FACE LANDMARK-BASED SPEAKER-INDEPENDENT AUDIO-VISUAL SPEECH ENHANCEMENT IN MULTI-TALKER ENVIRONMENTS

Giovanni Morrone, Luca Pasa, Vadim Tikhanoff, Sonia Bergamaschi, Luciano Fadiga, Leonardo Badino
University of Modena and Reggio Emilia, Istituto Italiano di Tecnologia

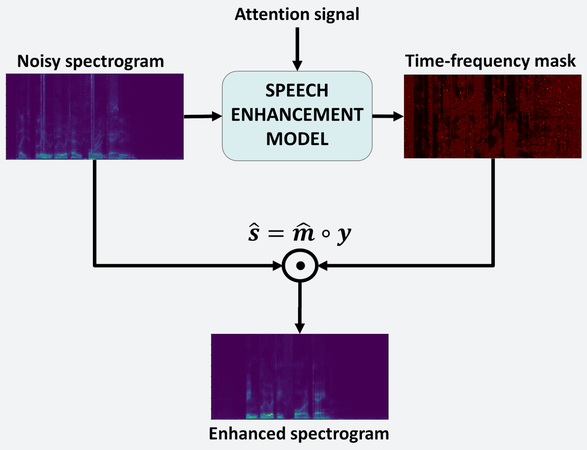
THE COCKTAIL PARTY PROBLEM

How to extract one target speaker's speech among many concurrent sounds.

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

PROBLEM DESCRIPTION

Input: mixture of two or more concurrent voices + "attention" signal (e.g. additional information about target speaker, prior knowledge about speech signal properties).
Output: time-frequency mask [2].



$\hat{\mathbf{s}}$: clean spectrogram $\hat{\mathbf{m}}$: time-frequency mask
$\mathbf{y}$: noisy spectrogram

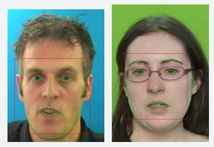
Project page

Contact information:
giovanni.morrone@unimore.it
luca.pasa@iit.it
leonardo.badino@iit.it

44th IEEE International Conference on Acoustics, Speech and Signal Processing
12-17 May, 2019 - Brighton, UK

AUDIO-VISUAL SPEECH ENHANCEMENT MODELS

"Attention" signal: motion of face landmarks of target speaker



- A pre-trained face landmarks extractor [3] is used.
- Our speech enhancement models do not have to learn useful visual features from raw pixels.
- Visual features extraction does not require parallel audio-visual datasets.

Targets: time-frequency masks

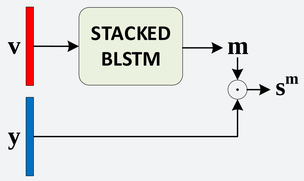
The models estimate two different masks to extract the spectrogram of a target speaker:

- **Target Binary Mask (TBM):** binary (1: speech; 0: noise/silence), acoustic context independent, approximated reconstruction.
- **Ideal Amplitude Mask (IAM):** real-valued, acoustic context dependent, perfect reconstruction.

Models

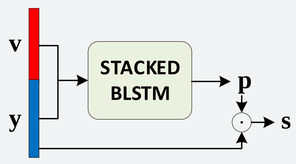
The models receive in input the target speaker's landmark motion vectors and the power-law compressed spectrogram of the single-channel mixed-speech signal. All models contain bi-directional LSTM layers of 250 hidden units.

Video-Landmark to Mask (VL2M)



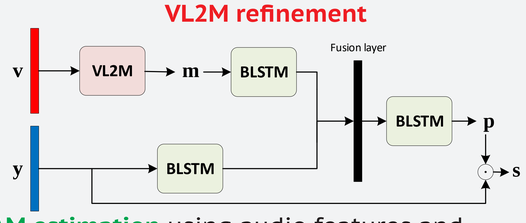
TBM estimation using visual features only.
Model: 5-layer BLSTM
Loss: $J_{vl2m} = \sum_{t=1}^{T}\sum_{f=1}^{d} -\mathbf{m}_t[f]\cdot\log(\hat{\mathbf{m}}_t[f]) - (1-\mathbf{m}_t[f])\cdot\log(1-\hat{\mathbf{m}}_t[f])$

Audio-Visual concat



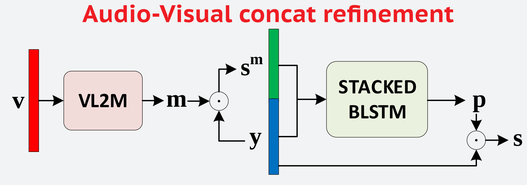
IAM estimation using audio-visual features concatenation.
Model: 3-layer BLSTM
Loss: $J_{mr} = \sum_{t=1}^{T}\sum_{f=1}^{d}(\hat{\mathbf{p}}_t[f]\cdot\mathbf{y}_t[f] - \mathbf{s}_t[f])^2$

VL2M refinement



IAM estimation using audio features and TBM estimated by VL2M component.

Audio-Visual concat refinement



IAM estimation using audio features and spectrogram denoised by VL2M operation.

Loss: J_{mr}
Training strategy: 1. Pre-training using oracle TBM
2. Fine-tuning replacing oracle TBM with VL2M component (VL2M weights are frozen)

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

EXPERIMENTAL RESULTS

The models are trained with mixture of two speakers in a speaker-independent setting.

GRID	2 Speakers SDR	2 Speakers PESQ	3 Speakers SDR	3 Speakers PESQ
Noisy	0.21	1.94	−5.34	1.43
VL2M	3.02	1.81	−2.03	1.43
VL2M-ref	6.52	2.53	2.83	2.19
AV concat	7.37	2.65	3.02	2.24
AV c-ref	**8.05**	**2.70**	**4.02**	**2.33**

TCD-TIMIT	2 Speakers SDR	2 Speakers PESQ	3 Speakers SDR	3 Speakers PESQ
Noisy	0.21	2.22	−3.42	1.92
VL2M	2.88	2.25	−0.51	1.99
VL2M-ref	9.24	2.81	5.27	2.44
AV concat	9.56	2.80	5.15	2.41
AV c-ref	**10.55**	**3.03**	**5.37**	**2.45**

- A successful mask generation has to depend on the acoustic context.
- Mask refinement is more effective when it directly refines the estimated clean spectrogram with TBM.

CONCLUSION

- The proposed models are the first trained and evaluated on the limited size GRID and TCD-TIMIT datasets that accomplish **speaker-independent speech enhancement in multi-talker setting**.
- Experiments show that **face landmark motion** features are very effective.
- Our models need a **small amount of training data** to achieve very good results.

REFERENCES

[1] Cherry, E. C. (1953) Some Experiments on the Recognition of Speech, with One and with Two Ears. The Journal of the Acoustical Society of America, 25(5), 975–979.
[2] Yuxuan Wang, Narayanan, A., and DeLiang Wang (December, 2014) On Training Targets for Supervised Speech Separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 22(12), 1849–1858.
[3] Kazemi, V. and Sullivan, J. (June, 2014) One Millisecond Face Alignment with an Ensemble of Regression Trees. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).