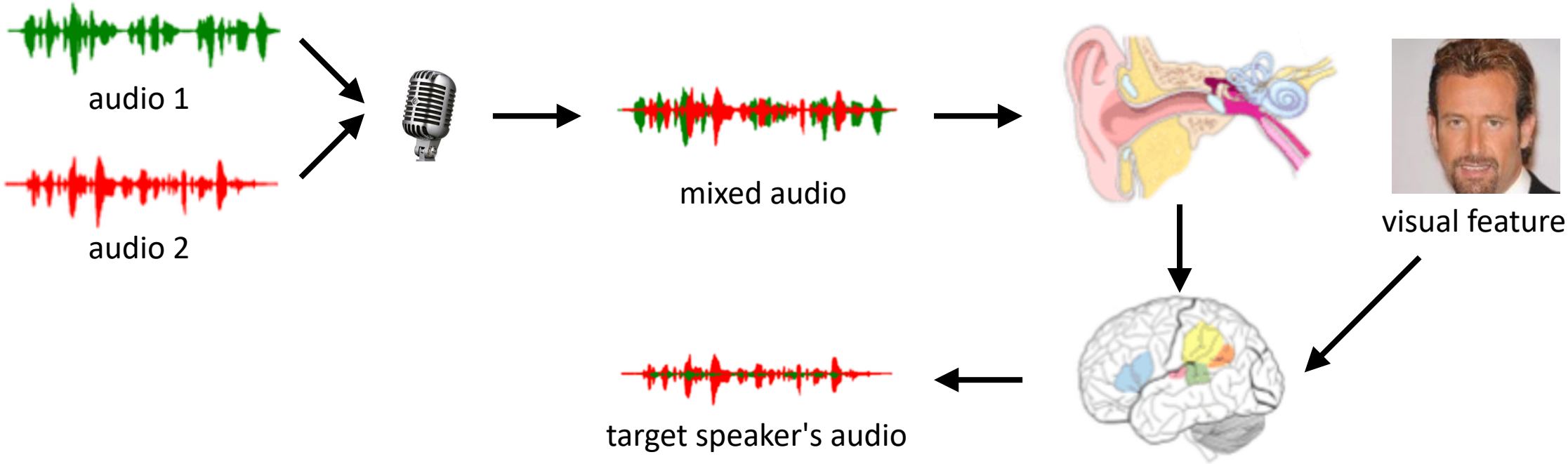
Multi-Modal Multi-Correlation Learning for Audio-Visual Speech Separation

Xiaoyu Wang[1,2], Xiangyu Kong[2], Xiulian Peng[2], Yan Lu[2]
[1]Xi'an Jiaotong University
[2]Microsoft Research Asia

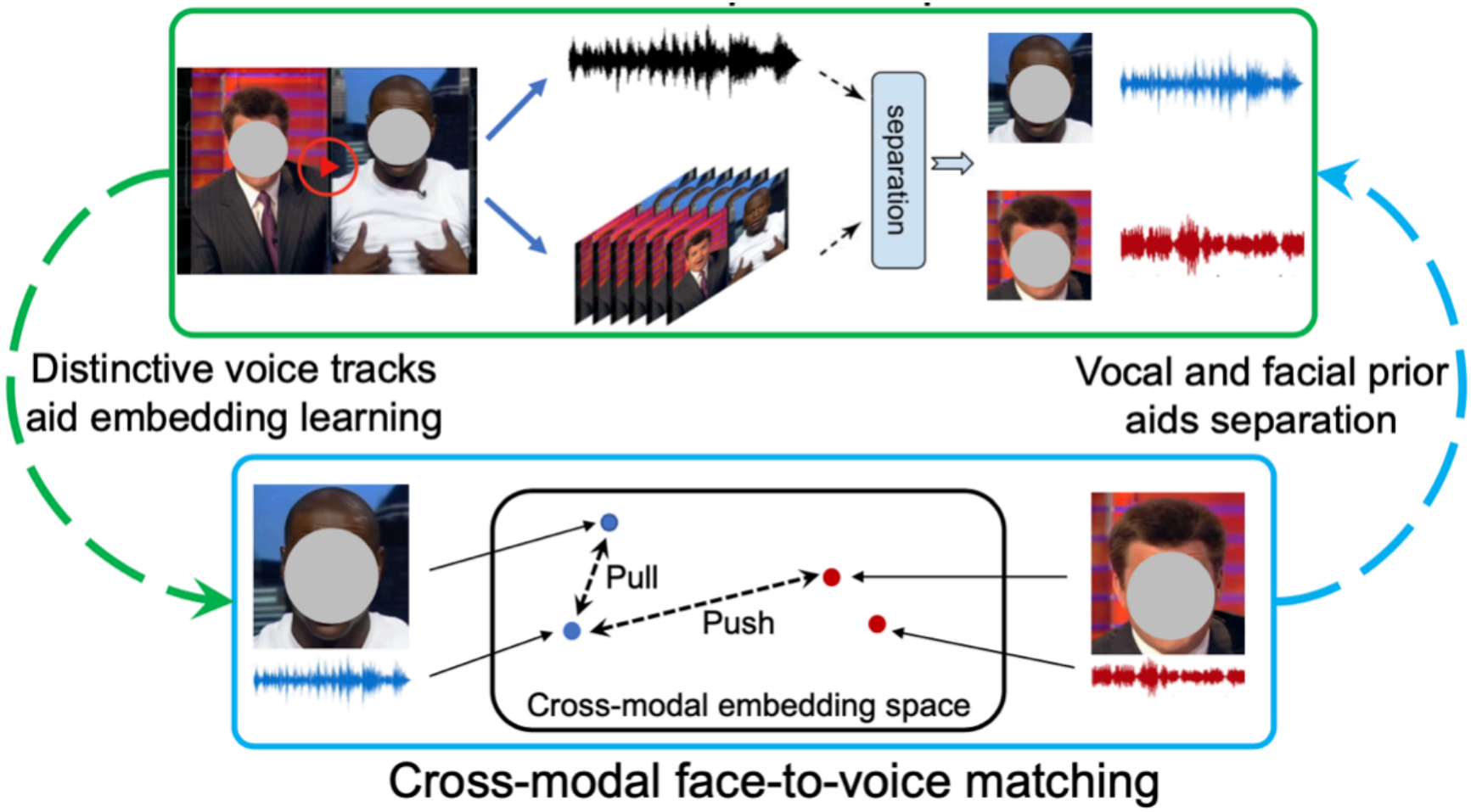
Audio-Visual Speech Separation

- Given the mixed audio and target speaker's visual features, our goal is to separate the target speaker's voice



Previous Work

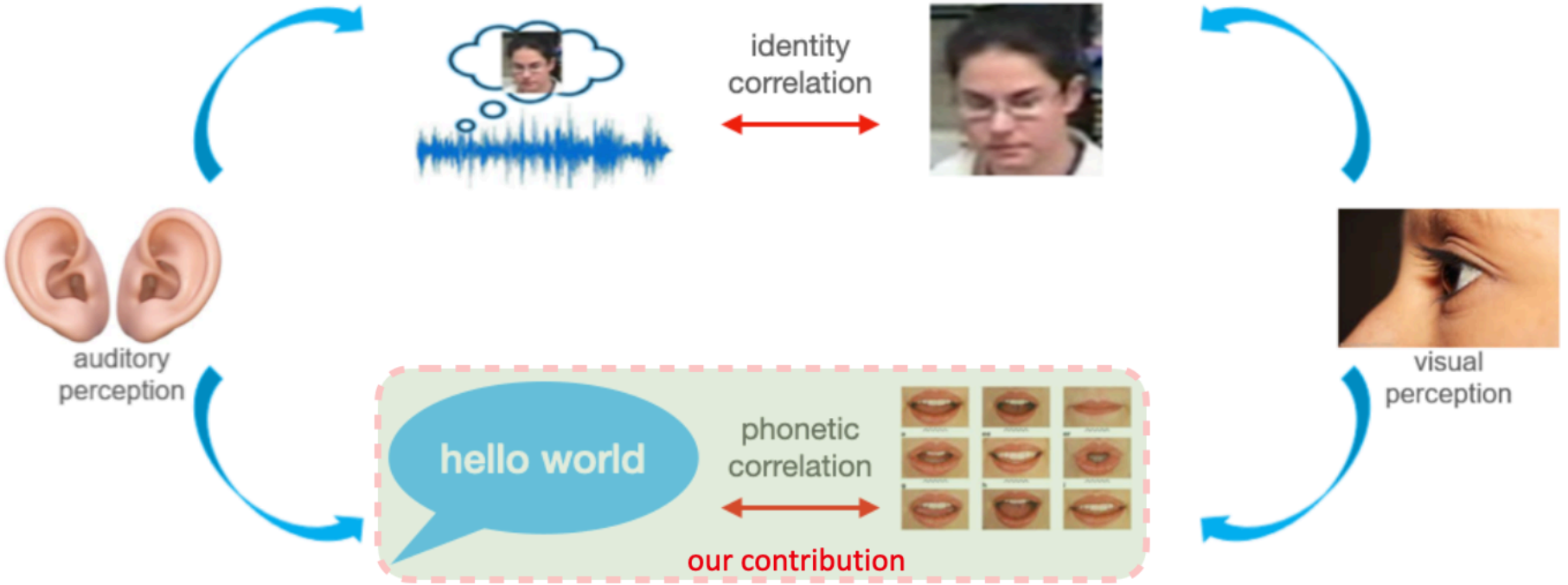
- VisualVoice[1]:explicitly model the audio-visual identity correlation



[1] R. Gao and K. Grauman. "VisualVoice: Audio-Visual Speech Separation with Cross-Modal Consistency". In CVPR 2021.

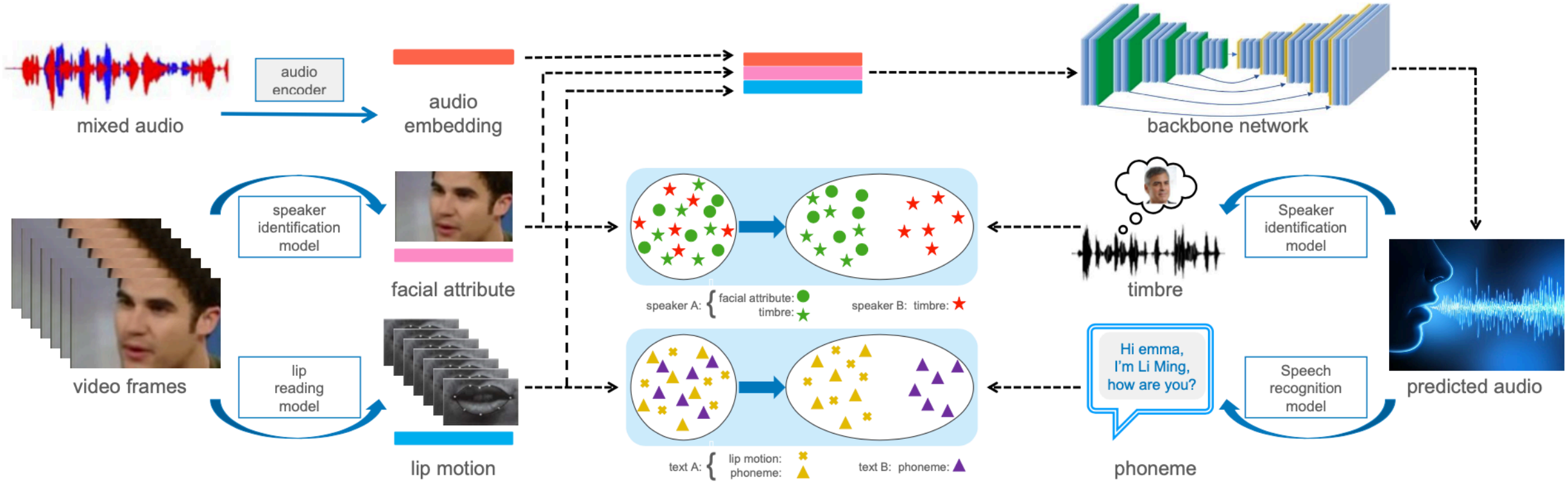
Audio-Visual Correlations

- Besides modeling the speaker identity, we propose to explicitly model the phonetic correlation between the audio (phoneme) and video (lip motion)



Pipeline

- Correlation enhancement in embedding space



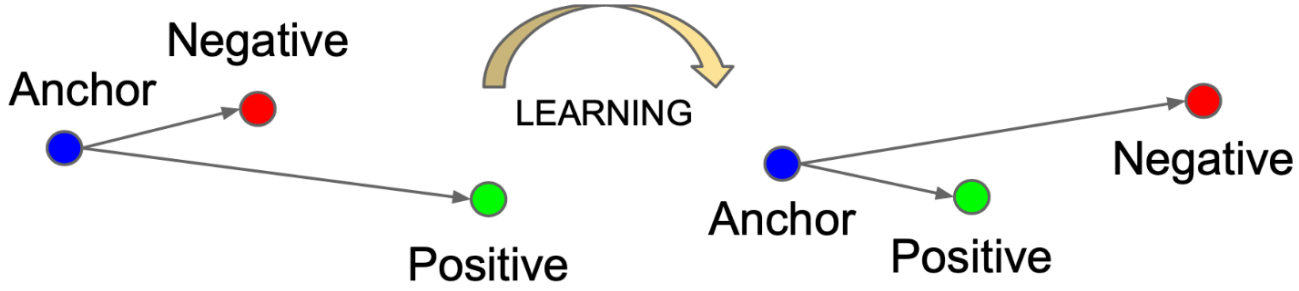
Learning Methods

- Contrastive learning

Triplet loss:
Define a triplet (positive, anchor, negative),
pull the positive and anchor closer,
and push the negative and anchor farther away.

$$\mathcal{L}_1 = \max\{d(\mathbf{i}^a_{\mathcal{A}_1}, \mathbf{i}^v_{\mathcal{A}_2}) - d(\mathbf{i}^a_{\mathcal{A}_1}, \mathbf{i}^v_{\mathcal{B}}) + m, 0\}$$

$$\mathcal{L}_2 = \max\{d(\mathbf{p}^a_{\mathcal{A}}, \mathbf{p}^v_{\mathcal{A}}) - d((\mathbf{p}^a_{\mathcal{A}}, \mathbf{p}^v_{\mathcal{B}}) + m, 0\}$$

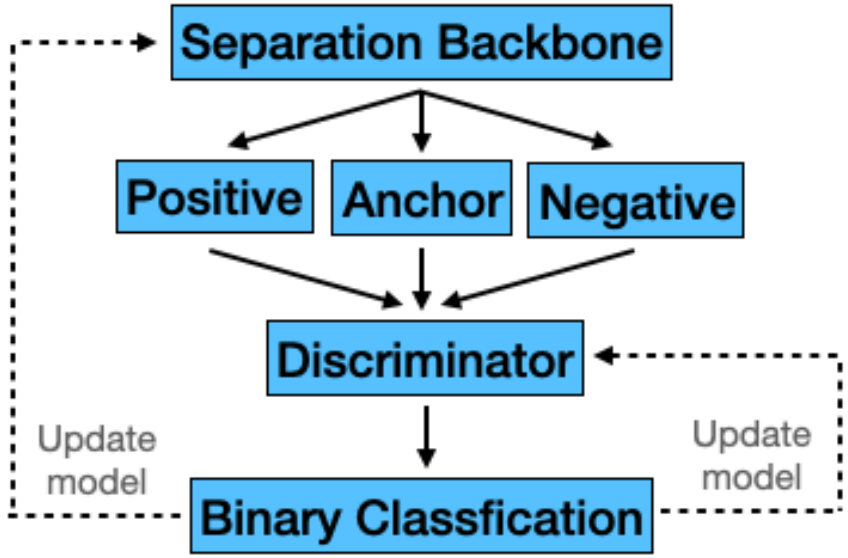


- Adversarial learning

Limitation of triplet loss:
When using cosine distance, the magnitude of vectors is not taken into account, while merely their direction information is included.

$$\mathcal{L}_{\mathcal{G}} = \min_G \mathbb{E}_{\mathbf{x}\sim\mathbf{i}^v} \log(D(\mathbf{x})) + \mathbb{E}_{\mathbf{x}\sim\mathbf{i}^a} \log(1 - D(\mathbf{x}))$$

$$\mathcal{L}_{\mathcal{D}} = \max_D \mathbb{E}_{\mathbf{x}\sim\mathbf{i}^v} \log(D(\mathbf{x})) + \mathbb{E}_{\mathbf{x}\sim\mathbf{i}^a} \log(1 - D(\mathbf{x}))$$

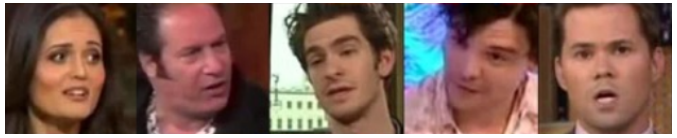


Experiment

- LRS3 Dataset[3]

	SDR	PESQ	STOI
[2](AV Baseline)	8.46	2.27	0.843
[2](CMC loss)	8.85	2.39	0.854
Ours(AV baseline)	9.392	2.536	0.851
Ours(triplet)	**9.623**	**2.545**	**0.855**
Ours(adversarial)	**9.982**	**2.584**	**0.861**

- VoxCeleb2 Dataset[4]

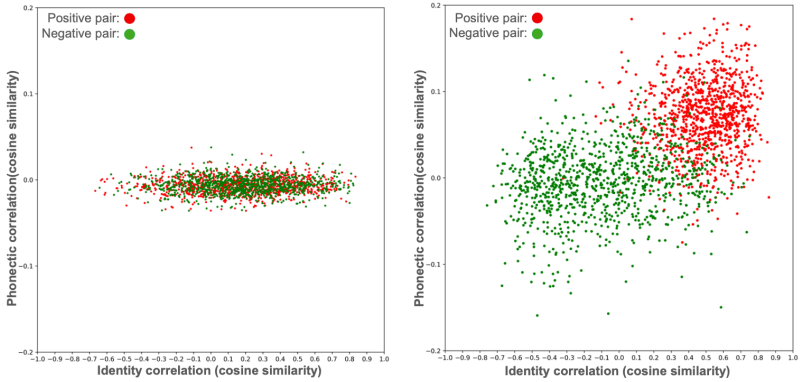


	SDR	SIR	SAR	PESQ	STOI	SI-SNR
[1](Reported)	10.2	17.2	11.3	2.83	0.87	-
[1](Released)	7.023	13.708	9.546	2.569	0.792	6.471
[1](Our impl.)	7.692	14.347	10.195	2.579	0.791	7.467
Ours(triplet)	**8.178**	**14.692**	**10.38**	**2.6**	**0.793**	**7.676**
Ours(adversarial)	**8.949**	**16.012**	**10.79**	**2.687**	**0.811**	**8.477**

[1] R. Gao and K. Grauman. "VisualVoice: Audio-Visual Speech Separation with Cross-Modal Consistency". In CVPR 2021.
[2] N. Makishima, M. Ihori, A. Takashima, T. Tanaka, S. Orihashi, and R. Masumura, “Audio-visual speech separation using cross- modal correspondence loss”. In *ICASSP 2021*.
[3] T. Afouras, J. S. Chung, A. Zisserman LRS3-TED: a large-scale dataset for visual speech recognition arXiv preprint arXiv:1809.00496
[4] J. S. Chung*, A. Nagrani*, A. Zisserman VoxCeleb2: Deep Speaker Recognition INTERSPEECH, 2018.

Analysis

- Identity&phonetic correlation before/after training



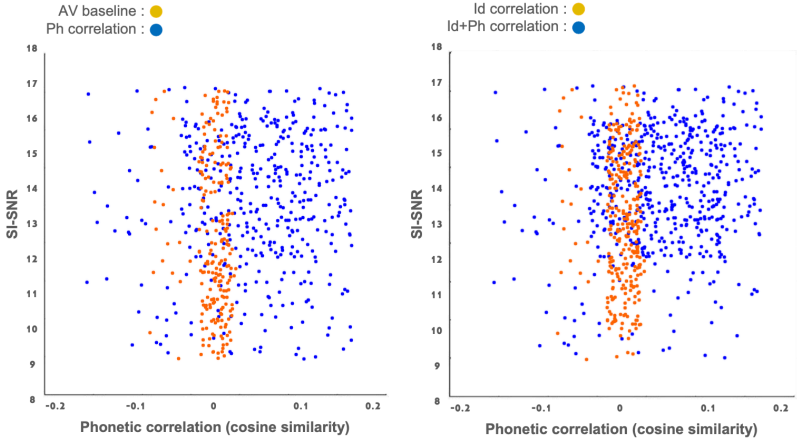
(a) AV baseline

(b) Our method

(a) AV baseline: without correlation learning.

(b) Our method: after joint identity & phonetic correlation learning.

- Separation metric after phonetic correlation learning



(a)

(b)

(a) AV baseline vs. learning phonetic correlation (Ph).

(b) Learning identity correlation (Id) vs. jointly learning both identity and phonetic correlation (Id+Ph).

Conclusion

- Contribution:
  - We explicitly model the phonetic correlation between audio (phoneme) and video (lip motion)
  - An adversarial training approach to learn identity and phonetic audio-visual correlation

- Future Work
  - We target at directly learning correlated audio-visual representations and apply it to downstream tasks

Thanks!