SEEING THROUGH NOISE: VISUALLY DRIVEN SPEAKER SEPARATION AND ENHANCEMENT

Aviv Gabbay *Ariel Ephrat* *Tavi Halperin* *Shmuel Peleg*

The Hebrew University of Jerusalem
Jerusalem, Israel

ABSTRACT

Isolating the voice of a specific person while filtering out other voices or background noises is challenging when video is shot in noisy environments. We propose audio-visual methods to isolate the voice of a single speaker and eliminate unrelated sounds. First, face motions captured in the video are used to estimate the speaker's voice, by passing the silent video frames through a video-to-speech neural network-based model. Then the speech predictions are applied as a filter on the noisy input audio. This approach avoids using mixtures of sounds in the learning process, as the number of such possible mixtures is huge, and would inevitably bias the trained model. We evaluate our method on two audio-visual datasets, GRID and TCD-TIMIT, and show that our method attains significant SDR and PESQ improvements over the raw video-to-speech predictions, and a well-known audio-only method.

Index Terms— visual speech processing, speech separation, cocktail party problem, speechreading

1. INTRODUCTION

Single channel speaker separation and speech enhancement have been extensively researched [1, 2]. Neural networks have recently been trained to separate audio mixtures into their sources [3]. These models were able to learn unique speech characteristics as spectral bands, pitches and chirps [4]. The main difficulty of audio-only approaches is their poor performance in separating similar human voices, such as same-gender mixtures.

We first describe the separation of a mixed speech of two speakers whose faces are visible in the video. We continue with the isolation of the speech of a single visible speaker from background sounds. This work builds upon recent advances in machine speechreading, generating speech from visible motion of the face and mouth [5, 6, 7].

Unlike other methods which utilize models trained on mixtures of speech and noise or two voices, our approach is speaker dependent and noise-invariant. This allows us to train models using far less data, and still obtain good results, even in cases of two overlapping voices of the same person.

1.1. Related work

Audio-only speech enhancement and separation Previous methods for single-channel, or monaural, speech enhancement and separation mostly use audio only input. The common *spectrographic masking* approach generates masking matrices containing time-frequency (TF) components dominated by each speaker [8, 9]. Huang *at al.* [10] are among the first to use a deep learning-based approach for speaker dependent speech separation.

Isik *et al.* [4] tackle the single-channel multi-speaker separation using *deep clustering*, in which discriminatively-trained speech embeddings are used as the basis for clustering and separating speech. Kolbaek *et al.* [11] introduce a simpler approach in which they use a permutation-invariant loss function which helps the underlying neural network discriminate between the different speakers.

Audio-visual speech processing Recent research in audio-visual speech processing makes extensive use of neural networks. The work of Ngiam *et al.* [12] is a seminal work in this area. Neural networks with visual input have been used for lipreading [13], sound prediction [14] and for learning unsupervised sound representations [15].

Work has also been done on audio-visual speech enhancement and separation [16, 17]. Kahn and Milner [18, 19] use hand-crafted visual features to derive binary and soft masks for speaker separation. Hou *et al.* [20] propose CNN based models to enhance noisy speech. Their network generates a spectrogram representing the enhanced speech.

2. VISUALLY-DERIVED SPEECH GENERATION

Several approaches exist for generation of intelligible speech from silent video frames of a person speaking [5, 6, 7]. In this work we rely on $vid2speech$ [6], briefly described in Sec. 2.1. It should be noted that these methods are *speaker dependent*, meaning a separate, dedicated model must be trained per speaker.

2.1. Vid2speech

In a recent paper, Ephrat *et al.* [6] present a neural network-based method for generating speech spectrograms from a se-

quence of silent video frames of a speaking person. Their model takes two inputs: (i) a video clip of K consecutive frames, and (ii) a "clip" of $(K-1)$ dense optical flow fields between consecutive frames. The network architecture consists of a dual-tower ResNet [21] which takes the aforementioned inputs and encodes them into a latent vector representing the visual features, which is subsequently fed into a series of two fully connected layers, generating mel-scale spectrogram predictions. This is followed by a post-processing network which aggregates multiple consecutive predictions and maps them to a linear-scale spectrogram representing the final speech prediction.

3. AUDIO-VISUAL SPEECH SEPARATION

We propose to examine the spectrogram of the audio input, a mixture of multiple sources, and to assign each time-frequency (TF) element to its respective source. The generated spectrograms are used to reconstruct the estimated individual source signals.

The above assignment operation is based on the estimated speech spectrogram of each speaker, as generated by a video-to-speech model from Sec. 2. Since the video-to-speech process does not generate perfect speech signals, we use them only as prior knowledge for separating the noisy mixture.

3.1. Speech separation of two speakers

In this case, two speakers (D_1, D_2) face a camera using a single microphone. We assume that the speakers are known, i.e. we train two separate video-to-speech networks (N_1, N_2) in advance, one for each speaker, where N_1 is trained using the audio-visual dataset of speaker D_1, and N_2 is trained on speaker D_2.

Given the video of speakers D_1 and D_2, whose sound track includes their mixed voices, the voice separation process is as follows:

1. The faces of speakers D_1 and D_2 are detected in the video using a face detection method [22].
2. Speech mel-scale spectrograms S_1 and S_2 of speakers D_1 and D_2 are predicted from the respective faces using networks N_1 and N_2.
3. A mixture mel-scale spectrogram C is generated from the input audio.
4. For each (t, f),
$$F_1(t,f) = \begin{cases} 1 & S_1(t,f) > S_2(t,f) \\ 0 & otherwise \end{cases} \quad (1)$$
$$F_2(t,f) = 1 - F_1(t,f) \quad (2)$$
5. Separated spectrograms P_i for each speaker are generated from the mixture spectrogram C by $P_i = C \odot F_i$, where $\odot$ denotes element-wise multiplication.
6. Separated speech signals are reconstructed from the spectrograms (P_1 or P_2), preserving the original phase of each isolated frequency.

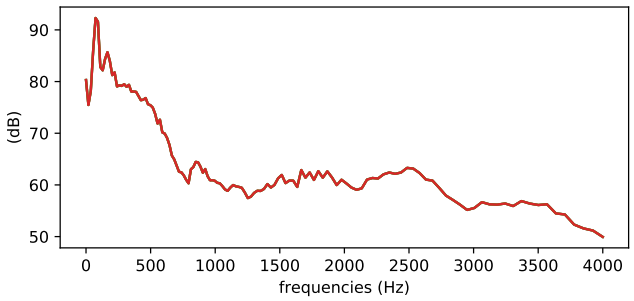


Fig. 1: Example of a thresholding function based on the Long-Term Speech Spectra (LTSS) of a male speaker. Here, for each frequency f, the threshold $\tau(f)$ is set to the 75% percentile of all seen magnitudes of f in the training data.

The binary separation in Step 4 above, where "winner takes all", can be modified to generate a ratio mask, which gives each TF bin a continuous value between 0 and 1, i.e. the generation of the two masks F_1 and F_2 can be done by:

$$F_i(t,f) = \left(\frac{S_i^2(t,f)}{S_1^2(t,f) + S_2^2(t,f)} \right)^{\frac{1}{2}}, \quad i = 1,2 \quad (3)$$

3.2. Speech enhancement of a single speaker

In the speech enhancement case one speaker (D) is facing a camera having a single microphone. Background noise, that may include voices of other (unseen) speakers, is also recorded. The task is to separate the speaker's voice from the background noise. As before, we assume that we train in advance a video-to-speech network (N) on an audio-visual dataset of this speaker. But unlike speech separation, only a single speech prediction is available.

As we assume that the speaker is previously known, we compute the Long-Term Speech Spectra (LTSS) from the speaker's training data, obtaining the distribution of each frequency in the speaker's voice. For each frequency f we pick a threshold $\tau(f)$, indicating when the frequency might come from this speaker's speech, and should be preserved when suppressing the noise. For example, the threshold for a given frequency can be set to the top X percentile (In this case X is a hyperparameter). An example of a thresholding function can be seen in Fig. 1.

Given a new video of same speaker, having a noisy sound track, the process to isolate the speaker's voice is as follows:

1. The thresholding function $\tau(f)$ is computed from the Long-Term Speech Spectra (LTSS) of the training data.
2. The face of speaker D is detected in the input video using a face detection method.
3. Speech mel-scale spectrogram S of speaker D is predicted from the detected face using network N.
4. The noisy mel-scale spectrogram C is generated from the noisy audio input.

5. A separation mask F is constructed using the threshold $\tau(f)$: For each (t, f) in the spectrogram, we compute:

$$F(t,f) = \begin{cases} 1 & S(t,f) > \tau(f) \\ 0 & otherwise \end{cases} \quad (4)$$

6. The noisy mel-scale spectrogram C is filtered by the following operation: $P = C \odot F$, where $\odot$ denotes element-wise multiplication.

7. The clean speech is reconstructed from the predicted mel-scale spectrogram P, preserving the original phase of each isolated frequency.

4. EXPERIMENTS

4.1. Datasets

GRID Corpus We performed experiments on the GRID audio-visual sentence corpus [23], a large dataset of audio and video (facial) recordings of 1,000 3-second sentences spoken by 34 people. A total of 51 different words are contained in the GRID corpus.

TCD-TIMIT We conducted additional experiments on the TCD-TIMIT dataset [24]. This dataset consists of 60 speakers with around 200 videos each, as well as three lipspeakers, people specially trained to speak in a way that helps lipreaders understand their visual speech. The speakers are recorded saying various sentences from the TIMIT dataset [25] using both front-facing and 30 degree cameras.

Mixing protocol For each one of the experiments, we synthesize audio mixtures from the speech signals of two speakers of the same gender. Given audio signals $s_1(t), s_2(t)$ their mixture is synthesized to be $s_1(t) + s_2(t)$, using the original, unnormalized gain of each source. The signals in all experiments are taken from data unseen when training the relevant *vid2speech* models.

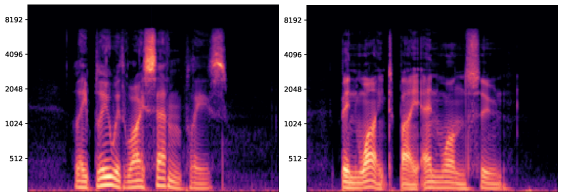
4.2. Performance evaluation

The results of our experiments are evaluated using objective source separation evaluation scores, including SDR, SIR and SAR [26] and PESQ [27]. In addition to these measurements, we assessed the intelligibility and quality of our results qualitatively using informal human listening. We strongly encourage readers to watch and listen to the supplementary video available on our project webpage[1], demonstrating the effectiveness of our approach.

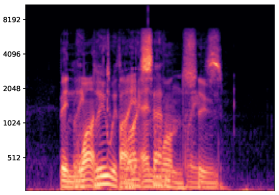
4.3. Results

Separation Table 1 shows the results of the separation experiments on synthesized mixtures of sentences spoken from the GRID and TCD-TIMIT datasets. The GRID experiment involved testing on random speech mixtures from two male speakers (*S2* and *S3*). The TCD-TIMIT experiment involved random speech mixtures of a female speaker (*lipspeaker 3*) with her own voice, emphasizing the capabilities of our approach. We present a comparison to results obtained by applying the audio-only method of Huang *et al.* [10]. In addition, we compare to the raw speech predictions generated by *vid2speech*, without applying any of our separation methods.

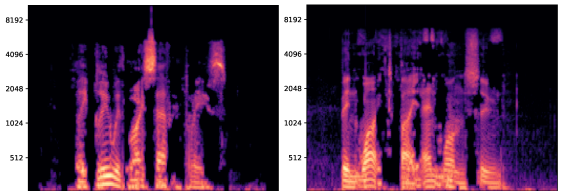
It can be seen that the raw speech predictions have reasonable quality (PESQ score) when dealing with a constrained-vocabulary dataset such as GRID. However, *vid2speech* generates low quality and mostly unintelligible speech predictions when dealing with a more complex dataset such as TCD-TIMIT, which contains sentences from a larger vocabulary. In this case, our separation methods have real impact, and the final speech signals sound much better than the raw speech predictions. We use the spectrograms of ground truth source signals to construct the ideal binary and ratio masks, and present their separation scores as a performance ceiling of our separation method. Examples of the separated spectrograms are shown in Figure 2.



(a) Spectrograms of two source voices.



(b) Spectrogram of the mixed voices.



(c) Spectrograms of the separated voices.

Fig. 2: Spectrograms for one segment in our separation testing data from the GRID dataset.

Enhancement Table 2 shows the results of enhancement experiments on synthesized mixtures of sentences spoken from the GRID and TCD-TIMIT datasets. The GRID experiment involved random speech mixtures of two male speakers (*S2* as target speaker and *S3* as background speaker). The TCD-TIMIT experiment involved random speech mixtures

[1]Examples of speech separation and enhancement can be found at `http://www.vision.huji.ac.il/speaker-separation`

	SDR	SIR	SAR	PESQ
GRID				
Noisy	0.04	0.05	40.6	2.1
Vid2speech [6]	-15.19	7.41	-14.2	1.91
Audio-only [10]	1.74	2.75	6.59	1.85
Ours - binary mask	5.1	**13.02**	6.41	2.07
Ours - ratio mask	**5.62**	8.83	**9.49**	**2.6**
Ideal binary mask	10.6	22.03	11.03	2.9
Ideal ratio mask	10.1	14.15	12.65	3.58
TCD-TIMIT				
Noisy	0.15	0.15	237.17	2.26
Vid2speech [6]	-12.99	13.53	-12.26	1.41
Audio-only [10]	2.91	4.62	9.04	2.16
Ours - binary mask	8.11	**17.8**	9.01	2.4
Ours - ratio mask	**8.68**	13.39	**11.04**	**2.71**
Ideal binary mask	15.49	28.76	15.88	3.4
Ideal ratio mask	15.19	21.61	16.6	3.86

Table 1: Comparison of the separation quality on the GRID and TCD-TIMIT datasets using binary and ratio masking, along with a comparison to the audio-only separation method of Huang *et al.* [10] and raw *vid2speech* [6] predictions.

of two female speakers (*lipspeaker 3* as target and *lipspeaker 2* as background). Here we also present a comparison to the raw speech predictions generated by *vid2speech*. We use the spectrograms of ground truth source signals as an 'oracle' to evaluate an upper bound for the performance of our method.

We also evaluated our enhancement method qualitatively on mixtures of speech and non-speech background noise, examples of which can been seen on our project webpage.

Speech separation of unknown speakers Attempts to predict speech of an unknown speaker using a model trained on a different speaker usually led to bad results. In this experiment, we attempted to separate the speech of two 'unknown' speakers. First, we trained a *vid2speech* network [5] on the data of a 'known' speaker (*S2* from GRID). The training data consisted of randomly selected sentences (40 minutes length in total). Before predicting the speech of each one of the 'unknown' speakers (*S3* and *S5* from GRID) as required in the separation method, we fine-tuned the network using a small amount of samples of the actual speaker (5 minutes length in total). Then, we applied the speech separation process to the synthesized mixtures of unseen sentences spoken by the unknown speakers. The results are summarized in Table 3.

5. CONCLUDING REMARKS

This work has shown that high-quality single-channel speech separation and enhancement can be performed by exploiting visual information. Compared to audio-only techniques mentioned in Sec. 1.1, our method is not affected by the issue of similar speech vocal characteristics as commonly observed in same-gender speech separation, since we gain the disambiguating power of visual information.

The work described in this paper can serve as a basis for several future research directions. These include using a less constrained audio-visual dataset consisting of real-world multi-speaker and noisy recordings. Another interesting point to consider is improving the performance of voice recognition systems using our enhancement methods. Implementing a similar speech enhancement system in an end-to-end manner may be a promising direction as well.

	SNR	PESQ
GRID		
Noisy	-0.63	1.83
Vid2speech [6]	-2.51	1.93
Ours	**2.11**	**1.97**
Ideal enhancement	2.82	2.4
TCD-TIMIT		
Noisy	0.97	2.19
Vid2speech [6]	-11.19	1.42
Ours	**4.52**	**2.1**
Ideal enhancement	9.28	2.41

Table 2: Evaluation of the enhancement quality using LTSS as the mask thresholding function.

	SDR	SIR	SAR	PESQ
Noisy	0.04	0.04	36.14	2.14
Vid2speech [6]	-16.37	6.55	-15.19	1.76
Ours - binary mask	1.85	**8.61**	4.06	1.74
Ours - ratio mask	**3.06**	5.86	**7.9**	**2.42**
Ideal binary mask	10.07	21.7	10.5	2.99
Ideal ratio mask	9.55	13.44	12.24	3.65

Table 3: Comparison of the separation quality of unknown speakers from GRID corpus using transfer learning.

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

6. REFERENCES

[1] Adelbert 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.

[2] Yariv Ephraim and David 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.

[3] Zhuo Chen, *Single Channel auditory source separation with neural network*, Ph.D. thesis, Columbia Univ., 2017.

[4] Yusuf Isik, Jonathan Le Roux, Zhuo Chen, Shinji Watanabe, and John R Hershey, "Single-channel multi-speaker separation using deep clustering," *arXiv:1607.02173*, 2016.

[5] Ariel Ephrat and Shmuel Peleg, "Vid2speech: speech reconstruction from silent video," in *ICASSP'17*, 2017.

[6] Ariel Ephrat, Tavi Halperin, and Shmuel Peleg, "Improved speech reconstruction from silent video," in *ICCV 2017 Workshop on Computer Vision for Audio-Visual Media*, 2017.

[7] Thomas Le Cornu and Ben Milner, "Generating intelligible audio speech from visual speech," in *IEEE/ACM Trans. Audio, Speech, and Language Processing*, 2017.

[8] Aarthi M Reddy and Bhiksha Raj, "Soft mask methods for single-channel speaker separation," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 6, pp. 1766–1776, 2007.

[9] Zhaozhang Jin and DeLiang Wang, "A supervised learning approach to monaural segregation of reverberant speech," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 625–638, 2009.

[10] Po-Sen Huang, Minje Kim, Mark Hasegawa-Johnson, and Paris Smaragdis, "Deep learning for monaural speech separation," in *ICASSP'14*, 2014.

[11] Morten Kolbaek, Dong Yu, Zheng-Hua Tan, and Jesper Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 25, pp. 1901–1913, 2017.

[12] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng, "Multimodal deep learning," in *ICML'11*, 2011, pp. 689–696.

[13] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman, "Lip reading sentences in the wild," *arXiv:1611.05358*, 2016.

[14] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H Adelson, and William T Freeman, "Visually indicated sounds," in *CVPR'16*, 2016.

[15] Yusuf Aytar, Carl Vondrick, and Antonio Torralba, "Soundnet: Learning sound representations from unlabeled video," in *NIPS'16*, 2016, pp. 892–900.

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

[17] Wenwu Wang, Darren Cosker, Yulia Hicks, S Saneit, and Jonathon Chambers, "Video assisted speech source separation," in *ICASSP'05*, 2005.

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

[19] Faheem Khan, *Audio-visual speaker separation*, Ph.D. thesis, University of East Anglia, 2016.

[20] Jen-Cheng Hou, Syu-Siang Wang, Ying-Hui Lai, Jen-Chun Lin, Yu Tsao, Hsiu-Wen Chang, and Hsin-Min Wang, "Audio-visual speech enhancement based on multimodal deep convolutional neural network," *arXiv:1703.10893*, 2017.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR'16*, 2016, pp. 770–778.

[22] Paul Viola and Michael J Jones, "Robust real-time face detection," *IJCV*, vol. 57, no. 2, pp. 137–154, 2004.

[23] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu 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.

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

[25] John S Garofolo, Lori F Lamel, William M Fisher, Jonathon G Fiscus, and David S Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *NISTIR 4930*, 1993.

[26] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte, "Performance measurement in blind audio source separation," *IEEE trans. audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[27] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *ICASSP'01*, 2001.