Audio-Visual Speech Enhancement

Solving the Cocktail Party Problem with Recurrent Neural Networks

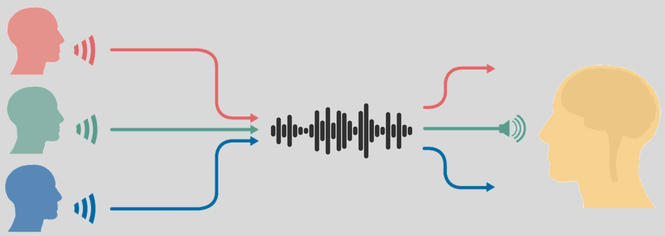
Joint work with Luca Pasa and Leonardo Badino (IIT)

Giovanni Morrone (giovanni.morrone@unimore.it)

PhD “Enzo Ferrari” in Industrial and Environmental Engineering (E4E Doctorate School), Cycle XXXIII, Industry 4.0 Curriculum, Tutor: Prof. Sonia Bergamaschi

The Cocktail Party Problem

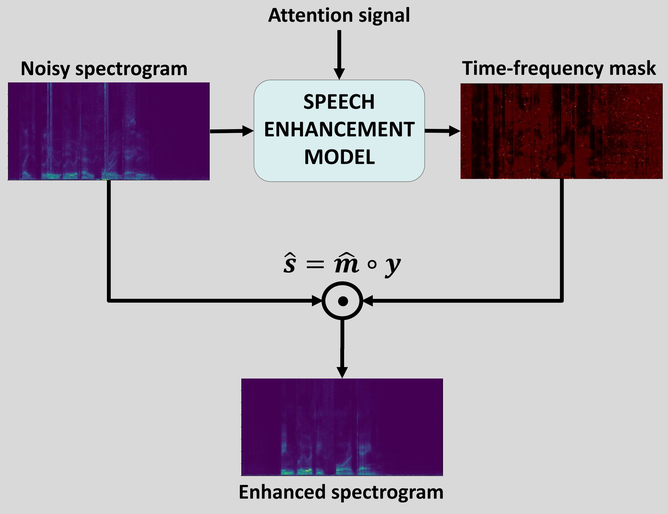
How to amplify one voice among many sound sources.



In the context of speech perception, this ability of human brain is called **cocktail party effect** [1].

Speech Enhancement

Speech enhancement aims at improving speech quality of degraded/noisy speech through computational methods. In our case, the noisy speech is composed by a mixture of two or more concurrent voices.

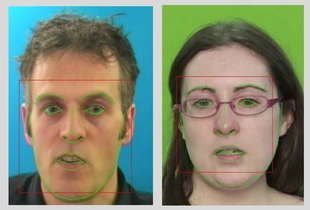


Task: time-frequency mask $\hat{\boldsymbol{m}}$ estimation from noisy audio spectrogram $\boldsymbol{y}$ and attention signal [2] (e.g. additional information about target speaker, prior knowledge about speech signal properties).

Our approach

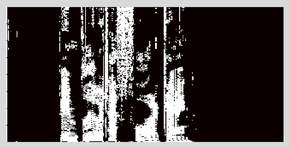
Visual features: face landmarks

Exploiting **facial movements of the speaker of interest** as attention signal. We use a pre-trained face landmarks extractor [3], so our models do not have to learn useful visual features from raw pixels and do not require huge audio-visual datasets.



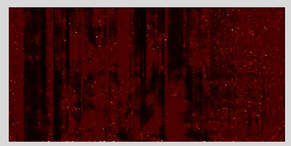
Targets: time-frequency masks

Target Binary Mask (TBM)



- Binary (1: speech; 0: noise/silence)
- Acoustic context independent
- Approximate reconstruction

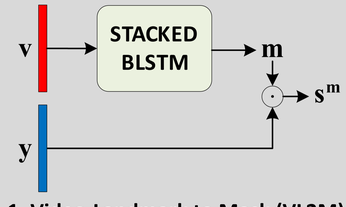
Ideal Binary Mask (IAM)



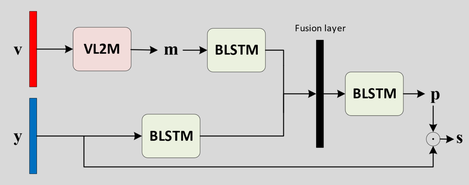
- Real-valued
- Acoustic context dependent
- Perfect reconstruction

Models

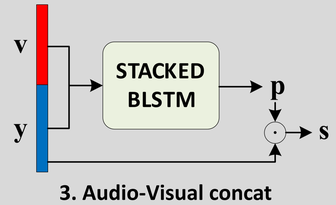
We experimented with four models [4]. All models receive in input the target speaker’s landmark motion vectors and the power-law compressed spectrogram of the single-channel mixed-speech signal.



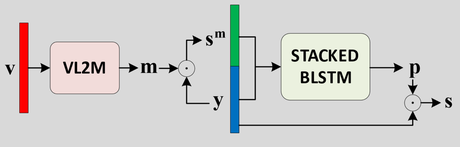
1. Video-Landmark to Mask (VL2M)



2. VL2M refinement



3. Audio-Visual concat



4. Audio-Visual concat-refinement

Legenda

v: video input **y**: noisy spectrogram **m**: TBM **p**: IAM
$\mathbf{s}^{\mathbf{m}}$: clean spectrogram TBM **s**: clean spectrogram IAM

Experimental results

The models are trained with mixture of two speakers in a speaker-independent setting (e.g. speakers in test set are unseen in training set). The models are evaluated both in 2-speakers and in 3-speakers scenario.

GRID	2 Speakers			3 Speakers		
	SDR	PESQ	ViSQOL	SDR	PESQ	ViSQOL
Noisy	0.21	1.94	2.58	-5.34	1.43	1.62
VL2M	3.02	1.81	1.70	-2.03	1.43	1.25
VL2M-ref	6.52	2.53	3.02	2.83	2.19	2.53
AV concat	7.37	2.65	3.03	3.02	2.24	2.49
AV c-ref	**8.05**	**2.70**	**3.07**	**4.02**	**2.33**	**2.64**

TCD-TIMIT	2 Speakers			3 Speakers		
	SDR	PESQ	ViSQOL	SDR	PESQ	ViSQOL
Noisy	0.21	2.22	2.74	-3.42	1.92	2.04
VL2M	2.88	2.25	2.62	-0.51	1.99	1.98
VL2M-ref	9.24	2.81	3.09	5.27	2.44	2.54
AV concat	9.56	2.80	3.09	5.15	2.41	2.52
AV c-ref	**10.55**	**3.03**	**3.21**	**5.37**	**2.45**	**2.58**

Conclusion

The proposed architectures are the first models trained and evaluated on the limited size GRID and TCD-TIMIT datasets that accomplish **speaker-independent speech enhancement in multi-talker setting**. We show that face landmark motion features are very effective features, and, most importantly, that huge audio-visual datasets are not a necessary requirement for this task.

References

[1]. Josh H McDermott, “The cocktail party problem,” *Current Biology*, vol. 19, no. 22, pp. R1024–R1027, 2009.

[2] Yuxuan Wang, Arun Narayanan, and DeLiang Wang, “On Training Targets for Supervised Speech Separation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849–1858, Dec. 2014.

[3] Vahid Kazemi and Josephine Sullivan, “One millisecond face alignment with an ensemble of regression trees,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[4] Giovanni Morrone, Luca Pasa, Vadim Tikhanoff, Sonia Bergamaschi, Luciano Fadiga, and Leonardo Badino, “Face Landmark-based Speaker-Independent Audio-Visual Speech Enhancement in Multi-Talker Environments,” in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019.