Multi-Task Joint Learning for Embedding Aware Audio-Visual Speech Enhancement

Chenxi Wang[1], *Hang Chen*[1], *Jun Du*[*1], *Baocai Yin*[2], *Jia Pan*[2]

[1]University of Science and Technology of China, Hefei, China
[2]iFlytek, Hefei, China

{cx_wang,ch199703}@mail.ustc.edu.cn, ✉jundu@ustc.edu.cn, {bcyin,jiapan}@iflytek.com

2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP) | 979-8-3503-9796-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISCSLP57327.2022.10038268

Abstract

In this paper, we propose a multi-task joint learning scheme to improve embedding aware audio-visual speech enhancement by adopting the phone and the articulation place together as the classification targets during the training of embedding extractor and enhancement network. Firstly, the multimodal embedding is extracted from noisy speech and lip frames, and supervised by the articulation place and the phone label levels together. Next, we train the embedding extractor and enhancement network jointly where the learning objects include the ideal ratio mask, the phone posteriori and the place posteriori. Experiments on the TCD-TIMIT corpus corrupted by simulated additive noises show that the proposed multimodal embedding at the multi-scale class level is more effective than the previous embedding at the place/phone level and the multi-task based joint learning framework further improves speech quality and intelligibility.

Index Terms: speech enhancement, audio-visual, multi-task

1. Introduction

Speech enhancement is a task aimed at improving the quality and intelligibility of noise-contaminated speech. Speech enhancement has a wide range of real-world applications, including hearing aids [1], speech recognition [2] and mobile communications [3], etc.

Traditional speech enhancement methods include spectral subtraction [4], Wiener filtering [5, 6] and minimum mean squared error (MMSE) estimation [7]. Recently, data-driven speech enhancement methods using deep neural networks have been extensively explored and demonstrated promising results [8]-[10]. Some researches [11]-[15] showed visual information such as facial/lip movements can help speech perception, and using visual data as input can have an auxiliary effect for speech enhancement in noisy environments. Inspired by these discoveries, the speech enhancement system utilizing both audio and visual modalities, which is also known as Audio-Visual Speech Enhancement (AVSE), has been developed [16]-[19]. However, since the acoustic information in the video is limited, how to obtain as much useful acoustic information as possible is particularly important for AVSE.

In recent years, researchers [20, 21] used a pre-trained isolated word recognition model to choose the most useful visual embedding while [22] found that using phone to pretrain visual embedding extractor rather than isolated word can get better performance. In previous works [23], we proposed a state-of-the-art AVSE model called the multimodal embedding aware speech enhancement (MEASE) model. The MEASE model includes a multimodal embedding extractor and an embedding aware enhancement network. The multimodal embedding extractor, which takes both audio and video as inputs and fuses them into multimodal embedding, is pretrained with the articulation place classification task. There is a high correlation between the low-resolution articulation place label and the limited acoustic information in the video, which is confirmed to be beneficial for visual embedding extraction. But for audio embedding extraction, more acoustic details in the audio ask for a classification target with finer granularity where phone is a more suitable label than articulation place. The lack of the phone information limits the performance of the MEASE model.

Multi-task learning has been successfully applied in the audio-only speech enhancement area. [24] used speech presence probability (SPP) estimation as a secondary task assisting the target estimation in the speech enhancement task. [25] proposed a multi-task network to estimate the magnitude spectrum of both the clean speech and the noise from the noisy speech. Our method is not a simple audio-video version of the above multi-task learning method. We consider the differences in audio and video modalities and design two focused auxiliary tasks respectively, namely phoneme classification and viseme classification.

According to the above analyses, we propose a two-stage multi-task joint learning scheme for the MEASE model. The first stage of the scheme is a multi-task learning method for the multimodal embedding extractor, i.e. using both the phone and the articulation place as training targets, to improve the effectiveness of the multimodal embedding. The second stage is a joint learning approach for the multimodal embedding extractor and the enhancement network where the learning targets consist of the ideal ratio mask, the phone posteriori and the place posteriori to further refine the multimodal embedding and improve the speech enhancement performance.

The rest of this paper is organized as follows. Section 2 introduces the details of the proposed multi-task joint learning scheme. Section 3 presents dataset, experimental setup and experimental results, and a conclusion is given in Section 4.

2. The proposed scheme

2.1. The proposed multi-task learning method

Our proposed multi-task learning method for the multimodal embedding extractor is shown in Fig. 1.

The multimodal embedding extractor takes filter bank (FBANK) features of noisy audio and lip frames cropped from video as inputs and outputs the multimodal embedding. It consists of an audio embedding extractor, a visual embedding extractor and a fusion module. The visual embedding extractor, which consists of a 3D convolutional layer with 64 kernels of 5×7×7 and a stride of 1×2×2, a batch normalization, a ReLU activation, a 3D max-pooling layer and a 18-layer ResNet [26], takes lip frames as input and outputs a 256-dimensional vector

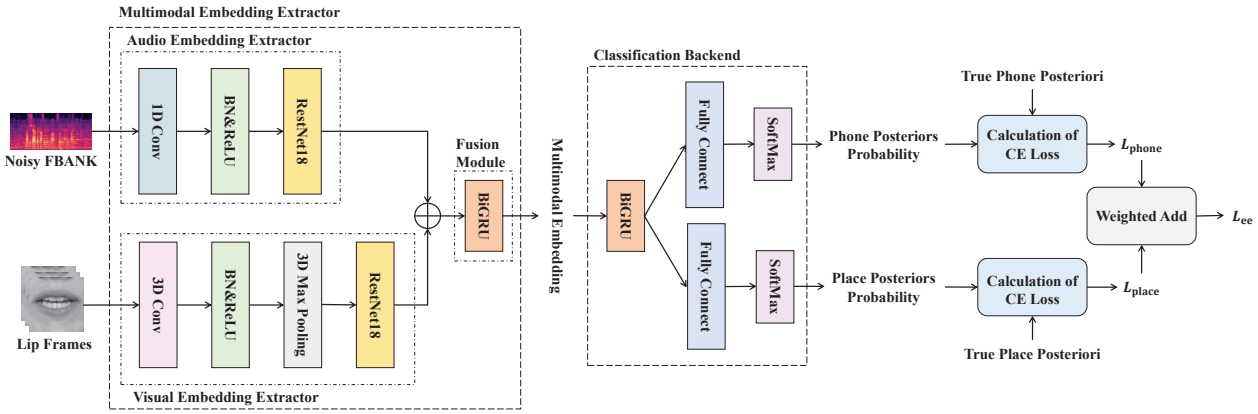

Figure 1: *The proposed multi-task learning method for the multimodal embedding extractor*

for each lip frame. The visual embedding is the composition of the vectors of all lip frames. The audio embedding extractor, which consists of a 1D convolutional layer with 64 kernels of 1 and a stride of 1, a batch normalization, a ReLU activation and a 18-layer variant of ResNet which all 2D convolutions are replaced with 1D convolutions, takes FBANK features as input and outputs a 256-dimensional vector for each time step. The audio embedding is the composition of the vectors of all time steps. The number of vectors in audio embedding N_{A} and the number of vectors in visual embedding N_{V} would not be equal if the sampling rates of the input audio and video are different. The solution is repeating each vector in the visual embedding $N_{\mathrm{A}}/N_{\mathrm{V}}$ times as shown in Eq. (1). We denote the visual embedding as E_{V} and the modified visual embedding as $\tilde{E}_{\mathrm{V}}$. Then the audio embedding and the modified visual embedding are concatenated over channel dimension and sent to the 2-layer Bidirectional Gated Recurrent Unit (BiGRU) [27] in the fusion module. The output is the multimodal embedding which consists of the 512-dimensional vector of all time step.

$$\tilde{E}_{\mathrm{V}} = \{\overbrace{E_{\mathrm{V}}^{0}, \cdots, E_{\mathrm{V}}^{0}}^{N_{\mathrm{A}}/N_{\mathrm{V}}}, \cdots, \overbrace{E_{\mathrm{V}}^{N_{\mathrm{V}}-1}, \cdots, E_{\mathrm{V}}^{N_{\mathrm{V}}-1}}^{N_{\mathrm{A}}/N_{\mathrm{V}}}\} \quad (1)$$

The multimodal embedding extractor is trained by a classification backend which consists of a 2-layer BiGRU, two fully connected layers followed by a SoftMax activation. The classification backend takes the multimodal embedding as input and outputs the posterior probabilities of each class of phone (P_{phone}) and place (P_{place}). There are 39 CI-phones and 10 classes of place as in [28, 29] used as the true labels named $P_{\mathrm{phone}}^{\mathrm{truth}}$ and $P_{\mathrm{place}}^{\mathrm{truth}}$. We calculate the cross entropy (CE) loss L_{phone} between P_{phone} and $P_{\mathrm{phone}}^{\mathrm{truth}}$ as in Eq. (2) and the cross entropy loss L_{place} between P_{place} and $P_{\mathrm{place}}^{\mathrm{truth}}$ as in Eq. (3).

$$L_{\mathrm{phone}} = -\sum P_{\mathrm{phone}}^{\mathrm{truth}} \log P_{\mathrm{phone}} \quad (2)$$

$$L_{\mathrm{place}} = -\sum P_{\mathrm{place}}^{\mathrm{truth}} \log P_{\mathrm{place}} \quad (3)$$

The total loss function is defined as follows, where α is a hyper-parameter that is tuned on the validation set.

$$L_{\mathrm{ee}} = \alpha \times L_{\mathrm{phone}} + (1-\alpha) \times L_{\mathrm{place}} \quad (4)$$

The structure of our proposed multi-task multimodal embedding aware speech enhancement (MTMEASE) model is the module marked by the blue dotted line in Fig. 2, which can be divided into two parts: the multimodal embedding extractor and the enhancement network. The input of the enhancement network is the log-power spectra (LPS) features [30] of noisy audio and the multimodal embedding. The enhancement network consists of an audio encoder, a multimodal encoder and a decoder, all of which are stacked by 1D convblocks. The 1D convblock includes a 1D convolution layer with residual connection, a ReLU activation and a batch normalization. The 1D convolution layer has 1536 channels, the kernel size is 5 and the stride is 5. The output of the audio encoder and the multimodal embedding are concatenated along channel dimension and then sent to the decoder. In the end, the ideal ratio mask (IRM) [31] of the noisy audio is predicted.

The multimodal embedding extractor is pretrained and kept frozen during the training of the enhancement network. We train the enhancement network by minimizing the mean square error (MSE) loss between the output IRM M and the target IRM M_{target}:

$$L_{\mathrm{en}} = \sum \|M - M_{\mathrm{target}}\|_2^2 \quad (5)$$

2.2. The proposed joint learning approach

As mentioned above, the multimodal embedding extractor is pretrained in a classification task and is frozen during the training of the enhancement network. We first proposed a model called u-MTMEASE which also optimizes the multimodal embedding extractor during the training of the enhancement network, but the performance of u-MTMEASE is not good. Thus we add constraints on the loss function, and propose a joint learning approach for the multimodal embedding extractor and the enhancement network in MTMEASE model. The schematic diagram of the approach is illustrated in Fig. 2. Compared with MTMEASE, a classification backend which is the same as that in Fig. 1 is added as indicated by the yellow dashed arrow. It predicts the phone and the articulation place while the enhancement network predicts IRM.

The loss function of the joint learning approach is shown in Eq. (6). It is the weighted sum of the loss of the classification backend L_{ee} as shown in Eq. (4) and the loss of the enhancement network L_{en} as shown in Eq. (5). W_{ee} and W_{en} are the hyper-parameters which are tuned on the validation set.

$$L_{\mathrm{total}} = W_{\mathrm{ee}} \times L_{\mathrm{ee}} + W_{\mathrm{en}} \times L_{\mathrm{en}} \quad (6)$$

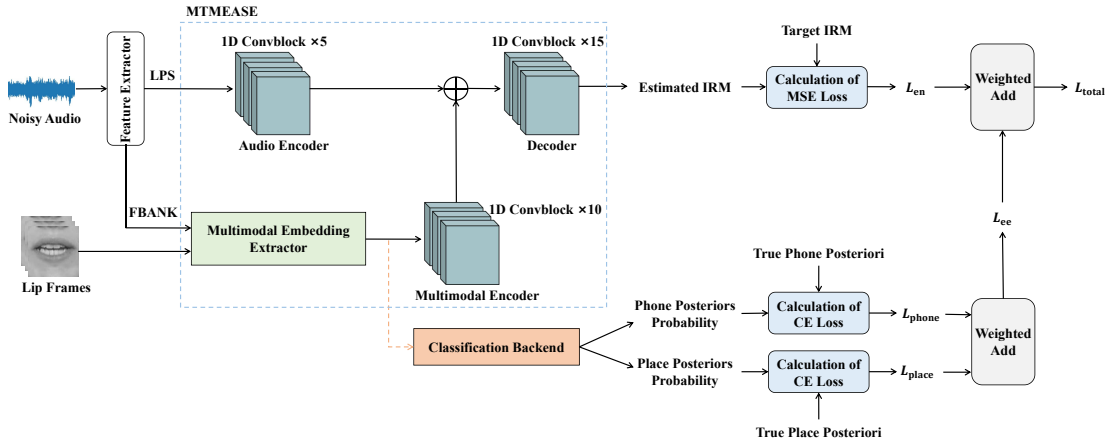

Figure 2: *The joint learning approach of the multimodal embedding extractor and the enhancement network*

3. Experiments

3.1. Dataset

The dataset used in experiments is the simulated dataset of noisy speech based on the TCD-TIMIT audio-visual corpus [32] which was created in [23]. It consists of 59 volunteer speakers with around 98 videos each. The 5643 utterances of 57 speakers (31 male and 26 female) were divided into 5600 utterances and 43 utterances as the train set and the validation set, respectively. The utterances of the remaining two speakers (1 male and 1 female) are used as the test set.

The 115 noise types which include 100 noise types in [33] and 15 home-made noise types were mixed with the train set at the SNRs of -5 dB, 0 dB, 5 dB, 10 dB and 15 dB. Three unseen noise types (Destroyer Operations, Factory2 and F16 Cockpit) at above-mentioned SNR levels are used to build a validation set. Three other unseen noises (Destroyer Engine, Factory1 and Speech Babble) at above-mentioned SNR levels are used for testing. All the above unseen noises were collected from the NOISEX92 corpus [34].

3.2. Experimental settings

All the speech utterances were resampled to 16 kHz. The audio frames were extracted by a 25-ms Hanning window and an overlap of 10-ms. Then a 400-point short-time Fourier transform was used to compute the spectra of each frame. 201-dimensional LPS features were calculated as the input of audio encoder and 40-dimensional FBANK features were calculated as the input of multimodal embedding extractor. The videos were resampled to 25 fps. We extracted 68 facial landmarks from every video frame by using Dlib [35] implementation of the face landmark estimator [36]. Then the lip frames of size 98×98 pixels were cropped by using the 20 mouth landmarks from 68 facial landmarks. PESQ [37] and STOI [38] are used as evaluate metrics.

The training of MTMEASE is divided into two parts. Firstly, we train the multimodal embedding extractor with Adam optimizer for 100 epochs. α used in Eq. (4) is set to 0.7. The mini-batch size is set to 64. The initial learning rate is 3e-4 and decreased on log scale after 30 epochs. Next, we train the enhancement network with Adam optimizer for 100 epochs. The batch size is set to 96. The initial learning rate is 1e-4 and halved if the loss on the validation set does not improve for three consecutive epochs.

We denote the MTMEASE model with the joint learning approach as JL-MTMEASE. We use the same optimization method as that of the enhancement network in MTMEASE except the initial learning rate is set to 1e-5 to minimize the loss during the joint learning of JL-MTMEASE. The values of W_{ee} and W_{en} in Eq. (6) are set to 1 and 0.1, respectively.

3.3. Results of the proposed MTMEASE model

To verify our points where the phone label and audio are highly correlated while the articulation place label and acoustic information in the video are highly correlated, we did experiments on the audio embedding aware speech enhancement (AEASE) model and the visual embedding aware speech enhancement (VEASE) model [23]. The differences among the model structures of AEASE, VEASE and MTMEASE are that AEASE has an audio embedding extractor, VEASE has a visual embedding extractor and MTMEASE has a multimodal embedding extractor. We compare the enhancement performance of AEASE-phone and AEASE-place, i.e. the AEASE model using the phone/the articulation place to train the audio embedding extractor. Similarly, we compare the performance of VEASE-phone and VEASE-place, i.e. the VEASE model using the phone/the articulation place to train the visual embedding extractor. We got results of AEASE-place, VEASE-phone, VEASE-place and MEASE from [23]. Table 1 presents evaluations for AEASE-phone, AEASE-place, VEASE-phone, VEASE-place, MEASE and our MTMEASE.

The first four rows in Table 1 show that the performance of AEASE-phone is better than that of AEASE-place while the performance of VEASE-place is better than that of VEASE-phone. The performance gain is more significant on the STOI metric. The results imply that there is a high correlation between the phone label and audio which is beneficial for audio embedding extraction and there is a high correlation between the articulation place label and visual acoustic information which is beneficial for visual embedding extraction.

Based on the results in Table 1, MTMEASE shows consistent improvements over MEASE across all evaluation metrics. For example, the PESQ of MTMEASE increased from 2.29 to 2.38 at -5 dB SNR and from 3.16 to 3.26 at 10 dB SNR. The results demonstrate our proposed multi-task learning method can improve the effectiveness of the multimodal embedding.

Table 1: *Average performance comparison of AEASE-phone, AEASE-place, VEASE-phone, VEASE-place, MEASE and our MTMEASE on the test set at different SNRs averaged over 3 unseen noise types.*

Model	PESQ					STOI (in %)				
SNR (in dB)	-5	0	5	10	15	-5	0	5	10	15
AEASE-place	2.09	2.39	2.69	2.98	3.27	60.84	72.24	81.58	88.39	92.76
AEASE-phone	2.10	2.40	2.71	3.00	3.29	61.20	72.67	81.93	88.56	92.87
VEASE-place	2.21	2.47	2.73	3.00	3.26	66.57	75.27	82.64	88.80	92.96
VEASE-phone	2.14	2.42	2.69	2.96	3.23	66.29	74.89	82.22	88.45	92.79
MEASE	2.29	2.59	2.88	3.16	3.42	68.96	77.64	84.43	89.99	93.64
MTMEASE	2.38	2.70	2.99	3.26	3.50	70.75	78.98	85.45	90.56	94.01

Table 2: *Average performance comparison of JL-AEASE-phone, JL-VEASE-place, u-MTMEASE, MTMEASE and JL-MTMEASE on the test set at different SNRs averaged over 3 unseen noise types.*

Model	PESQ					STOI (in %)				
SNR (in dB)	-5	0	5	10	15	-5	0	5	10	15
JL-AEASE-phone	2.12	2.47	2.79	3.08	3.35	62.45	74.62	83.29	89.47	93.46
JL-VEASE-place	2.24	2.52	2.77	3.04	3.29	66.67	75.53	83.07	89.10	93.07
MTMEASE	2.38	2.70	2.99	3.26	3.50	70.75	78.98	85.45	90.56	94.01
u-MTMEASE	2.32	2.62	2.91	3.19	3.45	70.48	78.58	85.22	90.45	93.95
JL-MTMEASE	2.39	2.71	3.00	3.27	3.51	71.02	79.25	85.64	90.67	94.08

3.4. Results on the joint learning approach

The joint learning framework can be generalized to AEASE and VEASE models [23] as training the audio/visual embedding extractor and enhancement network jointly. To examine the effectiveness of the proposed joint learning approach on enhancement performance, we not only train the JL-MTMEASE model, but also apply the approach to the AEASE-phone model and the VEASE-place model, which is the best AEASE/VEASE model in the previous experiments, respectively. We denote them as JL-AEASE-phone and JL-VEASE-place, respectively. The results of these models and the u-MTMEASE model we mentioned in the Section II are shown in Table 2.

Table 2 shows that the performance of u-MTMEASE, which simply trains the embedding extractor together with the enhancement network, degrades compared to MTMEASE. This may be attributed to the different training objectives of the embedding extractor and the enhancement network, resulting in the decline of effectiveness of embedding. Nevertheless, the results of JL-MTMEASE achieve a stable improvement in each SNR compared to MTMEASE. Besides, significant improvements were achieved on JL-AEASE-phone and JL-VEASE-place compared to AEASE-phone and VEASE-place in Table 1, respectively. The reason why the improvement of JL-MTMEASE is not obvious might be that the performance of MTMEASE is already excellent, and there is little room for improvement. In summary, the joint learning approach is robust and has sufficient generalization ability.

Fig. 3 shows the spectrogram comparison between MEASE and JL-MTMEASE. Compared with MEASE, JL-MTMEASE can remove most of the noise and reduce speech distortion as shown in the white box of Fig. 3.

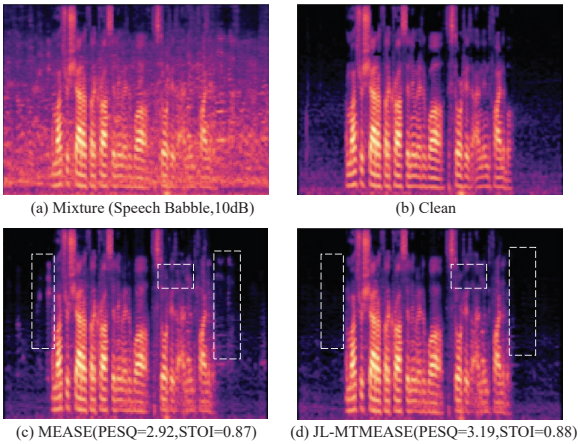

Figure 3: *The spectrogram comparison between MEASE and JL-MTMEASE*

4. Conclusions

In this study, we propose a multi-task joint learning scheme consisting of a multi-task training method on the multimodal embedding extractor which can get more effective multimodal embedding and an approach of joint learning between the embedding extractor and the enhancement network to make the multimodal embedding more suitable for speech enhancement task and further improve the enhancement performance. The experimental results show that our proposed scheme improves the PESQ and STOI metrics on the enhanced speech. In the future, we will try to apply the proposed scheme to other audio-visual speech enhancement models.

5. Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant No. 62171427 and the Strategic Priority Research Program of Chinese Academy of Sciences under Grant No. XDC08050200.

6. References

[1] L A. Passos, J. P. Papa, and A. Adeel, "Canonical cortical graph neural networks and its application for speech enhancement in future audio-visual hearing aids," *arXiv preprint arXiv:2206.02671*, 2022.

[2] J. Hong, M. Kim, D. Yoo, and Y. M. Ro, "Visual context-driven audio feature enhancement for robust end-to-end audio-visual speech recognition," *arXiv preprint arXiv:2207.06020*, 2022.

[3] M. Tammen, X. Li, S. Doclo, and L. Theverapperuma, "Dictionary-Based Fusion of Contact and Acoustic Microphones for Wind Noise Reduction," *arXiv preprint arXiv:2205.09017*, 2022.

[4] S. Boll, "FSuppression of acoustic noise in speech using spectral subtraction," in *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, no. 2, pp. 113-120, 1979.

[5] Jae Lim and A. Oppenheim, "All-pole modeling of degraded speech," in *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 26, no. 3, pp. 197-210, 1978.

[6] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," in *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586-1604, 1979.

[7] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," in *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 33, no. 2, pp. 443-445, 1985.

[8] X. -G. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising auto-encoder," in *Proc. Interspeech 2013*, pp.436-440, 2013.

[9] Y. Xu, J. Du, L. -R. Dai and C. -H. Lee, "Global variance equalization for improving deep neural network based speech enhancement," in *2014 IEEE China Summit & International Conference on Signal and Information Processing (ChinaSIP)*, pp. 71-75, 2014.

[10] Y. Xu, J. Du, L. -R. Dai and C. -H. Lee, "A regression approach to speech enhancement based on deep neural networks," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7-19, 2015.

[11] H. MCGURK, and J. MACDONALD, "Hearing lips and seeing voices." in *Nature*, vol. 264, no. 5588, pp. 746–748, 1976.

[12] A. MacLeod, and Q. Summerfield, "Quantifying the contribution of vision to speech perception in noise," in *British Journal of Audiology*, vol. 21, no. 2, pp. 131-141, 1987.

[13] L. E. Bernstein, and C. Benoit, "For speech perception by humans or machines, three senses are better than one," in *Proceeding of Fourth International Conference on Spoken Language Processing.ICSLP '96*, vol. 3, pp. 1477-1480, 1996.

[14] L. D. Rosenblum, "Speech perception as a multimodal phenomenon," in *Current Directions in Psychological Science*, vol. 17, no. 6, pp. 405–409, 2008.

[15] D. W. Massaro, and J. A. Simpson, "Speech perception by ear and eye: a paradigm for psychological inquiry," in *Psychology Press*, 2014.

[16] J. Hou, S. Wang, Y. Lai, Y. Tsao, H. Chang, and H. Wang, "Audio-visual speech enhancement using multimodal deep convolutional neural networks," in *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 117-128, 2018.

[17] A. Gabbay, A. Shamir, and S. Peleg, "Visual speech enhancement," in *Proc. Interspeech 2018*, pp.1170-1174, 2018.

[18] G. Morrone, S. Bergamaschi, L. Pasa, L. Fadiga, V. Tikhanoff, and L. Badino, "Face landmark-based speaker-independent audio-visual speech enhancement in multi-talker environments," in *ICASSP 2019*, pp. 6900-6904, 2019.

[19] X. Xu et al., "VSEGAN: visual speech enhancement generative adversarial network," in *ICASSP 2022*, pp. 7308-7311, 2022.

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

[21] E. Ideli, B. Sharpe, I. V. Bajić, and R. G. Vaughan, "Visually assisted time-domain speech enhancement," in *2019 IEEE global conference on signal and information processing (GlobalSIP)*, pp. 1–5, 2019.

[22] J. Wu et al., "Time domain audio visual speech separation," in *2019 IEEE automatic speech recognition and understanding workshop (ASRU),* pp. 667–673, 2019.

[23] H. Chen, J. Du, Y. Hu, L. -R. Dai, B. -C. Yin and C. -H. Lee, "Correlating subword articulation with lip shapes for embedding aware audio-visual speech enhancement," in *Neural Network*, vol. 143, pp: 171–182, 2021.

[24] L. Wang, J. Zhu and I. Kodrasi, "Multi-task Single Channel Speech Enhancement Using Speech Presence Probability As A Secondary Task Training Target," *2021 29th European Signal Processing Conference (EUSIPCO)*, pp. 296-300, 2021.

[25] J. Hou, S. Zhao and Y. An, "Single-channel Speech Enhancement Using Multi-Task Learning and Attention Mechanism," *2021 IEEE 6th International Conference on Signal and Image Processing (ICSIP)*, pp. 826-830, 2021.

[26] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[27] K. Cho et al., "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[28] S. M. Siniscalchi, and C. -H. Lee, "A study on integrating acoustic-phonetic information into lattice rescoring for automatic speech recognition," in *Speech Communication*, vol. 51, no. 11, pp. 1139–1153, 2009.

[29] C. -H. Lee, and S. M. Siniscalchi, "An information-extraction approach to speech processing: Analysis, detection, verification, and recognition," in *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1089–1115, 2013.

[30] J. Du and Q. Huo, "A speech enhancement approach using piecewise linear approximation of an explicit model of environmental distortions," in *Proc. INTERSPEECH 2018*, pp. 569–572, 2008.

[31] C. Hummersone, T. Stokes, and T. Brookes, "On the ideal ratio mask as the goal of computational auditory scene analysis," in *Blind source separation: advances in theory, algorithms and applications*, pp. 349–368, 2014.

[32] N. Harte, and E. Gillen, "Tcd-timit: An audio-visual corpus of continuous speech," in *IEEE Transactions on Multimedia*, vol. 17, nol. 5, pp. 603–615, 2015.

[33] G. Hu, and D. Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 2067–2079, 2010

[34] A. Varga, and H. J. M. Steeneken, "Assessment for automatic speech recognition: Ii. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," in *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[35] D. E. King, "Dlib-ml: A machine learning toolkit," in *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[36] V. Kazemi, and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *2014 IEEE conference on computer vision and pattern recognition*, pp. 1867–1874, 2014.

[37] 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 *2001 IEEE international conference on acoustics, speech, and signal processing. proceedings (Cat. No. 01CH37221)*, vol. 2, pp. 749–752, 2001.

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