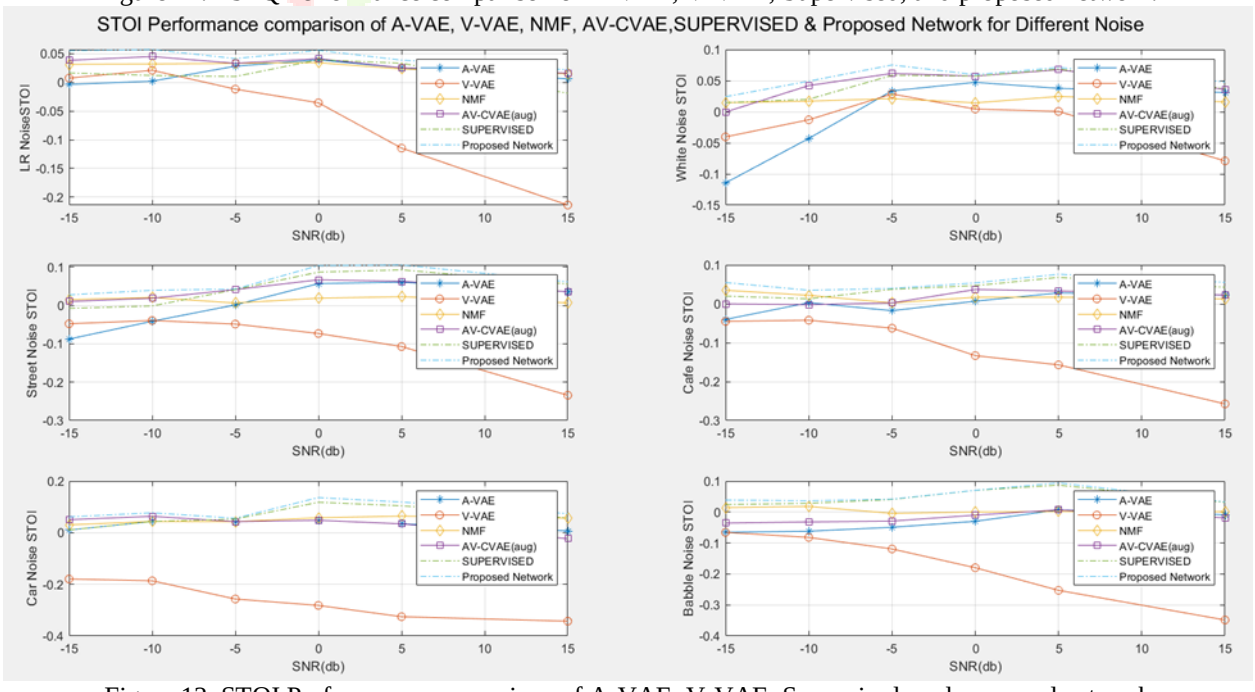

Figure 12. STOI Performance comparison of A-VAE, V-VAE, Supervised, and proposed network

5. Conclusion:

To model speech prior for speech enhancement, we proposed an audio-visual conditional VAE. We detailed several VAE architecture variants and provided information on how to estimate their parameters. This audiovisual speech prior model was combined with an audio mixture model and a noise variance model based on NMF. We developed an MCEM algorithm that infers the time-varying loudness of the speech input as well as the noise variance parameters. Finally, a probabilistic Wiener filter is used to reconstruct speech. Extensive empirical testing validates the efficacy of the proposed methodology for fusing audio and visual inputs for speech enhancement. The visual modality, specifically video frames of moving lips, has been shown to improve performance, especially when the audio modality is heavily corrupted with noise. Future work will investigate computationally efficient inference algorithms as well as the use of recurrent and convolutional layers to model temporal dependencies between audio and visual frames.

References:

[1] Jae Soo Lim, Speech enhancement, Prentice-Hall Englewood Cliffs, NJ, 1983.
[2] Jacob Benesty, Shoji Makino, and Jingdong Chen, Speech enhancement, Springer Science & Business Media, 2006.
[3] Philipos C. Loizou, Speech enhancement: theory and practice, CRC press, 2007.
[4] William Sumby and Irwin Pollack, “Visual contribution to speech intelligibility in noise,” The Journal of the Acoustical Society of America, vol. 26, no. 2, pp. 212–215, 1954.
[5] Norman Erber, “Auditory-visual perception of speech,” Journal of Speech and Hearing Disorders, vol. 40, no. 4, pp. 481–492, 1975.
[6] Alison MacLeod and Quentin Summerfield, “Quantifying the contribution of vision to speech perception in noise,” British Journal of Audiology, vol. 21, no. 2, pp. 131–141, 1987.
[7] Laurent Girin, Gang Feng, and Jean-Luc Schwartz, “Noisy speech enhancement with filters estimated from the speaker’s lips,” in Proc. European Conference on Speech Communication and Technology (EUROSPEECH), Madrid, Spain, 1995, pp. 1559–1562.
[8] Laurent Girin, Jean-Luc Schwartz, and Gang Feng, “Audio-visual enhancement of speech in noise,” The Journal of the Acoustical Society of America, vol. 109, no. 6, pp. 3007–3020, 2001.
[9] John W. Fisher III, Trevor Darrell, William T. Freeman, and Paul A. Viola, “Learning joint statistical models for audio-visual fusion and segregation,” in Proc. Advances in Neural Information Processing Systems (NIPS), 2001, pp. 772–778.
[10] Sabine Deligne, Gerasimos Potamianos, and Chalapathy Neti, “Audiovisual speech enhancement with AVCDCN (audio-visual codebook dependent cepstral normalization),” in Proc. IEEE International Workshop on Sensor Array and Multichannel Signal Processing, 2002, pp. 68–71.
[11] Roland Goecke, Gerasimos Potamianos, and Chalapathy Neti, “Noisy audio feature enhancement using audio-visual speech data,” in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2002, pp. II–2025–2028.
[12] John R. Hershey and Michael Casey, “Audio-visual sound separation via hidden Markov models,” in Proc. Advances in Neural Information Processing Systems (NIPS), 2002, pp. 1173–1180.
[13] Ahmed Hussen Abdelaziz, Steffen Zeiler, and Dorothea Kolossa, “Twin- HMM-based audio-visual speech enhancement,” in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013, pp. 3726–3730.
[14] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman, “The conversation: Deep audio-visual speech enhancement,” in Proc. Conference of the International Speech Communication Association (INTERSPEECH), 2018, pp. 3244–3248.