Neural Networks 143 (2021) 171–182

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

2021 Special Issue

Correlating subword articulation with lip shapes for embedding aware audio-visual speech enhancement

Hang Chen[a], Jun Du[a,*], Yu Hu[a], Li-Rong Dai[a], Bao-Cai Yin[b], Chin-Hui Lee[c]

[a] *National Engineering Laboratory for Speech and Language Information Processing, University of Science and Technology of China, Hefei, Anhui, China*
[b] *iFlytek Research, iFlytek Co., Ltd., Hefei, Anhui, China*
[c] *School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA*

ARTICLE INFO

Article history:
Available online 8 June 2021

Keywords:
Speech enhancement
Audio-visual
Representation learning
Deep learning
Universal attribute recognition

ABSTRACT

In this paper, we propose a visual embedding approach to improve embedding aware speech enhancement (EASE) by synchronizing visual lip frames at the phone and place of articulation levels. We first extract visual embedding from lip frames using a pre-trained phone or articulation place recognizer for visual-only EASE (VEASE). Next, we extract audio-visual embedding from noisy speech and lip frames in an information intersection manner, utilizing a complementarity of audio and visual features for multi-modal EASE (MEASE). Experiments on the TCD-TIMIT corpus corrupted by simulated additive noises show that our proposed subword based VEASE approach is more effective than conventional embedding at the word level. Moreover, visual embedding at the articulation place level, leveraging upon a high correlation between place of articulation and lip shapes, demonstrates an even better performance than that at the phone level. Finally the experiments establish that the proposed MEASE framework, incorporating both audio and visual embeddings, yields significantly better speech quality and intelligibility than those obtained with the best visual-only and audio-only EASE systems.

1. Introduction

Background noises considerably reduce the quality and intelligibility of the speech signal, thus limiting the performance of speech-related applications in real-world conditions (e.g. automatic speech recognition, dialog system and hearing aid, etc.). The objective of speech enhancement (Loizou, 2013) is to generate enhanced speech with better speech quality and clarity by suppressing background noise components in noisy speech.

Conventional speech enhancement approaches, such as spectral subtraction (Boll, 1979), Wiener filtering (Lim & Oppenheim, 1978), minimum mean squared error (MMSE) estimation (Ephraim & Malah, 1985), and the optimally-modified log-spectral amplitude (OM-LSA) speech estimator (Cohen, 2003; Cohen & Berdugo, 2001), have been extensively studied in the past. Recently, the application of deep learning technologies has been successful in speech enhancement (Narayanan & Wang, 2013; Wang & Chen, 2018; Xu, Du, Dai, & Lee, 2015).

Human auditory system can track a single target voice source in extremely noisy acoustic environment like a cocktail party, which is also known as the cocktail party effect (Cherry, 1953). This finding motivates us to design speech enhancement systems by drawing on the way humans perceive speech. McGurk Effect (McGurk & MacDonald, 1976) suggests a strong influence of vision on human speech perception. Other researches (Bernstein & Benoit, 1996; MacLeod & Summerfield, 1987; Massaro & Simpson, 2014; Rosenblum, 2008) have shown visual cues such as facial/lip movements can help speech perception, through supplementing the acoustic information related to the corresponding speaker, especially in noisy environments. Inspired by the aforementioned discoveries, the speech enhancement method utilizing both audio and visual signals, which is also known as audio-visual speech enhancement (AVSE), has been developed.

The AVSE methods can be traced back to Girin, Feng, and Schwartz (1995) and following work, e.g. Abdelaziz, Zeiler, and Kolossa (2013), Deligne, Potamianos, and Neti (2002), Fisher III, Darrell, Freeman, and Viola (2001), Girin, Schwartz, and Feng (2001), Goecke, Potamianos, and Neti (2002) and Hershey and Casey (2002). And recently numerous studies have attempted to build deep neural network-based AVSE models. Gabbay, Ephrat, Halperin and Peleg (2018) employed a video-to-speech method to construct T-F masks for speech enhancement. An encoder–decoder architecture was used in Gabbay, Shamir and Peleg (2018) and Hou et al. (2018). These methods were merely demonstrated under constrained conditions (e.g. the utterances

* Correspondence to: National Engineering Laboratory for Speech and Language Information Processing (NEL-SLIP), University of Science and Technology of China, No. 96, JinZhai Road, Hefei, Anhui, PR China.
E-mail address: jundu@ustc.edu.cn (J. Du).

https://doi.org/10.1016/j.neunet.2021.06.003
0893-6080/© 2021 Elsevier Ltd. All rights reserved.

consisted of a fixed set of phrases, or a small number of known speakers). Afouras, Chung, and Zisserman (2018) proposed a deep AVSE network consisting of the magnitude and phase subnetworks, which enhanced magnitude and phase, respectively. Ephrat et al. (2018) designed a model that hinged on the facial embedding of the source speaker and outputted the complex mask. Ideli, Sharpe, Bajić, and Vaughan (2019) proposed a time-domain AVSE framework based on Conv-tasnet (Luo & Mesgarani, 2019). These methods all performed well in the situations of unknown speakers and unknown noise types.

We briefly discuss the above-mentioned AVSE methods from the following two perspectives: visual embedding and audio-visual fusion method. Regrading the visual embedding, Gabbay, Ephrat et al. (2018), Gabbay, Shamir et al. (2018) and Hou et al. (2018) made use of the image sequences of the lip region. For discarding irrelevant variations between images, such as illumination, Ephrat et al. (2018) proposed using face embedding obtained from a pre-trained face recognizer and confirmed through ablation experiments that the lip area played the most important role for enhancement performance in the face area. Moreover, Afouras et al. (2018) and Ideli et al. (2019) chose lip embedding via the middle layer output in a pre-trained isolated word recognition model.

In recent work, Wu et al. (2019) adopted the phone as the classification target instead of isolated word and provided a more useful visual embedding for speech enhancement. In the term of audio-visual fusion method, most AVSE methods focus on audio-visual fusion that happens at the middle layer of the enhancement network in the fashion of channel-wise concatenation.

These pioneering works are the foundation of this research paper. A useful visual embedding should contain as much acoustic information in the video as possible. But the acoustic information in video is very limited, and there is also other information redundancy. In the current classification-based embedding extracting framework, we could yield a more robust and generalized visual embedding by both reducing the information redundancy and increasing the correlation between classification target and visual acoustic information. For the former, cutting out the lip area is useful, while for the latter, finding a classification target that is more relevant to lip movements will be of help.

The superset of speech information called speech attributes includes a series of fundamental speech sounds with their articulatory knowledge, linguistic interpretations, speaker characteristics, and emotional state, etc. (Lee et al., 2007). In contrast to phone models, a smaller number of universal attribute units are needed for a complete characterization of any spoken language (Li, Ma, & Lee, 2007). The place and manner of articulation are two speech attributes based on articulatory phonology, which were widely used in automatic speech recognition (Li, Tsao, & Lee, 2005), automatic spoken language recognition (Siniscalchi, Reed, Svendsen, & Lee, 2013) and non-native mispronunciation detection (Li, Siniscalchi, Chen, & Lee, 2016), etc. Early works (Livescu et al., 2007; Saenko, Darrell, & Glass, 2004) also proposed to use these articulatory features for visual and audio-visual speech recognition. We propose that the place and manner of articulation have a higher correlation with the visual acoustic information and can provide a more useful supervisory signal in the stage of visual embedding extractor training.

One consensus in multimodal learning is that the data of each mode obeys an independent distribution conditioned on the ground truth label (Blum & Mitchell, 1998; Dasgupta, Littman, & McAllester, 2002; Leskes, 2005; Lewis, 1998). Each mode captures features related to ground truth tags from different aspects, so the information extracted (labels excluded) is not necessarily related to the other. This shows that the ground truth can be seen as "information intersection" between all modes (Sun et al., 2020), i.e. the amount of agreement shared by all the modalities. Specifically, in AVSE, there is a mismatch between the information intersection and ground truth label. The intersection of audio modal (noisy speech) and video modal (lip video) is not clean speech which is ground truth.

In this paper, we extend the previous AVSE framework to the embedding aware speech enhancement (EASE) framework. The conventional AVSE methods are regarded as special EASE methods, which only utilize visual embedding extracted from lip frames, as known as visual embedding aware speech enhancement (VEASE) methods. In EASE framework, we propose a VEASE model using a novel visual embedding, which is the middle layer output in a pre-trained articulation place recognizer. We adopt the same dataset in the stages of embedding extractor training and enhancement network training. A more effective visual embedding is obtained by utilizing a high correlation between the designed classification target, i.e., the articulation place, and the visual acoustic information rather than additional video data. Moreover, we present a novel multimodal embedding aware speech enhancement (MEASE) model which extends the visual-only pre-trained embedding extractor to the audio-visual pre-trained embedding extractor and yields significantly better speech quality and intelligibility. In the MEASE model, the fusion of audio and visual embeddings occurs in the stage of embedding extractor training and is supervised by their information intersection at the articulation place label level, which is an early fusion. In order to better understand the effect of audio-visual fusion stage in a neural network on the enhancement performance, we also perform a series of experiments which make the fusion take place at different stages without changing the network structure. And we observe better speech enhancement performance in early fusion under the framework of a neural network.

The main contributions of this paper are:

(1) We explore the effectiveness of different visual embeddings pre-trained for various classification targets on enhancement performance. A novel classification target, i.e., the articulation place, is proposed for training visual embedding extractor. The visual embedding utilizing a high correlation between the articulation place and the acoustic information in video achieves the better enhancement performance with no additional data used.
(2) We verify the complementarity between audio and visual embeddings lies in different signal-to-noise ratio (SNR) levels, as well as different articulation places by ablation experiments. And based on the information intersection, we adopt a novel fusion method integrating visual and audio embeddings in the proposed MEASE model, as it performs better in all SNR levels and all articulation places.
(3) We design experiments to study the effect of the stage when audio-visual fusion occurs on the quality and intelligibility of enhanced speech under the framework of a neural network. And we observe that the early fusion of audio and visual embeddings delivers better enhancement performance.

Concurrently and independently from us, a number of groups have proposed various methods from visual embedding and audio-visual fusion for AVSE. Wang, Xing, Wang, Chen, and Sun (2020) observed serious performance degradations when these AVSE methods were applied with a medium or high SNR[1] and proposed a late fusion-based approach to safely combine visual

[1] Performance degradation in Wang et al. (2020) may result from the changes in the network structure, but we have indeed observed reduction in improvements from our results of comparative experiment, as will be discussed in Section 4.4.

knowledge in speech enhancement. This is the opposite of our work. Iuzzolino and Koishida (2020) proposed a new mechanism for audio-visual fusion. In this research, the fusion block was adaptable to any middle layers of the enhancement network. This kind of multiple fusion in the enhancement network was better than the standard single channel-wise concatenation. Lu, Duan, and Zhang (2019) proposed a novel audio-visual deep clustering model which employs a two-stage audio-visual fusion strategy during the speech separation training without the supervised pre-training. However, these two works differ from ours in that audio-visual integration still occurs in the middle of the enhancement/separation network. In Gu et al. (2020), audio and visual embeddings were extracted from video and noisy multi-channel speech in the pre-training stage, respectively and fused in speech enhancement training stage, which was supervised by the target speech. This is different from our research that fuses audio and visual embeddings in the pre-training stage, and where the audio-visual fusion is supervised by their information intersection at the articulation place label level. However, we also design a similar model in Section 3.3 and compare it with our method in Section 4.4.

The rest of the paper is structured as follows. In Section 2 we describe the proposed VEASE method. The proposed MEASE method is presented in Section 3. Section 4 has experimental setup including dataset, audio and video preprocessing as well as compares experimental results. Finally, we conclude this work and discuss future research directions in Section 5.

2. VEASE model utilizing articulation place label

In this section, we elaborate our proposed VEASE model, including two aspects, i.e. architecture and training process. The visual embedding extractor is an important part of the VEASE model, which takes a sequence of lip frames as input and outputs a compact vector for every lip frame, known as visual embedding. The VEASE model takes both noisy log-power spectra (LPS) features and visual embeddings as inputs, and outputs ideal ratio mask. The details of the visual embedding extractor and the VEASE model are elaborated as follows.

2.1. Overview of visual embedding extractor

The visual embedding extractor $f_{\mathrm{V}}(\cdot)$ has a similar structure to Petridis et al. (2018) and Stafylakis and Tzimiropoulos (2017), which is also used in previous AVSE studies (Afouras et al., 2018; Ideli et al., 2019). The extractor consists of a spatiotemporal convolution followed by an 18-layer ResNet (He, Zhang, Ren, & Sun, 2016a) which is the identity mapping version (He, Zhang, Ren, & Sun, 2016b), as shown in Fig. 1. A spatiotemporal convolution consists of a convolution layer with 64 3D-kernels of 5 × 7 × 7 (time/width/height), a batch normalization, a ReLU activation and a spatiotemporal max-pooling layer.

For a sequence of lip frames $V = \{V^t \in \mathbb{R}^{H\times W}; t = 0, 1, \ldots, T_{\mathrm{V}} - 1\}$, the feature maps is extracted by the spatiotemporal convolution. Then, the feature maps are passed through the 18-layer ResNet. The spatial dimensionality shrinks progressively in the ResNet until output becomes a L_{V}-dimensional vector per time step, known as the visual embedding E_{V}:

$$\begin{aligned} E_{\mathrm{V}} &= \{E_{\mathrm{V}}^t \in \mathbb{R}^{L_{\mathrm{V}}}; t = 0, 1, \ldots, T_{\mathrm{V}} - 1\} = f_{\mathrm{V}}(V) \\ &= \text{ResNet-18}_{\mathrm{V}}(\text{MaxPooling}_{3D}(\text{BN}(\text{ReLU}(\text{Conv}_{3D}(V))))) \end{aligned} \tag{1}$$

where T_{V}, H and W denote the number and the size of lip frames, respectively. In this study, we use $L_{\mathrm{V}} = 256$, $H = 98$ and $W = 98$ by default.

The visual embedding extractor is trained with a classification backend. E_{V} is fed to the classification backend and the posterior probability of each class P_{class} is outputted, where the class can be labeled as word, phone or place of articulation. We calculate the cross entropy (CE) loss $\mathcal{L}_{\mathrm{CE}}$ between P_{class} and the true distribution of class $P_{\mathrm{class}}^{\mathrm{truth}}$:

$$\mathcal{L}_{\mathrm{CE}} = \mathrm{CE}(P_{\mathrm{class}}^{\mathrm{truth}} \parallel P_{\mathrm{class}}) = -\sum P_{\mathrm{class}}^{\mathrm{truth}} \log P_{\mathrm{class}} \tag{2}$$

The objective function, $\mathcal{L}_{\mathrm{CE}}$, is minimized by using Adam optimizer (Kingma & Ba, 2015) for 100 epochs and the mini-batch size is set to 64. The initial learning rate is set to 0.0003 and is decreased on log scale after 30 epochs. Data augmentation is performed during training, by applying random cropping (±5 pixels) and horizontal flips, that is the same across all lip frames of a sequence. The best model is selected by the highest classification accuracy on the validation set.

2.2. Word based visual embedding extraction

Conventional AVSE techniques (Afouras et al., 2018; Ideli et al., 2019) often obtain the visual embedding extractor discussed earlier based on an isolated word classification task by using a lip reading dataset, such as the Lip Reading in the Wild (LRW).

We build our baseline model, known as VEASE-word that uses the LRW corpus consisting of up to 1000 audio-visual speech segments extracted from BBC TV broadcasts (News, Talk Shows, etc.), totaling around 170 h. There are 500 target words and more than 200 speakers. The LRW dataset provides a word-level label for each audio-visual speech segment, i.e. the real distribution of word $P_{\mathrm{word}}^{\mathrm{truth}}$.

The posterior probability of each class representing each segment of lip frames P_{word} is calculated by the classification backend $f_{\mathrm{C}}'(\cdot)$ on the right side of Fig. 1, consisting of a 2-layer BiGRU, a fully connected layer followed by a SoftMax activation:

$$P_{\mathrm{word}} = f_{\mathrm{C}}'(E_{\mathrm{V}}) = \mathrm{SoftMax}(\mathrm{Average}_T(\mathrm{FC}(\mathrm{BiGRU}(\mathrm{BiGRU}(E_{\mathrm{V}}))))) \tag{3}$$

2.3. Phone based visual embedding extraction

The isolated word classification task usually requires a word-level dataset which is not easy to collect on a large scale. To alleviate this problem, we propose that the same data is used during training visual embedding extractor and enhancement network with different labels. Under the guidance of results in Wu et al. (2019), we choose context-independent (CI) phones consisting of 39 units from CMU dictionary as classification labels, denoted as VEASE-phone.

The TCD-TIMIT dataset is a high quality audio-visual speech corpus labeled at both the phonetic and the word level. We can directly get the frame-level real distribution of CI-phone $P_{\mathrm{phone}}^{\mathrm{truth}}$.

E_{V} is fed to a classification backend $f_{\mathrm{C}}(\cdot)$ which outputs the posterior probability of each CI-phone for each specific time frame P_{phone}. $f_{\mathrm{C}}(\cdot)$ has a same structure as $f_{\mathrm{C}}'(\cdot)$ which only lacks in the average process along the temporal dimension:

$$P_{\mathrm{phone}} = f_{\mathrm{C}}(E_{\mathrm{V}}) = \mathrm{SoftMax}(\mathrm{FC}(\mathrm{BiGRU}(\mathrm{BiGRU}(E_{\mathrm{V}})))) \tag{4}$$

2.4. Articulation place based visual embedding extraction

As discussed earlier, we believe there is a high correlation between speech attributes and visual acoustic information. In order to verify our idea, we check the lip shapes belonging to different places and manners of articulation. We find that the influences of various articulation places on the change of lip shape vary, since the lip shape changes greatly in some utterance segments belonging to specific articulation place. An example is presented in Fig. 2. In contrast, we do not observe similar changes in the term of articulation manner. Consequently we propose

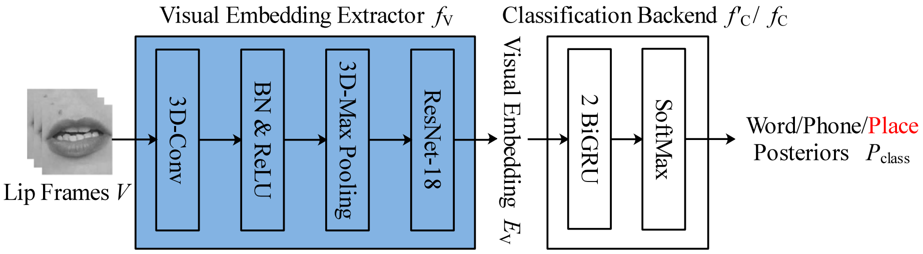


Fig. 1. Illustration of a visual embedding extractor (in color blue for ease of cross-referencing in Figs. 1, 3, 5 and 6). For every lip frame, the extractor outputs a compact vector. We train visual embedding extractor by using 3 different classification labels, i.e., word, phone and place.

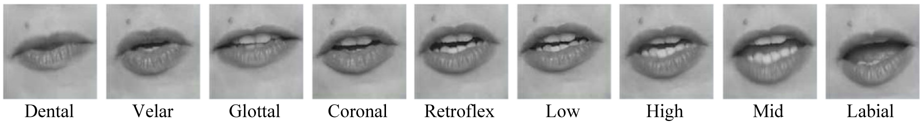


Fig. 2. 9 lip shapes corresponding to utterance segments representing 9 articulation positions: all lip shapes come from a single speaker starting with the lip closed. The lip shape changes greatly in High, Mid and Labial than Dental, Velar and Glottal.

Table 1
The mapping between articulation place classes and CI-phones as in Siniscalchi and Lee (2009).

Articulation place classes	CI-phones
Coronal	d, l, n, s, t, z
High	ch, ih, iy, jh, sh, uh, uw, y
Dental	dh, th
Glottal	hh
Labial	b, f, m, p, v, w
Low	aa, ae, aw, ay, oy
Mid	ah, eh, ey, ow
Retroflex	er, r
Velar	g, k, ng
Silence	sil

to train visual embedding extractor with the articulation place label in this study, denoted as VEASE-place. We adopt 10 units as in Lee and Siniscalchi (2013) and Siniscalchi and Lee (2009) for articulation place set.

Compared with the phone, the category granularity of articulation place is coarser. Thus, the classification model can achieve comparable performance with lower complexity. And the articulation place has fewer categories, which reduces the labeling costs. Moreover, the articulation place label is believed to be more language-independent than phones, which allows various languages to appear in training and testing.

$P^{\text{truth}}_{\text{phone}}$ is mapped into the frame-level real distribution of articulation place $P^{\text{truth}}_{\text{place}}$ by using Table 1. The same classification backend $f_C(\cdot)$ takes E_V as input and outputs the posterior probability of each articulation place class for each specific time frame P_{place}.

2.5. VEASE model

The VEASE model consists of three stacks of 1D-ConvBlocks and a frozen visual embedding extractor, as shown in Fig. 3. Each 1D-ConvBlock includes a 1D convolution layer with a residual connection, a ReLU activation, and a batch normalization, as in Afouras et al. (2018). Some of the blocks contain an extra up-sampling or down-sampling layer, because the number of audio frames is different from that of the video frames.

Visual embedding E_V is processed by the stack $s_E(\cdot)$ at the bottom left consisting of N_E 1D-ConvBlocks while noisy log-power spectra (LPS) features $A_{\text{LPS}} = \{A^t_{\text{LPS}} \in \mathbb{R}^F; t = 0, 1, \ldots, T_A - 1\}$ are processed by the stack $s_{\text{LPS}}(\cdot)$ at the bottom right consisting of N_{LPS} 1D-ConvBlocks:

$$R_E = s_E(E_V) = \overbrace{\text{ConvBlock}_{1D}(\cdots \text{ConvBlock}_{1D}}^{N_E}(E_V)) \tag{5}$$

$$R_{\text{LPS}} = s_{\text{LPS}}(A_{\text{LPS}}) = \overbrace{\text{ConvBlock}_{1D}(\cdots \text{ConvBlock}_{1D}(}^{N_{\text{LPS}}}A_{\text{LPS}})) \tag{6}$$

where T_A and F denote the number of time frames and frequency bins for spectrogram, respectively. In this study, we use $F = 201$ by default.

R_E and R_{LPS}, which denote outputs of different stacks, are then concatenated along the channel dimension and fed to the top stack $s_F(\cdot)$ consisting of N_F 1D-ConvBlocks. The last convolution layer in the top stack projects the output's dimension into the same one of noisy magnitude spectrogram. Then, the hidden representation is activated by a sigmoid activation to obtain a magnitude mask $M \in \mathbb{R}^{T_A \times F}$:

$$\begin{aligned} M &= \sigma(s_F([R_E, R_{\text{LPS}}])) \\ &= \sigma(\overbrace{\text{ConvBlock}_{1D}(\cdots \text{ConvBlock}_{1D}}^{N_F}([R_E, R_{\text{LPS}}]))) \end{aligned} \tag{7}$$

The values of M range from 0 to 1. In this study, we use $N_E = 10$, $N_{\text{LPS}} = 5$ and $N_F = 15$ by default if there are no special instructions.

To show the effectiveness of embedding on enhancement performance, we also design a competitive no-embedding version of the EASE model which lacks in the stack $s_E(\cdot)$ at the bottom left and the frozen visual embedding extractor, denoted as NoEASE model. The NoEASE model computes M only using the noisy LPS features as inputs:

$$M = \sigma(s_F(s_{\text{LPS}}(A_{\text{LPS}}))) \tag{8}$$

The ideal ratio mask (IRM) (Hummersone, Stokes, & Brookes, 2014) is employed as the learning target, which is widely used in monaural speech enhancement (Wang, Narayanan, & Wang, 2014). IRM $M_{\text{IRM}} \in \mathbb{R}^{T_A \times F}$ is calculated as follows:

$$M_{\text{IRM}} = \left(\frac{C_{\text{PS}}}{C_{\text{PS}} + D_{\text{PS}}}\right)^{\frac{1}{2}} \tag{9}$$

where $C_{\text{PS}} \in \mathbb{R}^{T_A \times F}$ and $D_{\text{PS}} \in \mathbb{R}^{T_A \times F}$ denote power spectrograms of clean speech and noise, respectively.

The mean square error (MSE) $\mathcal{L}_{\text{MSE}}$ between M and M_{IRM} is calculated as the loss function:

$$\mathcal{L}_{\text{MSE}} = \text{MSE}(M, M_{\text{IRM}}) = \sum \|M - M_{\text{IRM}}\|_2^2 \tag{10}$$

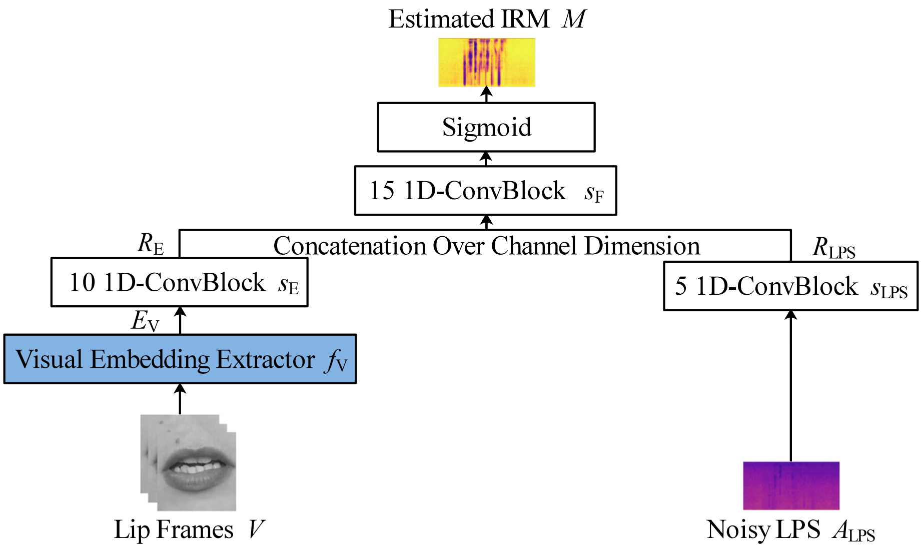

Estimated IRM M
Sigmoid
15 1D-ConvBlock s_{F}
R_{E}
Concatenation Over Channel Dimension
R_{LPS}
10 1D-ConvBlock s_{E}
5 1D-ConvBlock s_{LPS}
E_{V}
Visual Embedding Extractor f_{V}
Lip Frames V
Noisy LPS A_{LPS}

Fig. 3. Illustration of the VEASE model. The VEASE model takes the visual embeddings as the auxiliary inputs except regular noisy LPS features. The visual embedding extractor is pre-trained separately with classification backend, following the steps introduced in the above-mentioned sections. In the training of the VEASE model, the visual embedding extractor is kept frozen.

We use Adam optimizer to train for 100 epochs with early stopping when there is no improvement on the validation loss for 10 epochs. The batch size is 96. Initial learning rate is set to 0.0001, which is found by "LR range test" proposed in Smith (2017), and halved during training if there is no improvement for 3 epochs on the validation loss. The best model is selected by the lowest validation loss.

3. Proposed MEASE model

This section elaborates on the proposed MEASE model. The MEASE model takes the fused audio-visual embedding as the auxiliary input instead of the visual-only embedding. As described in Section 1, the MEASE model utilizes a complementarity of audio and visual features in an information intersection manner. In order to verify the complementarity between audio and visual embeddings, we design an EASE model that utilizes the audio embedding, which is denoted as AEASE model. For verifying the effectiveness of the information intersection-based audio-visual fusion manner on enhancement performance, we design an EASE model that utilizes the concatenation of audio and visual embeddings, which is denoted as cMEASE model. The details of the AEASE model, the MEASE model and the cMEASE model are elaborated in the following.

3.1. AEASE model

The AEASE model has a similar structure to the VEASE model as shown in Fig. 3, with the main difference of employing an audio embedding extractor, instead of the visual embedding extractor.

The audio embedding extractor $f_{\mathrm{A}}(\cdot)$ as shown in Fig. 4 has the similar structure as the visual embedding extractor $f_{\mathrm{V}}(\cdot)$ in Fig. 1. The 3D-kernels in spatiotemporal convolution are replaced by 1D-kernels meanwhile the 3D-MaxPooling layer is dropped in this case as the audio frame is a vector. We also use the ResNet-18 with the main difference of employing 1D-kernels instead of 2D-kernels. Given noisy Mel Filter Bank (FBANK) features $A_{\mathrm{FBANK}} \in \mathbb{R}^{T_{\mathrm{A}} \times F_{\mathrm{mel}}}$, the audio embeddings $E_{\mathrm{A}} \in \mathbb{R}^{T_{\mathrm{A}} \times L_{\mathrm{A}}}$ are calculated as follows:

$$\begin{aligned} E_{\mathrm{A}} &= \{E_{\mathrm{A}}^t \in \mathbb{R}^{L_{\mathrm{A}}}; t = 0, 1, \ldots, T_{\mathrm{A}} - 1\} = f_{\mathrm{A}}(A_{\mathrm{FBANK}}) \\ &= \text{ResNet-18}_{\mathrm{A}}(\mathrm{BN}(\mathrm{ReLU}(\mathrm{Conv}_{1D}(A_{\mathrm{FBANK}})))) \end{aligned} \tag{11}$$

where, F_{mel} and L_{A} are the number of triangular filters set for FBANK features and the length of E_{A}^t, respectively. In this study, $L_{\mathrm{A}} = L_{\mathrm{V}} = 256$ and $F_{\mathrm{mel}} = 40$ are used by default.

We use the same training process as training the visual embedding extractor in Section 2.1 to train the audio embedding extractor. Adam optimizer is used to minimize $\mathcal{L}_{\mathrm{CE}}$, which is calculated by Eq. (2). But P_{place} is computed by using E_{A}:

$$P_{\mathrm{place}} = f_{\mathrm{C}}(E_{\mathrm{A}}) \tag{12}$$

The AEASE model takes both A_{LPS} and E_{A} as inputs and outputs M:

$$M = \sigma(s_{\mathrm{F}}([s_{\mathrm{E}}(E_{\mathrm{A}}), s_{\mathrm{LPS}}(A_{\mathrm{LPS}})])) \tag{13}$$

The same optimization process as in Section 2.5 is also used to minimize $\mathcal{L}_{\mathrm{MSE}}$, which is calculated by Eq. (10).

3.2. MEASE model

The most significant change in the MEASE model is that the visual-only pre-trained embedding extractor evolves into the audio-visual pre-trained embedding extractor. The audio-visual embedding extractor takes not only lip frames but also noisy FBANK features as inputs and outputs the fused audio-visual embedding which is learned under the supervision of the information intersection, i.e., the articulation place label.

The audio-visual embedding extractor consists of visual, audio and fused streams, as shown at the bottom left part of Fig. 5. The visual stream has the same structure as the visual embedding extractor in Section 2.1 while the audio stream has the same structure as the audio embedding extractor in Section 3.1. V and A_{FBANK} are processed by visual and audio streams, respectively:

$$E_{\mathrm{AV}}^{\mathrm{V}} = f_{\mathrm{V}}(V) \tag{14}$$

$$E_{\mathrm{AV}}^{\mathrm{A}} = f_{\mathrm{A}}(A_{\mathrm{FBANK}}) \tag{15}$$

where $E_{\mathrm{AV}}^{\mathrm{V}} \in \mathbb{R}^{T_{\mathrm{V}} \times L_{\mathrm{V}}}$ and $E_{\mathrm{AV}}^{\mathrm{A}} \in \mathbb{R}^{T_{\mathrm{A}} \times L_{\mathrm{A}}}$ denote the outputs of visual and audio streams, respectively. The mismatch in the number of frames between $E_{\mathrm{AV}}^{\mathrm{V}}$ and $E_{\mathrm{AV}}^{\mathrm{A}}$, i.e. $T_{\mathrm{A}} \neq T_{\mathrm{V}}$, is solved by repeating a video frame for several audio frames:

$$\tilde{E}_{\mathrm{AV}}^{\mathrm{V}} = \{\overbrace{E_{\mathrm{AV}}^{\mathrm{V},0}, \ldots, E_{\mathrm{AV}}^{\mathrm{V},0}}^{T_{\mathrm{A}}/T_{\mathrm{V}}}, E_{\mathrm{AV}}^{\mathrm{V},1} \cdots\} \tag{16}$$

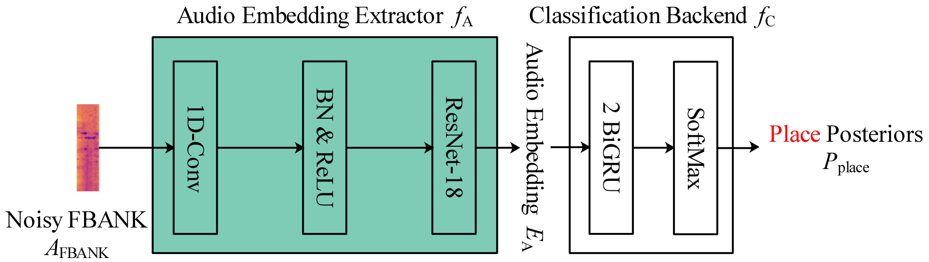

Audio Embedding Extractor f_A
Classification Backend f_C
1D-Conv
BN & ReLU
ResNet-18
Audio Embedding E_A
2 BiGRU
SoftMax
Noisy FBANK A_{FBANK}
Place Posteriors P_{place}

Fig. 4. Illustration of an audio embedding extractor (in color green for ease of cross-referencing in Figs. 4, 5 and 6). The audio embedding extractor has the similar structure and the same training process as the visual embedding extractor in Section 2.1.

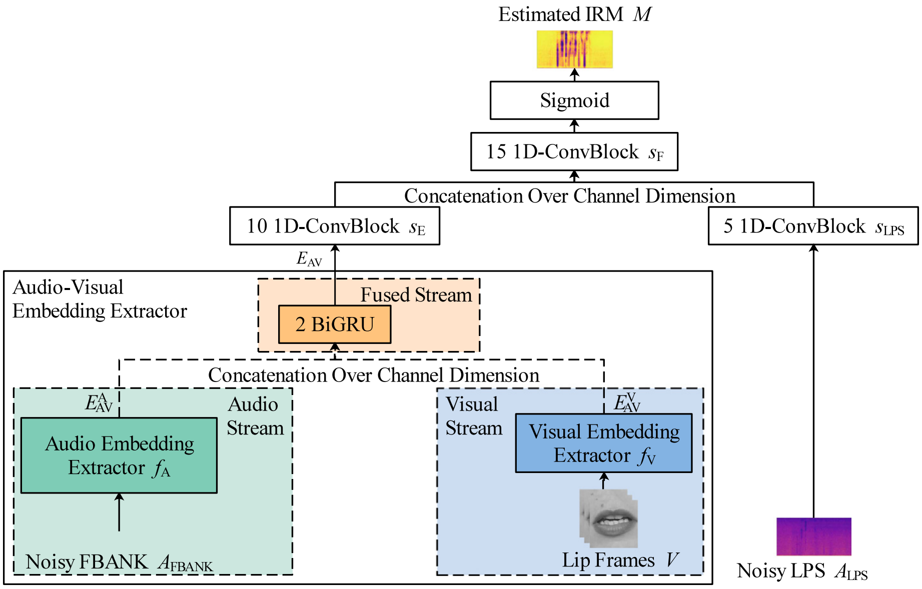


Fig. 5. Illustration of the proposed MEASE model. The previous visual embedding extractor evolves to the audio-visual embedding extractor, which consists of a visual stream (in blue), an audio stream (in green) and a fused stream (in orange). The audio-visual embedding extractor fuses the audio and visual embeddings in an information intersection manner.

The fused stream consisting of a 2-layers BiGRU at the top takes $\tilde{E}^V_{AV}$ and E^A_{AV} as inputs and outputs the audio-visual embedding $E_{AV} \in \mathbb{R}^{T_A \times L_{AV}}$:

$$E_{AV} = \{E^t_{AV}; t = 0, 1, \ldots, T_A - 1\} = \mathrm{BiGRU}(\mathrm{BiGRU}([\tilde{E}^V_{AV}, E^A_{AV}])) \tag{17}$$

where L_{AV} is the length of E^t_{AV}. In this paper, we use $L_{AV} = L_A + L_V = 512$ by default.

We also use the same training steps to minimize $\mathcal{L}_{CE}$, which is calculated by Eq. (2), as these in Section 2.1. But P_{place} is computed by using E_{AV}:

$$P_{place} = f_C(E_{AV}) \tag{18}$$

The audio-visual classification model can achieve a better and faster convergence, by initializing visual and audio streams with the independently pre-trained params, which is a significant finding.

The MEASE model takes both A_{LPS} and E_{AV} as inputs and outputs M:

$$M = \sigma(s_F([s_E(E_{AV}), s_{LPS}(A_{LPS})])) \tag{19}$$

We use the same optimization process as in Section 2.5 to minimize $\mathcal{L}_{MSE}$, which is calculated by Eq. (10).

3.3. cMEASE model

By ablating the fused stream in Fig. 5, another audio-visual embedding, $cE_{AV} \in \mathbb{R}^{T_A \times (L_A + L_V)}$, which is the concatenation of audio and visual embeddings, is designed:

$$cE_{AV} = [E_V, E_A] = [f_V(V), f_A(A_{FBANK})] \tag{20}$$

where f_A and f_V are trained independently, following the steps introduced in Sections 3.1 and 2.1, respectively.

The cMEASE model takes both A_{LPS} and cE_{AV} as inputs and outputs M:

$$M = \sigma(s_F([s_E(cE_{AV}), s_{LPS}(A_{LPS})])) \tag{21}$$

We use the same optimization process as in Section 2.5 to minimize $\mathcal{L}_{MSE}$, which is calculated by Eq. (10).

3.4. Fusion stage of audio and visual embeddings

To gain insight into the effect of the audio-visual fusion stage on enhancement performance under the framework of a neural network, we design a MEASE model that fuses visual and audio embeddings at the ith layer of the enhancement network, denoted as Middle-i model, as shown in Fig. 6. We change N_E with the fixed sum of N_E and N_F and use the same stack to process

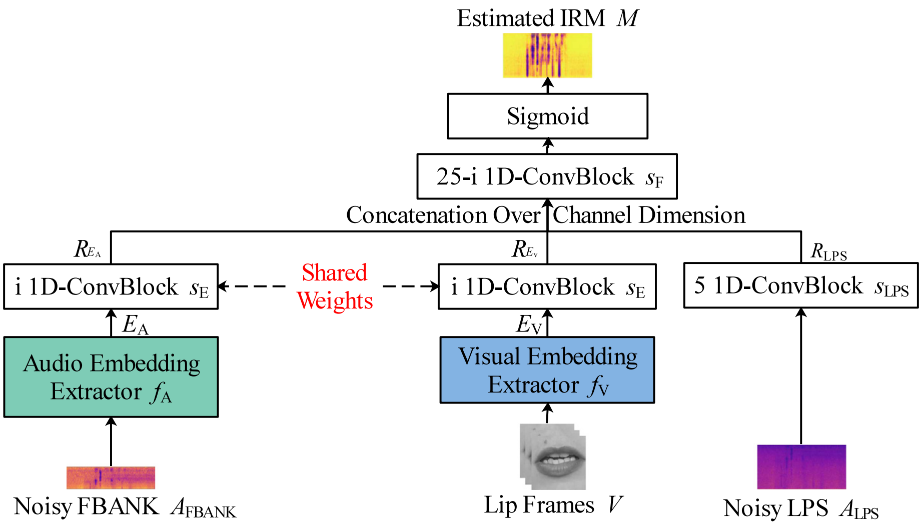


Fig. 6. Illustration of the MEASE model with different fusion stages of audio and visual embeddings.

audio and visual embeddings, respectively:

$$R_{E_V} = s_E(E_V) = \overbrace{\text{ConvBlock}_{1D}(\cdots \text{ConvBlock}_{1D}}^{N_E=i}(E_V)) \tag{22}$$

$$R_{E_A} = s_E(E_A) = \overbrace{\text{ConvBlock}_{1D}(\cdots \text{ConvBlock}_{1D}}^{N_E=i}(E_A)) \tag{23}$$

$$R_{LPS} = s_{LPS}(A_{LPS}) = \overbrace{\text{ConvBlock}_{1D}(\cdots \text{ConvBlock}_{1D}}^{N_{LPS}}(A_{LPS})) \tag{24}$$

$$\begin{aligned} M &= \sigma(s_F([R_{E_V}, R_{E_A}, R_{LPS}])) \\ &= \sigma(\overbrace{\text{ConvBlock}_{1D}(\cdots \text{ConvBlock}_{1D}}^{N_F=25-i}([R_{E_V}, R_{E_A}, R_{LPS}]))) \end{aligned} \tag{25}$$

where $s_E(\cdot)$ in Eq. (22) has the same params as that in Eq. (23), as well as E_A and E_V are extracted by using $f_A(\cdot)$ and $f_V(\cdot)$ trained independently. By modifying the value of i, we can make the fusion take places at different stages without changing the network structure.

4. Experiments

4.1. Dataset

To evaluate the performance of our proposed method, we created a simulation dataset of noisy speech based on the TCD-TIMIT audio-visual corpus (Harte & Gillen, 2015). The TCD-TIMIT consisted of 59 volunteer speakers with around 98 videos each, as well as 3 lipspeakers who specially were trained to speak in a way that helped the deaf understand their visual speech. The speakers were recorded saying various sentences from the TIMIT corpus (Garofolo, Lamel, Fisher, Fiscus, & Pallett, 1993) by using both front-facing and 30-degree cameras. However, the utterances of 3 lipspeakers and 30-degree videos were not used in this paper. With a view to test the robustness to unseen speaker condition, we divided these videos and audios into a *train–clean* set which consisted of 57 speakers (31 male and 26 female) and a *test–clean* set which consisted of 2 speakers (1 male and 1 female) who were not in the *train–clean* set.

We chose the TCD-TIMIT dataset for two main reasons:

(1) TCD-TIMIT was recorded in a controlled environment, and provided near-field signals collected by a microphone close to the mouth, which can ensure that the utterances do not contain background noise. While other large-scale in-the-wild audio-visual datasets, such as BBC-Oxford LipReading Sentences 2 (LRS2) dataset (Chung, Senior, Vinyals, & Zisserman, 2017), AVSpeech dataset (Ephrat et al., 2018), etc, were collected from real-world sources using automated pipeline, and none of them was checked whether background noise exists.[2] When testing an enhancement system, if the ground truth contains background noise, the metrics will be severely distorted and cannot well measure the performance of the system.

(2) The utterances consisted of various phrases in the TCD-TIMIT dataset, and thus they were more suitable for actual scenarios than the utterances consisting of a fixed set of phrases in the GRID dataset (Cooke, Barker, Cunningham, & Shao, 2006). The TCD-TIMIT dataset also contained phonetic-level transcriptions, which provided available labels for the embedding extractor training.

A total of 115 noise types, including 100 noise types in Hu and Wang (2010) and 15 homemade noise types, were adopted for training to improve the robustness to unseen noise types. The 5600 utterances from *train–clean* set were corrupted with the above-mentioned 115 noise types at five levels of SNRs, i.e. 15 dB, 10 dB, 5 dB, 0 dB and −5 dB, to build a 35-hour multi-condition training set consisting of pairs of clean and noisy utterances. The other 43 utterances from *train–clean* set were corrupted with 3 unseen noise types at above-mentioned SNR levels to build a validation set, i.e. Destroyer Operations, Factory2 and F-16 Cockpit. The 198 utterances from *test–clean* set were used to construct a test set for each combination of 3 other unseen noise type and above SNR levels, i.e. Destroyer Engine, Factory1 and Speech Babble. All unseen noise were collected from the NOISEX-92 corpus (Varga & Steeneken, 1993). The five levels of SNRs in the training set were also adopted for testing and validating.

For audio preprocessing, all speech signals were resampled to 16 kHz. A 400-point short-time Fourier transform was used to compute the spectra of each overlapping windowed frame. Here, a 25-ms Hanning window and a 10-ms window shift were adopted. In the experiments, 201-dimensional LPS vectors were generated to train the EASE network and 40-dimensional FBANK vectors were generated to train the embedding extractor, i.e.

[2] We manually listen to the test and verification sets of the LRS2 dataset. We find more than half of sentences can be clearly perceived as noisy.

Table 2
Average performance comparison of VEASE models with different visual embeddings 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
Noisy	1.70	1.97	2.26	2.56	2.86	54.34	65.11	75.33	84.48	90.88
NoEASE	2.07	2.34	2.64	2.92	3.21	58.79	70.29	80.24	87.83	92.57
VEASE-word	2.16	2.45	2.72	2.99	3.25	66.26	75.11	82.57	88.75	92.98
VEASE-phone	2.14	2.42	2.69	2.96	3.23	66.29	74.89	82.22	88.45	92.79
VEASE-place	2.21	2.47	2.73	3.00	3.26	66.57	75.27	82.64	88.80	92.96

$F = 201, F_{\mathrm{mel}} = 40$. Mean and variance normalizations were applied to the noisy LPS and FBANK vectors.

As for video preprocessing, a given video clip was downsampled from 29.97 fps to 25 fps, i.e. $T_A = 4 \times T_V$. For every video frame, 68 facial landmarks were extracted by using Dlib (King, 2009) implementation of the face landmark estimator described in Kazemi and Sullivan (2014), then we cropped a lip-centered window of size 98 × 98 pixels by using the 20 lip landmarks from the 68 facial landmarks. The frames were transformed to grayscale and normalized with respect to the overall mean and variance.

4.2. Evaluation protocol

In this experiment, we adopt Perceptual Evaluation of Speech Quality (PESQ) (Rix, Beerends, Hollier, & Hekstra, 2001) and Short-Time Objective Intelligibility (STOI) (Taal, Hendriks, Heusdens, & Jensen, 2011) as major means to evaluate models. Both metrics are commonly used to evaluate the performance of speech enhancement system. PESQ, which is a speech quality estimator, is designed to predict the mean opinion score of a speech quality listening test for certain degradations. Moreover, STOI was calculated to show the improvement in speech intelligibility. The STOI score is typically between 0 and 1, and the PESQ score is between −0.5 and 4.5. For both metrics, higher scores indicate better performance.

4.3. Results of VEASE models utilizing different visual embeddings

In Section 2, we proposed two VEASE models with different visual embeddings, i.e. VEASE-phone and VEASE-place. To compare their effectiveness with the baseline model, i.e. VEASE-word (LRW), on enhancement performance, a series of experiments was conducted for the unprocessed system denoted as Noisy, NoEASE, VEASE-word (LRW), VEASE-phone and VEASE-place. We present the learning curves of the MSEs among NoEASE, VEASE-word (LRW) and VEASE-phone and VEASE-place on the validation set in Fig. 7. The corresponding evaluation metrics are shown in Table 2. We evaluate the average performance of two measures at different SNRs across 3 unseen noise types.

Based on Fig. 7 and Table 2, we find following observations.

(1) The learning curves of the MSEs indicate that all VEASE models consistently generate smaller MSEs on the validation set than NoEASE. This result implies that the visual embedding is useful for speech enhancement. As shown in Table 2, all VEASE models yield improvements in PESQ and STOI over NoEASE in all SNRs. In particular, the improvement is more significant at low SNRs cases.

(2) VEASE-phone yields a slower convergence and similar MSE values than VEASE-word (LRW), which is consistent with the comparison of the objective evaluation metrics on the test set shown in Table 2. Even though the visual embedding extractor in VEASE-word (LRW) is trained with a large amount of additional video data (around 170h), it does not yield a significant improvement on the enhancement performance. The data mismatch between embedding learning and enhancement task brings information redundancy, which reduces the effectiveness of representation. This observation supports us to adopt the matched data to training the embedding extractor and the enhancement network.

(3) VEASE-place clearly achieves a better and faster convergence than VEASE-phone and VEASE-word (LRW). This implies that VEASE-place provides more useful and quick-fit visual embedding for speech enhancement. By comparing the evaluation metrics in Table 2, we also observe that VEASE-place not only yields remarkable gains over VEASE-phone across all evaluation metrics and on all SNR levels, but also outperforms VEASE-word (LRW) in most cases with only one exception for the STOI at 15 dB SNR. The results are still close, despite the situation being an exception. These results suggest that our proposed VEASE-place model achieves a better generalization capability, and at the same time reduces mismatch between embedding learning and enhancement task.

Overall, the high correlation between the articulation place label and the acoustic information in video is beneficial to the extraction of visual embedding, which is useful for speech enhancement, even if no requirement of additional data. Therefore, we select articulation place as the default classification target in all subsequent experiments and use VEASE to refer to VEASE-place in all subsequent sections.

4.4. Results of proposed MEASE model

In this section, the goal is to examine the effectiveness of the proposed MEASE model on enhancement performance, and obtain a better understanding about the contribution of different parts of the MESAE model. We present an average performance comparison between NoEASE, VEASE, AEASE, cMEASE and MEASE in Table 3. The last row of this table marked with "p-value" is the minimum value of p where a significant difference can be observed at the level of p in the statistical significance tests for MEASE and VEASE models. Here we adopt the "Matched Pair Test" method mentioned in Pallet, Fisher, and Fiscus (1990). The significance test is a two-tailed test with the null hypothesis that there is no performance difference between the two models. The smaller the "p-value" is, the bigger the significant differences between two models are.

Based on the results in Table 3, we can observe that MEASE shows significant improvements over VEASE across all evaluation metrics, and larger gains are observed at high SNRs. By comparing the results of VEASE with NoEASE, the improvement yielded by visual embedding decreases as SNR increases, for example, the PESQ of VEASE increased from 2.07 to 2.21 at −5 dB SNR and from 3.21 to 3.26 at 15 dB SNR. This observation is consistent with that in Wang et al. (2020). In contrast, MEASE shows stable improvements over NoEASE for high SNRs. For example, the PESQ of MEASE increased from 2.07 to 2.29 at −5 dB SNR and from 3.21 to 3.42 at 15 dB SNR. All these results indicate that MEASE is more robust against the change of noise level and yields better generalization capability than VEASE.

Table 3 also shows the results of AEASE. By comparing its results with NoEASE, we can observe that the improvement yielded by audio embedding increases as SNR grows, for example, the PESQ of AEASE increased from 2.07 to 2.09 at −5 dB SNR and from 3.21 to 3.27 at 15 dB SNR. This suggests that the complementarity between audio and visual embeddings lies in the variation tendencies of metric improvement with respect to SNR

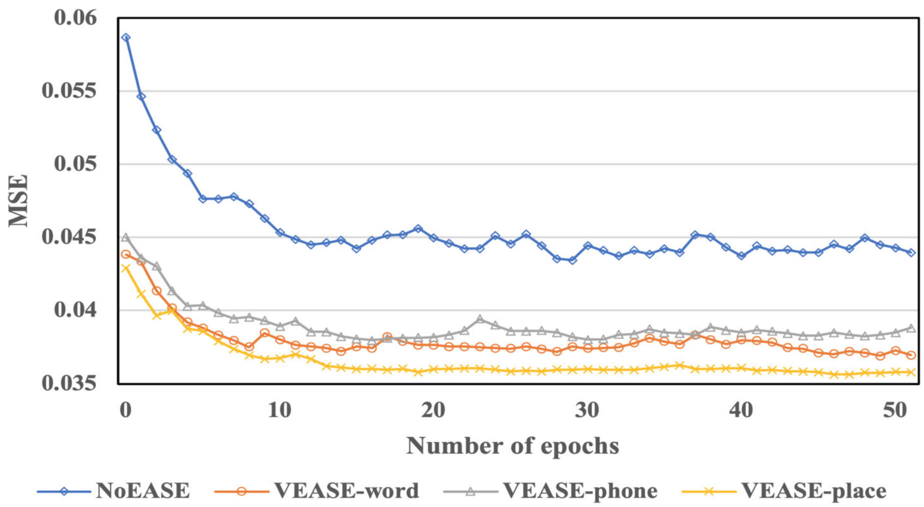


Fig. 7. A comparison of learning curves among NoEASE, VEASE-word (LRW), VEASE-phone and VEASE-place on the validation set.

Table 3
Average performance comparison of NoEASE model, VEASE model, AEASE model, cMEASE model and MEASE model on the test set at different SNRs averaged over 3 unseen noise types. The *p*-value row indicates significance test results between VEASE model and MEASE model.

Model	PESQ					STOI (in %)				
SNR (in dB)	−5	0	5	10	15	−5	0	5	10	15
NoEASE	2.07	2.34	2.64	2.92	3.21	58.79	70.29	80.24	87.83	92.57
VEASE	2.21	2.47	2.73	3.00	3.26	66.57	75.27	82.64	88.80	92.96
AEASE	2.09	2.39	2.69	2.98	3.27	60.84	72.24	81.58	88.39	92.76
cMEASE	2.27	2.55	2.81	3.08	3.34	67.60	76.26	83.26	89.13	93.12
MEASE	2.29	2.59	2.88	3.16	3.42	68.96	77.64	84.43	89.99	93.64
p-value	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001

level. Nevertheless by directly comparing AEASE and VEASE on the evaluation metrics as shown in Table 3, we cannot observe that AEASE performs better than VEASE at high SNRs, i.e. SNR = 5, 10 and 15 dB, especially at 5 dB SNR.

To further explore the complementarity between audio and visual embeddings, we present an average performance comparison between utterance segments belonging to different articulation places in Table 4. Because the utterance segment does not have actual semantics, we only examine the average performance of PESQ at different SNRs across 3 unseen noise types. Table 4 illustrates that VEASE and AEASE play a major role in different articulation places, respectively, at the same SNR level. Even at high SNRs, VEASE still yields improvement than AEASE in some articulation places. For example, VEASE's PESQ values are 2.23, 2.73, 2.42, while AEASE's PESQ values are 2.18, 2.72, 2.39 in Labial, Mid, High at 5 dB SNR level. This result explains why AEASE does not outperform VEASE at high SNR levels. Relating to the lip shapes belonging to different articulation places, as shown in Fig. 2, we find VEASE yields greater improvement at articulation places where the lip shapes change greatly, i.e. Labial, Mid and High, while AEASE is on the contrary. Therefore we can conclude that the complementarity between audio and visual embeddings lies in different SNR levels, as well as different articulation places. More specifically, in the cases where the SNR level is low and the articulation place has high visual correlation, visual embedding performs better. And audio embedding is better on articulation places with low visual correlation at high SNR levels. Based on these observations, our proposed MEASE model takes the advantages of visual and audio embeddings, and achieves the best performance in all SNRs and all articulation places.

The information intersection-based audio-visual fusion manner in the MEASE model is our another contribution. From Table 3, we can observe that MEASE consistently outperforms cMEASE at all SNR levels in terms of all 2 measures, especially at high SNRs. This observation demonstrates that the information intersection-based audio-visual fusion method has better information integration capability for audio and visual embeddings than channel-wise concatenation which is widely used in previous works.

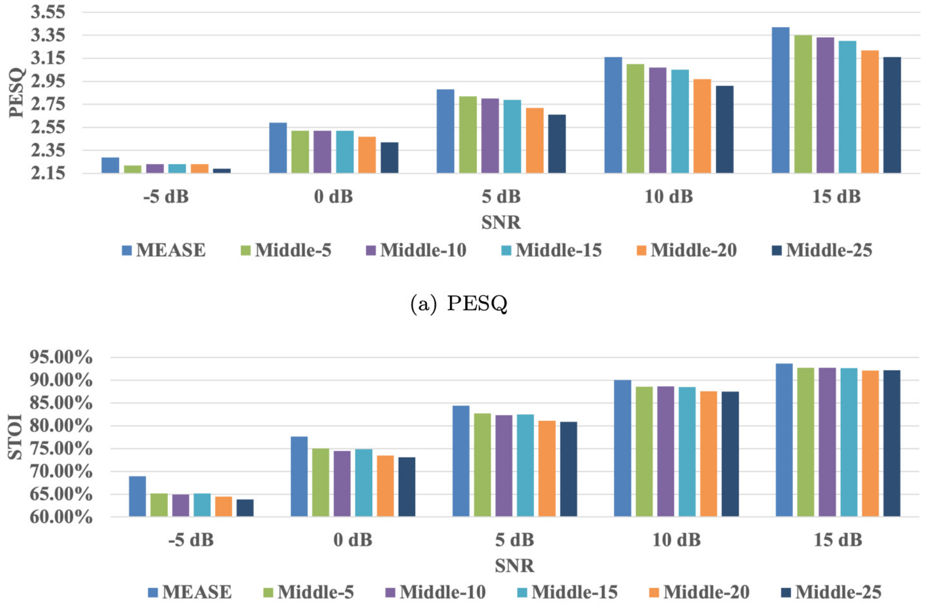
4.5. Results of different audio-visual fusion stages

One of the most significant differences between our method and previous methods is that the proposed MEASE model fuses audio and visual modes in the stage of embedding extractor training. It is an early fusion in contrast to previous methods that fuse audio and visual modes in the middle of the enhancement network, which is also known as the middle fusion. With the aim of verifying the effectiveness of the early fusion on enhancement performance under the framework of neural network, we design an experimental comparative study described in Section 3.4 and conduct a set of experiments using five different i, i.e. $i = 5, 10, 15, 20, 25$.

As we can see from Fig. 8, MEASE achieves the best results over all models utilizing the middle fusion across all evaluation metrics for all SNR levels. By comparing the results of different middle fusion-based models, the variation tendencies of all objective metrics with respect to different fusion stages get worse as the stage moves back. These results suggest that early fusion strategy can better integrate useful information for the neural network-based speech enhancement from both modalities than the standard fusion which happens at the middle layer of enhancement network.

5. Conclusion

In this study, we extend the previous audio-visual speech enhancement (AVSE) framework to embedding aware speech



(a) PESQ

(b) STOI

Fig. 8. Average performance comparison among different audio-visual fusion stages for the PESQ/STOI measures at different SNRs averaged over 3 unseen noise types. The top figure shows the PESQ measure. The bottom figure shows the STOI measure.

Table 4
Average performances of different models on the test set at different SNRs and different articulation places averaged over 3 unseen noise types.

SNR (in dB)	−5				0				5			
Place	Model											
	NoEASE	AEASE	VEASE	MEASE	NoEASE	AEASE	VEASE	MEASE	NoEASE	AEASE	VEASE	MEASE
Labial	1.28	1.38	1.58	1.76	1.57	1.75	1.81	2.06	2.05	2.18	2.23	2.50
Mid	1.54	1.68	1.86	2.02	2.03	2.21	2.29	2.45	2.58	2.72	2.73	2.96
High	1.38	1.52	1.65	1.81	1.79	1.95	1.99	2.17	2.28	2.39	2.42	2.62
Low	1.63	1.89	2.00	2.29	2.17	2.48	2.46	2.69	2.84	2.99	2.93	3.20
Retroflex	1.46	1.66	1.75	2.00	1.95	2.15	2.12	2.32	2.44	2.57	2.54	2.77
Coronal	1.59	1.74	1.80	1.93	1.92	2.07	2.05	2.23	2.30	2.39	2.35	2.56
Glottal	1.02	1.22	1.36	1.70	1.42	1.71	1.59	1.92	1.95	2.10	2.05	2.30
Velar	1.31	1.44	1.41	1.49	1.48	1.64	1.68	1.86	1.86	2.01	2.00	2.22
Dental	0.94	1.22	1.25	1.64	1.32	1.62	1.36	2.05	1.98	2.21	1.98	2.44

enhancement (EASE). We first propound visual embedding to enhance speech, leveraging upon the high correlation between articulation place labels and acoustic information in videos. Next, we propose multi-modal audio-visual embedding obtained by fusing audio and visual embeddings in the stage of embedding extractor training under the supervision of their information intersection at the articulation place label level.

Extensive experiments empirically validate that our proposed visual embedding consistently yields improvements over the conventional word-based approaches. And our proposed audio-visual embedding achieves even greater performance improvements by utilizing the complementarity of audio and visual embeddings in an information intersection-based way, with higher information integration capabilities and better speech enhancement performance in early fusion.

Future work will include exploring more modeling units for training embedding extractor, especially visme-based unit, and researching effective training methods for the joint training of the embedding extractor and enhancement network, e.g. (1) initialization with respective pre-trained parameters, and (2) multi-task learning. In addition, we plan to build a larger audio-visual dataset and conduct an evaluation for our method.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported by the Strategic Priority Research Program of Chinese Academy of Sciences under Grant No.XDC080-50200.

References

Abdelaziz, A. H., Zeiler, S., & Kolossa, D. (2013). Twin-HMM-based audio-visual speech enhancement. In *2013 IEEE international conference on acoustics, speech and signal processing* (pp. 3726–3730). http://dx.doi.org/10.1109/ICASSP.2013.6638354.

Afouras, T., Chung, J. S., & Zisserman, A. (2018). The conversation: Deep audio-visual speech enhancement. In *Proc. interspeech 2018* (pp. 3244–3248). http://dx.doi.org/10.21437/Interspeech.2018-1400.

Bernstein, L., & Benoit, C. (1996). 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 vol.3). http://dx.doi.org/10.1109/ICSLP.1996.607895.

Blum, A., & Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on computational learning theory* (pp. 92–100). http://dx.doi.org/10.1145/279943.279962.

Boll, S. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech and Signal Processing*, *27*(2), 113–120.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America*, *25*(5), 975–979.

Chung, J. S., Senior, A., Vinyals, O., & Zisserman, A. (2017). Lip reading sentences in the wild. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 3444–3453). http://dx.doi.org/10.1109/CVPR.2017.367.

Cohen, I. (2003). Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging. *IEEE Transactions on Speech and Audio Processing*, *11*(5), 466–475. http://dx.doi.org/10.1109/TSA.2003.811544.

Cohen, I., & Berdugo, B. (2001). Speech enhancement for non-stationary noise environments. *Signal Processing*, *81*(11), 2403–2418.

Cooke, M., Barker, J., Cunningham, S., & Shao, X. (2006). An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, *120*(5), 2421–2424. http://dx.doi.org/10.1121/1.2229005.

Dasgupta, S., Littman, M. L., & McAllester, D. A. (2002). Pac generalization bounds for co-training. In *Advances in neural information processing systems* (pp. 375–382).

Deligne, S., Potamianos, G., & Neti, C. (2002). Audio-visual speech enhancement with AVCDCN (audio-visual codebook dependent cepstral normalization). In *Sensor array and multichannel signal processing workshop proceedings, 2002* (pp. 68–71). http://dx.doi.org/10.1109/SAM.2002.1191001.

Ephraim, Y., & Malah, D. (1985). Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech and Signal Processing*, *33*(2), 443–445. http://dx.doi.org/10.1109/TASSP.1985.1164550.

Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., et al. (2018). Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation.. *ACM Transactions on Graphics*, *37*(4), 112:1–112:11.

Fisher III, J. W., Darrell, T., Freeman, W. T., & Viola, P. A. (2001). Learning joint statistical models for audio-visual fusion and segregation. In *Advances in neural information processing systems* (pp. 772–778).

Gabbay, A., Ephrat, A., Halperin, T., & Peleg, S. (2018). Seeing through noise: Visually driven speaker separation and enhancement. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 3051–3055). http://dx.doi.org/10.1109/ICASSP.2018.8462527.

Gabbay, A., Shamir, A., & Peleg, S. (2018). Visual speech enhancement. In *Proc. interspeech 2018* (pp. 1170–1174). http://dx.doi.org/10.21437/Interspeech.2018-1955.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., & Pallett, D. S. (1993). DARPA Timit acoustic-phonetic continous speech corpus CD-rom. NIST speech disc 1-1.1. *STIN*, *93*, 27403.

Girin, L., Feng, G., & Schwartz, J.-L. (1995). Noisy speech enhancement with filters estimated from the speaker's lips. In *Fourth european conference on speech communication and technology* (pp. 1559–1562).

Girin, L., Schwartz, J., & Feng, G. (2001). Audio-visual enhancement of speech in noise. *The Journal of the Acoustical Society of America*, *109*(6), 3007–3020.

Goecke, R., Potamianos, G., & Neti, C. (2002). Noisy audio feature enhancement using audio-visual speech data. In *2002 IEEE international conference on acoustics, speech, and signal processing, Vol. 2* (pp. II–2025–II–2028). http://dx.doi.org/10.1109/ICASSP.2002.5745030.

Gu, R., Zhang, S., Xu, Y., Chen, L., Zou, Y., & Yu, D. (2020). Multi-modal multi-channel target speech separation. *IEEE Journal of Selected Topics in Signal Processing*, *14*(3), 530–541. http://dx.doi.org/10.1109/JSTSP.2020.2980956.

Harte, N., & Gillen, E. (2015). Tcd-timit: An audio-visual corpus of continuous speech. *IEEE Transactions on Multimedia*, *17*(5), 603–615. http://dx.doi.org/10.1109/TMM.2015.2407694.

He, K., Zhang, X., Ren, S., & Sun, J. (2016a). Deep residual learning for image recognition. In *2016 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 770–778). http://dx.doi.org/10.1109/CVPR.2016.90.

He, K., Zhang, X., Ren, S., & Sun, J. (2016b). Identity mappings in deep residual networks. In *European conference on computer vision* (pp. 630–645).

Hershey, J., & Casey, M. (2002). Audio-visual sound separation via hidden Markov models. In T. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing systems, Vol. 14*. MIT Press.

Hou, J.-C., Wang, S.-S., Lai, Y.-H., Tsao, Y., Chang, H.-W., & Wang, H.-M. (2018). Audio-visual speech enhancement using multimodal deep convolutional neural networks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, *2*(2), 117–128. http://dx.doi.org/10.1109/TETCI.2017.2784878.

Hu, G., & Wang, D. (2010). A tandem algorithm for pitch estimation and voiced speech segregation. *IEEE Transactions on Audio, Speech, and Language Processing*, *18*(8), 2067–2079. http://dx.doi.org/10.1109/TASL.2010.2041110.

Hummersone, C., Stokes, T., & Brookes, T. (2014). 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).

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

Iuzzolino, M. L., & Koishida, K. (2020). AV (SE) 2: Audio-visual squeeze-excite speech enhancement. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 7539–7543). http://dx.doi.org/10.1109/ICASSP40776.2020.9054528.

Kazemi, V., & Sullivan, J. (2014). One millisecond face alignment with an ensemble of regression trees. In *2014 IEEE conference on computer vision and pattern recognition* (pp. 1867–1874). http://dx.doi.org/10.1109/CVPR.2014.241.

King, D. E. (2009). Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, *10*, 1755–1758.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *3rd international conference on learning representations, ICLR 2015*.

Lee, C.-H., Clements, M. A., Dusan, S., Fosler-Lussier, E., Johnson, K., Juang, B.-H., et al. (2007). An overview on automatic speech attribute transcription (ASAT). In *Eighth annual conference of the international speech communication association* (pp. 1825–1828).

Lee, C.-H., & Siniscalchi, S. M. (2013). An information-extraction approach to speech processing: Analysis, detection, verification, and recognition. *Proceedings of the IEEE*, *101*(5), 1089–1115. http://dx.doi.org/10.1109/JPROC.2013.2238591.

Leskes, B. (2005). The value of agreement, a new boosting algorithm. In *International conference on computational learning theory* (pp. 95–110). Springer Berlin Heidelberg.

Lewis, D. D. (1998). Naive (Bayes) at forty: The independence assumption in information retrieval. In *European conference on machine learning* (pp. 4–15). Springer Berlin Heidelberg.

Li, H., Ma, B., & Lee, C.-H. (2007). A vector space modeling approach to spoken language identification. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(1), 271–284. http://dx.doi.org/10.1109/TASL.2006.876860.

Li, W., Siniscalchi, S. M., Chen, N. F., & Lee, C. (2016). Improving non-native mispronunciation detection and enriching diagnostic feedback with DNN-based speech attribute modeling. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 6135–6139). http://dx.doi.org/10.1109/ICASSP.2016.7472856.

Li, J., Tsao, Y., & Lee, C.-H. (2005). A study on knowledge source integration for candidate rescoring in automatic speech recognition. In *Proceedings. (ICASSP '05). IEEE international conference on acoustics, speech, and signal processing, 2005., Vol. 1* (pp. I/837–I/840). http://dx.doi.org/10.1109/ICASSP.2005.1415244, Vol. 1.

Lim, J., & Oppenheim, A. (1978). All-pole modeling of degraded speech. *IEEE Transactions on Acoustics, Speech and Signal Processing*, *26*(3), 197–210. http://dx.doi.org/10.1109/TASSP.1978.1163086.

Livescu, K., Cetin, O., Hasegawa-Johnson, M., King, S., Bartels, C., Borges, N., et al. (2007). Articulatory feature-based methods for acoustic and audio-visual speech recognition: Summary from the 2006 JHU summer workshop. In *2007 IEEE international conference on acoustics, speech and signal processing - ICASSP '07, Vol. 4* (pp. IV–621–IV–624). http://dx.doi.org/10.1109/ICASSP.2007.366989.

Loizou, P. C. (2013). *Speech enhancement: theory and practice*. CRC press.

Lu, R., Duan, Z., & Zhang, C. (2019). Audio–visual deep clustering for speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *27*(11), 1697–1712. http://dx.doi.org/10.1109/TASLP.2019.2928140.

Luo, Y., & Mesgarani, N. (2019). Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *27*(8), 1256–1266. http://dx.doi.org/10.1109/TASLP.2019.2915167.

MacLeod, A., & Summerfield, Q. (1987). Quantifying the contribution of vision to speech perception in noise. *British Journal of Audiology*, *21*(2), 131–141.

Massaro, D. W., & Simpson, J. A. (2014). *Speech perception by ear and eye: a paradigm for psychological inquiry*. Psychology Press.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746–748.

Narayanan, A., & Wang, D. (2013). Ideal ratio mask estimation using deep neural networks for robust speech recognition. In *2013 IEEE international conference on acoustics, speech and signal processing* (pp. 7092–7096). http://dx.doi.org/10.1109/ICASSP.2013.6639038.

Pallet, D., Fisher, W. M., & Fiscus, J. G. (1990). Tools for the analysis of benchmark speech recognition tests. In *International conference on acoustics, speech, and signal processing* (pp. 97–100). http://dx.doi.org/10.1109/ICASSP.1990.115546, vol.1.

Petridis, S., Stafylakis, T., Ma, P., Cai, F., Tzimiropoulos, G., & Pantic, M. (2018). End-to-end audiovisual speech recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 6548–6552). http://dx.doi.org/10.1109/ICASSP.2018.8461326.

Rix, A. W., Beerends, J. G., Hollier, M. P., & Hekstra, A. P. (2001). 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). http://dx.doi.org/10.1109/ICASSP.2001.941023.

Rosenblum, L. D. (2008). Speech perception as a multimodal phenomenon. *Current Directions in Psychological Science*, *17*(6), 405–409.

Saenko, K., Darrell, T., & Glass, J. R. (2004). Articulatory features for robust visual speech recognition. In *Proceedings of the 6th international conference on multimodal interfaces* (pp. 152–158). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/1027933.1027960.

Siniscalchi, S. M., & Lee, C.-H. (2009). A study on integrating acoustic-phonetic information into lattice rescoring for automatic speech recognition. *Speech Communication*, *51*(11), 1139–1153.

Siniscalchi, S. M., Reed, J., Svendsen, T., & Lee, C.-H. (2013). Universal attribute characterization of spoken languages for automatic spoken language recognition. *Computer Speech and Language*, *27*(1), 209–227. http://dx.doi.org/10.1016/j.csl.2012.05.001.

Smith, L. N. (2017). Cyclical learning rates for training neural networks. In *2017 IEEE winter conference on applications of computer vision (WACV)* (pp. 464–472). http://dx.doi.org/10.1109/WACV.2017.58.

Stafylakis, T., & Tzimiropoulos, G. (2017). Combining residual networks with LSTMs for lipreading. In *Proc. interspeech 2017* (pp. 3652–3656). http://dx.doi.org/10.21437/Interspeech.2017-85.

Sun, X., Xu, Y., Cao, P., Kong, Y., Hu, L., Zhang, S., et al. (2020). Tcgm: An information-theoretic framework for semi-supervised multi-modality learning. In *European conference on computer vision* (pp. 171–188).

Taal, C. H., Hendriks, R. C., Heusdens, R., & Jensen, J. (2011). An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(7), 2125–2136. http://dx.doi.org/10.1109/TASL.2011.2114881.

Varga, A., & Steeneken, H. J. M. (1993). Assessment for automatic speech recognition: Ii. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Communication*, *12*(3), 247–251. http://dx.doi.org/10.1016/0167-6393(93)90095-3.

Wang, D., & Chen, J. (2018). Supervised speech separation based on deep learning: An overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26*(10), 1702–1726. http://dx.doi.org/10.1109/TASLP.2018.2842159.

Wang, Y., Narayanan, A., & Wang, D. (2014). On training targets for supervised speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *22*(12), 1849–1858. http://dx.doi.org/10.1109/TASLP.2014.2352935.

Wang, W., Xing, C., Wang, D., Chen, X., & Sun, F. (2020). A robust audio-visual speech enhancement model. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 7529–7533). http://dx.doi.org/10.1109/ICASSP40776.2020.9053033.

Wu, J., Xu, Y., Zhang, S.-X., Chen, L.-W., Yu, M., Xie, L., et al. (2019). Time domain audio visual speech separation. In *2019 IEEE automatic speech recognition and understanding workshop (ASRU)* (pp. 667–673). http://dx.doi.org/10.1109/ASRU46091.2019.9003983.

Xu, Y., Du, J., Dai, L.-R., & Lee, C.-H. (2015). A regression approach to speech enhancement based on deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *23*(1), 7–19. http://dx.doi.org/10.1109/TASLP.2014.2364452.