Visual Hallucination Elevates Speech Recognition

Fang Zhang[1,2], Yongxin Zhu[1,2], Xiangxiang Wang[3], Huang Chen[3], Xing Sun[3], Linli Xu[1,2]
[1]School of Computer Science and Technology, University of Science and Technology of China
[2]State Key Laboratory of Cognitive Intelligence
[3]Tencent YouTu Lab
{fangzhang,zyx2016}@mail.ustc.edu.cn
{xenoswang,huaangchen,winfredsun}@tencent.com
linlixu@ustc.edu.cn

Abstract

Due to the detrimental impact of noise on the conventional audio speech recognition (ASR) task, audio-visual speech recognition (AVSR) has been proposed by incorporating both audio and visual video signals. Although existing methods have demonstrated that the aligned visual input of lip movements can enhance the robustness of AVSR systems against noise, the paired videos are not always available during inference, leading to the problem of the missing visual modality, which restricts their practicality in real-world scenarios. To tackle this problem, we propose a Discrete Feature based Visual Generative Model (DFVGM) which exploits semantic correspondences between the audio and visual modalities during training, generating visual hallucinations in lieu of real videos during inference. To achieve that, the primary challenge is to generate the visual hallucination given the noisy audio while preserving semantic correspondences with the clean speech. To tackle this challenge, we start with training the audio encoder in the Audio-Only (AO) setting, which generates continuous semantic features closely associated with the linguistic information. Simultaneously, the visual encoder is trained in the Visual-Only (VO) setting, producing visual features that are phonetically related. Next, we employ K-means to discretize the continuous audio and visual feature spaces. The discretization step allows DFVGM to capture high-level semantic structures that are more resilient to noise and generate visual hallucinations with high quality. To evaluate the effectiveness and robustness of our approach, we conduct extensive experiments on two publicly available datasets. The results demonstrate that our method achieves a remarkable 53% relative reduction (30.5%→12.9%) in Word Error Rate (WER) on average compared to the current state-of-the-art Audio-Only (AO) baselines while maintaining comparable results ($< 5\%$ difference) under the Audio-Visual (AV) setting even without video as input.

Introduction

Recognizing speech is essential for natural human-computer interactions, facilitating accessibility for individuals with disabilities, and advancing various applications such as virtual assistants, transcription services, and voice-controlled technologies. In recent years, end-to-end Automatic Speech Recognition (ASR) based on deep learning (Graves, Mohamed, and Hinton 2013; Hinton et al. 2012) has become the standard approach. Meanwhile, the quality and intelligibility of speech recognition are highly vulnerable to noise and may degrade dramatically with corrupted speech (Vincent et al. 2017; Kinoshita et al. 2020). Therefore, enhancing noise robustness is crucial for ASR systems. Motivated by the fact that invariant lip movements in a video are not affected by noisy environments, Audio Visual Speech Recognition (AVSR) is proposed to transcribe text from both audio and visual streams. Various studies have confirmed the significant superiority of Audio-Visual (AV) models over their Audio-only (AO) counterparts in diverse noisy scenarios (Son Chung et al. 2017; Petridis et al. 2018a,b), as well as in handling overlapping speech (Rose et al. 2021; Yu et al. 2020).

However, the paired visual input is not always accessible at inference time, which is quite common in practice. For instance, the speaker may step away or eat food, the audio and lips go out of sync, and the camera or recording devices can be turned off, etc. These significantly limit the applicability of these AVSR methods in real-life scenarios, as most of them are unable to handle the absence of the visual modality. The conventional approach to addressing the problem of missing modality generally involves modality translation, which reconstructs the absent modality by leveraging information from the available modalities. In the AVSR task, it becomes more challenging due to the corruption of the audio modality. Hegde et al. propose to generate accurate lip movements given noisy audio. By employing a pretrained speech-to-lip model called Wav2Lip (Prajwal et al. 2020) as a teacher network that generates precise lip movements from clean speech, a student network is subsequently trained to imitate the teacher's lip movements given noisy speech. However, a major limitation of this method is its disregard of the high-level semantic relationships between the audio and visual modalities, resulting in the generation of pseudo videos with low information density. Therefore, an additional visual encoder is required in (Hegde et al. 2021) to extract the semantic features with a higher correlation to the speech content.

In this paper, to effectively address the aforementioned challenges regarding the visual modality dropout, we directly model the semantic relationships between the audio and visual modalities in the discrete feature spaces rather than in the real feature space. Firstly, we pretrain the audio and visual encoders through Audio Speech Recognition (ASR) and Lip Reading Recognition tasks respectively. These tasks enable the encoders to generate continuous semantic features that are strongly associated with the phonetic and linguistic information (Shi et al. 2021). Subsequently, we apply K-means (Lloyd 1982) clustering to discretize the feature spaces into the audio and visual codebooks. The discrete encoding has the following advantages compared to the continuous embedding: Firstly, the audio codebook is inferred from the clean audio tracks which is noise-invariant. Intuitively, a noisy audio is not far from its clean counterpart in the feature space. Therefore, by finding the nearest neighbor in the codebook, the noisy audio can be partly converted to its clean counterpart in the discrete feature space, which reduces the noise to some extent. Furthermore, the discrete feature spaces facilitate capturing the high-level semantic structures and identifying the semantic correlations. After discretizing both video and audio features into sequences using codebooks, Discrete Feature based Visual Generative Model (DFVGM), employing an encoder-decoder architecture is trained to generate visual sequences based on the clean or noisy audio sequences in an auto-regressive manner. Furthermore, to strengthen the short-range dependencies between audio and visual tokens, we introduce a distance penalty to penalize the cross-attention logits where the visual tokens focus more on the audio tokens at a closer time step. Finally, we train the fusion module and decoder with an additional consistency loss based on the KL divergence to reduce the mismatch between the real visual tokens and pseudo visual tokens generated from our DFVGM. During inference, our model requires only audio signals while employing visual hallucinations for multimodal fusion instead, as illustrated in Figure 1. Our key contributions are summarized as follows:

- We investigate the scenario where the visual modality is completely missing during inference in audio visual speech recognition.
- We introduce a novel Discrete Feature based Visual Generative Model, which captures semantic correspondences between the visual and audio modalities during training in discrete spaces and generates visual hallucinations in lieu of real visual inputs during inference.
- Extensive experiments on two public datasets demonstrate that our visual hallucinations can be leveraged to improve the robustness and performance of AO models. Our approach outperforms state-of-the-art AO baselines by a large margin, achieving an average reduction of 53% in WER across different SNR ratios.
- By utilizing ground truth visual information as input, our approach achieves an absolute improvement of 1.2% WER reduction over other state-of-the-art AV baselines.

Related Work

Audio Visual Speech Recognition

Most existing AVSR systems share a similar architecture composed of an audio encoder, a visual encoder, a fusion module, and a decoder (Pan et al. 2022) as shown is Figure 2(a). Previous works have made improvements specifically for these four components. Typically, both the audio encoder and visual encoder consist of two components: a front-end and a back-end. For the front-end, Pan et al. observe that utilizing pre-trained models such as Wav2Vec (Baevski et al. 2020), and Moco (Chen et al. 2020) to initialize the parameters of front-ends could enhance the performance for AVSR. The purpose of the back-end is to model temporal relationships, where sequence models such as RNN, LSTM have been widely employed in previous works (Makino et al. 2019). Besides, earlier works (Ma, Petridis, and Pantic 2021; Burchi and Timofte 2023) demonstrate that the Conformer architecture (Gulati et al. 2020) can better capture the temporal information locally and globally by progressively down-sampling the temporal sequence and reducing the computational overhead.

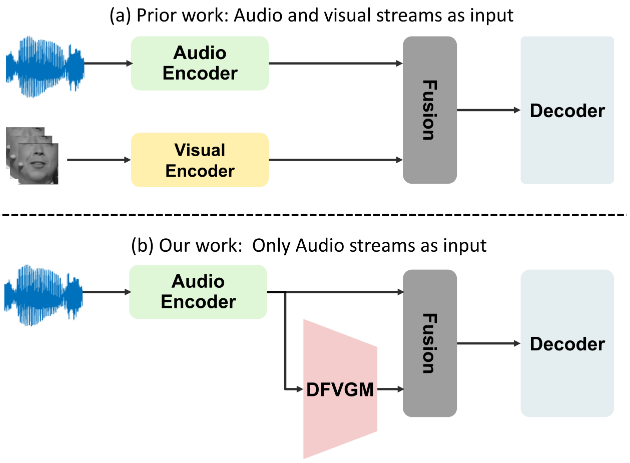
(a) Prior work: Audio and visual streams as input
Audio Encoder
Visual Encoder
Fusion
Decoder
(b) Our work: Only Audio streams as input
Audio Encoder
DFVGM
Fusion
Decoder

Figure 1: We propose DFVGM to generate visual features based on audio features. During Inference, the fusion module takes audio features and pseudo visual features as input. In comparison, prior AVSR approaches require multimodal inputs.

For the fusion model, the most common fusion strategy is concatenating two context vectors over the channel dimension (Afouras et al. 2018). However, it is pointed out that the straightforward concatenation of features fails to provide insight into the level of reliability of a particular data stream (Potamianos et al. 2003). To address that, multi-modality attention is proposed to adjust its modality attention towards the most reliable input modality (Zhou et al. 2019). Similarly, AV-RelScore (Hong et al. 2023) computes reliability scores for each time step, indicating how much the current audio features and visual features contribute to recognizing speech. Inspired by speech enhancement (Benesty, Makino, and Chen 2006), V-CAFE (Hong et al. 2022) introduces a noise reduction mask to encode audio features, aiming to reduce noise in audio representations. For the decoder, the Connectionist Temporal Classification (CTC) loss (Graves et al. 2006) and Sequence-to-Sequence loss (Sutskever, Vinyals, and Le 2014) based on cross-entropy are widely applied in end-to-end speech recognition systems. A hybrid CTC/attention architecture (Petridis et al. 2018b) has recently been proposed to address the limitations of both CTC and attention models. This architecture seeks to enforce monotonic alignments while also eliminating conditional independence. Besides, an additional language model which is trained separately on a large corpus of text data has been shown effective in augmenting the decoding process by incorporating linguistic contexts (Ma, Petridis, and Pantic 2021).

Missing Modality

The missing modality problem has received significant attention in the multimodal learning community, spanning diverse applications including emotion recognition (Zhao, Li, and Jin 2021), medical image segmentation (Azad et al. 2022), audio-visual expression recognition (Parthasarathy and Sundaram 2020), etc. Specifically, in the domain of AVSR, an initial endeavor (Chang et al. 2022) involves a cascaded model, wherein, for every video frame, the model follows the AV path if the frame is available and resorts to the AO path otherwise. On top of this, modality dropout (Shi et al. 2021; Shi, Hsu, and Mohamed 2022) is proposed to address the input discrepancy by masking the full features of one modality before fusing the audio and visual inputs. However, simply ignoring the visual modality and focusing more on the audio inputs would degrade performances in noisy settings.

Method

In this section, we introduce the pipeline of training stages. Initially, to obtain the continuous audio and visual feature spaces containing semantic information, we follow (Pan et al. 2022) to pretrain the audio encoder and visual encoder separately in audio-only (AO) and visual-only (VO) settings, where the audio front-end is initialized using Wav2Vec (Baevski et al. 2020), and the visual front-end is initialized by Moco V2 (Chen et al. 2020)[1]. We then discretize the continuous audio and visual feature spaces by applying K-means clustering, resulting in audio and visual tokens. After that, DFVGM is trained to learn the mapping between visual and audio token sequences. Finally, we further train the fusion module and decoder under the inputs of discrete audio and visual tokens. The complete training pipeline is illustrated in Figure 2.

Codebook

Suppose we are given a dataset with paired clean audio and visual recordings: $A = \{a_m\}_{m=1}^{N}, V = \{v_m\}_{m=1}^{N}$ where N is the number of pairs in the dataset. The audio recording a_m is processed by an audio encoder, producing the sequence $\{f_m^t\}_{t=1}^{t_m}$ where t_m is the length of the m-th audio feature sequence. We collect all the sequences for every audio clip in the dataset $F_A = \left\{f_1^1, f_1^2, \cdots, f_1^{1_m}, \cdots, f_m^1, \cdots, f_m^{t_m}, \cdots, f_N^1, \cdots, f_N^{t_N}\right\}$, to which K-means is applied to quantize the audio feature space and produce K_A clusters which constitute the audio codebook. We denote the audio codebook as $E_A = \left\{e_a^k\right\}_{k=1}^{K_A}$ where d is the dimension of each cluster. For the features of an audio recording, by finding the nearest neighbors in the audio codebook, we can obtain an audio discrete token sequence as $x = [x_1, \cdots, x_T]$ where $x_i \in \{1, \cdots, K_A\}$ is the index of its nearest audio cluster in E_A. Similarly we can obtain the visual codebook $E_V = \left\{e_v^k\right\}_{k=1}^{K_V}$ where K_A is not equal to K_V in general, and an visual discrete token sequence as $y = [y_1, \cdots, y_T]$ where $y_i \in \{1, \cdots, K_V\}$.

Discrete encoding has the following advantages compared to continuous encoding: Firstly, the audio and visual codebooks derived from clean inputs exhibit higher-level semantic structures. This characteristic facilitates the exploration of semantic relationships between two modalities. Additionally, in the continuous feature space, the features of noisy audio are not significantly distant from their clean counterparts. Intuitively, by identifying the nearest neighbor in the codebook, the noisy audio corresponds to the same cluster as its clean counterpart. Furthermore, discrete encoding enables the utilization of sequence-to-sequence generation with a cross-entropy loss, similar to natural language processing (NLP), which avoids the problem of continuous representations collapsing to the mean value.

DFVGM

The proposed DFVGM is a transformer-based model consisting of several encoder and decoder layers and is designed to generate visual hallucinations that have high semantic correspondences with the audio modality. We model the conditional probability distribution in an auto-regressive way as follows:

$$p(y \mid x) = \prod_{i=1}^{T} p(y_i \mid x) \tag{1}$$

The object function of DFVGM is computed as follows:

$$\mathcal{L} = \log p_{CE}(y \mid x) \tag{2}$$

Noise Augmented Training Previous works (Xu et al. 2020; Ma, Petridis, and Pantic 2021) add noise to the clean audio, sampled at a fixed or signal-to-noise ratio (SNR) to boost robustness to noise. We extend this noise augmented training to DFVGM with the assumption that both noisy and clean audio tokens should correspond to the same visual tokens. By adding diverse noise to an audio stream during training and then discretizing it, we can obtain the noisy audio sequence. Our DFVGM models the conditional probability based on the noisy or clean audio sequences, thereby narrowing the domain gap between the noisy and clean audio.

Distance Penalty The differences between the task of transforming audio tokens into visual tokens and machine translation can be observed in two aspects. Firstly, audio tokens and visual tokens should be of equal lengths considering that they are synchronized in time and need to be concatenated in the fusion module. Secondly, due to the down-

[1] The Moco V2 is firstly trained on the datastet LRW (Son Chung et al. 2017).

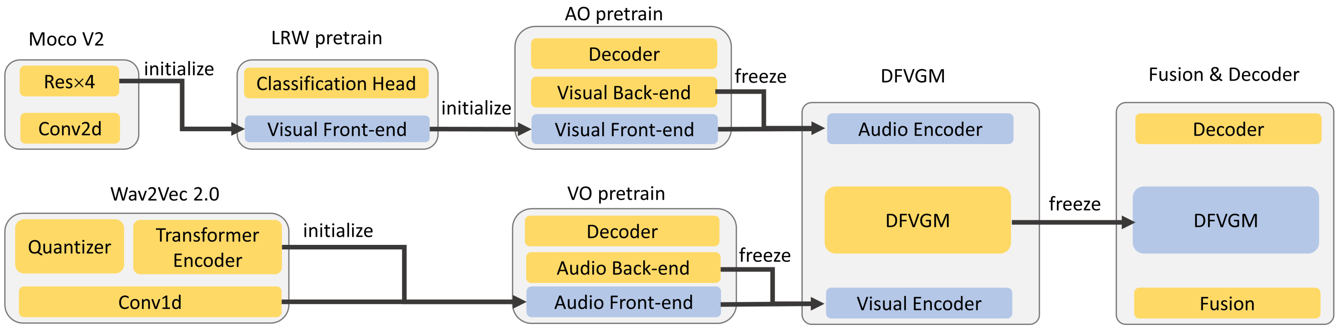

Figure 2: Training pipeline of our model. Yellow blocks represent newly initialized parameters, while blue blocks represent parameters that are inherited from the last stage.

sampling of the audio and visual encoders, x_i and y_i represent audio and visual features over a period of time and are strongly correlated, which requires the model to have stronger short-range modeling capabilities. As a result, the vanilla transformer architecture is likely to perform poorly on this task. We take the inspiration from the recent work on speech-to-speech translation (Di Gangi, Negri, and Turchi 2019) and apply a distance penalty to each head of its cross-attention layer in the decoder.

$$\text{CrossATTN} = \text{softmax}\left(\frac{QK^T}{\sqrt{d_{\text{model}}}} - \pi(D)\right)V \tag{3}$$

where d_{model} is the attention dimension, $K, V \in \mathbb{R}^{T \times d_{\text{model}}}$ are the key and value inputs from the encoder output, $Q \in \mathbb{R}^{T \times d_{\text{model}}}$ is the query inputs from the previous self-attention layer. $D \in \mathbb{R}^{T \times T}$ is the position distance matrix, representing the time difference between the audio tokens x_i and visual tokens y_j, i.e., $D_{ij} = |i - j|$. $\pi(\cdot)$ is based on a hard-coded function to penalize the attention logits as follows:

$$\pi(D) = \begin{cases} 0 & \text{if} \quad D_{ij} < R \\ \ln(R) & \text{otherwise} \end{cases} \tag{4}$$

where R is a hyper-parameter controlling the range of the local dependency. By doing so, the visual token y_j would pay more attention to the audio tokens $x_{j-R+1}, \cdot\cdot\cdot, x_j, \cdot\cdot\cdot, x_{j+R-1}$. For the audio tokens outside of the window, the logarithm function is applied to bias the long-dependence range but the penalty increases slowly with distance, which allows modeling global dependencies. During inference, to generate the pseudo visual sequence with the same length of the audio sequence, we set the number of inference steps to be the same as T which results in $\bar{y} = [\bar{y}_1, \cdot\cdot\cdot, \bar{y}_T]$.

$$\bar{y}_i = \underset{k \in \{1, \cdots, K_V\}}{\arg\max} \ \text{DFVGM}(\bar{y}_i = k \mid \bar{y}_{<i}, x) \quad i \le T \tag{5}$$

Training Objectives

Given the target $z = \{z_i\}_{i=1}^{T_z}$, our fusion module concatenates the audio cluster $e_a^{x_i}$ and visual cluster $e_v^{y_i}$ along the feature dimension which are then passed into the decoder. The objective function of the decoder is a hybrid CTC/attention loss (Petridis et al. 2018b) as follows:

$$\mathcal{L}_{\text{gt}} = \alpha \log p_{\text{CTC}}\left(z \mid x, y\right) + (1 - \alpha) \log p_{\text{CE}}\left(z \mid x, y\right) \tag{6}$$

where α controls the relative weight between the CTC loss and Seq2Seq loss and the subscript of $\mathcal{L}_{\text{gt}}$ indicates that the visual input comprises the ground-truth visual tokens. However, the fusion module takes the real visual discrete sequences during training but only has access to the pseudo visual sequences at inference. To reduce the mismatch, we firstly compute the output distribution based on the pseudo visual sequences by sharing the weights of the fusion module and decoder.

$$\mathcal{L}_{\text{pseudo}} = \alpha \log p_{\text{CTC}}\left(z \mid x, \bar{y}\right) + (1 - \alpha) \log p_{\text{CE}}\left(z \mid x, \bar{y}\right) \tag{7}$$

Then we propose a consistency loss to reduce the KL divergence between two conditional distributions.

$$\mathcal{L}_{\text{consistency}} = \left\{ \sum_{i=1}^{T_z} \mathcal{D}_{\text{KL}} \left[z_i^{\text{gt}} || z_i^{\text{pseudo}} \right] \right\} \tag{8}$$

where z_i^{gt} and z_i^{pseduo} are the next word logits predicted from the ground-truth visual tokens and pseudo visual tokens respectively. The final overall optimization objective of the decoder is computed as follows:

$$\mathcal{L}_{\text{overall}} = \mathcal{L}_{\text{gt}} + \mathcal{L}_{\text{pseudo}} + \mu \mathcal{L}_{\text{consistency}} \tag{9}$$

where μ controls the consistency between the two translation outputs.

Experiments

Datasets

Our methodology is assessed utilizing two comprehensive, publicly accessible audio-visual datasets, namely LRS2 (Son Chung et al. 2017) and LRS3 (Afouras, Chung, and Zisserman 2018). Both datasets pose considerable challenges due to vast variations in aspects such as head pose and illumination. The LRS2 dataset (Son Chung et al. 2017) is composed of 224 hours of video footage, featuring $144,482$ clips sourced from BBC videos. The training data consists of over 2 million word instances and a vocabulary exceeding $40,000$ words. The LRS3 dataset (Afouras, Chung, and Zisserman 2018), extracted from TED and TEDx presentations, includes $118,516$ utterances in the pre-training set (408 hours), $31,982$ in the training-validation set (30 hours), and $1,321$ in the test set (0.9 hours).

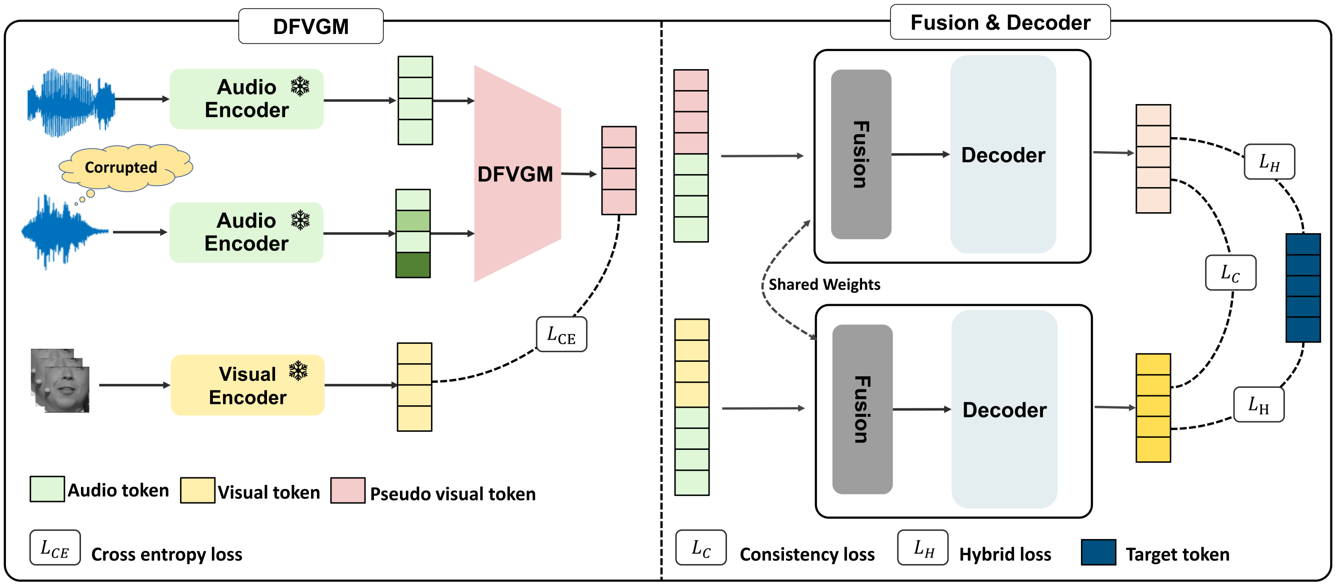

Figure 3: Overview of the architecture for our method. Left: The training procedure of our DFVGM. DFVGM takes both the clean and corrupted audio sequences to generate same visual token sequences. Right: The training procedure of the fusion module and decoder. The input of the fusion module comes from two streams of ground-truth and pseudo visual tokens respectively. During training, the other components are frozen.

Experimental Settings

Model Settings For our DFVGM, the model consists of 6 encoder layers and 6 decoder layers, where the number of attention heads is 4, the hidden dimension d_{model} is 256, and the feed-forward dimension is 1024. We set $K_a = 800, K_v = 1000$ for the codebooks. We train our DFVGM by an Adam (Kingma and Ba 2014) optimizer with hyperparameters $\beta_1 = 0.9$, $\beta_2 = 0.98$. The dropout is set to 0.2 and label smoothing weight is set to 0.1. The hyperparameter R is set to 2. The architecture of our audio encoder and visual encoder are inherited from the previous work (Pan et al. 2022). Our fusion module is simply concatenating on the feature dimension. Our decoder is composed of a Transformer seq2seq decoder and a CTC decoder. An additional pretrained language model is used during training and inference. For the final training, the relative weight μ is set to 0.5 tuned on the validation set. The fusion module and decoder are trained for 50 epochs, with an initial learning rate of $1e-5$. For the loss functions $\mathcal{L}_{\text{gt}}$ and $\mathcal{L}_{\text{pseudo}}$, the relative CTC weight is 0.1. During inference, we use beam search with a beam size of 10.

Data Processing To pre-process and augment the video frames and raw waveforms, 68 facial landmarks are detected and tracked using dlib. To remove differences related to face rotation and scale, the faces are aligned to a neural reference frame using a similarity transformation. To augment the data, random cropping with a size of 88×88 and horizontal flipping with a probability of 0.5 is applied to each video frame. Each raw audio waveform is normalized by subtracting its mean and dividing it by its standard deviation. We apply the same configuration that rejects a babble noise from the NOISEX corpus (Martinez et al. 2020) to the original audio clip. Babble noise is randomly added to the audio stream at 0 dB, -5dB and -10dB SNR with a probability of 0.25 for training our DFVGM and decoder.

Evaluation Our DFVGM model is trained in a sequence-to-sequence paradigm, and to evaluate its performance, we use BLEU (Papineni et al. 2002) as our evaluation metric which is widely used in machine translation and summarization. For the final experiments, word error rate (WER) is reported on LRS2 and LRS3 with clean and different SNR levels settings. A lower WER indicates better performance, while a lower SNR indicates a higher level of noise.

Evaluation and Analysis

Discrete Encoding Improves Robustness The results in Table 1 demonstrate the performance of our proposed DFVGM without noise augmented training. DFVGM, when tested with clean audio input, achieves a BLEU score of 72.3, indicating a strong ability to generate visual representations that are semantically consistent with the ground truth. In addition to the primary testing scenario, we also evaluated the model's performance when exposed to various types of unseen noise. Remarkably, DFVGM maintained comparable performances even under noise levels of 5 and 0 dB. We ascribe this robustness to the advantages of discrete spaces, where high-level semantic structures remain invariant to noise and semantic relationships between two modalities are strengthened.

This provides empirical support for our initial hypothesis that noisy audio can be effectively transformed into clean counterparts within discrete feature spaces. When the noise level comes to -5 dB, a significant deterioration in performance is observed. We hypothesize that the decline is due to the introduction of distortions and inconsistencies within the semantic structure by the noise, which subsequently impedes DFVGM's ability to generate consistent visual sequences.

Noise	clean	10	5	0	-5
babble	72.3	69.3	67.4	59.9	48.1
white	72.3	68.4	62.7	57.6	34.1
birds	72.3	64.9	60.4	50.1	27.5
jazz	72.3	65.4	62.3	56.1	44.6
train	72.3	59.4	55.1	42.3	31.5

Table 1: BLEU scores when DFVGM is evaluated on the LRS2 test dataset with different types of noise. In this paradigm, the babble noise is not accessible during training.

Impact of Noise Augmented Training The efficacy of integrating noise during the training of DFVGM is illustrated in Table 2. By incorporating babble noise into the training process, we observe a considerable relative improvement of 2.3% over the equivalent model trained without noise augmentation, which highlights the robustness and effectiveness of this approach.

Noise	clean	10	5	0	-5
babble$^\diamond$	74.2	72.4	71.5	62.7	53.1
white$^\diamond$	74.2	70.1	68.9	57.6	46.5
birds$^\diamond$	74.2	66.7	62.1	54.3	28.9
jazz$^\diamond$	74.2	67.8	65.7	62.0	47.6
train$^\diamond$	74.2	62.3	59.4	48.3	39.7

Table 2: BLEU scores when DFVGM is evaluated on the LRS2 test dataset with different types of noise. $\diamond$ indicates experiments being trained with babble noise.

Comparison with AO Models In this study, we conduct a comparative analysis between our proposed model and AO models, as depicted in Table 3. For this experiment, all models take solely an audio waveform contaminated by the same babble noise as input during inference. The results indicate that our model outperforms all AO models by a considerable margin, illustrating that visual hallucinations based on noisy audio signals can furnish supplementary semantic information to enhance speech content recognition. The advantages of incorporating visual hallucinations become more pronounced in challenging scenarios, where the WER degradation relative to the clean condition is large. Such results demonstrate the robustness and broad applicability of our proposed model. Upon comparison with AV-Hubert on the LRS3 dataset, which randomly masks the whole modality during training, our method obtains 77% (62.3% → 14.3%) and 58% (97.5% → 40.6%) WER relative reduction under noisy settings of -5dB and -10dB, respectively. This comparison highlights the limitations of the modality dropout mechanism. While it addresses the input discrepancy, it falls short in providing complementary information to enhance AO models as our method. For a previous work (Hegde et al. 2021) similar to ours, where pseudo videos are generated from noisy audio by an enhanced speech2lip model to facilitate speech enhancement, our approach achieves 63% (7.1% → 2.6%) and 66.7% (9.3% → 3.1%) relative WER reduction. We attribute this notable performance enhancement to the following two factors: 1) Compared to generating accurate pseudo videos with hundreds of frames in sync with audio, sequence-to-sequence generation based on discrete space is easier to achieve. 2) Utilizing features of a higher correlation with semantic information likely contributes to improving the AO model, while the video information density tends to be relatively low, necessitating an additional visual encoder to extract information from ambiguous videos, which could lead to the accumulation of errors.

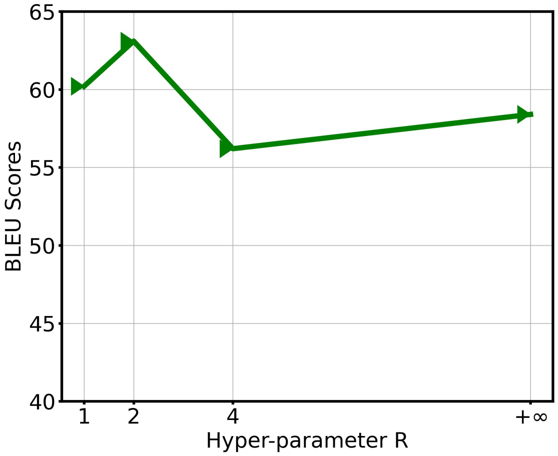

Figure 4: BLEU scores on the LRS3 dataset when changing R in the penalty distance.

Comparison with AV Models As illustrated in Table 4, our model's performance is marginally lower than the AV models, which is reasonable considering we do not incorporate visual information during inference. When compared to AV-RelScore, which utilizes noisy audio and corrupted videos as input, our model surpasses it across all SNR levels, even without access to the visual modality. By incorporating the ground-truth visual sequences as input, our model significantly outperforms other state-of-the-art AV baselines, more noticeably at higher SNR levels. It is important to note that all baseline models are trained in continuous spaces. This comparative performance substantiates our hypothesis that a discretized feature space is more robust to noise than its continuous counterpart. Furthermore, we conduct a comparative analysis of our model with LUSL, which shares a common architecture, the Conformer, along with identical audio and visual encoders. The 63% relative reduction in WER observed in our model underscores that its superiority does not stem from the architectural configuration. Instead, it is the unique training paradigm that our model employs that contributes to its enhanced performance. More-

Dataset	Input model	Method	clean	10	5	0	-5	-10	avg
LRS2	A	Conformer (Ma, Petridis, and Pantic 2021)	4.9	7.4	8.7	10.2	29.1	97.5	30.5
	A	AVEC (Burchi and Timofte 2023)	3.1	7.6	8.6	27.0	70.5	98.8	42.4
	A+PV	Pseudo Visual (Hegde et al. 2021)	7.1	9.3	17.0	39.5	78.7	96.3	48.2
	A+VH	Ours	**2.6**	**3.1**	**3.9**	**7.1**	**12.6**	**42.6**	**13.9**
LRS3	A	AVEC	2.3	4.1	9.3	32.4	75.9	97.4	43.8
	A	AV-Hubert (Shi, Hsu, and Mohamed 2022)	2.6	2.6	5.1	15.7	62.3	97.5	36.7
	A+PV	Pseudo Visual	4.4	6.6	12.4	35.6	78.2	98.0	46.2
	A+VH	Ours	**2.2**	**2.2**	**3.4**	**6.4**	**14.3**	**40.6**	**13.4**

Table 3: Comparison of WER(%) between state-of-the-art methods and our model in the audio-only environment with different noise levels of a babble noise, SNR(dB) on two datasets. A, V, VH, and PV denote audio, video, visual hallucination, and pseudo video respectively.

Dataset	Input model	Method	clean	10	5	0	-5	-10	avg
LRS2	A+V	LUSL (Pan et al. 2022)	2.6	5.3	6.4	24.5	34.9	52.8	24.7
	A+V	AVEC	2.6	2.8	3.4	**5.0**	9.7	30.4	10.3
	A+V	AV-RelScore (Hong et al. 2023)	4.1	4.3	5.2	6.3	11.3	-	-
	A+VH	Ours	2.6	3.1	3.9	7.1	12.6	42.6	13.9
	A+V	Ours*	**2.4**	**2.8**	**3.3**	5.2	**8.4**	**27.4**	**9.4**
LRS3	A+V	AVEC	**2.0**	2.5	3.1	4.9	11.2	34.9	11.3
	A+V	AV-Hubert	**2.0**	**2.1**	**2.6**	5.8	16.6	34.9	12.4
	A+V	AV-RelScore	2.8	2.8	3.2	4.8	**8.7**	-	-
	A+VH	Ours	2.2	2.2	3.4	6.4	14.3	40.6	13.4
	A+V	Ours*	**2.0**	2.2	2.9	**4.5**	10.3	**30.7**	**10.2**

Table 4: Comparison of WER(%) between state-of-the-art methods and our model in the audio-visual environment.

over, our model exhibits comparable performance whether using ground-truth visual sequences or visual hallucinations, which can partly be attributed to our consistency loss that reduces the semantic difference between them.

Ablation Study

Distance Penalty As depicted in Figure 4, the hyperparameter R in Equation 4 affects the flexibility of our distance penalty in modeling local range dependency. when R is set to $+\infty$, the cross-attention mechanism degrades to the vanilla since all D_{ij} values become zero. Generally, setting R to 2 yields the best performance, resulting in an improvement of 5 in the BLEU score compared to the vanilla cross attention model, demonstrating effectiveness of our distance penalty.

Additional ASR Data Augments the Performance Another advantage of our method is that we can use ASR datasets to construct fake AVSR datasets. Due to the strong ability of DFVGM to align audio tokens with visual tokens which has been validated in previous experiments, we can generate hallucinated visual sequences which are then fed into the fusion module with the audio sequences. By doing this, the fusion module and decoder are trained with $\mathcal{L}_{\text{pseudo}}$ in Equation 7. Table 5 shows the effect of additional ASR data. We utilize our model pretrained on LRS2 with 20% of the LRS3 Audio dataset which results in an absolute improvement of 1.4% and 1.7% WER reduction when using 20% and 40% additional data, respectively.

Method	Clean	5dB	0dB
Our model (LRS2)	2.4	7.1	12.6
+ LRS3 Audio dataset (20%)	2.3	6.5	11.2
+ LRS3 Audio dataset (40%	2.2	5.9	10.9

Table 5: WER metrics of our model with different proportions of additional ASR datasets.

Conclusion

In this paper, we address the problem of the missing visual modality in Audio Visual Speech Recognition. We propose to generate visual hallucinations from noisy audio as input during inference. To capture the semantic correspondence between the audio and visual modalities, we introduce a Discrete Feature based Visual Generative Model (DFVGM) that translates audio tokens into visual tokens in discrete feature spaces. The discrete feature spaces with high-level semantic structures are noise-invariant and beneficial to generating accurate visual hallucinations. Extensive experiments demonstrate that visual hallucinations can be leveraged to enhance the performance of speech recognition.

Acknowledgments

This research was supported by the National Key Research and Development Program of China (Grant No. 2022YFB3103100), the National Natural Science Foundation of China (Grant No. 62276245), and Anhui Provincial Natural Science Foundation (Grant No. 2008085J31).

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*, 44(12): 8717–8727.

Afouras, T.; Chung, J. S.; and Zisserman, A. 2018. LRS3-TED: a large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496*.

Azad, R.; Khosravi, N.; Dehghanmanshadi, M.; Cohen-Adad, J.; and Merhof, D. 2022. Medical image segmentation on mri images with missing modalities: A review. *arXiv preprint arXiv:2203.06217*.

Baevski, A.; Zhou, Y.; Mohamed, A.; and Auli, M. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33: 12449–12460.

Benesty, J.; Makino, S.; and Chen, J. 2006. *Speech enhancement*. Springer Science & Business Media.

Burchi, M.; and Timofte, R. 2023. Audio-visual efficient conformer for robust speech recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2258–2267.

Chang, O.; Braga, O.; Liao, H.; Serdyuk, D.; and Siohan, O. 2022. On Robustness to Missing Video for Audiovisual Speech Recognition. *Transactions on Machine Learning Research*.

Chen, X.; Fan, H.; Girshick, R.; and He, K. 2020. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*.

Di Gangi, M. A.; Negri, M.; and Turchi, M. 2019. Adapting transformer to end-to-end spoken language translation. In *Proceedings of INTERSPEECH 2019*, 1133–1137. International Speech Communication Association (ISCA).

Graves, A.; Fernández, S.; Gomez, F.; and Schmidhuber, J. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, 369–376.

Graves, A.; Mohamed, A.-r.; and Hinton, G. 2013. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, 6645–6649. Ieee.

Gulati, A.; Qin, J.; Chiu, C.-C.; Parmar, N.; Zhang, Y.; Yu, J.; Han, W.; Wang, S.; Zhang, Z.; Wu, Y.; et al. 2020. Conformer: Convolution-augmented Transformer for Speech Recognition. *Proc. Interspeech 2020*, 5036 − −5040.

Hegde, S. B.; Prajwal, K.; Mukhopadhyay, R.; Namboodiri, V. P.; and Jawahar, C. 2021. Visual speech enhancement without a real visual stream. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1926–1935.

Hinton, G.; Deng, L.; Yu, D.; Dahl, G. E.; Mohamed, A.-r.; Jaitly, N.; Senior, A.; Vanhoucke, V.; Nguyen, P.; Sainath, T. N.; et al. 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6): 82–97.

Hong, J.; Kim, M.; Choi, J.; and Ro, Y. M. 2023. Watch or Listen: Robust Audio-Visual Speech Recognition with Visual Corruption Modeling and Reliability Scoring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18783–18794.

Hong, J.; Kim, M.; Yoo, D.; and Ro, Y. M. 2022. Visual Context-driven Audio Feature Enhancement for Robust End-to-End Audio-Visual Speech Recognition. *arXiv preprint arXiv:2207.06020*.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kinoshita, K.; Ochiai, T.; Delcroix, M.; and Nakatani, T. 2020. Improving noise robust automatic speech recognition with single-channel time-domain enhancement network. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7009–7013. IEEE.

Lloyd, S. 1982. Least squares quantization in PCM. *IEEE transactions on information theory*, 28(2): 129–137.

Ma, P.; Petridis, S.; and Pantic, M. 2021. End-to-end audio-visual speech recognition with conformers. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7613–7617. IEEE.

Makino, T.; Liao, H.; Assael, Y.; Shillingford, B.; Garcia, B.; Braga, O.; and Siohan, O. 2019. Recurrent neural network transducer for audio-visual speech recognition. In *2019 IEEE automatic speech recognition and understanding workshop (ASRU)*, 905–912. IEEE.

Martinez, B.; Ma, P.; Petridis, S.; and Pantic, M. 2020. Lipreading using temporal convolutional networks. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6319–6323. IEEE.

Pan, X.; Chen, P.; Gong, Y.; Zhou, H.; Wang, X.; and Lin, Z. 2022. Leveraging Unimodal Self-Supervised Learning for Multimodal Audio-Visual Speech Recognition. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 4491–4503.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Parthasarathy, S.; and Sundaram, S. 2020. Training strategies to handle missing modalities for audio-visual expression recognition. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*, 400–404.

Petridis, S.; Stafylakis, T.; Ma, P.; Cai, F.; Tzimiropoulos, G.; and Pantic, M. 2018a. End-to-end audiovisual speech recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 6548–6552. IEEE.

Petridis, S.; Stafylakis, T.; Ma, P.; Tzimiropoulos, G.; and Pantic, M. 2018b. Audio-visual speech recognition with a hybrid ctc/attention architecture. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, 513–520. IEEE.

Potamianos, G.; Neti, C.; Gravier, G.; Garg, A.; and Senior, A. W. 2003. Recent advances in the automatic recognition of audiovisual speech. *Proceedings of the IEEE*, 91(9): 1306–1326.

Prajwal, K.; Mukhopadhyay, R.; Namboodiri, V. P.; and Jawahar, C. 2020. 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*, 484–492.

Rose, R.; Siohan, O.; Tripathi, A.; and Braga, O. 2021. End-to-End Audio-Visual Speech Recognition for Overlapping Speech. In *Interspeech*, 3016–3020.

Shi, B.; Hsu, W.-N.; Lakhotia, K.; and Mohamed, A. 2021. Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction. In *International Conference on Learning Representations*.

Shi, B.; Hsu, W.-N.; and Mohamed, A. 2022. Robust self-supervised audio-visual speech recognition. *arXiv preprint arXiv:2201.01763*.

Son Chung, J.; Senior, A.; Vinyals, O.; and Zisserman, A. 2017. Lip reading sentences in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6447–6456.

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Vincent, E.; Watanabe, S.; Nugraha, A. A.; Barker, J.; and Marxer, R. 2017. An analysis of environment, microphone and data simulation mismatches in robust speech recognition. *Computer Speech & Language*, 46: 535–557.

Xu, B.; Lu, C.; Guo, Y.; and Wang, J. 2020. Discriminative multi-modality speech recognition. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 14433–14442.

Yu, J.; Zhang, S.-X.; Wu, J.; Ghorbani, S.; Wu, B.; Kang, S.; Liu, S.; Liu, X.; Meng, H.; and Yu, D. 2020. Audio-visual recognition of overlapped speech for the lrs2 dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6984–6988. IEEE.

Zhao, J.; Li, R.; and Jin, Q. 2021. Missing modality imagination network for emotion recognition with uncertain missing modalities. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2608–2618.

Zhou, P.; Yang, W.; Chen, W.; Wang, Y.; and Jia, J. 2019. Modality attention for end-to-end audio-visual speech recognition. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6565–6569. IEEE.