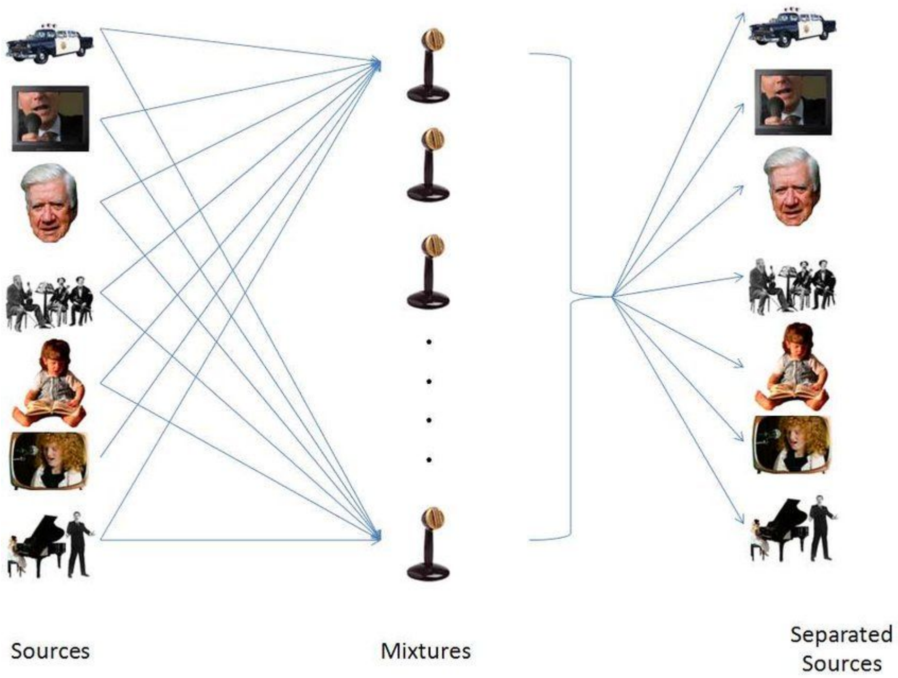
Audio-Visual Speech Separation

Kranti Kumar Parida

Feb. 8, 2022

Source Separation

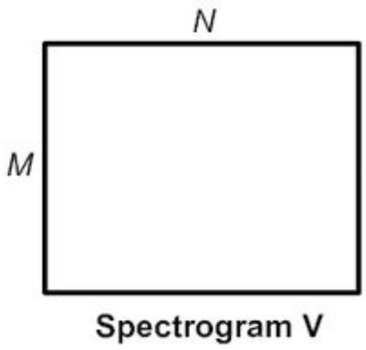
Cocktail Party Problem



https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.researchgate.net%2Ffigure%2FIllustration-du-cocktail-party-problem_fig1_281534016&psig=AOvVaw1Y52VwELFMBjbRjwRLYRUw&ust=1644216902631000&source=images&cd=vfe&ved=0CAsQjRxqFwoTCKCV5bu_6vUCFQAAAAAdAAAAABAK

Audio Only Methods

Non-Negative Matrix Factorization



Permutation Invariant Training

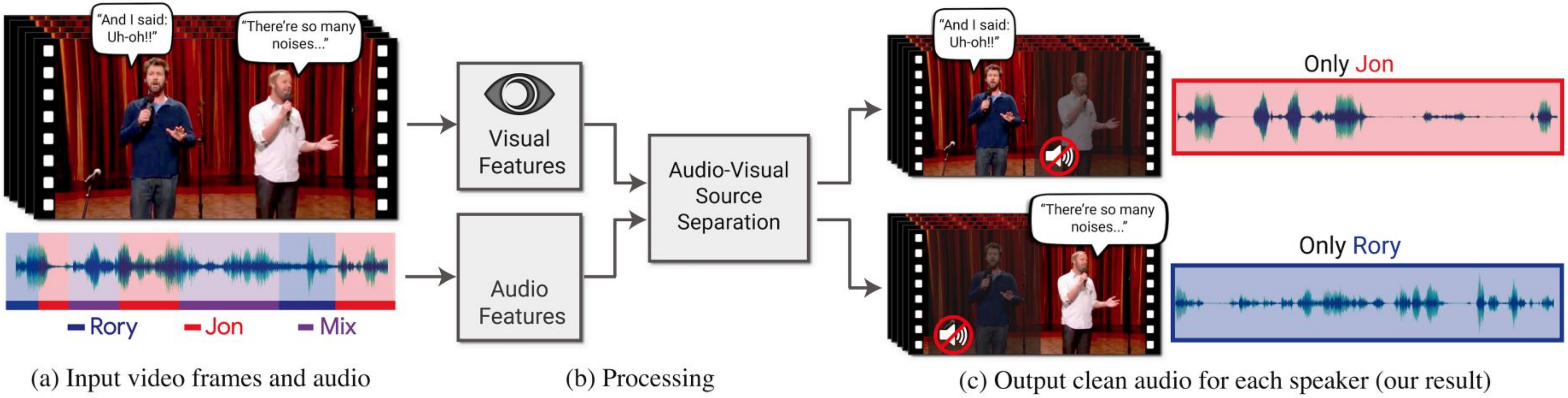
Looking to Listen at the Cocktail Party: A Speaker-Independent Audio-Visual Model for Speech Separation

ARIEL EPHRAT, Google Research and The Hebrew University of Jerusalem, Israel
INBAR MOSSERI, Google Research
ORAN LANG, Google Research
TALI DEKEL, Google Research
KEVIN WILSON, Google Research
AVINATAN HASSIDIM, Google Research
WILLIAM T. FREEMAN, Google Research
MICHAEL RUBINSTEIN, Google Research

ACM Transactions on Graphics (TOG). Proc. SIGGRAPH 2018
https://arxiv.org/pdf/1804.03619.pdf

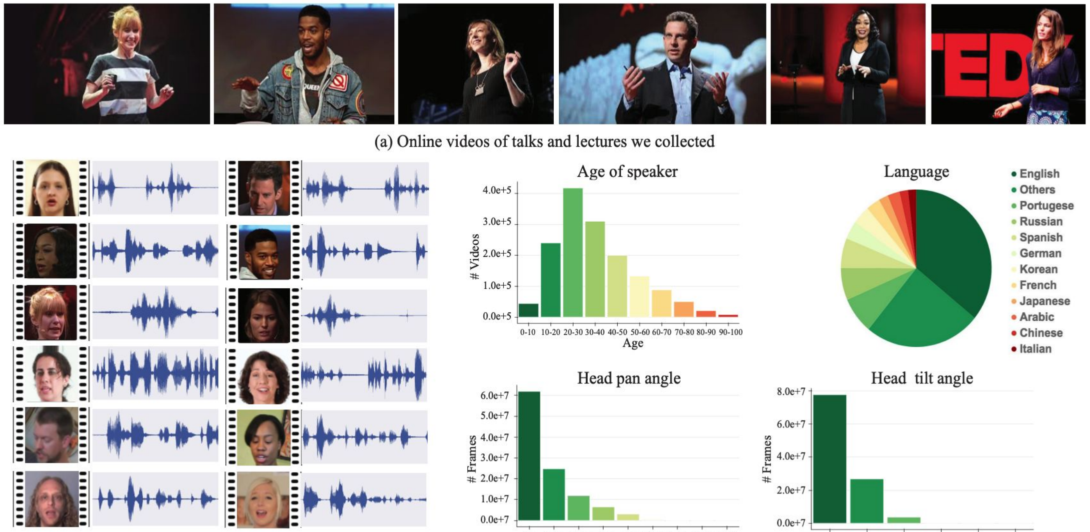
Introduction

- Isolating a single speech signal from a mixture of sounds
- Humans are capable, Better when looking at the person speaking
- Existing approaches - speaker dependent



AVSpeech Dataset

- 290k high quality lectures and TED videos - visible speakers, clean speech
- Dataset Creation Pipeline
  - Face Tracking for videos
  - Discard videos less than threshold SNR



(a) Online videos of talks and lectures we collected

(b) Video segments with localized speakers and clean speech (which comprise our dataset)

(c) Dataset statistics

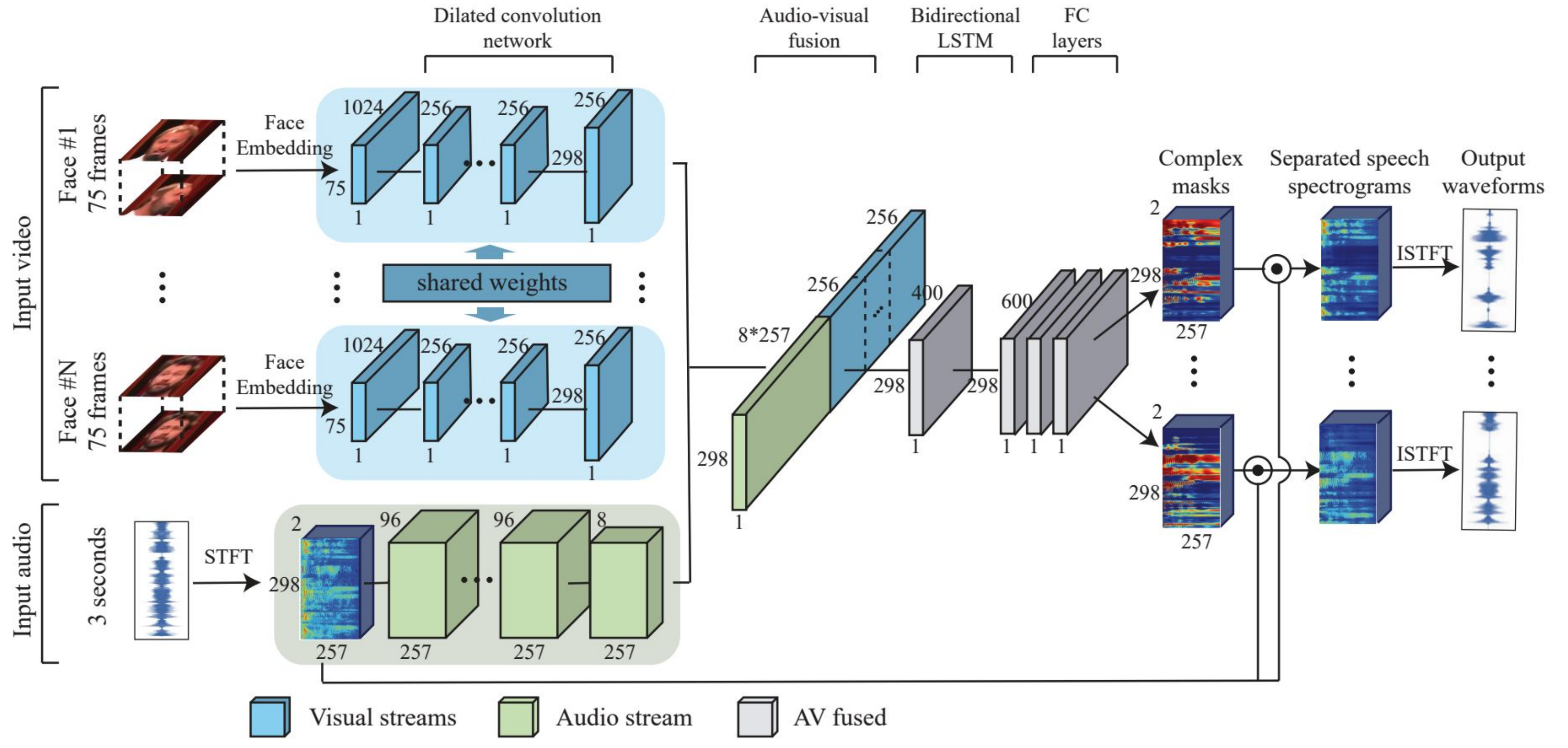
Training Data

- Self Supervised data generation
  - Mix and separate

$$\{A_1, A_2\} \Longrightarrow A_{mix} = A_1 + A_2$$

$$A_{mix} \Longrightarrow \{\hat{A}_1, \hat{A}_2\}$$

Approach



Dilated convolution network
Audio-visual fusion
Bidirectional LSTM
FC layers
Input video
Face #1 75 frames
Face #N 75 frames
Face Embedding
1024
256
298
75
1
shared weights
Complex masks
Separated speech spectrograms
Output waveforms
8*257
400
600
2
257
ISTFT
Input audio
3 seconds
STFT
96
8
Visual streams
Audio stream
AV fused

Results

Table 5. **Comparison with existing audio-visual speech separation work.** We compare our speech separation and enhancement results on several datasets to those of previous work, using the evaluation protocols and objective scores reported in the original papers. Note that previous approaches are *speaker-dependent*, whereas our results are obtained by using a general, *speaker-independent* model.

Mandarin (Enhancement)

	Gabbay et al. [2017]	Hou et al. [2018]	Ours
PESQ	2.25	2.42	**2.5**
STOI	-	0.66	**0.71**
SDR	-	2.8	**6.1**

TCD-TIMIT (Separation)

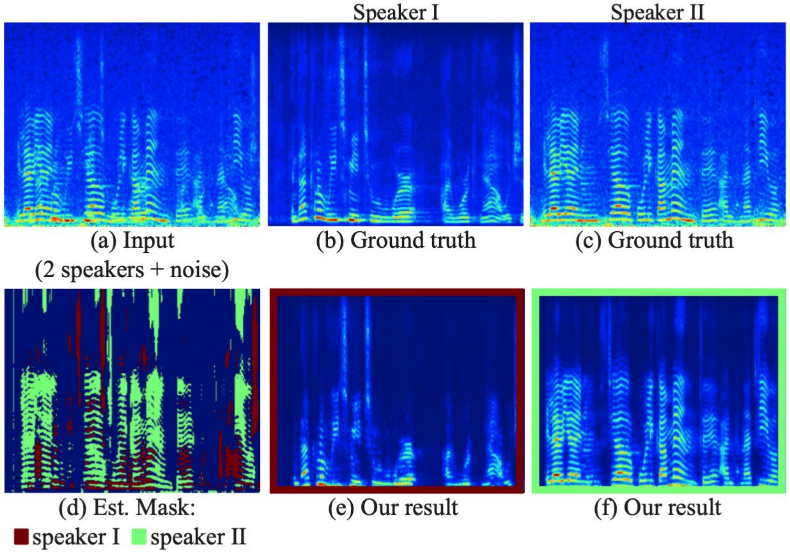
	Gabbay et al. [2017]	Ours
SDR	0.4	**4.1**
PESQ	2.03	**2.42**

CUAVE (Separation)

	Casanovas et al. [2010]	Pu et al. [2017]	Ours
SDR	7	6.2	**12.6**

Table 3. **Quantitative analysis and comparison with audio-only speech separation and enhancement:** Quality improvement (in SDR, see Section A in the Appendix) as function of the number of input visual streams using different network configurations. First row (audio-only) is our implementation of a state-of-the-art speech separation model, and shown as a baseline.

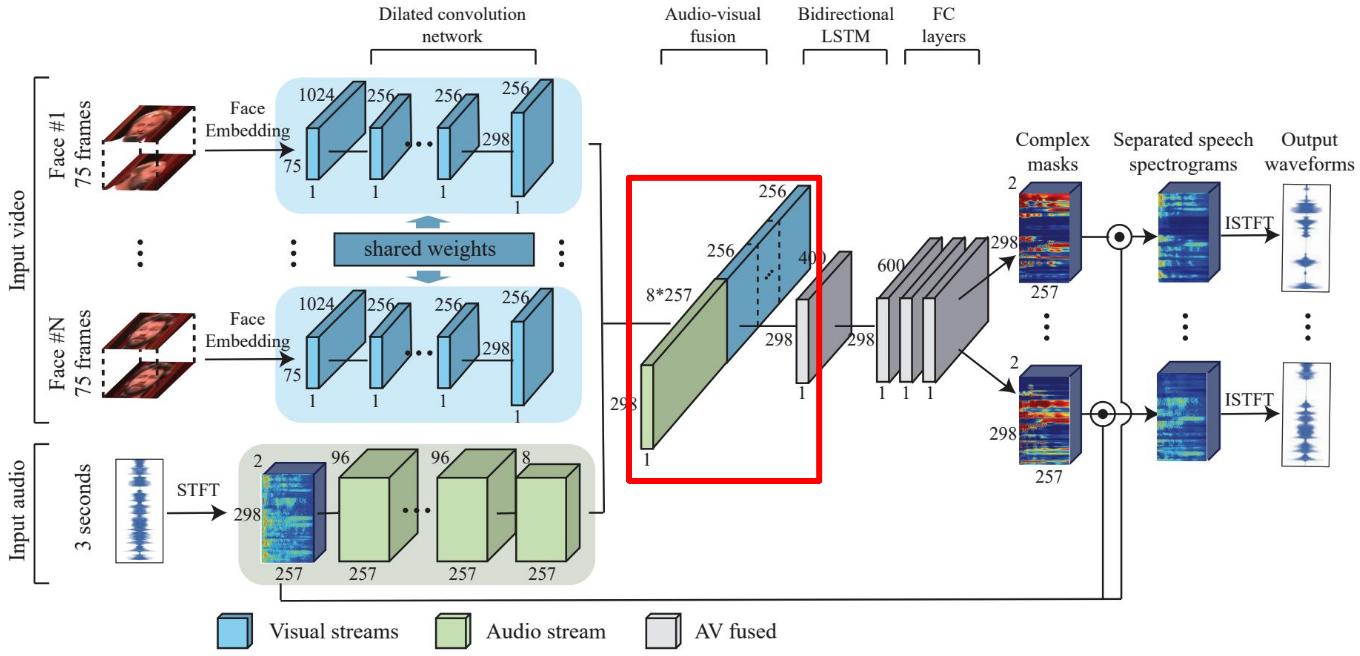
	1S+Noise	2S clean	2S+Noise	3S clean
AO [Yu et al. 2017]	**16.0**	8.6	10.0	8.6
AV - 1 face	**16.0**	9.9	10.1	9.1
AV - 2 faces	-	**10.3**	**10.6**	9.1
AV - 3 faces	-	-	-	**10.0**



Qualitative Results

Drawback

- Architecture different when different no. of speakers



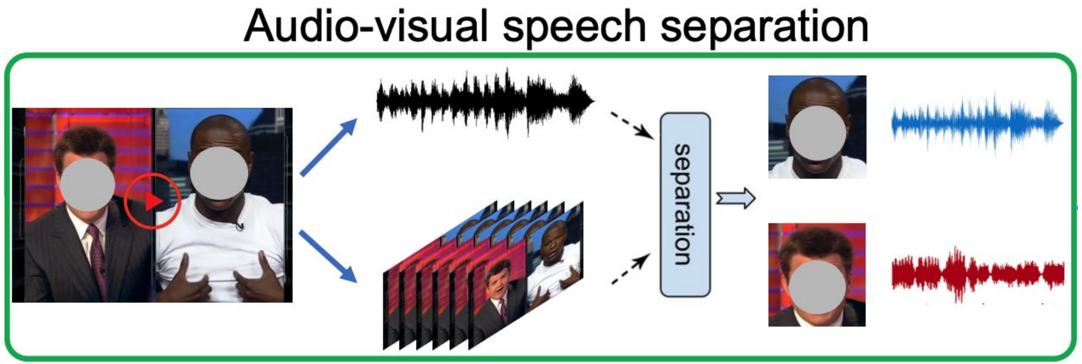
VISUALVOICE: Audio-Visual Speech Separation with Cross-Modal Consistency

Ruohan Gao[1,2] Kristen Grauman[1,3]
[1]The University of Texas at Austin [2]Stanford University [3]Facebook AI Research
rhgao@cs.stanford.edu, grauman@fb.com

CVPR 2021
https://vision.cs.utexas.edu/projects/VisualVoice/gao2021VisualVoice.pdf

Introduction

Goal: Extract speech in spite of background noise/other speaker



Motivation

- Force audio and visual features to be close to each other
- Focus on Lip region

Problem

Video: V

Multiple Speakers: $x(t) = \sum_{k=1}^{K} s_k(t)$

Estimate the individual audio: $s_k(t)$

Training Data

- Getting ground truth data is hard
- Self Supervised data generation - Mix and Separate

$$\text{video } V_{\mathcal{A}} \text{ for speaker } \mathcal{A} \quad s_{\mathcal{A}_1}(t),\ s_{\mathcal{A}_2}(t)$$

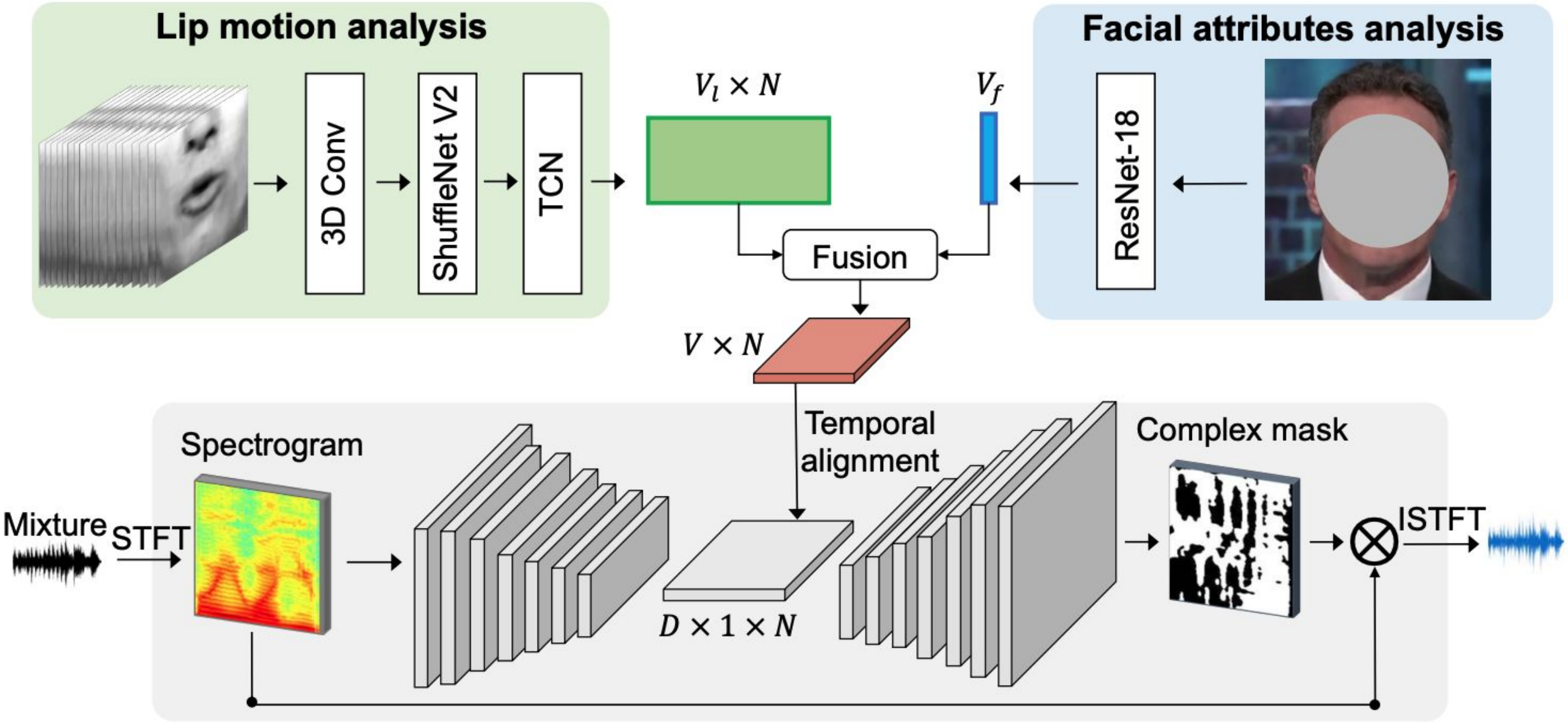
$$\text{video } V_{\mathcal{B}} \text{ for speaker } \mathcal{B} \quad s_{\mathcal{B}}(t)$$

$$x_1(t) = s_{\mathcal{A}_1}(t) + s_{\mathcal{B}}(t), \quad x_2(t) = s_{\mathcal{A}_2}(t) + s_{\mathcal{B}}(t)$$

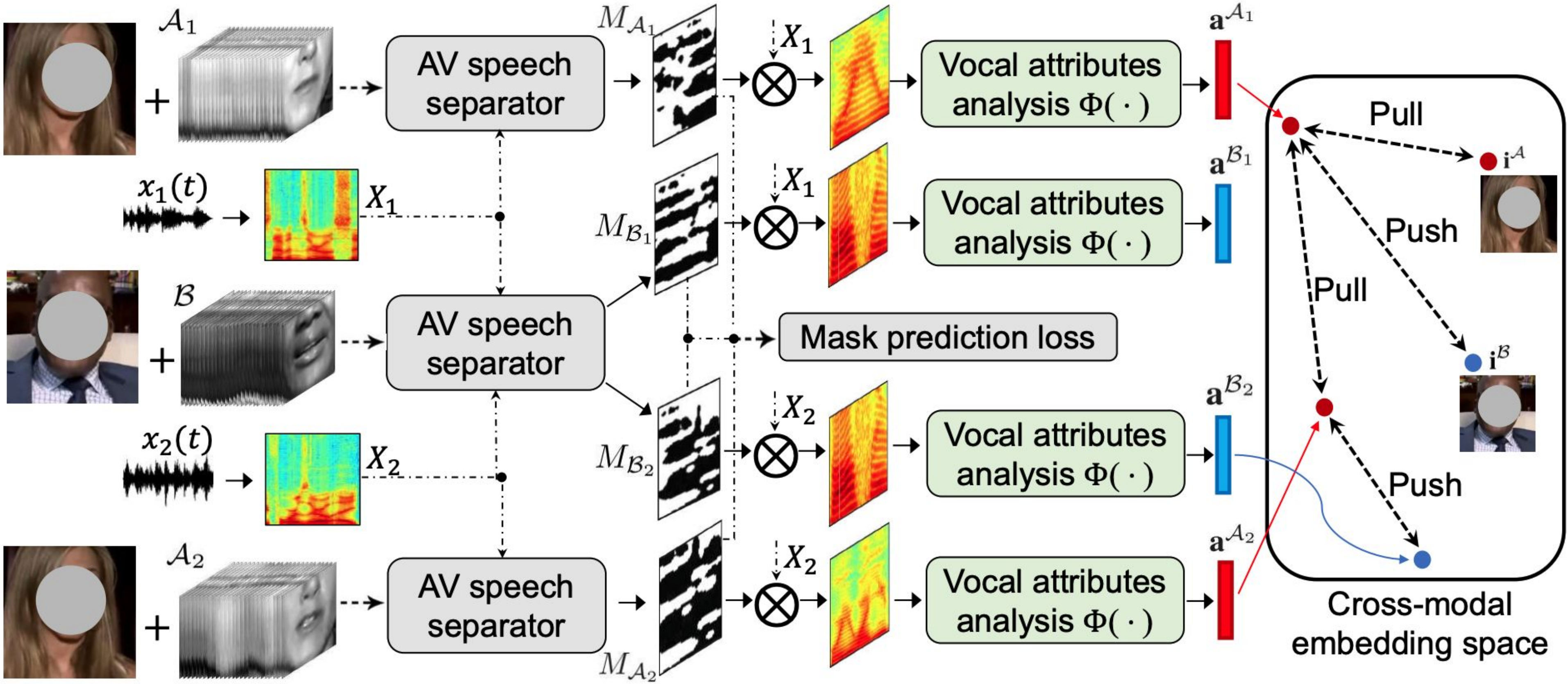
- Trained with spectrograms

$$S_{\mathcal{A}_i} = X_i * M_{\mathcal{A}_i},\ \ S_{\mathcal{B}_i} = X_i * M_{\mathcal{B}_i},\ \ i \in \{1, 2\}$$

AV Speech Separator



Approach



$\mathcal{A}_1$
AV speech separator
$x_1(t)$
X_1
$\mathcal{B}$
AV speech separator
$x_2(t)$
X_2
$\mathcal{A}_2$
AV speech separator
$M_{\mathcal{A}_1}$
$M_{\mathcal{B}_1}$
$M_{\mathcal{B}_2}$
$M_{\mathcal{A}_2}$
X_1
X_1
X_2
X_2
Mask prediction loss
Vocal attributes analysis $\Phi(\cdot)$
Vocal attributes analysis $\Phi(\cdot)$
Vocal attributes analysis $\Phi(\cdot)$
Vocal attributes analysis $\Phi(\cdot)$
$\mathbf{a}^{\mathcal{A}_1}$
$\mathbf{a}^{\mathcal{B}_1}$
$\mathbf{a}^{\mathcal{B}_2}$
$\mathbf{a}^{\mathcal{A}_2}$
Pull
$\mathbf{i}^{\mathcal{A}}$
Push
Pull
$\mathbf{i}^{\mathcal{B}}$
Push
Cross-modal embedding space

Loss Function

$$L = L_{mask\text{-}prediction} + \lambda_1 L_{cross\text{-}modal} + \lambda_2 L_{consistency}$$

$$L_{mask\text{-}prediction} = \sum_{i \in \{\mathcal{A}_1, \mathcal{A}_2, \mathcal{B}_1, \mathcal{B}_2\}} \| M_i - \mathcal{M}_i \|_2$$

$$\begin{aligned} L_{cross\text{-}modal} &= L_t(\mathbf{a}^{\mathcal{A}_1}, \mathbf{i}^{\mathcal{A}}, \mathbf{i}^{\mathcal{B}}) + L_t(\mathbf{a}^{\mathcal{A}_2}, \mathbf{i}^{\mathcal{A}}, \mathbf{i}^{\mathcal{B}}) \\ &+ L_t(\mathbf{a}^{\mathcal{B}_1}, \mathbf{i}^{\mathcal{B}}, \mathbf{i}^{\mathcal{A}}) + L_t(\mathbf{a}^{\mathcal{B}_2}, \mathbf{i}^{\mathcal{B}}, \mathbf{i}^{\mathcal{A}}). \end{aligned}$$

$$L_{consistency} = L_t(\mathbf{a}^{\mathcal{A}_1}, \mathbf{a}^{\mathcal{A}_2}, \mathbf{a}^{\mathcal{B}_1}) + L_t(\mathbf{a}^{\mathcal{A}_1}, \mathbf{a}^{\mathcal{A}_2}, \mathbf{a}^{\mathcal{B}_2})$$

Results

- Improved performance for both speech enhancement and source separation

	Reliable lip motion					Unreliable lip motion				
	SDR	SIR	SAR	PESQ	STOI	SDR	SIR	SAR	PESQ	STOI
Audio-Only [79]	7.85	13.7	9.97	2.61	0.82	7.85	13.7	9.97	2.61	0.82
AV-Conv [2]	8.91	14.8	11.2	2.73	0.84	7.23	11.4	9.98	2.51	0.80
Ours (static face)	7.21	12.0	10.6	2.52	0.80	7.21	12.0	10.6	2.52	0.80
Ours (lip motion)	9.95	16.9	11.1	2.80	0.86	7.57	12.7	10.0	2.54	0.81
Ours	**10.2**	**17.2**	**11.3**	**2.83**	**0.87**	**8.53**	**14.3**	**10.4**	**2.64**	**0.84**

Table 1: Audio-visual speech separation results on the VoxCeleb2 dataset. We show the performance separately for testing examples where the lip motion is reliable (left) or unreliable (right). See text for details. Higher is better for all metrics.

	Reliable lip motion					Unreliable lip motion				
	SDR	SIR	SAR	PESQ	STOI	SDR	SIR	SAR	PESQ	STOI
Audio-Only [79]	3.56	10.9	5.71	2.00	0.66	3.56	10.9	5.71	2.00	0.66
AV-Conv [2]	5.32	11.9	7.52	2.20	0.71	3.99	9.43	6.92	2.02	0.67
Ours (static face)	3.48	8.43	6.91	1.96	0.68	3.48	8.43	6.91	1.96	0.68
Ours (lip motion)	6.31	13.3	7.72	2.32	0.76	4.21	9.78	6.85	2.03	0.69
Ours	**6.55**	**13.7**	**7.84**	**2.34**	**0.77**	**4.95**	**11.0**	**7.02**	**2.12**	**0.72**

Table 2: Audio-visual speech enhancement results on the VoxCeleb2 dataset with audios from AudioSet used as non-speech background noise. Higher is better for all metrics.

	Gabbay *et al.* [21]	Hou *et al.* [35]	Ephrat *et al.* [19]	Ours
PESQ	2.25	2.42	2.50	**2.51**
STOI	–	0.66	0.71	**0.75**
SDR	–	2.80	6.10	**6.69**

(a) Results on Mandarin dataset.

	Gabbay *et al.* [21]	Ephrat *et al.* [19]	Ours
SDR	0.40	4.10	**10.9**
PESQ	2.03	2.42	**2.91**

(b) Results on TCD-TIMIT dataset.

	Casanovas *et al.* [12]	Pu *et al.* [60]	Ephrat *et al.* [19]	Ours
SDR	7.0	6.2	12.6	**13.3**

(c) Results on CUAVE dataset.

	Afouras *et al.* [2]	Afouras *et al.* [4]	Ours
SDR	11.3	10.8	**11.8**
PESQ	3.0	3.0	**3.0**

(d) Results on LRS2 dataset.

	Chung *et al.* [15]	Ours (static face)	Ours
SDR	2.53	7.21	**10.2**

(e) Results on VoxCeleb2 dataset.

Qualitative Results