Cross-Modal Mutual Learning for Audio-Visual Speech Recognition and Manipulation

Chih-Chun Yang Wan-Cyuan Fan Cheng-Fu Yang Yu-Chiang Frank Wang
National Taiwan University & ASUS Intelligent Cloud Services, Taiwan

Introduction

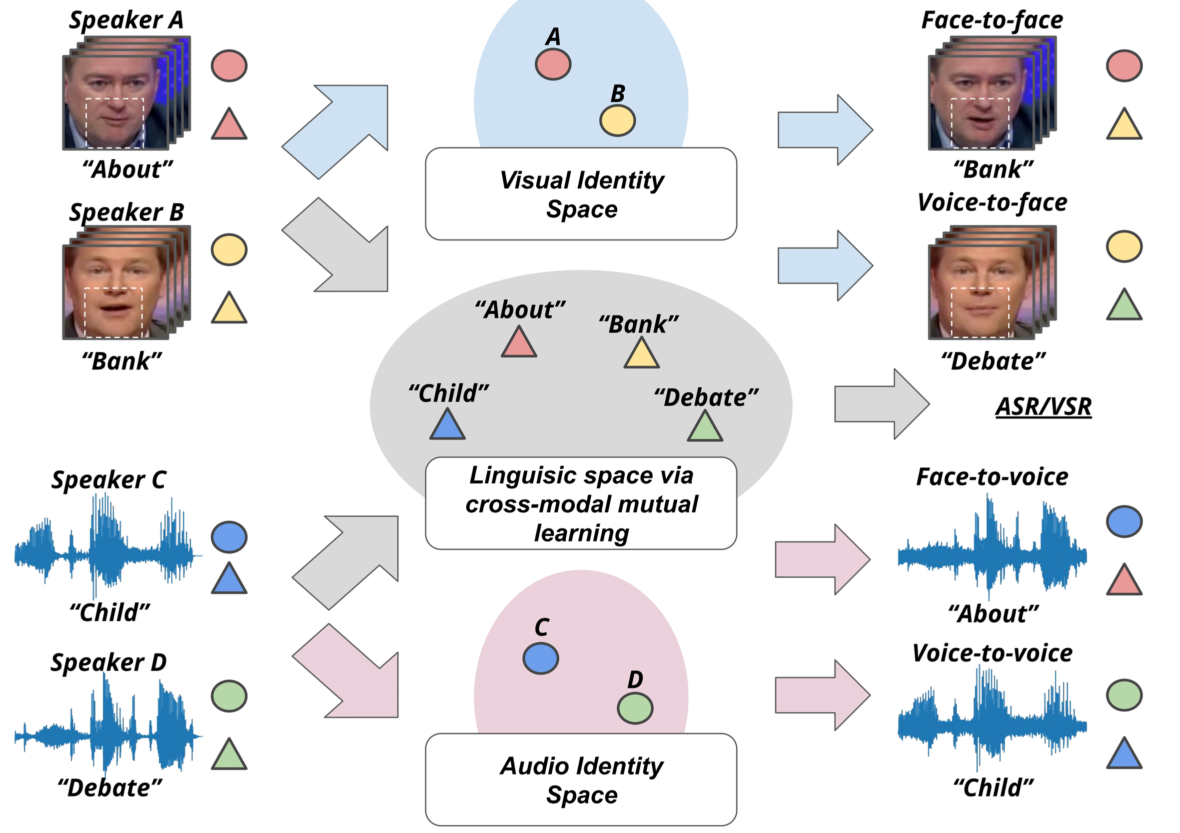

Figure 1: Illustration of joint audio-visual speech recogntion and manipulation.

- Audio-visual speech synthesis is an extension of audio-visual speech recognition, aiming at generating realistic talking face video or audio outputs based on the desirable identity and linguistic information.
- We present **a unified framework** for jointly addressing the above six different intra/cross-modality **synthesis** and **recognition** tasks.

Contribution

- We present **a unified framework** for joint audio-visual speech recognition and synthesis.
- To transfer linguistic knowledge across modalities, we advance **cross-modal mutual learning** which **aligns** cross-modality data, producing **modality-agnostic linguistic representation** for AVSR.
- Our framework allows manipulation of visual and/or audio speech data, conditioned on the desirable **linguistic** or **subject identity** information of the inputs from the **same** or **distinct modalities**.

Approach

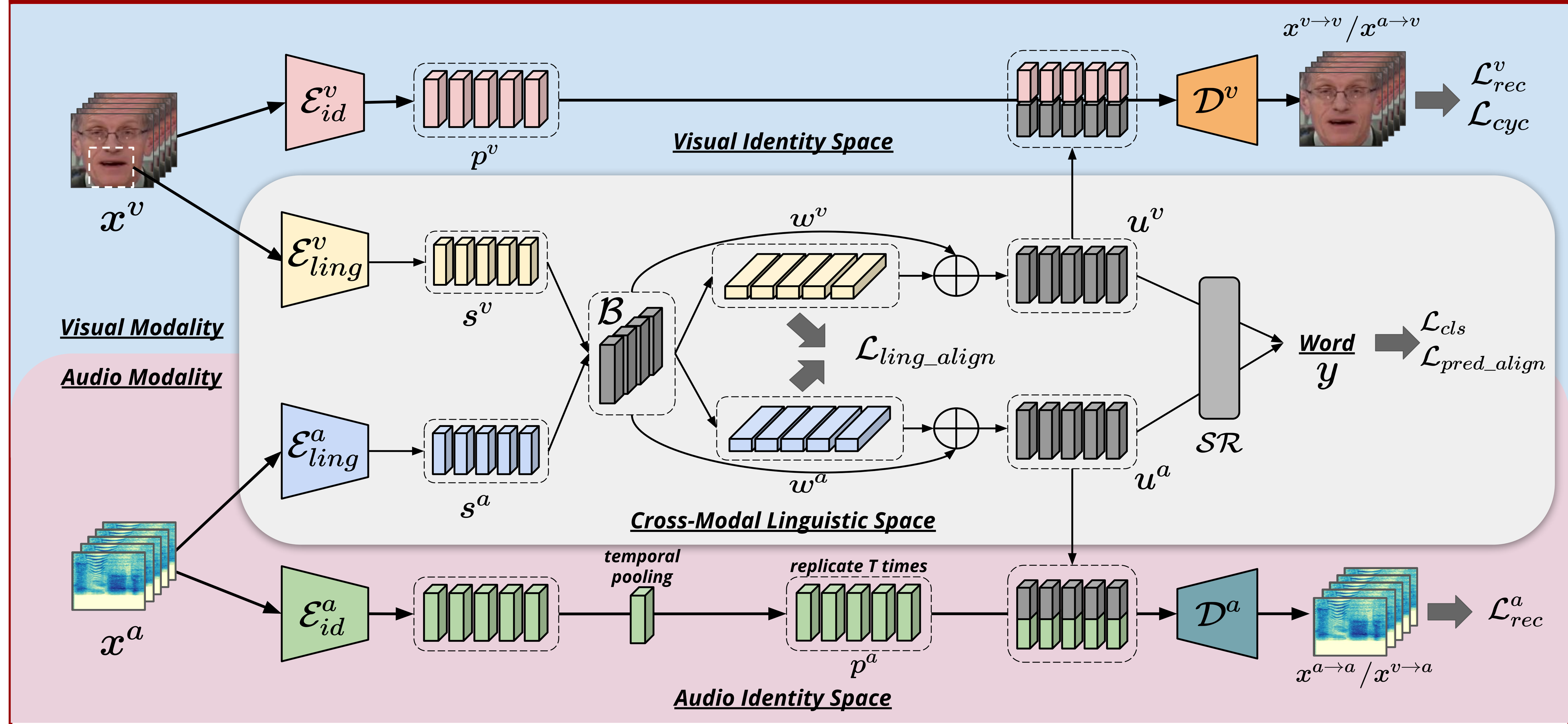

Figure 2: Our proposed framework for audio-visual speech recognition and manipulation

Cross-Modal Mutual Learning

$$\mathcal{L}_{ling\_align} = \mathcal{KL}(w^a||w^v) + \mathcal{KL}(w^v||w^a)$$

$$\mathcal{L}_{pred\_align} = \mathcal{KL}(d^a||d^v) + \mathcal{KL}(d^v||d^a)$$

$$\mathcal{L}_{mml} = \mathcal{L}_{cls} + \lambda_l \mathcal{L}_{ling\_align} + \lambda_p \mathcal{L}_{pred\_align}$$

- To **suppress** the modality information, we have s^v and s^a described as a linear combination of each modality-invariant codeword/basis in $\mathcal{B}$.
- To **relate** visual and audio data, we **mutually align** the basis weight and the word prediction distribution across visual and audio modalities.

Analysis on the Modality-Invariant Codebook

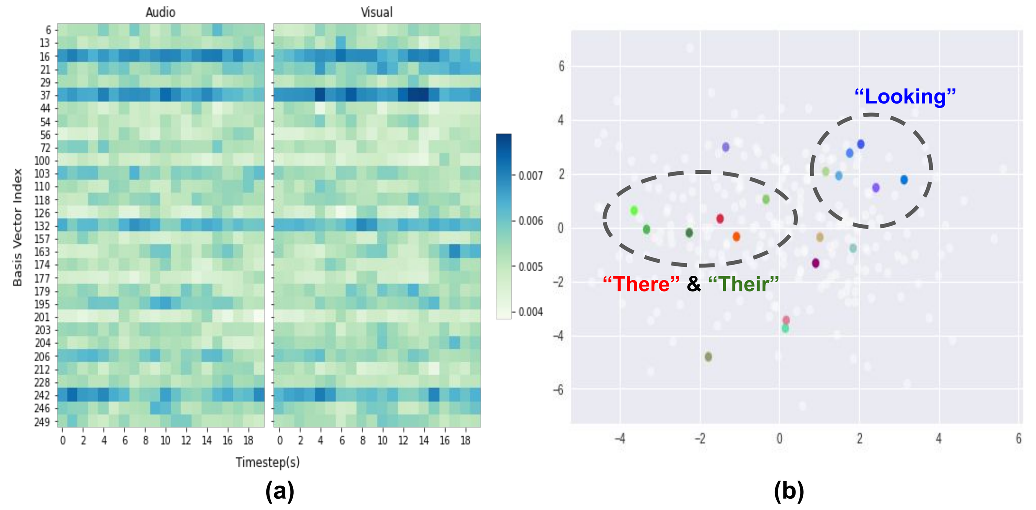

Figure 3: (a) Distributions of top-30 basis weights from audio and visual modalities of word "about". (b) 2D visualization via PCA of basis vectors for words with similar/dissimilar pronunciations.

- In Figure 3 (a), we see that the derived basis weights of audio/visual modality **share similar distributions**.
- In Figure 3 (b), we see that words with **similar pronunciation** tend to select **similar basis vectors**.

Experiments & Results

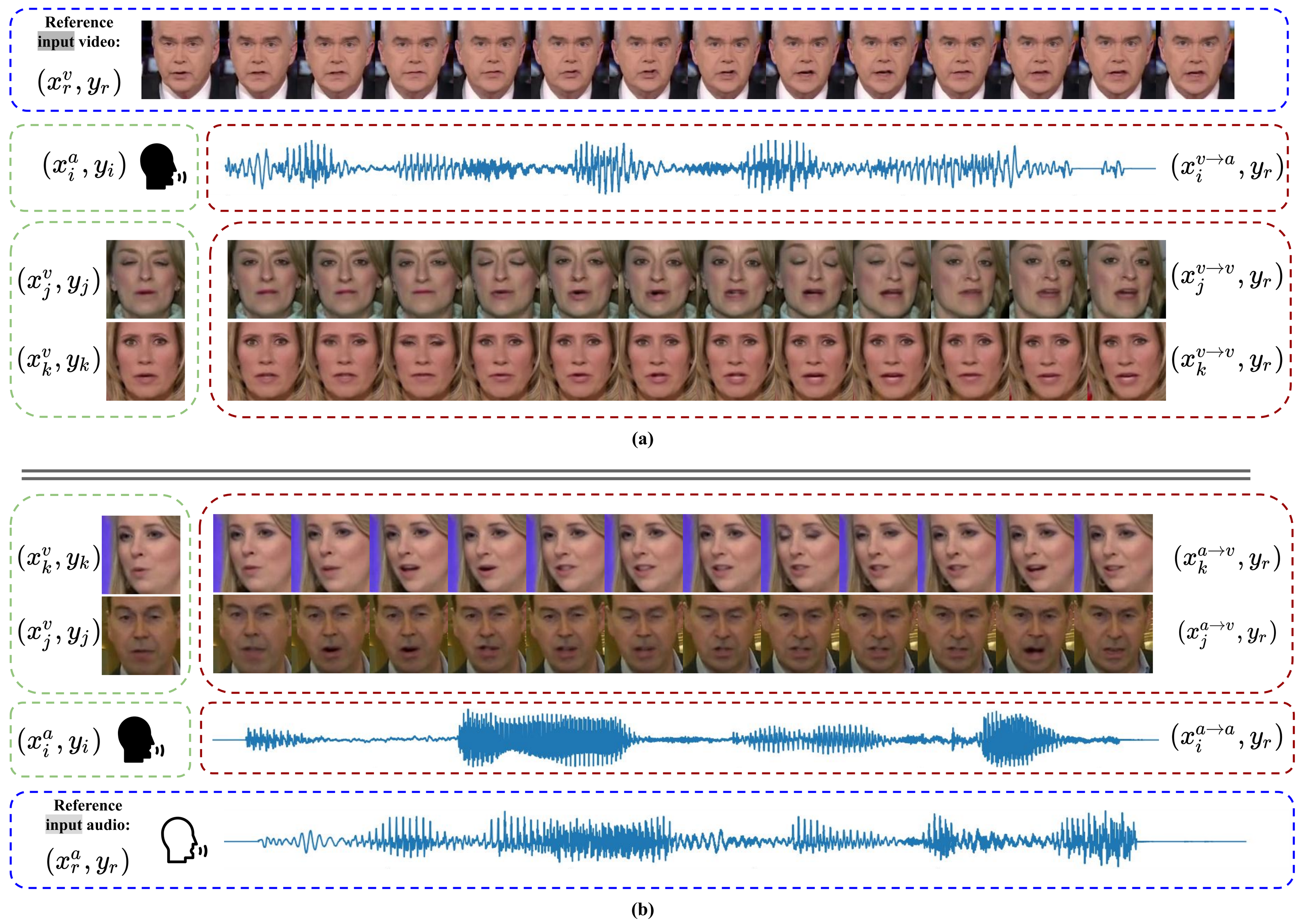

Figure 4: Example of **intra/cross-modality synthesis**: (a) face-to-voice & face-to-face synthesis, and (b) voice-to-face and voice-to-voice synthesis.

Method	Task	PSNR	SSIM	LSA.
DAVS	Intra	26.8	0.88	12.2
Ours		**33.4**	**0.96**	**22.1**
DAVS	Cross	26.7	0.88	10.7
ATVGNet		30.9	0.81	12.3
LipGAN		**33.4**	**0.96**	11.3
Wav2Lip		31.2	0.93	23.2
Ours		32.46	0.95	**27.7**

Table 1: Quantitative evaluation of **talking face generation**.

- Talking face video of satisfactory quality and **much higher lip sync accuracy (LSA.)**

Methods	Rec. Backbone	LRW		LRW-1000
		Visual	Audio	Visual
DAVS	None	67.5	91.8	-
Bi-LSTM	LSTM	84.3	-	-
MSTCN	ResNet	85.3	98.5	41.4
DSTCN	SEDenseNet	88.4	-	43.7
Bi-GRU	GRU	85.0	-	48.0
(Ren et al. 2021)	Transformer	85.7	-	-
Ours w/o syn.	ResNet	**88.4**	**98.5**	**50.5**
Ours	ResNet	**88.5**	98.4	**50.3**

Table 3: Quantitative evaluation of **speech recognition**.

- **Accurately recognize** audio/visual speech content

Method	Task	STOI	ESTOI	PESQ
VQ-VAE	Intra	0.852	0.720	1.943
Ours		**0.866**	**0.746**	**2.248**
Lip2Wav	Cross	0.543	0.344	1.197
Ours		**0.571**	**0.363**	**1.540**

Table 2: Quantitative evaluation of **voice generation**.

- Human voice of **better quality**

Experiment Setting	LRW		LRW-1000	
	Visual	Audio	Visual	Audio
Baseline	87.85	98.45	49.24	84.34
Ours (+ $\mathcal{B}$)	88.14	98.46	49.51	84.84
Ours (+ $\mathcal{B}$ + L_{align})	88.45	98.48	50.46	84.97
Ours (+ $\mathcal{B}$ + L_{align} + L_{rec})	88.47	98.40	50.32	84.84

Table 4: **Ablation studies** of our model design on speech recognition.

- Each module in cross-modal mutual learning **benefits ASR and VSR**