Interspeech 2018
2-6 September 2018, Hyderabad

The Conversation: Deep Audio-Visual Speech Enhancement

Triantafyllos Afouras, Joon Son Chung, Andrew Zisserman

Visual Geometry Group, Department of Engineering Science,
University of Oxford, UK
{afourast,joon,az}@robots.ox.ac.uk

Abstract

Our goal is to isolate individual speakers from multi-talker simultaneous speech in videos. Existing works in this area have focussed on trying to separate utterances from known speakers in controlled environments. In this paper, we propose a deep audio-visual speech enhancement network that is able to separate a speaker's voice given lip regions in the corresponding video, by predicting both the magnitude and the phase of the target signal. The method is applicable to speakers unheard and unseen during training, and for unconstrained environments. We demonstrate strong quantitative and qualitative results, isolating extremely challenging real-world examples.

Index Terms: speech enhancement, speech separation

1. Introduction

In the film *The Conversation* (dir. Francis Ford Coppola, 1974), the protagonist, played by Gene Hackman, goes to inordinate lengths to record a couple's converservation in a crowded city square. Despite many ingenious placements of microphones, he did not use the lip motion of the speakers to suppress speech from others nearby. In this paper we propose a new model for this task of audio-visual speech enhancement, that he could have used.

More generally, we propose an audio-visual neural network that can isolate a speaker's voice from others, using visual information from the target speaker's lips: Given a noisy audio signal and the corresponding speaker video, we produce an enhanced audio signal containing only the target speaker's voice with the rest of the speakers and background noise suppressed.

Rather than synthesising the voice from scratch, which would be a challenging task, we instead predict a mask that filters the noisy spectrogram of the input. Many speech enhancement approaches focus on refining only the magnitude of the noisy input signal and use the noisy phase for the signal reconstruction. This works well for high signal-to-noise-ratio scenarios, but as the SNR decreases, the noisy phase becomes a bad approximation of the ground truth one [1]. Instead, we propose correction modules for both the magnitude and phase. The architecture is summarised in Figure 1. In training, we initialize the visual stream with a network pre-trained on a word-level lipreading task, but after this, we train from unlabelled data (Section 3.1) where no explicit annotation is required at the word, character or phoneme-level.

There are many possible applications of this model; one of them is automatic speeech recognition (ASR) – while machines can recognise speech relatively well in noiseless environments, there is a significant deterioration in performance for recognition in noisy environments [2]. The enhancement method we propose could address this problem, and improve, for example, ASR for mobile phones in a crowded environment, or automatic captioning for YouTube videos.

The performance of the model is evaluated for up to five simultaneous voices, and we demonstrate both strong qualitative and quantitative performance. The trained model is evaluated on unconstrained 'in the wild' environments, and for speakers and languages unseen at training time. To the best of our knowledge, we are the first to achieve enhancement under such general conditions. We provide supplementary material with interactive demonstrations on http://www.robots.ox.ac.uk/~vgg/demo/theconversation.

1.1. Related works

Various works have proposed methods to isolate multi-talker simultaneous speech. The majority of these are based on methods that only use the audio, *e.g.* by using voice characteristics of a known speaker [3, 4, 5, 6, 7]. Compared to audio-only methods, we not only separate the voices but also properly assign them to the speakers, by using the visual information.

Speech enhancement methods have traditionally only dealt with filtering the spectral magnitudes, however many approaches have been recently been proposed for jointly enhancing the magnitude and phase spectra [1, 8, 9, 10, 11, 12, 13]. The prevalent method for estimating phase spectra from given magnitudes in speech synthesis is the one proposed by Griffin and Lim [14].

Prior to deep learning, a large number of previous works have been developed for audio-visual speech enhancement by predicting masks [15, 16] or otherwise [17, 18, 19, 20, 21, 22, 23], with an overview of audio-visual source separation is provided in [24]. However, we will concentrate from hereon on methods that have built on these using a deep learning framework.

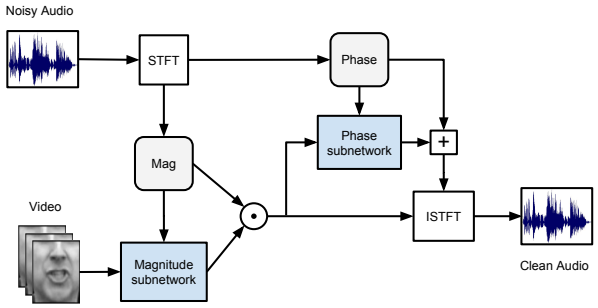


Figure 1: *Audio-visual enhancement architecture overview. It consists of two modules: a magnitude sub-network and a phase sub-network. The first sub-network receives the magnitude spectrograms of the noisy signal and the speaker video as inputs and outputs a soft mask. We then multiply the input magnitudes element-wise with the mask to produce a filtered magnitude spectrogram. The magnitude prediction, along with the phase spectrogram obtained from the noisy signal are then fed into the second sub-network, which produces a phase residual. The residual is added to the noisy phase, producing the enhanced phase spectrograms. Finally the enhanced magnitude and phase spectra are transformed back to the time domain, yielding the enhanced signal.*

10.21437/Interspeech.2018-1400

In [25] a deep neural network is developed to generate speech from silent video frames of a speaking person. This model is used in [26] for speech enhancement, where the predicted spectrogram serves as a mask to filter the noisy speech. However, the noisy audio signal is not used in the pipeline, and the network is not trained for the task of speech enhancement. In contrast, [27] synthesizes the clean signal conditioning on both the mixed speech input and the input video. [28] also use a similar audio-visual fusion method, trained to both generate the clean signal and to reconstruct the video. Both papers use the phase of the noisy input signal as an approximation for the clean phase. However, these methods are limited in that they are only demonstrated under constrained conditions (*e.g.* the utterances consist of a fixed set of phrases in [28]), or for a small number of speakers that have been seen during training.

Our method differs from these works in several ways: (i) we do not treat the spectrograms as images but as temporal signals with the frequency bins as channels; this allows us to build a deeper network with a large number of parameters that trains fast; (ii) we generate a soft mask for filtering instead of directly predicting the clean magnitudes, which we found to be more effective; (iii) we include a phase enhancing sub-network; and, finally, (iv) we demonstrate on previously unheard (and unseen) speakers and on in-the-wild videos.

In concurrent and independent work, [29] develop a similar system, based on dilated convolutions and a bidirectional LSTM, demonstrating good results in unconstrained environments, while [30] train a network for audio-visual synchronisation and successfully use its features for speech separation.

The enhancement method proposed here is complementary to lip reading [31, 32, 33], which has also been shown to improve ASR performance in noisy environments [34, 35].

2. Architecture

This section describes the input representations and architectures for the audio-visual speech enhancement network. The network ingests continuous clips of the audio-visual data. The model architecture is given in detail in Figure 2.

2.1. Video representation

Visual features are extracted from the input image frame sequence with a spatio-temporal residual network similar to the one proposed by [33], pre-trained on a word-level lip reading task. The network consists of a 3D convolution layer, followed by a 18-layer ResNet [36]. For every video frame the network outputs a compact 512 dimensional feature vector f_0^v (where the subscript 0 refers to the layer number in the audio-visual network). Since we train and evaluate on datasets with pre-cropped faces, we do not perform any extra pre-processing, besides conversion to grayscale and an appropriate scaling.

2.2. Audio representation

The acoustic representation is extracted from the raw audio waveforms using Short Time Fourier Transform (STFT) with a Hann window function, which generates magnitude and phase spectrograms. STFT parameters are computed in a similar manner to [27], so that every video frame of the input sequence corresponds to four temporal slices of the resulting spectrogram. Since the videos are at 25fps (40ms per frame), we select a hop length of 10ms with a window length of 40ms at a sample rate of 16Khz. The resulting spectrograms have frequency resolution $F = 321$, representing frequencies from 0 to 8 kHz, and time resolution $T \approx \frac{T_s}{hop}$, where T_s is the duration of the signal in seconds. The magnitude and phase spectrograms are represented as $T \times 321$ and $T \times 642$ tensors respectively, with the real and imaginary components concatenated along the frequency axis for the latter. We convert the magnitudes to mel-scale spectrograms, with 80 frequency bins before feeding them to the magnitude subnetwork, however we conduct the filtering on the original, linear-scale spectrograms.

2.3. Magnitude sub-network

The visual feature sequence f_0^v is processed by a residual network of 10 convolutional blocks. Every block consists of a temporal convolution with kernel width 5 and stride 1, preceded by ReLU activation and batch normalization. A shortcut connection adds the block's input to the result of the convolution. A similar stack of 5 convolutional blocks is employed for processing the audio stream. The convolutions are performed along the temporal dimension, with the frequencies of the noisy input spectrogram M_n viewed as the channels. Two of the intermediate blocks perform convolutions with stride 2, overall down-sampling the temporal dimension by 4, in order to bring it down to the video stream resolution. The skip connections of those layers are down-sampled by average pooling with stride 2. The audio and visual streams are then concatenated over the channel dimension: $f_0^{av} = [f_{10}^v; f_5^a]$. The fused tensor is passed through another stack of 15 temporal convolution blocks. Since we want the output mask to have the same temporal resolution as the input magnitude spectrogram, we include two transposed convolutions, each up-sampling the temporal dimension by a factor of 2, resulting in a factor of 4 in total. The fusion output is projected through position-wise convolutions onto the original magnitude spectrogram dimensions and passed through sigmoid activation in order to output a mask with values between 0 and 1. The resulting tensor is multiplied with the noisy magnitude spectrogram element-wise to produce the enhanced magnitudes:

$$\hat{M} = \sigma(W_m^T f_{15}^{av}) \odot M_n$$

2.4. Phase sub-network

Our intuition for the design of the phase enhancement subnetwork is that there is structure in speech that induces a correlation between the magnitude and phase spectrograms. As with the magnitudes, instead of trying to predict the clean phase from scratch, we only predict a residual that refines the noisy phase. The phase sub-network is therefore conditioned on both the noisy phase and the magnitude predictions. These two inputs are fused together through linear projection and concatenation and then processed by a stack of 6 temporal convolution blocks, with 1024 channels each. The phase residual is formed by projecting the result onto the dimensions of the phase spectrogram and is added to the noisy phase. The clean phase prediction is finally obtained by L_2-normalizing the result:

$$\phi_6 = \underbrace{ConvBlock(\dots ConvBlock([W_{m\phi}^T \hat{M}; W_{n\phi}^T \Phi_n]))}_{\times 6}$$

$$\hat{\Phi} = \frac{(W_\phi^T \phi_6 + \Phi_n)}{||(W_\phi^T \phi_6 + \Phi_n)||_2}$$

In training, the weights of the layers are initialized with small values and zero biases, so that the initial residuals are nearly zero and the noisy phase is propagated to the output.

2.5. Loss function

The magnitude subnetwork is trained by minimizing the L_1 loss between the predicted magnitude spectrogram and the ground

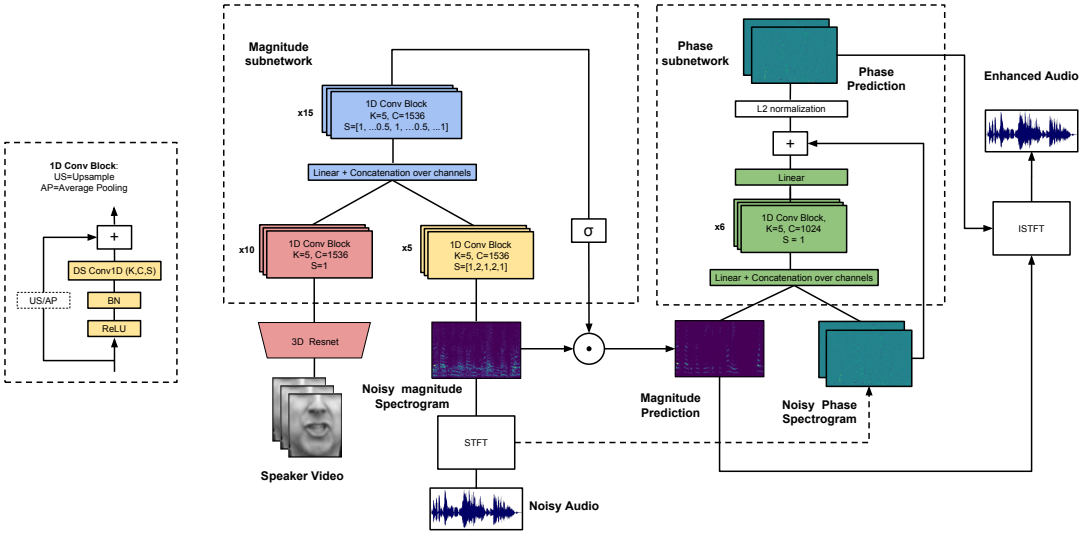


Figure 2: *Audio-visual enhancement network.* ***BN****: Batch Normalization,* **C***: number of channels;* **K***: kernel width;* **S***: strides – fractional ones denote transposed convolutions. The network consists of a magnitude and a phase sub-network. The basic building unit is the temporal convolutional block with pre-activation [37] shown on the left. Identity skip connections are added after every convolution layer (and speed up training). All convolutional layers have* 1536 *channels in the magnitude sub-network and* 1024 *in the phase subnetwork. Depth-wise separable convolution layers [38] are used, which consist of a separate convolution along the time dimension for every channel, followed by a position-wise projection onto the new channel dimensions (equivalent to a convolution with kernel width 1).*

truth. The phase subnetwork is trained by maximizing the cosine similarity between the phase prediction and ground truth, scaled by the ground truth magnitudes. The overall optimisation objective is:

$$\mathcal{L} = ||\hat{M} - M^*||_1 - \lambda \frac{1}{TF} \sum_{t,f} M^*_{tf} < \hat{\Phi}_{tf}, \Phi^*_{tf} > \quad (1)$$

3. Experiments

3.1. Datasets

The model is trained on two datasets: the first is the BBC-Oxford Lip Reading Sentences 2 (LRS2) dataset [34, 39], which contains thousands of sentences from BBC programs such as Doctors and EastEnders; the second is VoxCeleb2 [40], which contains over a million utterances spoken by over 6,000 different speakers.

The LRS2 dataset is divided into training and test sets by broadcast date, in order to ensure that there is no overlapping video between the sets. The dataset covers a large number of speakers, which encourages the trained model to be speaker agnostic. However, since no identity labels are provided with the dataset, there may be some overlapping speakers between the sets. The ground truth transcriptions are provided with the dataset, which allows us to perform quantitative tests on the intelligibility of the generated audio.

The VoxCeleb2 dataset lacks the text transcriptions, however the dataset is divided into training and test sets by identity, which allows us to test the model explicitly for speaker-independent performance.

The audio and video on these datasets are properly synchronized. Evaluation on videos where this is not the case (*e.g.* TV broadcast), is possible by preprocessing with the pipeline described in [41] to detect and track active speakers and synchronize the video and the audio.

3.2. Experimental setup

We examine scenarios where we add 1 to 4 extra interference speakers on the clean signal, therefore we generate signals with 2 to 5 speakers in total. It should be noted that the task of separating the voice of multiple speakers with equal average "loudness" is more challenging than separating the speech signal from background babble noise.

3.3. Evaluation protocol

We evaluate the enhancement performance of the model in terms of perceptual speech quality using the blind source separation criteria described in [42] (we use the implementation provided by [43]). The Signal to Interference Ratio (SIR) measures how well the unwanted signals have been suppressed, the Signal to Artefacts Ratio (SAR) accounts for the introduction of artefacts by the enhancement process, and the Signal to Distortion Ratio (SDR) is an overall quality measure, taking both into account. We also report results on PESQ [44], which measures the overall perceptual quality and STOI [45], which is correlated with the intelligibility of the signal. From the metrics presented above, PESQ has been shown to be the one correlating best with listening tests that account for phase distortion[46].

Additionally, we use an ASR system to test for the intelligibility of the enhanced speech. For this, we use the Google Speech Recognition interface, and report the Word Error Rates (WER) on the clean, mixed and generated audio samples.

3.4. Training

We pre-train the spatio-temporal visual front-end on a word-level lip reading task, following [33]. This proceeds in two stages: first, training on the LRW dataset [31], which covers near-frontal poses; and then on an internal multi-view dataset of a similar size. To accelerate the subsequent training process, we freeze the front-end, pre-compute and save the visual features for all the videos, and also compute and save the magnitude and phase spectrograms for both the clean and noise audio.

Training takes place in three phases: first, the magnitude prediction sub-network is trained, following a curriculum which starts with high SNR inputs (i.e. only one additional speaker) and then progressively moves to more challenging examples with a greater number of speakers; second, the magnitude sub-network is frozen, and only the phase network is trained ; finally, the whole network is fine-tuned end-to-end. We did not experiment with the hyperparameter balancing the magnitude and phase loss terms, but set it to $\lambda = 1$.

To generate training examples we first select a reference pair of visual and audio features (v_r, a_r) by randomly sampling a 60-frame clean segment, making sure that the audio and visual features correspond and are correctly aligned. We then sample N noise spectrograms $x_n, n \in [1, N]$, and mix them with the

Mag	Φ \ # Spk.	SIR (dB) 2	3	4	5	SDR (dB) 2	3	4	5	PESQ 2	3	4	5	WER (%) 2	3	4	5
Mix	**Mix**	–	–	–	–	-0.3	-3.4	-5.4	-6.7	1.73	1.47	1.37	1.21	93.1	99.5	99.9	100
Pr	**GT**	10.8	13.2	13.8	13.7	15.7	13.0	10.8	9.5	3.41	3.05	2.93	2.80	9.4	12.0	16.7	21.5
Pr	**GL**	0.9	2.5	3.6	4.0	-2.9	-2.8	-2.9	-2.7	2.98	2.71	2.52	2.35	10.5	13.7	20.3	27.8
Pr	**Mix**	1.6	2.7	2.5	2.0	10.5	7.8	5.9	4.8	3.02	2.70	2.49	2.33	10.8	14.9	22.0	31.9
Pr	**Pr**	3.9	5.4	5.4	4.8	11.8	9.1	7.1	5.8	3.08	2.79	2.56	2.43	9.7	13.8	20.3	28.9

Table 1: *Evaluation of speech enhancement performance on the LRS2 dataset, for scenarios with different number of speakers (denoted by # Spk). The magnitude (Mag) and phase (Φ) columns specify if the spectrograms used for the reconstructions are predicted or are obtained directly from the mixed or ground truth signal:* **Mix:** *Mixed;* **Pr:** *Predicted;* **GT:** *Ground Truth;* **GL:** *Griffin-Lim;* **SIR:** *Signal to Interference Ratio;* **SDR:** *Signal to Distortion Ratio;* **PESQ:** *Perceptual Evaluation of Speech Quality, varies between 0 and 4.5; (higher is better for all three);* **WER:** *Word Error Rate from off-the-shelf ASR system (lower is better). The WER on the ground truth signal is* 8.8%.

reference spectrogram in the frequency domain by summing up the complex spectra, obtaining the mixed spectrogram a_m. This is a natural way to augment our training data since a different combination of noisy audio signals is sampled every time. Before adding in the noise samples, we normalize their energy to have the reference signal's one: $a_m = a_r + \sum_n \frac{rms(x_r)}{rms(a_n)} a_n$.

3.5. Results

LRS2. We summarize our results on the test set of the LRS2 dataset in Table 1. The performance under the different metrics is listed for the following signal types: The mixed signal which serves as a baseline, and the reconstructions that are obtained using the magnitudes predicted by our network and either the ground truth phase, the phase approximated with the Griffin Lim algorithm, the mixed signal phase or the predicted phase. The signal reconstructed from predicted magnitudes and phases is what we consider the final output of our network.

The evaluation when using the ground truth phase is included as an upper bound to the phase prediction. As can be seen from all measures on the mixed signal, the task becomes increasingly difficult as more speakers are added. In general both the BSS metrics and PESQ correlate well with our observations. It is interesting to note that while more speakers are added, the SIR stays roughly the same, however more overall distortion is introduced. The model is very effective in suppressing cross-talk in the output, however it does so with a trade-off in the quality of the target voice.

The phase predicted by our network performs better than the mixed phase. Even though the improvement is relatively small in numbers, the difference in speech quality is noticeable as the "robotic" effect of having off-sync harmonics is significantly reduced. We encourage the reader to listen to the samples in the supplementary material, where those differences can be understood better. However, the considerable gap with the performance of the ground truth phase shows that there is much room for improvement in the phase network.

The transcription results using the Google ASR are also in line with these findings. In particular, it is noteworthy that our model is able to generate highly intelligible results from noisy audio that is incomprehensible by a human or an ASR system.

Although the content is mainly carried by the magnitude, we see major improvement in terms of WER when using a better phase approximation. It is interesting to note that, although the phase obtained using the Griffin Lim (GL) algorithm achieves significantly worse performance on the objective measures, it demonstrates relatively strong WER results, even slightly surpassing the predicted phase by a small margin in the case of 5 simultaneous speakers.

VoxCeleb2. In order to explicitly assess whether our model can generalize to speakers unseen during training, we also fine-tune and test on VoxCeleb2, using train and test sets that are disjoint in terms of speaker identities. The results are summarized in Table 2, where we showcase an experiment for the 3-speaker scenario. We additionally include evaluation using the SAR and STOI metrics. Overall the performance is comparable to, but slightly worse than, on the LRS2 dataset – which is in line with the qualitative performance. This can be attributed to the visual features not being fine-tuned, and the presence of a lot of other background noise in VoxCeleb2. The results confirm that the method can generalize to unseen (and unheard) speakers.

The last column of the table shows the PESQ evaluation for the original model trained on LRS2, without any fine-tuning on VoxCeleb. The performance is worse than that of the fine-tuned model, however it clearly works. Since LRS2 is constrained to English speakers only, but VoxCeleb2 contains multiple languages, this demonstrates that the model learns to generalise to languages not seen during training.

Mag	Φ	SIR	SAR	SDR	STOI	PESQ	PESQ-NF
Mix	**Mix**	-	-1.59	-2.99	0.34	1.58	1.58
Pr	**GT**	11.43	16.41	10.30	0.77	3.02	2.79
Pr	**GL**	2.05	3.49	-2.42	0.65	2.59	2.39
Pr	**Mix**	1.72	13.54	6.71	0.65	2.59	2.41
Pr	**Pr**	5.02	13.77	7.91	0.67	2.67	2.45

Table 2: *Evaluation of speech enhancement performance on the VoxCeleb2 dataset, for 3 simultaneous speakers, Notations are described in the caption of Table 1. Additional metrics used here:* **SAR:** *Signal to Artefacts Ratio;* **STOI:** *Short-Time Objective Intelligibility, varies between 0 and 1;* **PESQ-NF:** *PESQ score with a model that has not been fine-tuned on VoxCeleb; Higher is better for all.*

3.6. Discussion

Phase refinement. Training our whole network end-to-end decreases the phase loss and this might suggest that the inclusion of visual features also improves the phase enhancement. However, a thorough investigation to determine if, and to what extent, this is true is left to future work.

AV synchronization. Our method is very sensitive to the temporal alignment between the voice and the video. We use SyncNet for the alignment, but since the method can fail under extreme noise, we need to build some invariance in the model. In future work this will be incorporated in the model.

4. Conclusion

In this paper, we have proposed a method to separate the speech signal of a target speaker from background noise and other speakers using visual information from the target speaker's lips. The deep network produces realistic speech segments by predicting both the phase and the magnitude of the target signal; we have also demonstrated that the network is able to generate intelligible speech from very noisy audio segments recorded in unconstrained 'in the wild' environments.

Acknowledgements. Funding for this research is provided by the UK EPSRC CDT in Autonomous Intelligent Machines and Systems, the Oxford-Google DeepMind Graduate Scholarship, and the EPSRC Programme Grant Seebibyte EP/M013774/1.

5. References

[1] S.-W. Fu, T.-Y. Hu, Y. Tsao, and X. Lu, "Complex spectrogram enhancement by convolutional neural network with multi-metrics learning," *arXiv preprint arXiv:1704.08504*, 2017.

[2] M. Anusuya and S. K. Katti, "Speech recognition by machine, a review," *arXiv preprint arXiv:1001.2267*, 2010.

[3] A. M. Reddy and B. Raj, "Soft mask methods for single-channel speaker separation," *IEEE Transactions on Audio, Speech, and Language Processing*, 2007.

[4] Z. Jin and D. Wang, "A supervised learning approach to monaural segregation of reverberant speech," *IEEE Transactions on Audio, Speech, and Language Processing*, 2009.

[5] M. H. Radfar and R. M. Dansereau, "Single-channel speech separation using soft mask filtering," *IEEE Transactions on Audio, Speech, and Language Processing*, 2007.

[6] S. Makino, T.-W. Lee, and H. Sawada, *Blind speech separation*. Springer, 2007.

[7] D. Wang and J. Chen, "Supervised speech separation based on deep learning: an overview," *arXiv preprint arXiv:1708.07524*, 2017.

[8] P. Mowlaee and J. Kulmer, "Phase estimation in single-channel speech enhancement: Limits-potential," *Trans. Audio, Speech and Lang. Proc.*, 2015.

[9] P. Mowlaee, R. Saeidi, and Y. Stylianou, "Advances in phase-aware signal processing in speech communication," *Speech Communication Elsevier*, 2016.

[10] J. Fahringer, T. Schrank, J. Stahl, P. Mowlaee, and F. Pernkopf, "Phase-aware signal processing for automatic speech recognition," in *Interspeech*, 2016.

[11] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for joint enhancement of magnitude and phase," in *ICASSP*, 2016.

[12] H.-G. Hirsch and M. Gref, "On the influence of modifying magnitude and phase spectrum to enhance noisy speech signals," in *Interspeech*, 2017.

[13] M. L. Dubey, G. T. Kenyon, N. Carlson, and A. Thresher, "Does phase matter for monaural source separation?" *CoRR*, vol. abs/1711.00913, 2017.

[14] D. Griffin and J. S. Lim, "Signal estimation from modified short-time fourier transform," in *ICASSP*, 05 1984.

[15] Q. Liu, W. Wang, P. J. Jackson, M. Barnard, J. Kittler, and J. Chambers, "Source separation of convolutive and noisy mixtures using audio-visual dictionary learning and probabilistic time-frequency masking," *IEEE Transactions on Signal Processing*, 2013.

[16] F. Khan and B. Milner, "Speaker separation using visually-derived binary masks," in *AVSP*, 2013.

[17] W. Wang, D. Cosker, Y. Hicks, S. Saneit, and J. Chambers, "Video assisted speech source separation," in *ICASSP*, 2005.

[18] L. Girin, J.-L. Schwartz, and G. Feng, "Audio-visual enhancement of speech in noise," *The Journal of the Acoustical Society of America*, 2001.

[19] S. Deligne, G. Potamianos, and C. Neti, "Audio-visual speech enhancement with avcdcn (audio-visual codebook dependent cepstral normalization)," in *Sensor Array and Multichannel Signal Processing Workshop Proceedings, 2002*, 2002.

[20] J. R. Hershey and M. Casey, "Audio-visual sound separation via hidden markov models," in *NIPS*, 2002.

[21] J. Hershey, H. Attias, N. Jojic, and T. Kristjansson, "Audio-visual graphical models for speech processing," in *Proc. ICASSP*, 2004.

[22] I. Almajai and B. P. Milner, "Effective visually-derived wiener filtering for audio-visual speech processing," in *AVSP*, 2009.

[23] R. Goecke, G. Potamianos, and C. Neti, "Noisy audio feature enhancement using audio-visual speech data," May 2002.

[24] B. Rivet, W. Wang, S. M. Naqvi, and J. A. Chambers, "Audio-visual speech source separation: An overview of key methodologies," *IEEE Signal Processing Magazine*, 2014.

[25] A. Ephrat, T. Halperin, and S. Peleg, "Improved speech reconstruction from silent video," in *ICCV 2017 Workshop on Computer Vision for Audio-Visual Media*, 2017.

[26] A. Gabbay, A. Ephrat, T. Halperin, and S. Peleg, "Seeing through noise: Visually driven speaker separation and enhancement," *arXiv preprint arXiv:1708.06767*, 2017.

[27] A. Gabbay, A. Shamir, and S. Peleg, "Visual Speech Enhancement using Noise-Invariant Training," *arXiv preprint arXiv:1711.08789*, 2017.

[28] J.-C. Hou, S.-S. Wang, Y.-H. Lai, Y. Tsao, H.-W. Chang, and H.-M. Wang, "Audio-Visual Speech Enhancement Using Multimodal Deep Convolutional Neural Networks," *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2018.

[29] 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," *CoRR*, vol. abs/1804.03619, 2018.

[30] A. Owens and A. A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," *CoRR*, vol. abs/1804.03641, 2018.

[31] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *Proc. ACCV*, 2016.

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

[33] T. Stafylakis and G. Tzimiropoulos, "Combining Residual Networks with LSTMs for Lipreading," in *Interspeech*, 2017.

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

[35] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic, "End-to-end audiovisual speech recognition," *CoRR*, vol. abs/1802.06424, 2018.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015.

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. ECCV*, 2016.

[38] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. CVPR*, 2017.

[39] J. S. Chung and A. Zisserman, "Lip reading in profile," in *Proc. BMVC.*, 2017.

[40] J. S. Chung, A. Nagrani, , and A. Zisserman, "VoxCeleb2: Deep speaker recognition," *arXiv preprint arXiv:1001.2267*, 2018.

[41] J. S. Chung and A. Zisserman, "Out of time: automated lip sync in the wild," in *Workshop on Multi-view Lip-reading, ACCV*, 2016.

[42] C. Févotte, R. Gribonval, and E. Vincent, "BSS EVAL toolbox user guide," *IRISA Technical Report 1706. http://www.irisa.fr/metiss/bss eval/.*, 2005.

[43] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Transactions on Audio, Speech and Language Processing*, 2011.

[44] 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*, 2001.

[45] C. Taal, R. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Transactions on Audio, Speech and Language Processing*, 2011.

[46] P. Mowlaee, "On speech intelligibility estimation of phase-aware single-channel speech enhancement," in *ICASSP*, 2015.