Interspeech 2018
2-6 September 2018, Hyderabad

Visual Speech Enhancement

Aviv Gabbay, Asaph Shamir, Shmuel Peleg

School of Computer Science and Engineering
The Hebrew University of Jerusalem
Jerusalem, Israel

Abstract

When video is shot in noisy environment, the voice of a speaker seen in the video can be enhanced using the visible mouth movements, reducing background noise. While most existing methods use audio-only inputs, improved performance is obtained with our visual speech enhancement, based on an audio-visual neural network.

We include in the training data videos to which we added the voice of the target speaker as background noise. Since the audio input is not sufficient to separate the voice of a speaker from his own voice, the trained model better exploits the visual input and generalizes well to different noise types.

The proposed model outperforms prior audio visual methods on two public lipreading datasets. It is also the first to be demonstrated on a dataset not designed for lipreading, such as the weekly addresses of Barack Obama.

Index Terms: speech enhancement, visual speech processing

1. Introduction

Speech enhancement aims to improve speech quality and intelligibility when audio is recorded in noisy environments. Applications include telephone conversations, video conferences, TV reporting and more. Speech enhancement can also be used in hearing aids [1], speech recognition, and speaker identification [2, 3]. Speech enhancement has been the subject of extensive research [4, 5, 6], and has recently benefited from advancements in machine lip reading [7, 8] and speech reading [9, 10, 11].

We propose an audio-visual end-to-end neural network model for separating the voice of a visible speaker from background noise. Once the model is trained on a specific speaker, it can be used to enhance the voice of this speaker. We assume a video showing the face of the target speaker is available along with the noisy soundtrack, and use the visible mouth movements to isolate the desired voice from the background noise.

While the idea of training a deep neural network to differentiate between the unique speech or auditory characteristics of different sources can be very effective in several cases, the performance is limited by the variance of the sources, as shown in [12, 13]. We show that using the visual information leads to significant improvement in the enhancement performance in different scenarios. In order to cover cases where the target and background speech can not be totally separated using the audio information alone, we add to the training data videos with synthetic background noise taken from the voice of the target speaker. With such videos in the training data, the trained model better exploits the visual input and generalizes well to different noise types.

We evaluate the performance of our model in different speech enhancement experiments. First, we show better performance compared to prior art on two common audio-visual datasets: GRID corpus [14] and TCD-TIMIT [15], both designed for audio-visual speech recognition and lip reading. We also demonstrate speech enhancement on public weekly addresses of Barack Obama.

1.1. Related work

Traditional speech enhancement methods include spectral restoration [16, 5], Wiener filtering [17] and statistical model-based methods [18]. Recently, deep neural networks have been adopted for speech enhancement [19, 20, 12], generally outperforming the traditional methods [21].

1.1.1. Audio-only deep learning based speech enhancement

Previous methods for single-channel speech enhancement mostly use audio only input. Lu *et al.* [19] train a deep auto-encoder for denoising the speech signal. Their model predicts a mel-scale spectrogram representing the clean speech. Pascual *et al.* [22] use generative adversarial networks and operate at the waveform level. Separating mixtures of several people speaking simultaneously has also become possible by training a deep neural network to differentiate between the unique speech characteristics of different sources e.g. spectral bands, pitches and chirps, as shown in [12, 13]. Despite their decent overall performance, audio-only approaches achieve lower performance when separating similar human voices, as commonly observed in same-gender mixtures [12].

1.1.2. Visually-derived speech and sound generation

Different approaches exist for generation of intelligible speech from silent video frames of a speaker [11, 10, 9]. In [10], Ephrat *et al.* generate speech from a sequence of silent video frames of a speaking person. Their model uses the video frames and the corresponding optical flow to output a spectrogram representing the speech. Owens *et al.* [23] use a recurrent neural network to predict sound from silent videos of people hitting and scratching objects with a drumstick.

1.1.3. Audio-visual multi-modal learning

Recent research in audio-visual speech processing makes extensive use of neural networks. The work of Ngiam *et al.* [24] is a seminal work in this area. They demonstrate cross modality feature learning, and show that better features for one modality (e.g., video) can be learned if both audio and video are present at feature learning time. Multi-modal neural networks with audio-visual inputs have also been used for lip reading [7], lip sync [25] and robust speech recognition [26].

10.21437/Interspeech.2018-1955

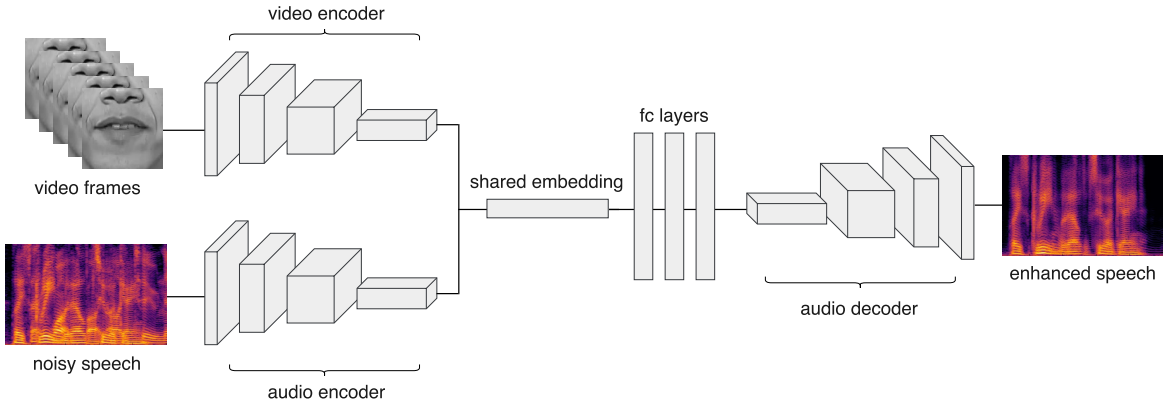

Figure 1: *Illustration of our encoder-decoder model architecture. A sequence of 5 video frames centered on the mouth region is fed into a convolutional neural network creating a video encoding. The corresponding spectrogram of the noisy speech is encoded in a similar fashion into an audio encoding. A single shared embedding is obtained by concatenating the video and audio encodings, and is fed into 3 consecutive fully-connected layers. Finally, a spectrogram of the enhanced speech is decoded using an audio decoder.*

1.1.4. Audio-visual speech enhancement

Work has also been done on audio-visual speech enhancement and separation [27]. Kahn and Milner [28, 29] use hand-crafted visual features to derive binary and soft masks for speaker separation. Hou *et al.* [30] propose convolutional neural network model to enhance noisy speech. Their network gets a sequence of frames cropped to the speaker's lips region and a spectrogram representing the noisy speech, and outputs a spectrogram representing the enhanced speech. Gabbay *et al.* [31] feed the video frames into a trained speech generation network [10], and use the spectrogram of the predicted speech to construct masks for separating the clean voice from the noisy input.

2. Neural Network Architecture

The speech enhancement neural network model gets two inputs: (i) a sequence of video frames showing the mouth of the speaker; and (ii) a spectrogram of the noisy audio. The output is a spectrogram of the enhanced speech. The network layers are stacked in encoder-decoder fashion (Fig. 1). The encoder module consists of a dual tower Convolutional Neural Network which takes the video and audio inputs and encodes them into a shared embedding representing the audio-visual features. The decoder module consists of transposed convolutional layers and decodes the shared embedding into a spectrogram representing the enhanced speech. The entire model is trained end-to-end.

2.1. Video encoder

The input to the video encoder is a sequence of 5 consecutive gray scale video frames of size 128×128, cropped and centered on the mouth region. While using 5 frames worked well, other number of frames might also work. The video encoder has 6 consecutive convolution layers described in Table 1. Each layer is followed by Batch Normalization, Leaky-ReLU for non-linearity, max pooling, and Dropout of 0.25.

2.2. Audio encoder

Both input and output audio are represented by log mel-scale spectrograms having 80 frequency intervals between 0 to 8kHz and 20 temporal steps spanning 200 ms.

layer	video encoder		audio encoder		
	# filters	kernel	# filters	kernel	stride
1	128	5×5	64	5×5	2×2
2	128	5×5	64	4×4	1×1
3	256	3×3	128	4×4	2×2
4	256	3×3	128	2×2	2×1
5	512	3×3	128	2×2	2×1
6	512	3×3			

Table 1: *Detailed architecture of the video and audio encoders. Pooling size and stride used in the video encoder are always* 2×2*, for all six layers.*

As previously done in several audio encoding networks [32, 33], we design our audio encoder as a convolutional neural network using the spectrogram as input. The network consists of 5 convolution layers as described in Table 1. Each layer is followed by Batch Normalization and Leaky-ReLU for non-linearity. We use strided convolutions instead of max pooling in order to maintain temporal order.

2.3. Shared representation

The video encoder outputs a feature vector having 2,048 values, and the audio encoder outputs a feature vector of 3,200 values. The feature vectors are concatenated into a shared embedding representing the audio-visual features, having 5,248 values. The shared embedding is then fed into a block of 3 consecutive fully-connected layers, of sizes 1,312, 1,312 and 3,200, respectively. The resulting vector is then fed into the audio decoder.

2.4. Audio decoder

The audio decoder consists of 5 transposed convolution layers, mirroring the layers of the audio encoder. The last layer is of the same size as the input spectrogram, representing the enhanced speech.

2.5. Optimization

The network is trained to minimize the mean square error (l_2) loss between the output spectrogram and the target speech spectrogram. We use Adam optimizer with an initial learning rate of $5e^{-4}$ for back propagation. Learning rate is decreased by 50% once learning stagnates i.e. the validation error does not improve for 5 epochs.

3. Multi-Modal Training

Neural networks with multi-modal inputs can often be dominated by one of the inputs [34]. Different approaches have been considered to overcome this issue in previous work. Ngiam *et al.* [24] proposed to occasionally zero out one of the input modalities (e.g., video) during training, and only have the other input modality (e.g., audio). This idea has been adopted in lip reading [7] and speech enhancement [30]. In order to enforce using the video features, Hou *et al.* [30] adds an auxiliary video output that should resemble the input.

We enforce the exploitation of visual features by introducing a new training strategy. We include in the training data samples where the added noise is the voice of the same speaker. Since separating two overlapping sentences spoken by the same person is impossible using audio only information, the network is forced to exploit the visual features in addition to the audio features. We show that a model trained using this approach generalizes well to different noise types, and is capable to separate target speech from indistinguishable background speech.

4. Implementation Details

4.1. Video pre-processing

In all our experiments video is resampled to 25 fps. The video is divided to non-overlapping segments of 5 frames (200 ms) each. From every frame we crop a mouth-centered window of size 128×128 pixels, using the 20 mouth landmarks from the 68 facial landmarks suggested by [35]. The video segment used as input is therefore of size $128\times128\times5$. We normalize the video inputs over the entire training data by subtracting the mean video frame and dividing by the standard deviation.

4.2. Audio pre-processing

The corresponding audio signal is resampled to 16 kHz. Short-Time-Fourier-Transform (STFT) is applied to the waveform signal. The spectrogram (STFT magnitude) is used as input to the neural network, and the phase is kept aside for reconstruction of the enhanced signal. We set the STFT window size to 640 samples, which equals to 40 milliseconds and corresponds to the length of a single video frame. We shift the window by hop length of 160 samples at a time, creating an overlap of 75%. Log mel-scale spectrogram is computed by multiplying the spectrogram by a mel-spaced filterbank. The log mel-scale spectrogram comprises 80 mel frequencies from 0 to 8000 Hz. We slice the spectrogram to pieces of length 200 milliseconds corresponding to the length of 5 video frames, resulting in spectrograms of size 80×20: 20 temporal samples, each having 80 frequency bins.

4.3. Audio post-processing

The speech segments are inferred one by one and concatenated together to create the complete enhanced spectrogram. The waveform is then reconstructed by multiplying the mel-scale spectrogram by the pseudo-inverse of the mel-spaced filterbank, followed by applying the inverse STFT. We use the original phase as of the noisy input signal.

Figure 2: *Sample frames from GRID (top-left), TCD-TIMIT (top-right), Mandarin speaker (bottom-left) and Obama (bottom-right) datasets. Bounding boxes in red mark the mouth-centered crop region. The Obama videos have varied background, illuminations, resolutions, and lighting.*

5. Datasets

Sample frames from the datasets are shown in Fig. 2.

5.1. GRID Corpus and TCD-TIMIT

We perform our experiments on speakers from the GRID audio-visual sentence corpus [14], a large dataset of audio and video (facial) recordings of 1,000 sentences spoken by 34 people (18 male, 16 female). We also perform experiments on the TCD-TIMIT dataset [15] which consists of 60 volunteer speakers with around 200 videos each, as well as three lipspeakers, people specially trained to speak in a way that helps the deaf understand their visual speech. The speakers are recorded saying various sentences from the TIMIT dataset [36].

5.2. Mandarin Sentences Corpus

Hou *et al.* [30] prepared an audio-visual dataset containing video recordings of 320 utterances of Mandarin sentences spoken by a native speaker. Each sentence contains 10 Chinese characters with phoneme designed to distribute equally. The length of each utterance is approximately 3-4 seconds. The utterances were recorded in a quiet room with sufficient light, and the speaker was captured from frontal view.

5.3. Obama Weekly Addresses

We assess our model's performance in more general conditions compared to datasets specifically prepared for lip-reading. For this purpose we use a dataset containing weekly addresses given by Barack Obama. This dataset consists of 300 videos, each of 2-3 minutes long. The dataset varies greatly in scale (zoom), background, lighting and face angle, as well as in audio recording conditions, and includes an unbounded vocabulary.

6. Experiments

We evaluate our model on several speech enhancement tasks using the four datasets mentioned in Sec. 5. In all cases, background speech is sampled from the LibriSpeech [37] dataset. For the ambient noise we use different types of noise such as rain, motorcycle engine, basketball bouncing, etc. The speech and noise signals are mixed with SNR of 0 dB both for training and testing, except of the Mandarin experiment where the same protocol of [30] is used. In each sample, the target speech is mixed with background speech, ambient noise, or another speech of the target speaker. We call the latter case *self* mixtures.

We report an evaluation using two objective scores: SNR for measuring the noise reduction and PESQ for assessing the improvement in speech quality [38]. Since listening to audio samples is essential to understand the effectiveness of speech enhancement methods, supplementary material is available on our project web page [1].

6.1. Baselines and previous work

We show the effectiveness of using the visual information by training a competitive audio-only version of our model which has a similar architecture (stripping out the visual stream). Training this baseline does not involve *self* mixtures since audio-only separation in this case is ill-posed. It can be seen that the audio-only baseline consistently achieves lower performance than our model, especially in the *self* mixtures where no improvement in speech quality is obtained at all. In order to validate our assumption stating that using samples of the target speaker as background noise makes the model robust to different noise types as well as cases where the background speech is indistinguishable from the target speech, we train our model once again without *self* mixtures in the training set. It is shown that this model is not capable of separating speech samples of the same voice, although it has access to the visual stream. Detailed results are presented in Table 2.

We show that our model outperforms the previous work of Vid2speech [10] and Gabbay *et al.* [31] on the two audio-visual datasets of GRID and TCD-TIMIT, as well as achieves significant improvements in SNR and PESQ on the new dataset of Obama. It can be seen that the results are somewhat less convincing on the TCD-TIMIT dataset. One possible explanation might be the smaller amount of clean speech (20 minutes) in the training data compared to other experiments (40-60 minutes). In the Mandarin experiment, we follow the protocol of Hou *et al.* [30], and train our model on their proposed dataset containing speech samples mixed with car engine ambient noise and other interfering noise types, in different SNR configurations. Table 3 shows that our model achieves slightly better performance on their proposed test set, while it should be noted that PESQ is not accurate on Chinese [39].

7. Concluding remarks

An end-to-end neural network model, separating the voice of a visible speaker from background noise, has been presented. Also, an effective training strategy for audio-visual speech enhancement was proposed - using as noise overlapping sentences spoken by the same person. Such training builds a model that is robust to similar vocal characteristics of the target and noise speakers, and makes an effective use of the visual information.

The proposed model consistently improves the quality and intelligibility of noisy speech, and outperforms previous methods on two public benchmark datasets. Finally, we demonstrated for the first time audio-visual speech enhancement on a general dataset not designed for lipreading research. Our model is compact, and operates on short speech segments, and thus suitable for real-time applications. On average, enhancement of 200 ms segment requires 36 ms of processing (using NVIDIA Tesla M60 GPU).

We note that our method fails when one input modality is missing, since during training both audio and video were used.

The field of combined audio-visual processing is very active. Recent work showing much progress, that appeared just before the camera ready deadline, includes [40], [41], [42], [43].

Acknowledgment. This research was supported by Israel Science Foundation and Israel Ministry of Science and Technology.

	SNR (dB)			PESQ		
Noise type:	speech (other)	amb-ient	speech (*self*)	speech (other)	amb-ient	speech (*self*)
GRID						
Noisy	0.07	0.26	0.05	1.91	2.08	2.15
Vid2speech [10]	-2.4	-2.4	-2.4	2.14	2.14	2.14
Gabbay [31]	4.23	3.68	1.94	2.18	2.15	2.19
Audio-only	4.61	4.43	2.03	2.57	2.58	1.96
Ours without *self*	**5.66**	**5.65**	2.81	2.85	2.88	2.20
Ours with *self*	**5.66**	**5.65**	**4.05**	**2.86**	**2.92**	**2.67**
TCD-TIMIT						
Noisy	0.01	0.03	0.01	2.09	2.28	2.21
Vid2speech [10]	-14.25	-14.25	-14.25	1.27	1.27	1.27
Gabbay [31]	3.88	3.84	2.62	1.71	1.81	1.82
Audio-only	**5.16**	5.44	1.73	2.48	2.68	1.91
Ours without *self*	5.12	**5.46**	3.28	**2.62**	**2.70**	**2.22**
Ours with *self*	4.54	4.81	**3.38**	2.53	2.61	**2.22**
Obama						
Noisy	0	0.01	0	2.00	2.12	2.31
Audio-only	5.06	5.7	1.84	2.44	2.56	2.11
Ours without *self*	5.71	6.38	3.6	2.61	2.72	2.33
Ours with *self*	**6.1**	**6.78**	**5.21**	**2.67**	**2.75**	**2.56**

Table 2: *Evaluation of our model, with a comparison to baselines and previous work. Our model achieves significant improvement both in noise reduction and speech quality in the different noise types. See text for further discussion.*

	SNR (dB)		PESQ	
Noise type:	from [30]	speech (*self*)	from [30]	speech (*self*)
Input	-3.82	0.01	2.01	2.38
Hou et al. [30]	3.7	-	2.41	-
Audio-only	3.23	2.09	2.30	2.24
Ours without *self*	**4.13**	3.41	**2.45**	2.38
Ours with *self*	3.99	**4.02**	2.43	**2.47**

Table 3: *Evaluation of our model on the* **Mandarin** *dataset, along with a comparison to baselines and Hou* et al. *[30], where noise type is speech and ambient.*

[1] Examples of speech enhancement can be found at *http://www.vision.huji.ac.il/visual-speech-enhancement*

8. References

[1] L.-P. Yang and Q.-J. Fu, "Spectral subtraction-based speech enhancement for cochlear implant patients in background noise," *J. of the Acoustical Society of America*, vol. 117, no. 3, pp. 1001–1004, 2005.

[2] D. Yu, L. Deng, J. Droppo, J. Wu, Y. Gong, and A. Acero, "A minimum-mean-square-error noise reduction algorithm on mel-frequency cepstra for robust speech recognition," in *ICASSP'08*, 2008, pp. 4041–4044.

[3] A. L. Maas, Q. V. Le, T. M. O'Neil, O. Vinyals, P. Nguyen, and A. Y. Ng, "Recurrent neural networks for noise reduction in robust asr," in *13th Conf. of the International Speech Communication Association*, 2012.

[4] A. W. Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," *Acta Acustica united with Acustica*, vol. 86, no. 1, pp. 117–128, January 2000.

[5] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. ASSP*, vol. 32, no. 6, pp. 1109–1121, 1984.

[6] P. C. Loizou, *Speech Enhancement: Theory and Practice*, 2nd ed. CRC Press, Inc., 2013.

[7] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *CVPR'17*, 2017.

[8] Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas, "Lipnet: Sentence-level lipreading," *arXiv:1611.01599*, 2016.

[9] T. L. Cornu and B. Milner, "Generating intelligible audio speech from visual speech," in *IEEE/ACM Trans. Audio, Speech, and Language Processing*, 2017.

[10] A. Ephrat, T. Halperin, and S. Peleg, "Improved speech reconstruction from silent video," in *ICCV'17 Workshop on Computer Vision for Audio-Visual Media*, 2017.

[11] A. Ephrat and S. Peleg, "Vid2speech: speech reconstruction from silent video," in *ICASSP'17*, 2017.

[12] Y. Isik, J. L. Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," in *ICASSP'16*, 2016.

[13] Z. Chen, "Single channel auditory source separation with neural network," Ph.D. dissertation, Columbia Univ., 2017.

[14] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *J. Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[15] N. Harte and E. Gillen, "Tcd-timit: An audio-visual corpus of continuous speech," *IEEE Trans. Multimedia*, vol. 17, no. 5, pp. 603–615, 2015.

[16] P. Scalart *et al.*, "Speech enhancement based on a priori signal to noise estimation," in *ICASSP'96*, vol. 2, 1996, pp. 629–632.

[17] J. Lim and A. Oppenheim, "All-pole modeling of degraded speech," *IEEE Trans. ASSP*, vol. 26, no. 3, pp. 197–210, 1978.

[18] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proc. of the IEEE*, vol. 80, no. 10, pp. 1526–1555, 1992.

[19] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder."

[20] S. Parveen and P. Green, "Speech enhancement with missing data techniques using recurrent neural networks," in *ICASSP'04*, vol. 1, 2004, pp. I–733.

[21] M. Kolbæk, Z.-H. Tan, and J. Jensen, "Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 153–167, 2017.

[22] S. Pascual, A. Bonafonte, and J. Serrà, "Segan: Speech enhancement generative adversarial network," *arXiv:1703.09452*, 2017.

[23] A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, and W. T. Freeman, "Visually indicated sounds," in *CVPR'16*, 2016.

[24] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *ICML'11*, 2011, pp. 689–696.

[25] J. S. Chung and A. Zisserman, "Out of time: automated lip sync in the wild," in *ACCV'16*, 2016, pp. 251–263.

[26] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, "Audio-visual speech recognition using deep learning," *Applied Intelligence*, vol. 42, no. 4, pp. 722–737, 2015.

[27] L. Girin, J. L. Schwartz, and G. Feng, "Audio-visual enhancement of speech in noise," *J. of the Acoustical Society of America*, vol. 109 6, pp. 3007–20, 2001.

[28] F. Khan and B. Milner, "Speaker separation using visually-derived binary masks," in *Auditory-Visual Speech Processing (AVSP)*, 2013.

[29] F. Khan, "Audio-visual speaker separation," Ph.D. dissertation, University of East Anglia, 2016.

[30] J.-C. Hou, S.-S. Wang, Y.-H. Lai, J.-C. Lin, Y. Tsao, H.-W. Chang, and H.-M. Wang, "Audio-visual speech enhancement based on multimodal deep convolutional neural network," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 117–128, 2018.

[31] A. Gabbay, A. Ephrat, T. Halperin, and S. Peleg, "Seeing through noise: Speaker separation and enhancement using visually-derived speech," in *ICASSP'18*, 2018.

[32] J. Engel, C. Resnick, A. Roberts, S. Dieleman, D. Eck, K. Simonyan, and M. Norouzi, "Neural audio synthesis of musical notes with wavenet autoencoders," in *ICML'17*, 2017.

[33] E. M. Grais and M. D. Plumbley, "Single channel audio source separation using convolutional denoising autoencoders," in *GlobalSIP'17*, 2017.

[34] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *CVPR'16*, 2016, pp. 1933–1941.

[35] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *CVPR'14*, 2014, pp. 1867–1874.

[36] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *NISTIR 4930*, 1993.

[37] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *ICASSP'15*, 2015, pp. 5206–5210.

[38] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *ICASSP'01*, vol. 2, 2001, pp. 749–752.

[39] F. L. Chong, I. V. McLoughlin, and K. Pawlikowski, "A methodology for improving PESQ accuracy for Chinese speech," in *TENCON'05*, 2005.

[40] T. Afouras, J. Son Chung, and A. Zisserman, "The Conversation: Deep Audio-Visual Speech Enhancement," *arXiv:1804.04121*, 2018.

[41] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *arXiv:1804.03619*, 2018.

[42] A. Owens and A. A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," *arXiv:1804.03641*, 2018.

[43] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. H. McDermott, and A. Torralba, "The sound of pixels," *arXiv: 1804.03160*, 2018.