VSET: A MULTIMODAL TRANSFORMER FOR VISUAL SPEECH ENHANCEMENT

Karthik Ramesh[†*] *Wupeng Wang*[†] *Chao Xing*[†] *Dong Wang*[⋆⋆] *Xiao Chen*[†]

[†] Huawei Noah's Ark Lab
[⋆⋆] Center for Speech and Language Technology (CSLT), Tsinghua University

ABSTRACT

The transformer architecture has shown great capability in learning long-term dependency and works well in multiple domains. However, transformer has been less considered in audio-visual speech enhancement (AVSE) research, partly due to the convention that treats speech enhancement as a short-time signal processing task. In this paper, we challenge this common belief and show that an audio-visual transformer can significantly improve AVSE performance, by learning the long-term dependency of both intra-modality and inter-modality. We test this new transformer-based AVSE model on the GRID and AVSpeech datasets, and show that it beats several state-of-the-art models by a large margin.

Index Terms— Transformer, Audio-Visual Speech Enhancement, Attention Mechanism

1. INTRODUCTION

Audio visual speech enhancement (AVSE) is inspired by speech perception studies [1, 2, 3, 4]. These studies showed that with assistance from facial information, humans could do significantly better on auditory perception tasks than if only audio signals were available. Most of the modern AVSE approaches are based on deep neural networks (DNNs) due to their great potential in integrating multiple information sources, thereby simulating the multi-modality processing capability of humans.

Almost all existing DNN-based AVSE models assume a local correspondence (alignment) between the audio and visual streams. For example, Gabbay et al. [5] proposed an encoder-decoder architecture where 200 ms of audio signal and the corresponding visual signal are mapped to a shared embedding space through different encoders and the decoder generates clean speech based on the concatenated embedding of the two streams. Hou et.al [6] built a similar model that maps the audio and video features through CNNs, however the decoder reconstructs not only the clean speech but also the mouth images. Ephrat et.al [7] introduced a Looking-to-Listen model, which produces embeddings of audio and visual signals and concatenates them to feed into a bidirectional LSTM to predict complex masks that derive clean speech. Gogate [8] followed the same idea and presented an architecture called CochleaNet. Wang et.al [9] introduced an online visual speech enhancement architecture that has a loose coupling connection between the audio and visual components and uses a stacked unidirectional LSTM to predict the ideal ratio masks(IRMs) that is applied to noisy speech to derive clean speech.

In all of the above AVSE models, the enhancement is based on local information from both audio and visual streams, and long-term information, e.g., linguistic knowledge or prosody patterns, is simply ignored. This is certainly not ideal and is not the way that humans do. From daily experience, it is clear that we heavily rely on long-term information in perception tasks. More concrete evidence is from controlled perceptual studies. For example, Kalikow et al. [10] showed that words in a syntactic and semantic constrained context can be more easily identified by humans than in an unconstrained context. The same conclusion was found by Boothroyd et al. [11], where a mathematical relation was discovered between the recognition performance in constrained and unconstrained contexts. Minematsu et al. [12] found that people use prosody features to perceive spoken words.

The importance of long-term information has been noticed by researchers in the field of audio-only speech enhancement. Early research employs recurrent neural networks(RNN [13]) to capture this information [14]. Recently, transformer [15] was proposed to increase the perceptual field [16, 17]. Compared to RNN, the transformer model employs a self-attention mechanism to select the most important context in parallel, therefore avoiding the limited remembrance problem suffered by autoregressive models such as RNN.

In this paper, we employ the transformer technique to audio-visual speech enhancement. Our argument is that in the audio-visual condition, long-term dependence is more important than in the audio-only condition. One reason is that the audio and visual modalities are naturally asynchronous. For example, Schwartz et al. [18] showed clear asynchrony between audio and visual frames, varying between 20ms audio lead to 70ms audio lag. This means that we have to look into a broader context in order to integrate the corresponding visual knowledge. Moreover, as has been shown in the previous study [9], local visual information is often weak and ambiguous, and one needs a sequence of visual frames to infer the pronunciation action. Following this motivation, we present a Visual Speech Enhancement Transformer(VSET) model to exploit the long-term information from both audio and vi-

[*] Work done during co-op at Huawei

sual signals. The model involves three components: an audio transformer and a visual transformer that capture long-term information from the audio and visual modalities respectively, and a multimodal transformer based on a multi-head attention that selects the desired information from the output of the audio and visual transformers.

We evaluate our VSET model on the Grid dataset and the AVSpeech dataset in terms of PESQ score. Based on our studies, we show that our model outperforms several state-of-the-art models, particularly in conditions with unknown noise and unknown speakers.

2. MODEL ARCHITECTURE

The proposed VSET architecture is shown in Fig. 1. We will introduce the Transformer model we used in the audio and visual encoders, and then present more details of the components in the architecture.

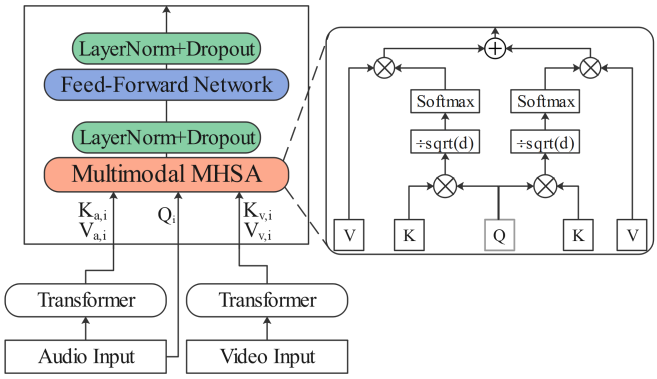


Fig. 1: Architecture of the VSET model.

2.1. Transformer Block

We follow the NEZHA structure [19] to build our audio and visual transformers. The structure is shown in Fig. 2. For the k-th transformer, the i-th self-attention head is formulated as follows:

$$\mathbf{Q}_i = Query_i(\mathbf{O}_{k-1})$$
$$\mathbf{K}_i = Key_i(\mathbf{O}_{k-1}) + \mathbf{P}$$
$$\mathbf{V}_i = Value_i(\mathbf{O}_{k-1}) + \mathbf{P}$$
$$\mathbf{S}_i = softmax(\frac{\mathbf{Q}_i\mathbf{K}_i}{\sqrt{d}})$$
$$\mathbf{Y}_i = \mathbf{V}_i\mathbf{S}_i^T$$

where $\mathbf{O}_{k-1}$ is the output from the previous transfromer, $Query_i$, Key_i, and $Value_i$ are three linear transforms to extract the queries, keys and values for the input features, $\mathbf{P}$ is the matrix derived by positional encoding [15], and $\mathbf{Y}_i$ is the output of the i-th self-attention head. S is the annotation

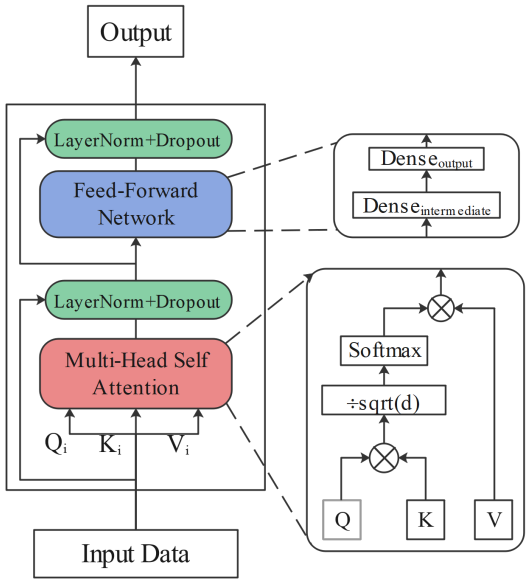


Fig. 2: Architecture of the audio and visual transformers.

matrix, for which the element $s(i,j)$ denotes the annotation weight on the j-th feature when processing at position i. Note that d is the dimension of the input feature. The output of the self-attention layer will be fed to the Feed Forward Network(FNN) layers to produce the output of the k-th transformer, formulated by:

$$\mathbf{O}_k = FFN(f(\mathbf{Y}_1, \mathbf{Y}_2, ...)).$$

Note that for the first transformer, the input $\mathbf{O}_{k-1}$ is just the input features $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_T]$.

2.2. Audio-based component

The audio-based component contains a fully-connected(FC) layer and a stack of 6 transformers proposed in the previous section. The input to the FC layer is a sequence of log power spectrum (LPS) features of the noisy speech. The feature dimension is 257, and the output dimension of the FC layer is 768.

2.3. Visual-based component

Table 1: Architecture of the visual feature extractor.

Layer	#Filters	Kernel size	Stride
SeparableConv1	64	3x3	2x2
SeparableConv2	64	3x3	2x2
SeparableConv3	128	3x3	2x2
SeparableConv4	128	3x3	2x2
SeparableConv5	256	3x3	2x2
SeparableConv6	256	3x3	2x2
FC-768			

The visual-based component contains a visual feature extractor and a stack of 6 transformers presented in Section 2.1.

The feature extractor is based on a separable convolution neural network [20] shown in Table 1. The input is a sequence of images that contain the mouth region of the speaker present in the audio input, and output feature is of 768 dimensions.

2.4. Audio-Visual fusion component

The audio-visual fusion part of the VSET model is based on the multi-modal Transformer block, as shown in Fig. 1. The main part of the transformer is a multi-modal multi-head attention (MMA) layer, which fuses the output features from the audio-based component ($\mathbf{A}$) and the visual-based component ($\mathbf{V}$). Denoting the position matrices derived for the audio and visual streams by $\mathbf{P}_a$ and $\mathbf{P}_v$ respectively, for the k-th transformer, the MMA block is formulated as follows:

$$\mathbf{Q}_i = Query(\mathbf{O}_{k-1})$$
$$\mathbf{K}_{a,i} = Key_{a,i}(\mathbf{A}) + \mathbf{P}_a \qquad \mathbf{V}_{a,i} = Value_{a,i}(\mathbf{A}) + \mathbf{P}_a$$
$$\mathbf{S}_{a,i} = softmax(\frac{\mathbf{Q}_i\mathbf{K}_{a,i}}{\sqrt{d}}) \qquad \mathbf{Y}_{a,i} = \mathbf{V}_{a,i}\mathbf{S}_{a,i}^T$$
$$\mathbf{K}_{v,i} = Key_{v,i}(\mathbf{V}) + \mathbf{P}_v \qquad \mathbf{V}_{v,i} = Value_{v,i}(\mathbf{V}) + \mathbf{P}_v$$
$$\mathbf{S}_{v,i} = softmax(\frac{\mathbf{Q}_i\mathbf{K}_{v,i}}{\sqrt{d}}) \qquad \mathbf{Y}_{v,i} = \mathbf{V}_{v,i}\mathbf{S}_{v,i}^T$$

The output of the MMA layer is fed to the rest of the layers and produce the output of the k-th transformer:

$$O_k = FFN(f(\mathbf{Y}_{a,1}, \mathbf{Y}_{v,1}, Y_{a,2}, Y_{v,2}, ...)).$$

There are 6 multi-modal transformer blocks in total. For the first block, the query is based on the noisy input speech LPS features.

Note that the visual positional encoding should consider the different frame rates of the audio and visual streams. Specifically:

$$P_V(a, v, k) = g((a - v * r)/(10000^{\frac{2\lfloor k/2 \rfloor}{d}}))$$

where k indexes the dimension of the embedding d_i, and r refers to the ratio of audio frame rate to visual frame rate. Function g corresponds to sin or cos when the dimension is even or odd respectively. We show that with the multimodal positional encoding the VSET learns better alignment for the visual signals.

2.5. Decoding component

The decoding component is an FC layer that predicts the IRM which is computed as the ratio of Power Spectra(PS) of the clean and noisy speech in each TF-bin.

$$IRM = \frac{PS_{clean}}{PS_{clean} + PS_{noise}}$$

Once the IRM is predicted, it can be multiplied with the noisy power spectrum to produce the clean spectrogram. Furthermore, reusing the phase of the noisy speech, we can recover the waveform of the clean speech.

3. EXPERIMENTS

3.1. Datasets

- **GRID:** The Grid Audio-Visual Corpus[21] consists of 32 speakers each speaking 1000 English sentences. We divide the data following [9]. The training set consists of 900 utterances from 30 speakers. There are two test sets: *Test(S)* set consists of 100 utterances from the same 30 speakers present in the Training set; *Test(U)* set consists of the complete utterances from the remaining 2 speakers. The *Test(S)* data can be used to test performance on in-domain speakers while the *Test(U)* can be used to test out-of-domain speakers.
- **AVSpeech:** The AVSpeech[7] dataset contains over 4700 hours of audio-visual data comprising over 150k unique speakers. We take the top 2000 speakers that have the largest amount of speaking data as our training set. The following 300 speakers in the same order comprises our testing set. The testing set consists of out-of-domain speakers.
- **Chime and Human Noise:** We use 80% of the data from the Chime noise to perturb the training data. The remaining is used to perturb the testing data. The entire Human noise collection is used to perturb the testing data. The Chime noise can be considered as in-domain noise while the Human noise is out-of-domain noise.

3.2. Methodology

The audio is sampled at 16kHZ and the spectrogram is extracted using STFT with parameters as follows: window size of 400, hop size of 160 and FFT length of 512. The audio data for training is prepared by mixing noises of SNRs [-5, 0, 5] into the clean speech. We test the model by mixing noises in the range of SNRs [-5, 0, 5, 10, 15, 20]. The mixing of clean and noise data is done in the time-domain with the noise clipped to the same length as clean speech. The LPS feature is used as input to the audio component after normalization using the speech present in the training set.

The mouth boundaries are detected using MTCNN[22] and an image of size 160x160 is extracted that contains the mouth at the center. The image is used as input to the video component after normalization.

We train the model using a learning rate scheduler - we warmup to 1e-4 over 1 epoch, keep the learning rate steady for 1 epoch and then decay to 1e-7 over 18 epochs.

3.3. Results on Grid Dataset

We compare our model (VSET) with the Looking to Listen (L2L) [7], Visual Speech Enhancement (VSE) [5], Online Visual Speech Enhancement (OVSE) [9] and the Audio-Only (AO) model from [9]. From Table 2, we can conclude that VSET outperforms all of the models on almost all SNR conditions when tested on the in-domain Chime noise that the models are trained with. When we view the results on Table

Table 2: Results for the models trained on GRID dataset

GRID Training																				
SNR	Chime(S)					Chime(U)					Human(S)					Human(U)				
	AO	VSE	L2L	OVSE	VSET	AO	VSE	L2L	OVSE	VSET	AO	VSE	L2L	OVSE	VSET	AO	VSE	L2L	OVSE	VSET
-5	2.25	**2.43**	2.36	2.20	2.35	2.05	2.05	2.03	1.98	**2.06**	1.89	**2.28**	2.10	1.89	2.14	1.85	**2.03**	1.96	1.84	2.02
0	2.84	2.81	2.81	2.81	**2.88**	2.65	2.54	2.59	2.59	**2.68**	2.38	**2.69**	2.55	2.36	2.60	2.34	2.48	2.46	2.34	**2.51**
5	3.17	3.03	3.08	3.16	**3.19**	3.06	2.82	2.92	2.99	**3.08**	2.76	**2.96**	2.87	2.75	2.94	2.76	2.78	2.81	2.74	**2.88**
10	3.41	3.19	3.27	3.39	**3.42**	3.34	3.01	3.16	3.29	**3.37**	3.07	3.15	3.13	3.06	**3.21**	3.09	2.99	3.07	3.08	**3.18**
15	3.58	3.30	3.43	3.56	**3.61**	3.56	3.14	3.36	3.52	**3.60**	3.33	3.29	3.33	3.31	**3.44**	3.35	3.14	3.29	3.35	**3.43**
20	3.71	3.40	3.55	3.70	**3.77**	3.71	3.24	3.51	3.69	**3.78**	3.53	3.39	3.49	3.52	**3.63**	3.57	3.25	3.45	3.57	**3.64**
AVG	3.16	3.03	3.08	3.14	**3.20**	3.06	2.80	2.93	3.01	**3.10**	2.83	2.96	2.91	2.82	**3.00**	2.82	2.78	2.84	2.82	**2.94**

Table 3: Results for the models trained on AVSpeech dataset

AVSpeech Training								
SNR	AVSpeech Test Chime				Grid Chime Unseen			
	AO	VSE	L2L	VSET	AO	VSE	L2L	VSET
-5	2.20	2.18	1.97	**2.33**	2.05	**2.27**	1.75	2.19
0	2.57	2.51	2.25	**2.69**	2.52	2.63	2.11	**2.67**
5	2.92	2.80	2.56	**3.03**	2.83	2.87	2.45	**2.99**
10	3.25	3.05	2.89	**3.34**	3.08	3.05	2.76	**3.25**
15	3.54	3.25	3.23	**3.61**	3.32	3.20	3.05	**3.49**
20	3.79	3.41	3.54	**3.84**	3.55	3.33	3.33	**3.71**
AVG	3.04	2.87	2.74	**3.14**	2.89	2.89	2.58	**3.05**

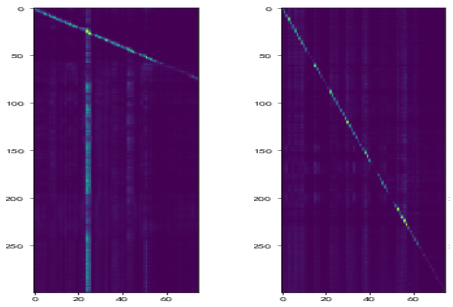
2 for the Human noise (which was not a part of the training data) perturbed Test sets, VSET falls below the VSE model on the lower SNRs. We hypothesize that this is due to the over-reliance of the VSE model on the visual modality - input images do not change when we introduce a different noise into the test set - thereby making a visual-trusting model like VSE more effective. However, over-reliance on visual data may lead to a disadvantage when the visual information is relatively weak. This hypothesis can be further corroborated by the observation that under higher SNR conditions, the VSE model falls below the AO model, as the visual modality is unable to provide sufficient information to compete with a relatively clean audio. Furthermore, under no condition does the VSET model falls below the AO model showing the robustness of our approach in introducing the visual modality to the speech enhancement task.

3.4. Results on AVSpeech

The results on this large dataset setting also demonstrate that our VSET model can add visual information effectively on top of the AO model thereby leading to a robust visual-boosted speech enhancement network whereas the L2L and VSE models perform below the AO model in most of the scenarios. Comparison between Table 2(Chime(U)) and Table 3 shows an increasing gap between VSET and L2L when there are more training data, indicating that the L2L model cannot utilize large datasets as well as the VSE model.

3.5. Case Study

We investigate how the multi-head attention fuses the audio and visual information. The attention heat maps with two positional encoding schemes are shown in Fig. 3: the vanilla schemes from [19] and the multi-modality scheme that accounts for the frame rate mismatch between audio and visual modalities. It clearly shows that the multi-modality scheme can obtain a reasonable fusion plan by discovering the correct audio-visual correspondence, while the vanilla scheme cannot.



(a) (b)

Fig. 3: Attention heat maps with (a) the vanilla positional encoding scheme [19] and (b) the multi-modality positional encoding scheme.

4. CONCLUSION

We presented a visual speech transformer architecture for audio-visual speech enhancement. This architecture is able to integrate knowledge of long-term spans from both the audio and visual streams, and perform effective multi-modality information fusion. We test this model on Grid and AVSpeech datasets. The results show that the new model outperforms several stat-of-the-art models by a large margin. For future work, we aim to integrate the local information along with the global information through transformers. Along with that we will explore multi-modal transformers in AV domains other than AVSE.

5. REFERENCES

[1] William H Sumby and Irwin Pollack, "Visual contribution to speech intelligibility in noise," *The journal of the acoustical society of america*, vol. 26, no. 2, pp. 212–215, 1954.

[2] Harry McGurk and John MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–748, 1976.

[3] Sarah Partan and Peter Marler, "Communication goes multimodal," *Science*, vol. 283, no. 5406, pp. 1272–1273, 1999.

[4] Elana Zion Golumbic, Gregory B Cogan, Charles E Schroeder, and David Poeppel, "Visual input enhances selective speech envelope tracking in auditory cortex at a 'cocktail party'," *Journal of Neuroscience*, vol. 33, no. 4, pp. 1417–1426, 2013.

[5] Aviv Gabbay, Asaph Shamir, and Shmuel Peleg, "Visual speech enhancement," *arXiv preprint arXiv:1711.08789*, 2017.

[6] Jen-Cheng Hou, Syu-Siang Wang, Ying-Hui Lai, Yu Tsao, Hsiu-Wen Chang, and Hsin-Min Wang, "Audio-visual speech enhancement using multimodal deep convolutional neural networks," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 117–128, 2018.

[7] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *arXiv preprint arXiv:1804.03619*, 2018.

[8] Mandar Gogate, Kia Dashtipour, Ahsan Adeel, and Amir Hussain, "Cochleanet: A robust language-independent audio-visual model for speech enhancement," *arXiv preprint arXiv:1909.10407*, 2019.

[9] Wupeng Wang, Chao Xing, Dong Wang, Xiao Chen, and Fengyu Sun, "A robust audio-visual speech enhancement model," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7529–7533.

[10] Daniel N Kalikow, Kenneth N Stevens, and Lois L Elliott, "Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability," *The Journal of the acoustical society of America*, vol. 61, no. 5, pp. 1337–1351, 1977.

[11] Arthur Boothroyd and Susan Nittrouer, "Mathematical treatment of context effects in phoneme and word recognition," *The Journal of the Acoustical Society of America*, vol. 84, no. 1, pp. 101–114, 1988.

[12] Nobuaki Minematsu and Keikichi Hirose, "Role of prosodic features in the human process of perceiving spoken words and sentences in japanese," *Journal of the Acoustical Society of Japan (E)*, vol. 16, no. 5, pp. 311–320, 1995.

[13] Yoshua Bengio, Patrice Simard, and Paolo Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE transactions on neural networks*, vol. 5, no. 2, pp. 157–166, 1994.

[14] Lei Sun, Jun Du, Li-Rong Dai, and Chin-Hui Lee, "Multiple-target deep learning for lstm-rnn based speech enhancement," in *2017 Hands-free Speech Communications and Microphone Arrays (HSCMA)*. IEEE, 2017, pp. 136–140.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," 2017.

[16] Jaeyoung Kim, Mostafa El-Khamy, and Jungwon Lee, "Transformer with gaussian weighted self-attention for speech enhancement," *arXiv preprint arXiv:1910.06762*, 2019.

[17] Szu-Wei Fu, Chien-Feng Liao, Tsun-An Hsieh, Kuo-Hsuan Hung, Syu-Siang Wang, Cheng Yu, Heng-Cheng Kuo, Ryandhimas E Zezario, You-Jin Li, Shang-Yi Chuang, et al., "Boosting objective scores of speech enhancement model through metricgan post-processing," *arXiv preprint arXiv:2006.10296*, 2020.

[18] Jean-Luc Schwartz and Christophe Savariaux, "No, there is no 150 ms lead of visual speech on auditory speech, but a range of audiovisual asynchronies varying from small audio lead to large audio lag," *PLoS Comput Biol*, vol. 10, no. 7, pp. e1003743, 2014.

[19] Junqiu Wei, Xiaozhe Ren, Xiaoguang Li, Wenyong Huang, Yi Liao, Yasheng Wang, Jiashu Lin, Xin Jiang, Xiao Chen, and Qun Liu, "Nezha: Neural contextualized representation for chinese language understanding," *arXiv preprint arXiv:1909.00204*, 2019.

[20] François Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[21] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao, "An audio-visual corpus for speech perception and automatic speech recognition (l)," *The Journal of the Acoustical Society of America*, vol. 120, pp. 2421–4, 12 2006.

[22] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.