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

**Chih-Chun Yang[1], Wan-Cyuan Fan[2*], Cheng-Fu Yang [2*], and Yu-Chiang Frank Wang [2,3]**

[1]Department of Computer Science and Information Engineering, National Taiwan University, Taiwan, R.O.C.
[2]Graduate Institute of Communication Engineering, National Taiwan University, Taiwan, R.O.C.
[3]ASUS Intelligent Cloud Services, Taiwan, R.O.C.
{r08922050, r09942092, cfyang58, ycwang}@ntu.edu.tw

## Abstract

As a key characteristic in audio-visual speech recognition (AVSR), relating linguistic information observed across visual and audio data has been a challenge, benefiting not only audio/visual speech recognition (ASR/VSR) but also for manipulating data within/across modalities. In this paper, we present a feature disentanglement-based framework for jointly addressing the above tasks. By advancing cross-modal mutual learning strategies, our model is able to convert visual or audio-based linguistic features into modality-agnostic representations. Such derived linguistic representations not only allow one to perform ASR, VSR, and AVSR, but also to manipulate audio and visual data output based on the desirable subject identity and linguistic content information. We perform extensive experiments on different recognition and synthesis tasks to show that our model performs favorably against state-of-the-art approaches on each individual task, while ours is a unified solution that is able to jointly tackle the aforementioned audio-visual learning tasks.

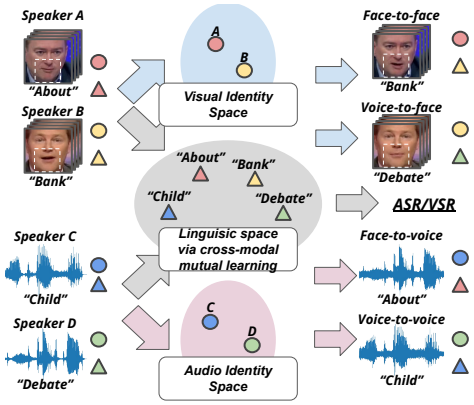


Figure 1: Illustration of joint audio-visual speech recogntion and manipulation. With decoupled linguistic and identity spaces, we aim to perform six different intra/cross-modality tasks, as noted in Table 1.

## Introduction

Audio-visual speech recognition (AVSR) is the task to perform speech recognition, with the aid of the observed visual information (e.g., lip motion). On the other hand, audio-visual speech synthesis can be viewed as an extension of AVSR, aiming at generating realistic talking face video or audio outputs. Such manipulated data outputs are conditioned on either audio or visual information observed from particular inputs (e.g., subjects), and the learning tasks such as face-to-face, face-to-voice, voice-to-voice, and voice-to-face conversion (Chen et al. 2018, 2019; van den Oord, Vinyals, and kavukcuoglu 2017; KR et al. 2019; Prajwal et al. 2020a,b; Song et al. 2019; Zhou et al. 2019) can be viewed as the applications of audio-visual speech synthesis.

It can be seen that, for both audio-visual speech recognition and synthesis, one needs to extract representative features from cross-modality (i.e., audio vs. visual) input data. While extracting linguistic representation would be necessary to realize the task of AVSR, modality-preserving information such as subject identity needs to be preserved for data recovery/synthesis purposes. With the above two types of representations derived, one can perform the aforementioned intra- or cross/inter-modality synthesis tasks such as multi-speaker speaking synchronization (Zhou et al. 2019), visual audio lip synchronization(Prajwal et al. 2020b), automatic voice acting (Prajwal et al. 2020a), voice conversion (Ding and Gutierrez-Osuna 2019), and audio-visual speech separation (Gao and Grauman 2021). However, most existing works typically focus on addressing only one or few selected tasks. For such cross-modality learning tasks, it would be desirable to advance multi-task learning strategies to utilize inputs across modalities for solving the above diverse yet related learning tasks.

To extract linguistic features from given input data, techniques of adversarial training, vector quantization (VQ), or instance normalization (IN) (Chou, Yeh, and Lee 2019; Ding and Gutierrez-Osuna 2019; van den Oord, Vinyals, and kavukcuoglu 2017; Zhou et al. 2019) have been proposed. However, previous studies (Ding and Gutierrez-Osuna 2019; Zhang, Song, and Qi 2018) suggest that such techniques might suffer from training instability or the degraded synthesis data quality due to the design of the information bottleneck. Furthermore, performing audio-visual speech synthesis requires learning from cross-modality data; how to per-

*indicates equal contribution

| Methods | Face to Face | Face to Voice | Face to Text | Voice to Voice | Voice to Face | Voice to Text |
|---|---|---|---|---|---|---|
| MSTCN | - | - | ✓ | - | - | ✓ |
| DSTCN | - | - | ✓ | - | - | ✓ |
| (Ren et al. 2021) | - | - | ✓ | - | - | ✓ |
| VAE+AdaIN | - | - | - | ✓ | - | - |
| Grouped VQ-VAE | - | - | - | ✓ | - | - |
| Lip2Wav | - | ✓ | - | - | - | - |
| (Chen et al. 2018) | - | - | - | - | ✓ | - |
| ATVGNet | - | - | - | - | ✓ | - |
| LipGAN | - | - | - | - | ✓ | - |
| Wav2Lip | - | - | - | - | ✓ | - |
| DAVS | ✓ | - | ✓ | - | ✓ | ✓ |
| Ours | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

Table 1: Comparisons with recent audio-visual learning recognition and synthesis models. Note that Face to Text and Voice to Text denote the tasks of visual and audio-based speech recognition, respectively.

form feature disentanglement across data modalities remains a challenging task.

For example, one of the challenges in AVSR or synthesis tasks would be the need to handle homophenes, which describe the fact that multiple sounds (phonemes) are auditorily distinguished from each other, but with correspondence to some identical lip shapes (viseme); this is due to diverse styles of speaking intonation, emotion, and stress. In other words, modeling the lip-and-voice correspondence is considered among the obstacles, which requires one to handle the ambiguity between audio and visual clues. Most existing works learn a one-way mapping between visual and audio data. For example, (Prajwal et al. 2020a) applies sequence-to-sequence learning to map talking face video to voice data for each speaker. (Zhou et al. 2019) applies the contrastive loss to align visual and audio speech representation. Extensions by (Chen et al. 2018; KR et al. 2019; Prajwal et al. 2020b) are realized by applying an extra discriminator to improve the synchronization of generated talking face and input voice. Although the promising voice-to-face result has been achieved, these methods only deal with uni-direction cross-modality synthesis, which cannot handle synthesis or manipulation across visual-audio modality or across multiple speaker identities.

In this paper, we propose a unified learning framework, which can be applied to jointly address the tasks of audio-visual speech recognition and manipulation (i.e., intra- and cross-modality synthesis), as depicted in Figure 1. We advance feature disentanglement learning strategies, followed by a linguistic module that extracts and transfers knowledge across modalities via *cross-modal mutual learning*. This allows us to extract linguistic and identity information from cross-modality input data, while the linguistic representation would be *modality agnostic* realizing the task of AVSR. Since we do not require adversarial learning techniques during training, our model does not suffer from learning instability problems. With the ability to perform both recognition and manipulation tasks within and across data modalities, we summarize and compare with recent audio-visual learning models in Table 1.

The contributions of this paper are highlighted below:

- We present a unified framework for joint audio-visual speech recognition and synthesis. The former takes inputs from either modality for speech recognition, while the latter allows one to manipulate intra-/cross-modality outputs with desirable information.
- To transfer linguistic knowledge between visual and audio modalities, we advance cross-modal mutual learning and learn a codebook 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 observed from the same or distinct modalities.

## Related Works

**Synthesis of Talking Faces**: A number of approaches have been proposed to perform talking face video generation, conditioned on particular facial or audio inputs. For example, (Zhou et al. 2019) adopts disentangle strategy to encode lip movements and the appearance of the speaker with word and identity labels as the guidance. To ensure each feature without impurity, a strong information bottleneck, which obstructs certain information from the other feature space, is employed by adversarial training. Nevertheless, for lip movement extraction, other unrelated features such as head movement and facial expression are usually drawn with only word labels as adversarial training targets. This leads to inaccurate speech knowledge transfer since the speech feature does not actually focus on lips dynamics only. On the other hand, (KR et al. 2019; Prajwal et al. 2020b) directly combine audio and visual identity features to generate videos, with an additional discriminator deployed to ensure the synchronization between audio and video data. As for (Chen et al. 2019), it chooses to generate facial landmarks for synthesizing talking face videos, based on the input audio as the prior knowledge and guidance.

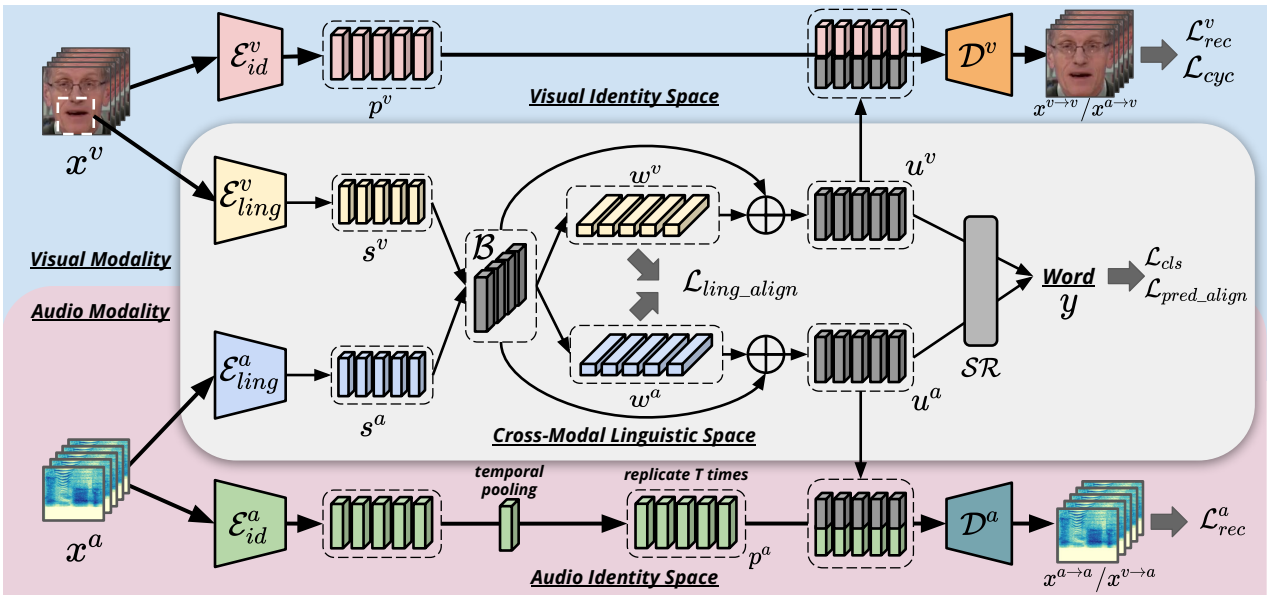


Figure 2: Our proposed framework for audio-visual speech recognition and manipulation. Note that A/VSR is performed on the cross-modal linguistic space formulated by linguistic encoders and the modality-sharing linguistic codebook and speech recognizer. Together with modality-specific modules in identity space, desirable intra/cross-modality manipulation can be achieved.

**Voice Synthesis**: Recent approaches have demonstrated promising results on voice-to-voice style transfer and text-to-voice generation. For the task of voice-to-voice style transfer, existing works (Chou, Yeh, and Lee 2019; Ding and Gutierrez-Osuna 2019; van den Oord, Vinyals, and kavukcuoglu 2017) utilize techniques of vector quantization (VQ) and instance normalization (IN) as the speaker style information bottleneck, which decouples linguistic information from the audio signal while the speaker style information is additionally embedded. However, such designs do not consider the speaking style which might undermine the quality of the signal thus requiring additional modules for refinement. For the text-to-voice generation task, sequence-to-sequence learning with attention mechanisms is applied to synthesize mel-spectrograms in an auto-regressive manner (Li et al. 2019; Shen et al. 2018). Despite impressive voice-and-text results that have been shown, tasks of talking face to voice are still under investigation due to the barrier posed by homophenes. Recently, (Prajwal et al. 2020a) builds a model for each individual speaker with similar architecture as the text-to-voice model but conditions on the face sequence instead of text. Nevertheless, this framework design requires a large amount of training data for each individual speaker, and thus the models cannot be easily extended to perform voice synthesis for multiple speakers.

**Audio/Visual Speech Recognition**: Previous studies have shown remarkable performance on audio speech recognition (ASR) while visual speech recognition (VSR) is a more challenging task due to the variety and ambiguity of lip movements across speakers. For word-level speech recognition, (Feng et al. 2020; Stafylakis, Khan, and Tzimiropoulos 2018) adopt recurrent module for input sequence. However, one typically needs to tackle the convergence issue of the associated recurrent networks. Recent studies like (Ma et al. 2021; Martinez et al. 2020) adopt temporal convolution networks as an alternative to improve learning efficiency. As for sentence-level speech recognition, previous works (Afouras et al. 2018; Chan et al. 2016; Garcia et al. 2019; Zhang, Cheng, and Wang 2019) aim at recognizing the character or word sequence in a full sentence by sequence-to-sequence learning. Regarding the difficulty of learning discriminative linguistic features from long videos, a pretraining feature extractor with word-level methods is adopted. Additionally, due to improved performance of ASR over VSR, previous works like (Ren et al. 2021; Zhao et al. 2020) distill knowledge from models learned from audio or audio-visual data to guide the VSR ones. However, existing methods typically do not consider leveraging information from VSR to ASR models.

## Approach

### Notations and Problem Formulation

We first define the notations to be used in this paper. Given a visual-speech dataset of $N$ videos $\mathcal{D} = \{(x_i^v, x_i^a, y_i)\}_{i=1}^N$, we have $x_i^v$, $x_i^a$ denote the talking face video and voice data pair, and $y_i$ as the corresponding word label. Our goal is to learn linguistic representation from cross-modality data (i.e., $x_i^v$ and $x_i^a$), while modality-preserving information can be extracted in visual and audio domains. The former can be applied for visual-speech recognition by observing either visual or audio inputs, while the latter allows one to recover or manipulate desirable visual and/or audio data outputs.

Figure 2 depicts our proposed framework. As shown in this figure, we have encoders $\mathcal{E}_f^m$ deployed for each modality

and feature type (i.e., $m \in \{\text{video, audio}\}$ and $f \in \{\text{identity, linguistic}\}$) and modality-specific decoders $\mathcal{D}^m$ to produce outputs for each modality. Note that the feature $s^m$ describes the *modality-preserving linguistic representation* from either modality, while $p^m$ indicates the *modality-preserving identity feature*. To extract linguistic knowledge from visual and audio data, we introduce a specialized cross-modal linguistic module, which consists of a modality-invariant linguistic codebook $\mathcal{B}$ and an audio-visual speech recognizer $\mathcal{SR}$. The former produces *modality-agnostic linguistic representation* $u^m$, while the latter is applied for audio/visual speech recognition. It is worth noting that $u^m$ would not contain any modality-specific information; the superscript $m$ simply indicates the origin modality where $u$ is derived.

With the above-learned model, intra-modality and cross-modality data synthesis can be performed. Take voice-to-face synthesis for example. To manipulate the linguistic knowledge conveyed in the form of lip movement in the talking face video, we extract the modality-specific identity feature $p^v$ from $\mathcal{E}^v_{ID}$ and the modality-agnostic linguistic representation $u^a$ from $\mathcal{B}$, followed by visual decoder $\mathcal{D}^v$ for producing the desirable video output. As for audiovisual speech recognition, we have $\mathcal{SR}$ take either $u^v$ or $u^a$ for predicting the associated word-level labels.

## Linguistic Knowledge Extraction via Cross-Modal Mutual Learning

As depicted in Figure 2, we have a unique cross-modal linguistic module in our framework, which aims at extracting linguistic features from either data modality for audio-visual speech recognition. Together with the identity features encoded by $\mathcal{E}^v_{ID}$ or $\mathcal{E}^a_{ID}$, intra-/cross-modality data recovery and manipulation can also be achieved. In order to relate the visual and audio inputs during the extraction of the above linguistic features, we particularly learn a linguistic codebook $\mathcal{B}$ in this module. This codebook is shared by cross-modality data while allowing a unified linguistic representation derived from either modality for recognition and synthesis purposes. We now detail the design of this module.

**Linguistic Encoder** $\mathcal{E}_{ling}$: For each data modality $m \in \{\text{video, audio}\}$, we have the encoder $\mathcal{E}^m_{ling}$ to disentangle linguistic information from the input data. For visual modality, we focus on the lip region of the talking face video $x^v$ and feed it to convolutional neural networks to extract visual-based linguistic representation $s^v$, as a visual-modality preserving linguistic feature. As for audio modality, we adopt a fully 1D convolution network to process audio signals with arbitrary lengths, resulting in the audio-based linguistic representation $s^a$ (as an audio-modality preserving linguistic feature).

**Modality-Invariant Linguistic Codebook** $\mathcal{B}$: To perform AVSR with input data from either modality, one needs to extract *modality-agnostic representation* from the aforementioned $s^v$ and $s^a$, so that only word-level information would be observed. Inspired by (Ding and Gutierrez-Osuna 2019), we propose to learn a linguistic codebook $\mathcal{B}$ shared by the visual-audio modality pair, with the goal to associate and describe the linguistic information across features $s^v$ and $s^a$. Once this codebook is obtained, one would be able to suppress the modality information presented in $s^v$ and $s^a$, with the resulting linguistic representation to be viewed as modality agnostic.

We now detail the learning of this modality-invariant codebook $\mathcal{B}$ and the derivation of modality-agnostic linguistic representation $u$. For $m \in \{\text{video, audio}\}$, we have the aforementioned linguistic feature $s^m$ described as a linear combination of each modality-invariant codeword/basis in $\mathcal{B}$. To be more precise, we learn the linguistic codebook $\mathcal{B} = \{b_j\}^n_{j=1}$, where $b_j \in R^d$ and $n$ is the total number of bases. To represent $s^m$, the weight $w^m \in R^{n\times 1}$ is calculated by the pairwise similarity between $s^m$ and each basis $b_j$ of $\mathcal{B}$. And, the above similarity is measured by the Euclidean distance with an extra softmax layer for normalization. This then produces the soft-quantized modality-agnostic linguistic representation $u^m$:

$$u^m = \mathcal{B}w^m. \tag{1}$$

It is worth noting that, in order to encourage $\mathcal{B}$ to encode linguistic knowledge shared by both modalities, we calculate and suppress the Kullback-Leibler Divergence (KLD) between the basis weights derived by visual and audio modalities. This is to enforce that cross-modality inputs would have a shared linguistic representation, allowing the subsequent speech recognition or visual/audio data synthesis. In other words, we introduce the *cross-modality linguistics alignment loss* by calculating:

$$\mathcal{L}_{ling\_align} = \mathcal{KL}(w^a||w^v) + \mathcal{KL}(w^v||w^a), \tag{2}$$

where $w^a$ and $w^v$ denote the similarity weights derived for audio and visual modalities, respectively.

**Audio-Visual Speech Recognizer** $\mathcal{SR}$: We deploy an audio visual speech recognizer $SR$ to perform either visual or audio-based speech recognition on $u^v$ or $u^a$. Using the classification loss $\mathcal{L}_{cls}$ as the objective, the learning of $\mathcal{SR}$ also guides the learning of encoders $\mathcal{E}^m_{ling}$ and the modality-invariant linguistic codebook $\mathcal{B}$.

Since modality-agnostic linguistic representations from visual and audio data are expected to realize the same task of speech recognition, we further align the associated prediction distributions across these two modalities. Inspired by the work of (Zhang et al. 2018), we choose to calculate the following *prediction alignment loss*:

$$\mathcal{L}_{pred\_align} = \mathcal{KL}(d^a||d^v) + \mathcal{KL}(d^v||d^a), \tag{3}$$

where $d^a$ and $d^v$ are the word prediction distributions of audio and visual modality. It can be seen that, this prediction alignment term also serves as a guidance for syncing and relating visual and audio data. Together with the cross-modality linguistics alignment loss, the learned codebook $\mathcal{B}$ and modality-agnostic linguistic representation $u^m$ would better align the visual and audio features, which not only benefit audio visual speech recognition but also cross-modality data synthesis. Thus, we have $\mathcal{L}_{mml}$ sum up $\mathcal{L}_{cls}$, $\mathcal{L}_{ling\_align}$, and $\mathcal{L}_{pred\_align}$ as the objective for cross-modal linguistic mutual learning.

## Intra/Inter-Modality Visual-Speech Synthesis

In addition to learning linguistic features for audio visual speech recognition, our learning model also performs intra-/inter-modality visual-speech synthesis with the deployed modality-specific identity encoders and decoders. We now discuss the design and learning of these modules.

**Modality-Specific Identity Encoder** $\mathcal{E}_{ID}$: The identity encoder in Figure 2 is to encode identity information from the input data of either modality. Take the visual data input for example. $\mathcal{E}^v_{ID}$ disregards any linguistic information (e.g., speech information carried by lip movements) while preserving the visual identity information. To learn this encoder, we consider two types of visual inputs: the talking face video with lip region masked as head pose prior, and the first frame in the video sequence as the lip appearance prior. By jointly passing the above visual inputs into $\mathcal{E}^v_{ID}$, the visual-preserving identity features of visual modality $p^v$ would be extracted. When encoding the identity information (e.g., speaker style) from audio data, we feed $x^a$ into a fully convolutional network $\mathcal{E}^a_{ID}$, followed by temporal pooling to extract the feature. This is to enforce that the resulting feature $p^a$ only contains audio identity information.

**Modality-Specific Decoder** $\mathcal{D}$: For both visual and audio modalities, we deploy a decoder for each to generate the output data with the desirable identity and linguistic information. Take visual modality for example. We concatenate $p^v$ with the linguistic representation $u^v$ processed by our linguistic module, which jointly serves as the input to the visual decoder $\mathcal{D}^v$, which is a 2D deconvolutional network. As for producing audio outputs, we have a 1D deconvolution network as the audio decoder to produce mel-spectrogram.

During the training of audio and video synthesis tasks, we apply L1 reconstruction loss; moreover, an additional linguistic cycle consistency is calculated for the visual modality, which is to ensure that the synthesized linguistic representation would be consistent with the referenced input video/audio. Taking face synthesis as an example, we calculated the following loss functions:

$$\begin{aligned}\mathcal{L}^v_{rec} &= \mathbb{E}_{x\sim D}||\mathcal{D}^v(u^v_i, p^v_i) - x^v_i||_1 \\ &+\mathbb{E}_{x\sim D}||\mathcal{D}^v(u^a_i, p^v_i) - x^v_i||_1,\end{aligned} \tag{4}$$

$$\mathcal{L}_{cyc} = ||\mathcal{E}^v_s(\mathcal{D}^v(u^m_j, p^v_i)) - s^v_j||_1. \tag{5}$$

As for audio synthesis, the objective function considered is:

$$\begin{aligned}\mathcal{L}^a_{rec} &= \mathbb{E}_{x\sim D}||\mathcal{D}^a(u^a_i, p^a_i) - x^a_i||_1 \\ &+\mathbb{E}_{x\sim D}||\mathcal{D}^a(u^v_i, p^a_i) - x^a_i||_1.\end{aligned} \tag{6}$$

We have $\mathcal{L}_{syn}$ sum up the above losses as the final objective function for visual-speech synthesis. Thus, the full learning objective for our proposed framework is defined as:

$$\mathcal{L} = \mathcal{L}_{mml} + \mathcal{L}_{syn}. \tag{7}$$

Once the learning of our framework is complete, intra- and cross-modality synthesis can be performed using input data of the desirable modality.

# Experiments

## Datasets

**LRW** (Chung and Zisserman 2016): Known for its variety of speaking styles and head poses across subjects, LRW is an English-speaking video dataset collected from BBC programs with more than 1000 speakers. The vocabulary size is 500 and each video is 1.16 sec long (29 frames) with target word and context before and after involved.
**LRW-1000** (Yang et al. 2019): LRW-1000 is a Mandarin-speaking video dataset collected from more than 2,000 subjects with 1,000 vocabulary size. The videos provided are of various lengths and only focus on the lip region.

## Implementation Details

We implement our model using Pytorch. For visual modality, we adopt a simplified ResNet-18 as the identity encoder and a modified ResNet-18 as the speech encoder with the first 2D convolution layer replaced by a 3D convolution layer to better capture temporal dynamics. The decoder is a fully deconvolutional network to reconstruct the whole input video, with skip-connection from the identity encoder to the decoder deployed to improve the video quality. For the audio modality, the identity encoder is a fully 1D convolution network, and the speech encoder is similar to the identity encoder but without temporal pooling. The amount of basis vectors in the modality-invariant module is 256. For the speech recognizer, we adopt the same architecture as (Martinez et al. 2020), a variant of temporal convolution network which adopts multiple kernel sizes to increase the receptive field so as to capture different level temporal dependencies, to learn the linguistic knowledge efficiently. AdamW is used as the optimizer for training with weight decay $5 \times 10^{-4}$ as regularization. For the linguistic and synthesis modules, initial learning rates of $3 \times 10^{-4}$ and $1 \times 10^{-4}$ with a schedule of reduction are applied, respectively. We follow the pre-processing procedures of (Prajwal et al. 2020a; Chen et al. 2019; Feng et al. 2020) for the input audio and visual data, while we change the window and hop length to 20 ms and 5 ms for mel-spectrogram extraction for video synchronization. Griffin-Lim algorithm (Griffin and Lim 1984) is adopted to convert mel-specrogram back to a waveform.

## Evaluation Metrics

We take the following metrics to evaluate the intra-/cross-modality synthesis tasks. As for audio/visual speech recognition tasks, top-1 accuracy is considered.
**Peak Signal-to-Noise Ratio:** PSNR is an image quality measurement, computes the ratio between the maximum possible power of a signal and the power of noise.
**Structural Similarity:** SSIM reflects the perceived quality of the reconstructed image by measuring the similarity between the original and generated image.
**(Extended) Short-Time Objective Intelligibility :** STOI is correlated with the intelligibility of the audio signal via a simple time-frequency-decomposition while ESTOI functions the same as STOI but does not assume mutual independence between frequency bands during calculation.

| 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 2: Quantitative evaluation of talking face video generation. Note that Intra and Cross indicate face-to-face and voice-to-face generation, respectively.

| 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 3: Quantitative evaluation of voice generation. Note that Intra and Cross denote voice-to-voice and face-to-voice generation, respectively.

**Perceptual Evaluation of Speech Quality:** PSEQ measures the perceptual quality of audio via analyzing specific audio parameters such as variable delays and transcoding.
**Lip Sync Accuracy:** To evaluate if linguistic knowledge is well-preserved after voice/face-to-face synthesis, we randomly sample audio/visual-visual pairwise data and perform synthesis followed by VSR with a pretrained classifier released by (Martinez et al. 2020). The recognition result is expected to be the same as the one of the reference audio/visual input. We denote this metric as LSA in Table 2.

## Quantitative Evaluation

With six different tasks (i.e., face-to-face, face-to-audio, audio-to-audio, audio-to-face, and audio/visual speech recognition) considered, we only compare our proposed framework with SOTA methods for each due to page limitation. We consider the LRW dataset for audio-visual speech recognition and intra-/cross-modality synthesis, while the LRW-1000 dataset is only for AVSR since it only contains videos frames of lip regions.

**Intra-Modality Synthesis**: Face-to-face conversion can be achieved via passing the specified visual identity and modality-agnostic linguistic features into the visual decoder. To assess the quality of the output visual data, two standard quality metrics are considered: PSNR and SSIM. Table 2 lists and compares the performances of our method and DAVS, which disentangles linguistic knowledge from the identity feature via adversarial learning to achieves the face-to-face conversion and is the most related work which focused on the transfer of linguistic knowledge as ours. From this table, we observe that our model consistently produced improved visual quality over DAVS, which tends to produce outputs with diverse quality due to its adversarial learning design. Furthermore, we see that our video with derived linguistic features achieved satisfactory performances in lip accuracy, outperforming those derived by DAVS.

Voice-to-voice conversion can be achieved following the same procedure as face-to-face conversion but using audio data. To evaluate the synthesis quality, three metrics are considered: STOI, ESTOI, and PESQ. We compare the linguistic representation extracted by VQ-VAE on the generated audio quality. From the results shown in Table 3, we see that our model is able to produce audio data without deteriorating the auditory information. This is mainly due to our derivation of soft-quantized modality agnostic representation in our linguistic module; this also explains why our model performed favorably against the standard vector quantization based method VQ-VAE.

**Cross-Modality Synthesis**: To generate talking face video condition on particular voice data, we pass the visual identity feature and modality-agnostic linguistic feature from the audio modality to the visual decoder. We evaluate the synthesis video on the LRW dataset following the same metrics for video quality considered in previous works, and we present the results in the lower part of Table 2. In addition, we repeat the evaluation protocol in face-to-face synthesis on the extracted linguistic features; this is to verify whether the linguistic content of the synthesized video is consistent with that of the referenced audio input. From this table, we see that our model produced satisfactory voice-to-face video outputs. It is worth noting that although recent approaches were also able to produce realistic output, they cannot achieve comparable lip accuracy as ours did. This verifies our learning and transfer of modality-agnostic linguistic knowledge across modalities.

Similar to the voice to face cross-modality synthesis, face-to-voice generation can be achieved via combining the audio identity feature and modality-agnostic linguistic representation from visual modality. We evaluate on LRW with the same metrics introduced in the previous work, Lip2Wav, and report the results in Table 3. As demonstrated, the quality of voice generated with our framework was above those reported by recent works. We refer such improvements to the learning of our modality-invariant linguistic codebook, which allows the output data to be based on the related and representative linguistic features. Besides, we also found that the linguistics alignment loss would be a key factor, as we later verify in our ablation studies.

**Audio/Visual Speech Recognition**: To evaluate the proposed framework for visual and audio speech recognition, we adopt top-1 accuracy as the metric and report the results in Table 4. In this table, we compare our method with previous works using data from either single modality or multi-modalities. Note that DAVS is the only multi-task learning model (as ours is) while the rest focused on individual task only. Besides, (Ren et al. 2021) is also a distillation-based method which learns from a master module taking both audio and visual data as input. From the results listed

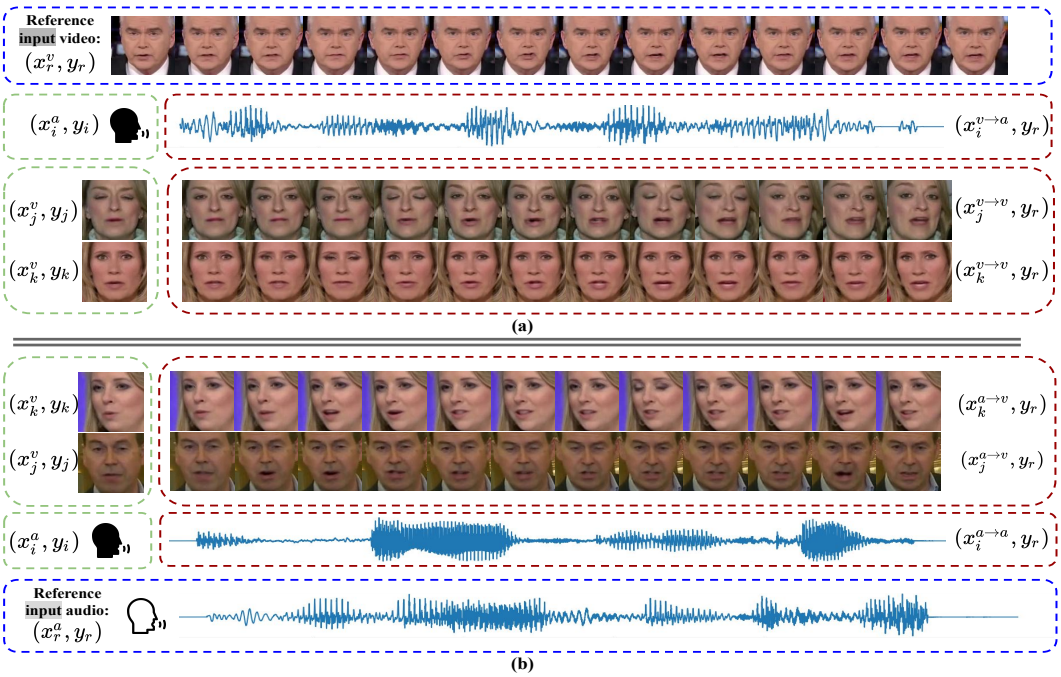


Figure 3: Example of intra/cross-modality synthesis: (a) face-to-voice & face-to-face synthesis, and (b) voice-to-face and voice-to-voice synthesis. Note that i, j, and k denote the indices of subjects of interest. Taking face-to-voice generation in (a) for example, $(x_r^v, y_r)$ denotes the reference video of subject $r$ speaking word $y_r$ and $(x_i^a, y_i)$ as audio with subject $i$ of interest, and we have $(x_i^{v\to a}, y_r)$ denote the synthesized audio of the same subject $i$ while speaking word $y_r$.

| 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 4: ASR/VSR comparisons in terms of Top-1 Accuracy. Note that Rec. Backbone denotes the architecture adopted in the speech recognizer.

in this table, it can be seen the model-agnostic linguistic representation learned by proposed cross-modal mutual learning achieves promising recognition performances. Additionally, we see that the enforcement of the loss for synthesis can probably degrade the audio recognition performance slightly. This is expected (as a tradeoff), since adding such a data recovery loss allows our model to perform intra-/inter-modality synthesis tasks. We note that ablation studies can be found in the supplementary material to show the effectiveness of our proposed framework. It is worth repeating that, from the above experiments, we confirm that our model not only produces high-quality talking face video and voice output, but also achieves high ASR/VSR accuracy and performs favorably against most previous works.

## Qualitative Evaluation

We now consider the intra/cross-modality synthesis tasks, including face-to-face, face-to-voice, voice-to-voice, and voice-to-face conversion. In Fig. 3, we demonstrate the qualitative result of our approach, especially for face-to-face and voice-to-face synthesis in a frame-by-frame manner. From the face sequence in Fig. 3, we can observe that the duration and extent of the mouth opening are consistent with the reference audio/visual input. Please refer to our supplementary material for a more comprehensive demonstration.

## Conclusion

In this paper, we proposed a unified framework for audio-visual speech recognition and synthesis. We advance cross-modal mutual learning for aligning linguistic information across visual and audio data, resulting in modality-agnostic representation for ASR/VSR. By preserving modality-specific identity features from either modality, our model can be applied to manipulate intra-/cross-modality data outputs with desirable audio or visual information. Extensive experiments were conducted on the challenging LRW and LRW-1000 benchmark datasets, which qualitatively and quantitatively demonstrated the effectiveness of our model over state-of-the-art audio-visual learning approaches in recognition and synthesis tasks.

## Acknowledgements

This work is supported in part by the Ministry of Science and Technology of Taiwan under grants MOST 110-2221-E-002-121 and 110-2634-F-002-052.

## References

Afouras, T.; Chung, J. S.; Senior, A.; Vinyals, O.; and Zisserman, A. 2018. Deep Audio-visual Speech Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP: 1–1.

Chan, W.; Jaitly, N.; Le, Q.; and Vinyals, O. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4960–4964.

Chen, L.; Li, Z.; Maddox, R. K.; Duan, Z.; and Xu, C. 2018. Lip Movements Generation at a Glance. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Chen, L.; Maddox, R. K.; Duan, Z.; and Xu, C. 2019. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7832–7841.

Chou, J.-c.; Yeh, C.-c.; and Lee, H.-y. 2019. One-shot Voice Conversion by Separating Speaker and Content Representations with Instance Normalization. *arXiv preprint arXiv:1904.05742*.

Chung, J. S.; and Zisserman, A. 2016. Lip Reading in the Wild. In *Asian Conference on Computer Vision*.

Ding, S.; and Gutierrez-Osuna, R. 2019. Group Latent Embedding for Vector Quantized Variational Autoencoder in Non-Parallel Voice Conversion. In *Proc. Interspeech 2019*, 724–728.

Feng, D.; Yang, S.; Shan, S.; and Chen, X. 2020. Learn an Effective Lip Reading Model without Pains. *arXiv preprint arXiv:2011.07557*.

Gao, R.; and Grauman, K. 2021. VisualVoice: Audio-Visual Speech Separation with Cross-Modal Consistency. *arXiv preprint arXiv:2101.03149*.

Garcia, B.; Shillingford, B.; Liao, H.; Siohan, O.; de Pinho Forin Braga, O.; Makino, T.; and Assael, Y. 2019. Recurrent Neural Network Transducer for Audio-Visual Speech Recognition. In *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop*.

Griffin, D.; and Lim, J. 1984. Signal estimation from modified short-time Fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(2): 236–243.

KR, P.; Mukhopadhyay, R.; Philip, J.; Jha, A.; Namboodiri, V.; and Jawahar, C. 2019. Towards automatic face-to-face translation. In *Proceedings of the 27th ACM International Conference on Multimedia*, 1428–1436.

Li, N.; Liu, S.; Liu, Y.; Zhao, S.; and Liu, M. 2019. Neural Speech Synthesis with Transformer Network. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01): 6706–6713.

Ma, P.; Wang, Y.; Shen, J.; Petridis, S.; and Pantic, M. 2021. Lip-Reading With Densely Connected Temporal Convolutional Networks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2857–2866.

Martinez, B.; Ma, P.; Petridis, S.; and Pantic, M. 2020. Lipreading using Temporal Convolutional Networks. In *ICASSP*.

Prajwal, K. R.; Mukhopadhyay, R.; Namboodiri, V. P.; and Jawahar, C. 2020a. Learning Individual Speaking Styles for Accurate Lip to Speech Synthesis. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Prajwal, K. R.; Mukhopadhyay, R.; Namboodiri, V. P.; and Jawahar, C. 2020b. A Lip Sync Expert Is All You Need for Speech to Lip Generation In the Wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, 484–492. New York, NY, USA: Association for Computing Machinery. ISBN 9781450379885.

Ren, S.; Du, Y.; Lv, J.; Han, G.; and He, S. 2021. Learning From the Master: Distilling Cross-Modal Advanced Knowledge for Lip Reading. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13325–13333.

Shen, J.; Pang, R.; Weiss, R. J.; Schuster, M.; Jaitly, N.; Yang, Z.; Chen, Z.; Zhang, Y.; Wang, Y.; Skerrv-Ryan, R.; Saurous, R. A.; Agiomvrgiannakis, Y.; and Wu, Y. 2018. Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4779–4783.

Song, Y.; Zhu, J.; Li, D.; Wang, A.; and Qi, H. 2019. Talking Face Generation by Conditional Recurrent Adversarial Network. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, 919–925. International Joint Conferences on Artificial Intelligence Organization.

Stafylakis, T.; Khan, M. H.; and Tzimiropoulos, G. 2018. Pushing the boundaries of audiovisual word recognition using Residual Networks and LSTMs. *Computer Vision and Image Understanding*, 176-177: 22–32.

van den Oord, A.; Vinyals, O.; and kavukcuoglu, k. 2017. Neural Discrete Representation Learning. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Yang, S.; Zhang, Y.; Feng, D.; Yang, M.; Wang, C.; Xiao, J.; Long, K.; Shan, S.; and Chen, X. 2019. LRW-1000: A Naturally-Distributed Large-Scale Benchmark for Lip Reading in the Wild. In *2019 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019)*, 1–8.

Zhang, X.; Cheng, F.; and Wang, S. 2019. Spatio-Temporal Fusion Based Convolutional Sequence Learning for Lip Reading. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Zhang, Y.; Xiang, T.; Hospedales, T. M.; and Lu, H. 2018. Deep Mutual Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, Z.; Song, Y.; and Qi, H. 2018. Decoupled Learning for Conditional Adversarial Networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 700–708. Los Alamitos, CA, USA: IEEE Computer Society.

Zhao, Y.; Xu, R.; Wang, X.; Hou, P.; Tang, H.; and Song, M. 2020. Hearing Lips: Improving Lip Reading by Distilling Speech Recognizers. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04): 6917–6924.

Zhou, H.; Liu, Y.; Liu, Z.; Luo, P.; and Wang, X. 2019. Talking Face Generation by Adversarially Disentangled Audio-Visual Representation. In *AAAI Conference on Artificial Intelligence (AAAI)*.