Audio-Visual Speech Enhancement using Deep Neural Networks

Jen-Cheng Hou[*], Syu-Siang Wang[*†], Ying-Hui Lai[§], Jen-Chun Lin[‡], Yu Tsao[*], Hsiu-Wen Chang[₪],
and Hsin-Min Wang[‡]

[*]Research Center for Information Technology Innovation, Academia Sinica, Taiwan, E-mail: coolkiu@citi.sinica.edu.tw
[†]Graduate Institute of Communication Engineering, National Taiwan University, Taiwan, E-mail: d02942007@ntu.edu.tw
[§]Department of Electrical Engineering, Yuan Ze University, Taiwan, E-mail: yhlai@ee.yzu.edu.tw
[₪]Department of Audiology and Speech language pathology, Mackay Medical College, Taiwan, E-mail: hsiuwen@mmc.edu.tw
[‡]Institute of Information Science, Academia Sinica, Taiwan, E-mail: whm@iis.sinica.edu.tw

Abstract— **This paper proposes a novel framework that integrates audio and visual information for speech enhancement. Most speech enhancement approaches consider audio features only to design filters or transfer functions to convert noisy speech signals to clean ones. Visual data, which provide useful complementary information to audio data, have been integrated with audio data in many speech-related approaches to attain more effective speech processing performance. This paper presents our investigation into the use of the visual features of the motion of lips as additional visual information to improve the speech enhancement capability of deep neural network (DNN) speech enhancement performance. The experimental results show that the performance of DNN with audio-visual inputs exceeds that of DNN with audio inputs only in four standardized objective evaluations, thereby confirming the effectiveness of the inclusion of visual information into an audio-only speech enhancement framework.**

I. INTRODUCTION

The primary goal of speech enhancement is to reduce the noise components of noisy speech signals to reconstruct clean speech signals, accordingly increasing the signal-to-noise ratio (SNR) and the quality of noise-corrupted speech. In a wide range of speech-related applications, such as automatic speech recognition (ASR) [1-3], speaker recognition [4,5] speech coding [6,7], and hearing aids [8,9], speech enhancement processing serves as a key component. In the past decades, numerous and diverse speech enhancement methods have been proposed and proven to provide satisfactory performance. One of these well-known approaches named spectral restoration estimates a gain function (based on the statistics of noise and speech components) to suppress noise components in the frequency domain to obtain a clean speech spectrum from the noisy speech input [10-15]. Another successful class of noise reduction (NR) approaches is the subspace-based methods, which adopt transformations to obtain enhanced speech given noisy speech. These transformations are estimated by minimizing the speech distortion with the constraint of a predetermined level of residual noise [16-18].

More recently, approaches based on machine learning have attracted considerable attention in the speech enhancement research field. Among these approaches, the speech enhancement method based on non-negative matrix factorization (NMF) and its extensions have been extensively investigated [19-22]. The NMF-based methods prepare the spectral bases for clean speech and noise using the corresponding training samples. Given noisy speech, the prepared bases are used to extract the clean speech portion. Another notable class of speech enhancement methods is based on deep learning. Generally, speech enhancement methods based on deep learning compute a mapping function, which aims to reconstruct clean signals from noisy input signals. Effective examples include deep neural networks (DNNs) [23-25], recurrent neural networks [26-28], convolutional neural networks [29,30], and deep denoising autoencoder (DDAE) [31,32] models. When a sufficient amount of training data becomes available, these methods based on deep learning have been proven to provide comparable or even more effective noise reduction capability than traditional speech enhancement methods. In this study, we focus our attention on and intend to enhance its capability.

In addition to speech signals, visual information carries important information in human-human or human-machine interaction. A study of the McGurk effect [33] indicated that the motion of the mouth or lips can play an important role in speech processing. Accordingly, audio-visual multimodality has been adopted in numerous speech-processing fields [34-38]. These results showed that the visual modality indeed enhances the performance of speech processing compared to the counterpart that only uses audio modality. In this study, we propose to integrate the audio and visual information to form joint input features for the DNN model for speech enhancement. The experimental results show that the fused audio-visual input feature vector is capable of outperforming an audio-only input feature vector in terms of several standard evaluation metrics, including hearing-aid speech quality index (HASQI) [39], hearing-aid speech perception index (HASPI) [40], speech distortion index (SDI) [41] and segmental signal-to-noise ratio improvement (SSNRI) [10], confirming the effectiveness of incorporating the visual information into the DNN speech enhancement framework.

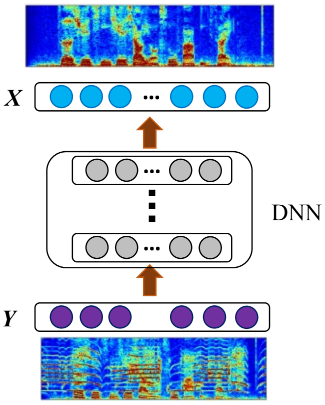

Fig. 1 Structure of the DNN-based speech enhancement approach.

The structure of this paper is organized as follows: Section II introduces the DNN model and describes the incorporation of the joint audio-visual features in DNN for speech enhancement. Section III depicts the feature extraction procedure in both the audio and visual channels. Section IV shows our experimental setup and results, and Section V provides the concluding remarks of this study.

II. Audio-Visual DNN for Speech Enhancement

A. *The DNN Model for Speech Enhancement*

This section reviews the conventional DNN-based speech enhancement system, in which only audio features are considered. Fig. 1 shows the structure of the DNN-based speech enhancement approach. The DNN-based speech enhancement procedure can be divided into training and testing phases. In the training phase, a set of noisy-clean speech pairs is prepared. The noisy-clean speech signals are converted into the frequency domain by applying the short time Fourier transform first, and then taking the logarithmic to the noisy-clean magnitude spectra. Finally, Mel-filter banks are performed to the noisy-clean log-spectra to form the noisy ($\boldsymbol{Y}$) and clean ($\boldsymbol{X}$) Mel-filter-bank features, respectively, for the input and output feature vectors of the DNN model. For the m-th frame, the input vector contains the Mel-filter bank features of the noisy spectrum: $\widetilde{\boldsymbol{Y}}_m=[\boldsymbol{Y}_{1,m-\tau}\cdots\ \boldsymbol{Y}_{l,m}\cdots\ \boldsymbol{Y}_{L,m+\tau}]'$, where τ is the length of the context window and is setting to one in this work. Then, the output vector is $\widetilde{\boldsymbol{X}}_m=[\boldsymbol{X}_{1,m-\tau}\cdots\ \boldsymbol{X}_{l,m}\cdots\ \boldsymbol{X}_{L,m+\tau}]'$, which is obtained from the clean speech, where $\boldsymbol{Y}_{l,m}$ and $\boldsymbol{X}_{l,m}$ are the Mel-filter-bank features of the noisy and clean spectra, respectively, at the l-th frequency bin at the m-th frame. For a DNN model with J hidden layers, we have

$$\begin{gathered} h^1(\widetilde{\boldsymbol{Y}}_m)=\sigma(\boldsymbol{W}^1\widetilde{\boldsymbol{Y}}_m+\boldsymbol{b}^1),\\ \vdots\\ h^J(\widetilde{\boldsymbol{Y}}_m)=\sigma(\boldsymbol{W}^{J-1}h^{J-1}(\widetilde{\boldsymbol{Y}}_m)+\boldsymbol{b}^{J-1}),\\ \widehat{\boldsymbol{X}}_m=\boldsymbol{W}^J h^J(\boldsymbol{Y}_m)+\boldsymbol{b}^J, \end{gathered} \tag{1}$$

where $\{\boldsymbol{W}^1\ldots\boldsymbol{W}\}$ denote the weighting matrices, $\{\boldsymbol{b}^1\ldots\boldsymbol{b}^J\}$ are the bias vectors, and $\widehat{\boldsymbol{X}}_m$ is the vector containing the Mel-filter-bank features of restored speech corresponding to the noisy counterpart $\widetilde{\boldsymbol{Y}}_m$. The nonlinear function $\sigma(.)$ of a hidden neuron is performed by logistic function as

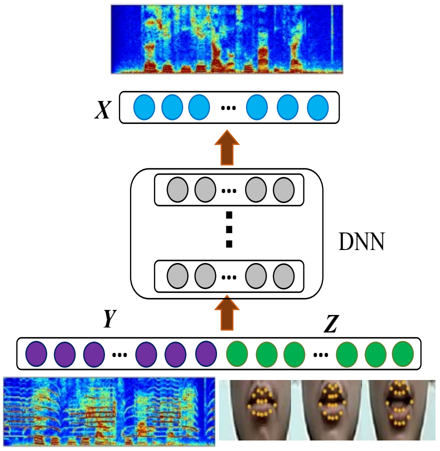

Fig. 2 Training audio-visual DNN neurons with pairs of noisy speech features concatenated with visual features and corresponding clean speech features.

$$\sigma(t)=1/(1+\exp(-t)). \tag{2}$$

The parameters are determined by optimizing the following objective function:

$$\begin{gathered} \Lambda^*=\arg\min_{\theta}\left(F(\Lambda)+\eta^1\|\boldsymbol{W}^1\|_F^2+\ldots+\eta^L\|\boldsymbol{W}^L\|_F^2\right),\\ F(\Lambda)=\frac{1}{M}\sum_{m=1}^{M}\|\widetilde{\boldsymbol{X}}_m-\widehat{\boldsymbol{X}}_m\|_2^2, \end{gathered} \tag{3}$$

where $\Lambda=\{\boldsymbol{W}^1\ldots\boldsymbol{W}^J;\boldsymbol{b}^1\ldots\boldsymbol{b}^J\}$ is the parameter set of the DNN model, and M is the total number of training samples. In addition, $\{\eta^1\ldots\eta^L\}$ are the regularization terms.

In the testing phase, the Mel-filter-bank features of noisy speech signals are input into the trained DNN model to obtain the Mel-filter-bank features of enhanced speech signals as the output. Similar to spectral restoration approaches, the phases of the noisy speech are borrowed as the phases for the enhanced speech. The DNN-enhanced Mel-filter-bank features and the phase information can then be used to synthesize the enhanced speech.

B. *Audio-Visual DNN for Speech Enhancement*

The proposed audio-visual DNN for speech enhancement is shown in Fig 2. The input layer is now composed by concatenating noisy speech and visual feature vectors. Thus, for the m-th frame, we have the input feature vector as: $\boldsymbol{U}_m=[\widetilde{\boldsymbol{Y}}_m{}'\ \boldsymbol{Z}_m{}']'$, where $\boldsymbol{Z}_m$ denotes the visual features. Meanwhile, the output vector is $\widetilde{\boldsymbol{X}}_m$, which is the same as that used in the DNN model with only audio features. The same training procedures as shown in Eqs. (1)-(3) are carried out to train the parameters of the audio-visual DNN.

In the testing phase, the audio-visual features are first extracted and then input into the trained DNN model to obtain the Mel-filter-bank features of enhanced speech signals as the output. The enhanced speech can be synthesized by borrowing the phases of the noisy speech, same as in the DNN model.

III. Feature Extraction

In this section, we provide details of the dataset and the feature extraction processes for both audio and visual channels.

A. *Datasets*

The prepared dataset contains audio-visual recordings of 40 utterances of Mandarin sentences by a native male speaker. The length of each utterance is around 3-4 seconds. The recordings were carried out in a quiet room with sufficient light. The audio signals were recorded at a sampling rate of 48 kHz while variable video frame rates were used for recording the video part with 320 × 240 resolutions. The feature extraction procedures were simplified by resampling the recordings into fixed 8 kHz and 25 frame per second (fps) in the audio and visual channels, respectively. One-fourth (10 utterances) of the dataset was used for testing with the remaining 30 utterances used as the training set.

B. *Audio Feature Extraction*

In the audio channel, the speech signals were first processed into a sequence of frames. Each frame was 32 milliseconds long, and the sliding frame rate was 62.5%, which was 20 milliseconds. For each speech frame, 256-point fast Fourier transform was applied to convert the signal from time to frequency domains to form the 256-point complex spectra. Since the first half of the frequency range (from zero to the Nyquist frequency) is sufficient to represent the spectrum information, the magnitude spectra with 129 frequency bins were calculated, and then the logarithmic process was performed. Finally, 40 Mel-scaled band-pass filters were designed to filter the 129-element log-magnitude vector into 40 dimensional Mel-filter-bank features, which has been proved to provide satisfactory performance of a DNN-based enhancement systems [31,32,42].

C. *Visual Feature Extraction*

For the visual channel, we first performed face region detection using the Viola-Jones algorithm [43]. Next, the Gauss-Newton deformable part model (GN-DPM) was adopted for mouth shape extraction, as described in [44]. Fig. 3 (a) and (b) show the results of face detection and mouth-shape extraction, respectively. In this study, we adopted 18 points to represent the mouth region. The value of the distances between any two points of the 18 mouth-shape points (in both the x- and y-direction) were computed, and thus two sets of 153-dimensional vectors were generated. These vectors were then normalized by subtracting the mean and dividing by the standard deviation. We reduced the dimensionality while retaining lip motion information throughout an utterance by computing and sorting the variance of each element of the 153-dimensional vectors in an utterance. The top 20 elements with the highest variance values were selected to form a compact feature vectors. As a consequence, a 40-dimensional vector was used as the final visual feature vector for each image frame. Additionally, an upsampling process was carried out by interpolation from 25 fps to 50 fps to match the frame rate of the audio feature vector.

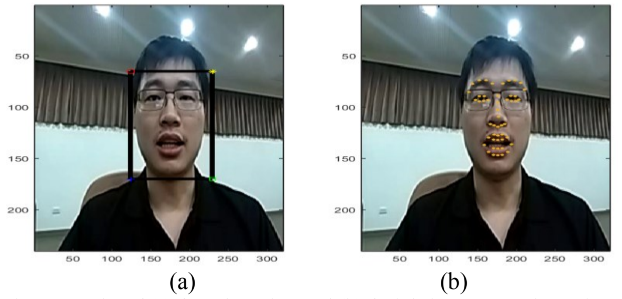

Fig. 3 Results of (a) face detection and (b) facial shape extraction using the Viola-Jones algorithm and the GN-DPM algorithm, respectively.

IV. Experimental Results

A. *Experimental Setup*

In this section, we evaluate the proposed audio-visual DNN on a speech enhancement task. As presented in Section II, the utterances recorded with both audio and visual data were used to form training and testing sets. We used a segment of baby crying sounds (around 5 minutes) as the source of noise, which was further divided into 3-minute and 2-minute segments. The 3-minute part was combined with the training data to produce six different SNR conditions (-10, -6, -2, 2, 6, and 10dB) to form the training set, and the remaining 2-minute part was combined with the testing data at -5, 0, and 5dB to form the test set. In the following sections, we first visually compare the difference between DNN with audio-only features and audio-visual features, which are termed A-DNN and AV-DNN, respectively, for simplicity. For A-DNN, the model structure is :{120, 100, 100, 100, 120}. For AV-DNN, the model structure is {240, 100, 100, 100, 120}. Next, four sets of objective evaluations are presented to validate the effectiveness of the proposed AV-DNN. The results of noisy speech are also presented for comparison, where the same signal-processing steps as presented in Section III are carried out for A-DNN and AV-DNN to ensure a fair comparison.

B. *Spectrogram Comparison*

Fig. 4 (a), (b), (c), and (d) demonstrate the spectrograms of clean speech, noisy speech (at 0dB SNR), speech enhanced by A-DNN, and speech enhanced by AV-DNN, respectively. By comparing (b), (c), and (d), we note that both A-DNN and AV-DNN clearly remove noise components (the baby crying sounds in the background), confirming the effectiveness of both A-DNN and AV-DNN. In addition, when comparing (c) with (d) in the figure, more detailed sound structures in the magnitude spectrograms are revealed by AV-DNN than those generated from A-DNN. In the next section, we present more objective evolutions to compare these two models.

C. *Objective Results*

We first compare A-DNN and AV-DNN using the SSNRI and SDI evaluation metrics, which are two important indicators to judge the performance of a speech enhancement method. SSNRI calculates the improvement in the SNR attained by the enhanced speech signals over the noisy speech signals in decibel, as shown in Eq. (4).

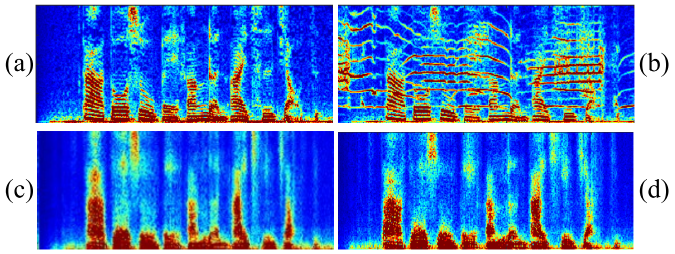

Fig. 4 Spectrograms: (a) clean speech, (b) noisy speech with 0dB SNR, (c) enhanced speech by A-DNN, and (d) enhanced speech by AV-DNN.

$$\text{SSNRI}=10\log_{10}\left(\frac{\text{P}_{\text{Enhanced}}}{\text{P}_{\text{Noisy}}}\right), \qquad (4)$$

where $\text{P}_{\text{Enhanced}}$ and P_{Noisy} denote power of enhanced and noisy speech respectively. In addition, SDI measures the distortion of an enhanced speech with respect to the referenced clean one. A lower SDI indicating a smaller difference between clean and enhanced speech signals. Eq. (5) gives the definition of SDI, where $\text{S}_{\text{Enhanced}}[n]$ and $\text{S}_{\text{Clean}}[n]$ represent enhanced and clean speech samples, respectively, with time index n. The results of the SSNRI and SDI calculations are presented in Fig. 5 (a) and (b), respectively.

$$\text{SDI}=\frac{\sum_n (\text{S}_{\text{Enhanced}}[n]\text{-}\text{S}_{\text{Clean}}[n])^2}{\sum_n \text{S}^2_{\text{Clean}}[n]}. \qquad (5)$$

From Fig. 5 (a), we first note that both A-DNN and AV-DNN significantly improve the SNR over the noisy inputs, especially for the -5dB SNR condition. In addition, the figure indicates that both A-DNN and AV-DNN provide considerable SDI reductions in the -5 dB and 0 dB SNRs. However, neither A-DNN nor AV-DNN were able to reduce SDI values on 5dB SNR condition. The reason for that is probably the insufficient training data available in this study (30 training utterances). Furthermore, AV-DNN achieves lower SDI than A-DNN consistently over all testing conditions.

Next, we present the HASQI and HASPI scores to compare the enhancement performance in terms of speech quality and intelligibility, respectively. The score range for both indices are {0 to 1} that the higher scores represent the better quality or perception of a sound. Evaluated results are listed in Table 1 and 2, respectively. From both tables, we again note that both A-DNN and AV-DNN clearly enhance speech signals, thereby proving improved quality and intelligibility. Moreover, AV-DNN achieves higher scores than A-DNN consistently, confirming that the effectiveness of incorporating audio and visual features in DNN for speech enhancement.

TABLE 1. AVERAGE HASQI SCORES OF A-DNN AND AV-DNN, WHERE THE SCORES OF NOISY SPEECH ARE ALSO PRESENTED.

SNR	-5dB	0dB	5dB
Noisy	0.1587	0.2431	0.3253
A-DNN	0.2925	0.3391	0.3736
AV-DNN	**0.3001**	**0.3505**	**0.4001**

TABLE 2. AVERAGE HASPI SCORES OF A-DNN AND AV-DNN, WHERE THE SCORES OF NOISY SPEECH ARE ALSO PRESENTED.

SNR	-5dB	0dB	5dB
Noisy	0.8827	0.9725	0.9900
A-DNN	0.9861	0.9937	0.9961
AV-DNN	**0.9893**	**0.9949**	**0.9973**

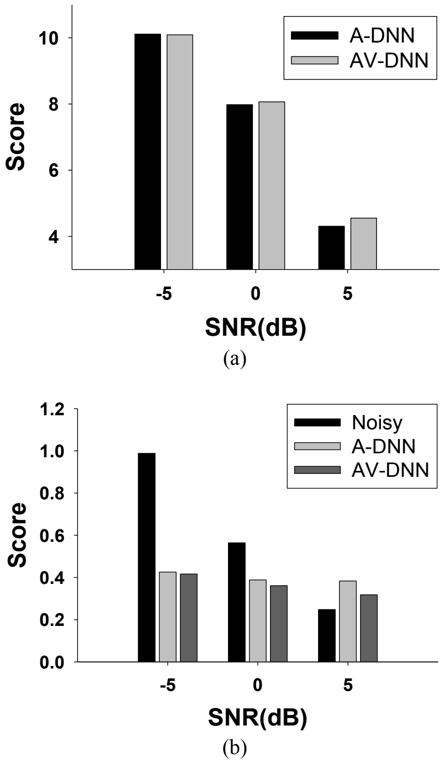

Fig. 5 Average scores of (a) SSNRI and (b) SDI. A-DNN and AV-DNN denote DNN using audio and audio-visual features, respectively.

V. CONCLUSIONS

In this paper, we have proposed a novel DNN-based system that incorporates audio and visual information for speech enhancement. Since the audio and visual information were used to form a new feature input, the proposed method can be considered as a feature-level-fusion method in multimodality research. The experimental results show that audio-visual DNN outperforms audio-only DNN. For our future work, we plan to investigate approaches based on model-level fusion or decision-level fusion and compare their performance. Meanwhile, we aim to collect additional audio-visual recordings with the view of constructing a multi-speaker audio-visual DNN approach to compare its speech enhancement performance.

REFERENCES

[1] J. Li, L. Deng, R. Haeb-Umbach, and Y. Gong, Robust Automatic Speech Recognition: A Bridge to Practical Applications. Academic Press, 2015.

[2] T. Virtanen, R. Singh, and B. Raj, Techniques for noise robustness in automatic speech recognition, John Wiley & Sons, 2012.

[3] B. Li, Y. Tsao, and K. C. Sim, "An investigation of spectral restoration algorithms for deep neural networks based noise robust speech recognition.", in *Proc. INTERSPEECH*, 2013, pp. 3002-3006.

[4] A. El-Solh, A. Cuhadar, and R. A. Goubran. "Evaluation of speech enhancement techniques for speaker identification in noisy environments. ", *in Proc. ISMW,* 2007, pp. 235-239.

[5] J.Ortega-Garcia and J.Gonzalez-Rodriguez, "Overview of speech enhancement techniques for automatic speaker recognition," in *Proc. ICSLP,* 1996, pp. 929-932.

[6] J. Li, L. Yang, Y. Hu, M. Akagi, P.C. Loizou, J. Zhang, and Y. Yan, "Comparative intelligibility investigation of single-channel noise reduction algorithms for Chinese, Japanese and English.", *Journal of the Acoustical Society of America.,* vol. 129, no. 5, pp. 3291–3301, 2011.

[7] J. Li, S. Sakamoto, S. Hongo, M. Akagi, Y. Suzuki, "Two-stage binaural speech enhancement with Wiener filter for high-quality speech communication," *Speech Communication*, vol. 53, no. 5, pp. 677–689, 2011.

[8] T. Venema, *Compression for Clinicians*, *Chapter 7*, Thomson Delmar Learning, 2006.

[9] H. Levitt, "Noise reduction in hearing aids: an overview.", *J. Rehab. Res. Dev.*, vol. 38, no. 1, pp.111–121, 2001.

[10] J. Chen, Fundamentals of Noise Reduction in Spring Handbook of Speech Processing, Chapter 43, Springer, 2008.

[11] P. Scalart and J. V. Filho, "Speech enhancement based on a priori signal to noise estimation," in *Proc. ICASSP,* 1996, pp. 629-632

[12] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, pp. 1109-1121, 1984.

[13] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Transactions on Speech and Audio Processing*, vol. 13, pp. 845-856, 2005.

[14] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 33, pp. 443-445, 1984.

[15] Y. Tsao and Y.-H. Lai, "Generalized maximum a posteriori spectral amplitude estimation for speech enhancement," *Speech Communication*, vol. 76, 2015.

[16] Y. Ephraim, and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, pp. 251-266, 1995.

[17] A. Rezayee, and S. Gazor, "An adaptive KLT approach for speech enhancement," *IEEE Transactions on Speech and Audio Processing,* vol. 9, no. 2, pp. 87-95, 2001.

[18] Y. Hu, and P.C. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 4, pp. 334-341, 2003.

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

[20] A. Ozerov and C. Fevotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation*,"* *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, 2010.

[21] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 21, no. 10, pp. 2140–2151, 2013.

[22] J.-T. Chien and P.-K. Yang, "Bayesian factorization and learning for monaural source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 1, pp. 185–195, 2016.

[23] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, pp. 65-68, 2014.

[24] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, pp. 7-19, 2015.

[25] D. Liu, P. Smaragdis, and M. Kim, "Experiments on deep learning for speech denoising," in *Proc. INTERSPEECH*, 2014, pp. 2685-2689.

[26] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. LeRoux, J. R. Hershey and B. Schuller, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," *Latent Variable Analysis and Signal Separation*, vol. 9237, pp. 91-99, 2015.

[27] K. Osako, R. Singh and B. Raj. "Complex recurrent neural networks for denoising speech signals," in *Applications of Signal Processing to Audio and Acoustics, 2015 IEEE Workshop on. IEEE*, 2015, pp. 1-5.

[28] F. Weninger, F. Eyben, and Bjorn Schuller, "Single-channel speech separation with memory-enhanced recurrent neural networks," in *Proc. ICASSP*, 2014, pp. 3709-3713.

[29] S.-W. Fu, Y. Tsao, and X. Lu, "SNR-Aware Convolutional Neural Network Modeling for Speech Enhancement," in *Proc. INTERSPEECH* , 2016.

[30] M. Zhao, D. Wang, Z. Zhang, and X. Zhang, "Music removal by denoising autoencoder in speech recognition," in *APSIPA*, 2015.

[31] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Proc. INTERSPEECH*, 2013, pp. 436-440.

[32] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Ensemble modeling of denoising autoencoder for speech spectrum restoration," in *Proc. INTERSPEECH*, 2014, pp. 885-889.

[33] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–748, 1976.

[34] D. G. Stork and M. E. Hennecke, *Speechreading by Humans and Machines*, Springer, 1996.

[35] G. Potamianos, C. Neti, G. Gravier, A. Garg, and Andrew W, "Recent Advances in the Automatic Recognition of Audio-Visual Speech," *Proceedings of IEEE*, vol. 91, no. 9, 2003.

[36] D. Kolossa, S. Zeiler, A. Vorwerk, and R. Orglmeister, "Audiovisual Speech Recognition with Missing or Unreliable Data," in *Proc. AVSP*, 2009, pp. 117-122.

[37] A. V. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy, "Dynamic Bayesian Networks for Audio-Visual Speech Recognition," *EURASIP Journal on Applied Signal Processing*, vol. 2002, no. 11, pp.1274–1288, 2002.

[38] A. H. Abdelaziz, S. Zeiler, and D. Kolossa, "Twin-HMM-based audio-visual speech enhancement," in *Proc. ICASSP,* 2013, pp. 3726-3730.

[39] J. M. Kates and K. H. Arehart, "The Hearing-Aid Speech Quality Index (HASQI)," *Journal of the Audio Engineering Society*, vol. 58, no. 5, pp. 363-381, 2010.

[40] J. M. Kates and K. H. Arehart, "The Hearing-Aid Speech Perception index (HASPI)," *Speech Communication*, vol. 65, pp. 75–93, 2014.

[41] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction Wiener filter," *IEEE/ACM Transactions on*

Audio, Speech, and Language Processing, vol. 14, pp. 1218-1234, 2006.

[42] S.-S Wang, H.-T Hwang, Y.-H Lai, Y. Tsao, X. Lu, H.-M Wang, and B. Su, "Improving denoising auto-encoder based speech enhancement with speech parameter generation algorithm," in Proc. *APSIPA* , 2015, pp. 365-369.

[43] P. Viola and M. J. Jones, "Robust Real-Time Face Detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137-154, 2004.

[44] G. Tzimiropoulos and M. Pantic, "Gauss-Newton Deformable Part Models for Face Alignment in-the-Wild," in *Proc. CVPR*, 2014, pp. 1851-1858.