Lip2Vec: Efficient and Robust Visual Speech Recognition via Latent-to-Latent Visual to Audio Representation Mapping

Yasser Abdelaziz Dahou Djilali[1,2]   Sanath Narayan[1]   Haithem Boussaid[1]

Ebtessam Almazrouei[1]   Merouane Debbah[1]

[1]Technology Innovation Institute, UAE   [2]Dublin City University, Ireland

Abstract

Visual Speech Recognition (VSR) differs from the common perception tasks as it requires deeper reasoning over the video sequence, even by human experts. Despite the recent advances in VSR, current approaches rely on labeled data to fully train or finetune their models predicting the target speech. This hinders their ability to generalize well beyond the training set and leads to performance degeneration under out-of-distribution challenging scenarios. Unlike previous works that involve auxiliary losses or complex training procedures and architectures, we propose a simple approach, named Lip2Vec that is based on learning a prior model. Given a robust visual speech encoder, this network maps the encoded latent representations of the lip sequence to their corresponding latents from the audio pair, which are sufficiently invariant for effective text decoding. The generated audio representation is then decoded to text using an off-the-shelf Audio Speech Recognition (ASR) model. The proposed model compares favorably with fully-supervised learning methods on the LRS3 dataset achieving 26 WER. Unlike SoTA approaches, our model keeps a reasonable performance on the VoxCeleb2-en test set. We believe that reprogramming the VSR as an ASR task narrows the performance gap between the two and paves the way for more flexible formulations of lip reading.

1. Introduction

The process of inferring visual cues from a speaker's facial expressions and lip movements to interpret speech in a silent setting is refereed to as lip-reading or visual speech recognition (VSR). VSR is mostly useful in environments where the speech is unclear or difficult to hear due to some

---

Code: https://github.com/YasserdahouML/Lip2Vec
Correspondence: yasser.djilali@tii.ae

confounding factors [8, 7]. Hearing and speech-impaired individuals also greatly benefit from VSR [57]. Albeit the small variations around the mouth area, the space of spoken words can be large due to the phonemes composition mechanism. This makes the task highly ambiguous as several phonemes incur similar visual characteristics. Moreover, VSR needs to be robust to variations w.r.t. multiple speakers, head pose movements, non-verbal facial expressions and imaging conditions. Furthermore, lip-reading requires the integration of visual features and contextual information (*i.e.*, topic, key words search, environment and place, *etc.*) [53, 33, 6]. Over the last few years, computational methods for VSR has seen a surge with the recent proposed datasets, and can be grouped into (*i*) word-level prediction that classifies a silent video segment into a pre-defined vocabulary of words; (*ii*) continuous visual speech recognition, which predicts sentences for varying length video sequences.

Most existing VSR approaches employ a common pipeline, where lip sequences are spatially encoded using a convolution-based backbone and passed to a contextual encoder (*i.e.*, transformer [59] or conformer [19]) to model temporal dependencies. Finally, auto-regressive transformer decoder cross-attends to these representations for predicting the text. Previous works focused on enhancing the video representations for better decoding, while early approaches pretrained the backbone on word-level LRW dataset [14] for better convergence on continuous VSR [1, 28]. In contrast, [30, 3] exploit audio information as an extra supervision for an auxiliary task. Recently, cross-modal self-supervised pretraining has been a dominant paradigm for a smoother supervised finetuning afterwards [50, 51, 20].

Alternatively, the audio latent space exhibits the properties of local smoothness between input and its representation, is temporally coherent over a sequence of observations, has simple dependencies among its factors and is sparsely activated for a specific input, leading to robust and performing models [5, 4, 42, 45]. Whereas the lip sequence

is more ambiguous, with complex dependencies over the sequences as the movements are only a partial observation of a larger system that includes tongue, and other facial muscles[18]. Thus, this highlights a fundamental question about supervised learning on lip-reading data that is likely to result in local generalization, while lacking robustness on out-of-distribution data. In this work, we study these questions, uncovering key representational analogies between audio and lip sequences, the ways in which these analogies can act as a robust support for downstream task transfer, allowing for reprogramming the VSR using off-the-shelf ASR models. Specifically, our contributions are:

- We propose Lip2Vec framework that simulates VSR as an ASR task by learning a prior network that maps lip sequence features to audio-like representations, which can then be decoded to text using an ASR model.

- Through extensive evaluation, we show that learning the prior network can be exploited for decoding text. Furthermore, it performs on par with fully-supervised methods on the LRS3 [2] test set and generalizes better on the VoxCeleb2-en [13] test set.

- Our approach addresses the generalization and robustness challenges encountered by VSR models. The design explicitly bridges the gap between the VSR and ASR performances, that is proportional to the quality of the learned prior network.

- Our approach benefits from CTC-only decoding of ASR models and is 4-10$\times$ faster compared to standard VSR approaches, which decode text auto-regressively.

2. Related Works

Here, we briefly discuss the works related to the task of visual speech recognition.

2.1. Visual Speech Recognition

Sentence-level VSR, also referred as continuous visual speech recognition is challenging due to unconstrained large corpus and complex dependencies across the sequence length with regards to the text target. Whilst we briefly overview the recent sentence-level VSR efforts, we refer to [48, 60, 17] for extensive reviews. Learning from scratch on VSR datasets [2, 1] raises serious optimization issues. This difficulty emerges as the decoder cross-attention is under-optimized in early training, resulting in noisy contextual information for the queries.

Several hypotheses have been proposed to account for this. The work of [29] proposed a curriculum learning approach, where shorter sequences are initially used for training followed by progressively adding longer ones. Differently, VTP [39] proposed sub-words learning scheme using

frame-word boundaries to crop out training samples for a better convergence. These training strategies are computationally demanding and hard to scale to larger datasets. The recent works of [30, 3] proposed to exploit the audio latent representations as part of a auxiliary task, where the network is optimized to predict pretrained ASR representations along with the target text, making the optimization more stable as it provides extra supervision. Intuitively, if the transformer encoder is able to match the audio features statistics, it has to adjust the attention weights for a better decoding. Another line of research leverages pretraining on larger datasets in a self-supervised way (SSL), then finetuning on labeled VSR data using video-text pairs [50, 51, 20, 61, 28]. AV-HuBERT [50] fused the masked audio-visual representations to predict the cluster assignments created from the audio features, thus, distilling knowledge from the audio stream features to model visual inputs. VATLM [61] attempted unifying both modalities using a one tower design, where a single network is optimized to construct a common representation space for video, audio and text. This is achieved by setting a unified tokenizer for all modalities, and then performing the masked prediction task over the unified tokens. The works of [49, 31] designed cross-modal self-supervised learning frameworks, by adopting contrastive learning [22] to learn discriminative visual representations that appear to improve VSR performance and generalization. Recently, RAVen [20] designed an asymmetric SSL framework to induce a one-way knowledge distillation, where the audio networks predict both audio and video representations, whereas the visual network is restricted to predict the audio features only. This forces the audio network to serve as a strong teacher, as it would adjust to both modalities at the same time. While the work of LiRA [28] sets a pretext task to align pretrained video backbone features to the PASE+ [44] audio features, its conformer encoder and transformer decoder are trained using video-text pairs during the finetuning stage. This is fundamentally different from our Lip2Vec approach, where the video encoder and the prior network never explicitly access the text labels during training (although the ASR's latents that can be seen as proxy to the labels) and decoding is performed using Wav2vec 2.0, solely designed for ASR.

In this work, we argue that, despite the remarkable results of SSL pretraining, its expressive power can be further exploited differently. One design choice is to freeze the learned representation for the downstream task of VSR. Unlike the classification setting, the common practice of linear probing [11] is not effective on the VSR datasets [28]. The contributions of this paper attempt to address this question.

2.2. Latent-to-Latent Models

Over the last few years, latent-to-latent approaches have attracted much attention especially in the cross-modal gen-

eration literature. The high-level idea aims to match representations from two manifolds unified by a unique generating process, where correspondences are recovered and knowledge from one domain is transferred to another. In fact, DALL-E[43] trained a prior network on large scale datasets to map text to image tokens so as to perform text-guided image generation using VQ-VAEs [58], while in the work of [62], latent-to-latent network is employed to map dense visual features to discrete music representation. The work of [24] manipulates the GAN's latent space by steering the representations to change facial attributes. Adversarial reprogramming [34] was taken out of the realm of adversarial attacks to repurpose an image classification to perform sequence classification tasks [34]. In this work, we take a step forward and extend latent-to-latent techniques to VSR task, which is more fine-grained and requires better temporal modeling.

3. Method

As mentioned earlier, audio encoders of ASR models learn to transform the audio inputs to well-structured latent representations that are sufficiently robust for the task of text decoding. Our approach takes advantage of these audio representations by utilizing them as targets for training a differentiable parametric function $f_\theta : \mathbf{z_v} \mapsto z_{asr}$, with parameters θ (e.g., a neural network). Such a prior network transforms video latent representations computed by a video encoder to synthetic audio representations, which are then input to the corresponding ASR decoder for predicting the text. Our prior network is optimized to model the joint distribution over the video and audio representations by maximizing the cosine similarity between the respective representations of the pairs.

3.1. Preliminaries

We call for a function $f_\omega : \mathbf{V}^{T \times W \times H} \mapsto z_\mathbf{v}$. This function is trained in a self-supervised way such that it encodes the lip sequences by explicitly capturing the characteristics of the lip motion (i.e., temporal smoothness, invariances of small and local changes in the lip sequences, etc.), while still being unconditioned by the text labels. For the audio modality, the goal is to learn a model $f_\gamma : \mathbf{A}^{T \times S} \mapsto y$, which maps the input audio signal to the corresponding text labels y.

Video encoder: We adopt the self-supervised learned model from AV-HuBERT [50] as our video encoder. It comprises a modified ResNet [21] as frontend, followed by a transformer encoder. The 2D frontend of ResNet is replaced with a 3D convolutional layer [38]. The AV-HuBERT model is pretrained to solve the masked-prediction task, where given the masked audio and video representations output by the ResNet, the goal is to predict the clusters assigned using acoustic features (e.g., MFCC). This is itera-
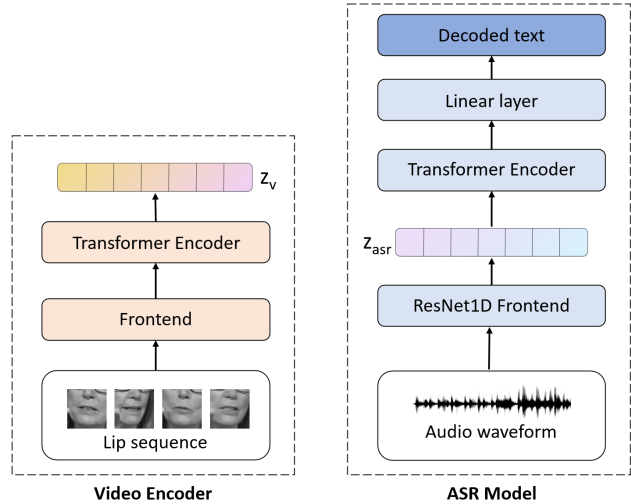

Figure 1. **On the left:** The video encoder takes a sequence of frames as input and computes the corresponding video representation $z_\mathbf{v}$. **On the right:** The frontend of the ASR model takes an audio input and obtains the audio representation $z_\mathbf{asr}$, which is the passed through a transformer encoder and linear layer for obtaining the text output.

tively refined using k-means clustering of features learned by the audio-visual encoder. Consequently, the encoder learns to better encode the characteristics of a video sequence. Given a video sequence in $\mathbb{R}^{T \times W \times H}$, the video encoder $f_\omega(\cdot)$ maps it to $z_\mathbf{v} \in \mathbb{R}^{T \times D}$. Figure 1 (left) shows the video encoder architecture for extracting $\mathbf{z_v}$ from a video sequence.

ASR model: While our framework can host any off-the-shelf ASR model, we leverage Wav2Vec2.0 [5] for its simplicity and generalization capacity. Its contrastive pretraining maximizes the mutual information between a set of anchors from contextualized representations output by the transformer encoder, and their positive pair samples from quantized representations of the ResNet features, while pushing away the set of negatives. Such a pretraining on 53k hours of unlabeled data promotes better temporal modeling and achieves a low WER of 4.8 on Librispeech [36] even when finetuning on just ten minutes of labeled data. The ASR model $f_\gamma(\cdot)$ maps an acoustic signal to audio representations z_{asr} using a feature extractor and projector. The z_{asr} is then contextualized by the transformer encoder and mapped to a vocabulary of 32 characters using a linear layer, making it faster compared to auto-regressive decoding techniques [55]. Figure 1 (right) shows the pipeline for decoding the text from an audio input.

3.2. Learning the Prior Network

We freeze the encoders ($f_\gamma(\cdot)$ and $f_\omega(\cdot)$) and learn the prior distribution over video and audio latents by maximiz-
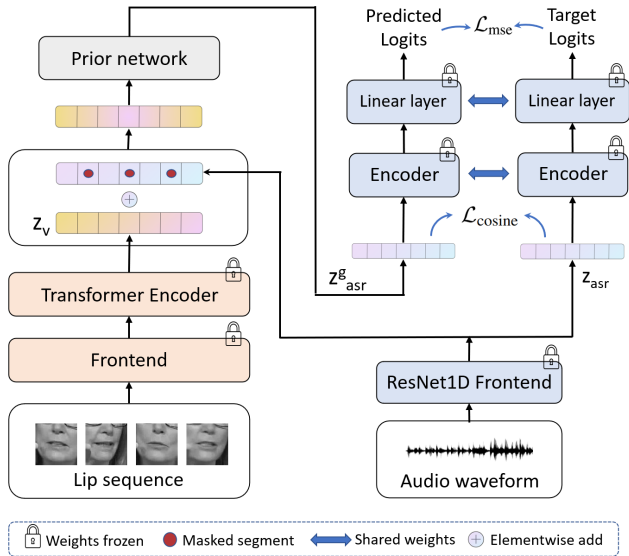
Figure 2. **Training pipeline of our Lip2Vec framework.** The video representations $\mathbf{z_v}$ are summed with the masked audio representations $\mathcal{M}(\mathbf{z_{asr}})$ and input to the prior network. The prior network generates corresponding synthetic audio representations $\mathbf{z_{asr}^g}$, which are compared with the original $\mathbf{z_{asr}}$ through a cosine similarity loss ($\mathcal{L}_{cosine}$). Furthermore, the representations $\mathbf{z_{asr}^g}$ and $\mathbf{z_{asr}}$ are passed independently through the transformer encoder and linear layer of the ASR model to output the predicted and target logits, respectively and aligned through an MSE loss ($\mathcal{L}_{mse}$). Note that the video encoder and the ASR model parameters are kept frozen throughout the training.

ing the cosine similarity between $\mathbf{z_v}$ and $\mathbf{z_{asr}}$ w.r.t. $\mathbf{p_\theta}$ [43]. To this end, we instantiate the prior network $\mathbf{f_\theta}(\cdot)$ as a standard transformer encoder [59].

Given a video-audio pair as inputs, we employ $\mathbf{f_\omega}(\cdot)$ to encode the lip sequence, whereas the audio signal is encoded with $\mathbf{f_\gamma}(\cdot)$ up to the ResNet level. The video representations $\mathbf{z_v}$ are summed with their corresponding masked audio features $\mathcal{M}(\mathbf{z_{asr}})$, where $\mathcal{M}(\cdot)$ denotes the time-masking operation. The resulting combined representation is modeled as a single data source to generate the synthetic audio representations $\mathbf{z_{asr}^g}$. The prior network is an encoder-only transformer model that exploits the expressive power of the self-attention to perform the manifold mapping from video to audio. Moreover, the task is to model the joint distribution over video and audio representations by associating the recurring patterns, compare their dependencies, and infer analogies on how the lip movements can be synthesized as an audio signal. Finally, the prior network $\mathbf{f_\theta}(\cdot)$ is optimized to predict the unmasked audio representations.

Avoiding collapse: Albeit representing the same target speech, the audio and video manifolds are likely disjoint and may not transport easily. In the process of maximizing the similarity between the respective representations, the task is to construct an input stream that achieves the op-
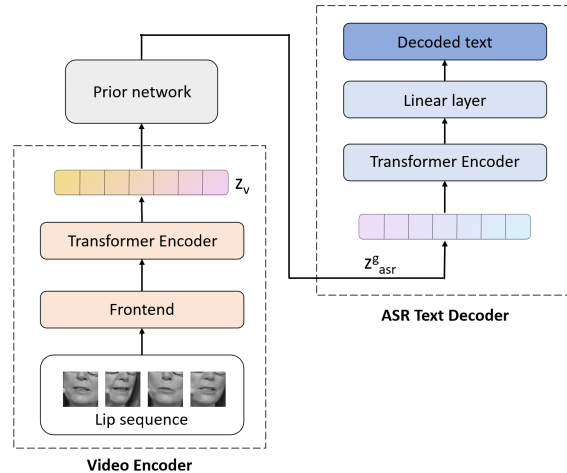

Figure 3. **Decoding text from video during inference.** The video representations $\mathbf{z_v}$ computed by the video encoder are input to our learned prior network, which synthesizes audio representations $\mathbf{z_{asr}^g}$. These representations are then passed through the encoder and linear layer of the ASR model to predict the text. Note that audio representations are not used at test time.

timization sweet-spot, thereby allowing the prior network to smoothly learn the mapping between the two manifolds. Furthermore, utilizing only video representations as input leads to degraded performance due to the difficulty in optimization resulting from missing informative features. In contrast, utilizing the audio representations summed with the video representations results in the prior network relying solely on the former for the prediction, while neglecting the latter completely and results in degraded VSR performance. To alleviate these issues of collapse, we opt for a masking schedule over the audio representations, where the mask proportion is linearly increased as the training progresses and ensures the input stream is video features only during the final epochs of training. Such a progressive masking of audio representations at the input of the prior network promotes smoother optimization during the early stage of training, and pushes the transformer to slowly learn the generalizable features for the VSR task.

3.3. Training and Inference

Training: For a pair of video and audio representations $\mathbf{z_v}$ and $\mathbf{z_{asr}}$, the prior network optimizes:

$$f_\theta : \mathbf{z_{in}} \mapsto \mathbf{z_{asr}^g}, \text{where} \quad \mathbf{z_{in}} = \mathbf{z_v} + \mathcal{M}(\mathbf{z_{asr}}).$$

We define $\mathcal{M}(\cdot)$ as the masking operation with a probability p that is a function of training steps. Given the audio input, the corresponding representations $\mathbf{z_{asr}}$ and logits $\mathbf{h_{asr}} \in \mathbb{R}^{T \times C}$ (with C being the vocabulary size of the ASR model) are utilized as targets for optimization. Here, $\mathbf{z_{asr}}$ is extracted at the ResNet level, while the logits are extracted

after the final linear projection layer of the Wav2Vec2.0. The training objective is to minimize the negative cosine similarity between representations summed over the temporal dimension, while maintaining a small distance with logits. Particularly, the losses are given by

$$\mathcal{L}_{\text{cosine}} = -\sum_{i=1}^{T} \mathbf{z}_{\mathbf{asr},i}^{\top} \mathbf{z}_{\mathbf{asr},i}^{\mathbf{g}}, \quad \text{and} \tag{1}$$

$$\mathcal{L}_{\text{mse}} = \frac{1}{T} \sum_{i=1}^{T} (\mathbf{h}_{\mathbf{asr},i}^{\mathbf{g}} - \mathbf{h}_{\mathbf{asr},i})^2. \tag{2}$$

The final objective function is given by

$$\mathcal{L} = \mathcal{L}_{\text{cosine}} + \alpha \mathcal{L}_{\text{mse}}, \tag{3}$$

where α is a hyperparameter for weighting the MSE loss.
Inference: At test time, a query video is input to the video encoder to obtain the video representation $\mathbf{z_v}$. The prior network takes this $\mathbf{z_v}$ and generates a corresponding audio representation $\mathbf{z_{asr}^g}$, which is then passed to the transformer encoder and linear layer of the ASR model to obtain the predicted text $\hat{y}$. Figure 3 shows our inference pipeline for decoding the text from a video-only input. Note that audio is not utilized at inference time for decoding the text.

4. Experiments

Datasets: We train the prior network using the video-audio pairs on: LRS3 [2] and VoxCeleb2-en [13]. The LRS3 dataset comprises a total of 433 hours of training videos from pretrain and trainval sets. From the multi-lingual Vox-Celeb2 dataset, a subset of 1326 hours of videos for the English language is selected as VoxCeleb2-en, as in [50]. We evaluate the prior network using two test sets of LRS3 and VoxCeleb2-en, as detailed below:

- LRS3: a small scale test set of around 1 hour in total, consisting of 1321 sequences. We leverage the 68 facial landmarks provided by [30] to crop the utterances around the mouth area.

- VoxCeleb-En: we randomly sample 5K videos from the VoxCeleb2-en test set, with the same duration statistics as the LRS3 test set. We use Whisper medium [42] as the labeling tool to obtain the text transcripts. Moreover, for efficiency reasons, we utilize Yolo5Face [41] to obtain the landmarks instead of relying on Reti-naFace [15, 9] face detector. We found that the resulting 5 facial landmarks are sufficient for cropping the mouth regions.

Evaluation metric: As in [30], we employ the word error rate (WER) to measure the matching between the predicted text and the ground truth transcript.

Table 1. **Supervised finetuning *vs*. latent-to-latent training.** Comparison in terms of WER on LRS3 test set is shown. The same pretrained video encoder from AV-HuBERT [50] is finetuned or utilized for the prior network. For the supervised learning, AV-HuBERT is trained with either linear layer (CTC) or a decoder (CE). Our Lip2Vec consistently improves the performance across different settings with simple CTC decoding.

Encoder	Pretrain	Finetune	Supervised [50]		Ours: Lip2Vec
			CTC	CE	CTC
Base	433h	30h	55.3	51.8	49.5
		433h	49.3	44.0	42.0
	1759h	30h	47.3	46.1	40.6
		433h	43.0	34.8	34.1
Large	433h	30h	48.4	44.8	55.4
		433h	44.3	41.6	50.1
	1759h	30h	40.7	32.5	31.2
		433h	38.6	28.6	26.0

Implementation details: We adopt the implementations of AV-HuBERT [50] and Wav2Vec2.0 [5] from the official fairseq repository[1]. For the prior network, we consider two configurations: BASE with 6 transformer layers and LARGE with 12 layers. The embedding dimension/feedforward dimension/attention heads in each transformer layer are 768/3072/12 for both variants. Furthermore, we employ a fully-connected layer and a temporal convolution upsampling to match the 50 fps of the audio representations. Base and large are trained on the low and high resource settings respectively. The prior network is implemented in PyTorch [37] and trained using 4 and 8 NVidia A100 40GB GPUs for base and large models, respectively. All the models are trained for 30 epochs using the AdamW [27] optimizer. We employ a warmup of 5 epochs and a cosine learning rate scheduler with maximum lr set to 10^{-3}.
On using labeled video-text data: It is worth mentioning that the prior network weights are not fine-tuned using labeled data containing video-text pairs. Both video encoder and ASR models are kept frozen when performing the latent to latent training. The main motivation is to set a robust evaluation procedure and to prevent the prior network from adapting its parameters to represent the video as an audio, but rather to semantically match their latent spaces.

4.1. Main Results

Finetuning *vs*. latent-to-latent: Table 1 shows the performance comparison between supervised finetuning and our proposed latent-to-latent training in terms of WER score on the LRS3 test set. For both settings, identical pretrained video encoder from [50] is utilized. The supervised finetuning using AV-HuBERT [50] is performed with either a linear layer (CTC) or a decoder (CE) using labeled video-

---

[1]https://github.com/facebookresearch/fairseq/tree/main/fairseq

Table 2. **Performance comparison on LRS3 test set in low-resource setting.** In this setting, only 30h of LRS3 trainval set is utilized for finetuning after pretraining on unlabeled data from either LRS3 (433h) or LRS3+VoxCeleb2-en (1759h) data. 'Base' and 'Large' denote the size of the pretrained video encoder employed. Our Lip2Vec achieves favorable gains across different settings. Furthermore, compared to other approaches that require an auto-regressive decoder (CE), our inference speed is significantly higher due to CTC decoding. † denotes that our method does not utilize labeled video-text data during finetuning, but uses unlabeled video-audio pairs for the same.

	Method	Unlabeled AV data	Labeled Data	Decoding	VSR
Base	AV-HuBERT [50]	433h	30h	CTC	55.3
	AV-HuBERT [50]	433h	30h	CE	51.8
	RAVen [20]	433h	30h	CTC+CE	47.0
	VATLM [61]	433h	30h	CE	48.0
	Ours: Lip2Vec	433h	30h†	CTC	49.5
	AV-HuBERT [50, 51]	1759h	30h	CTC	47.3
	AV-HuBERT [50, 51]	1759h	30h	CE	46.1
	RAVen [20]	1759h	30h	CTC+CE	40.2
	VATLM [61]	1759h	30h	CE	42.6
	Ours: Lip2Vec	1759h	30h†	CTC	40.6
Large	AV-HuBERT [50]	433h	30h	CTC	48.4
	AV-HuBERT [50]	433h	30h	CE	44.8
	Ours: Lip2Vec	433h	30h†	CTC	55.4
	AV-HuBERT [50, 51]	1759h	30h	CTC	40.7
	AV-HuBERT [50, 51]	1759h	30h	CE	32.5
	AV-HuBERT [50, 51] w/ self-training	1759h	30h	CE	28.6
	RAVen [20]	1759h	30h	CTC+CE	33.1
	RAVen [20] w/ self-training	1759h	30h	CTC+CE	24.8
	VATLM [61]	1759h	30h	CE	31.6
	Ours: Lip2Vec	1759h	30h†	CTC	31.2

text pairs. In contrast, our latent-to-latent training employs unlabeled video-audio pair for training the prior network alone while the pretrained video encoder and ASR decoder (Wav2Vec2.0) are kept frozen. We observe that our latent-to-latent approach obtains consistent improvements across different settings. However, we observe that when the large video encoder is pretrained only on LRS3 (433h), the supervised finetuning achieves better performance. This is likely due to the large encoder overfitting to the pretraining data while being less generalizable and being prone to change at the finetuning stage to fit the labeled video-text pairs. Since the latent-to-latent procedure does not involve training the video encoder, our approach suffers when the pretrained video encoder is not generalizable. Such an issue does not arise for the base video encoder or when pretraining is performed on LRS3+VoxCeleb2-en (1759h), which helps in obtaining robust video representations that are better suited for latent-to-latent learning. It is also worth mentioning that Wav2Vec2.0 achieves 6.2 WER on the LRS3 test set. Furthermore, when using the large encoder pretrained on LRS3+VoxCeleb-En (1759h) and finetuning on 433h, our approach achieves the best WER score of 26.0, with gains of 12.6 and 2.6 over the supervised CTC and CE finetuning, respectively. These results show the efficacy of our latent-to-latent learning approach for the VSR task.

State-of-the-art comparison: Here, we compare the Lip2Vec approach to SoTA VSR approaches on the LRS3 test set. Tables 2 and 3 show the performance comparison in terms of WER for the low-resource and high-resource settings, respectively. While the low-resource setting denotes that finetuning is performed with only 30h of LRS3 trainval data, the high-resource setting indicates finetuning with 433h of LRS3. Supervised methods using varying labeled data are also reported in Table 3 for comparison. We observe that our Lip2Vec performs favorably against existing approaches across different settings. Furthermore, the approach depends on the generalizability of the pretrained video encoder representations because training the Lip2Vec does not utilize labeled video-text pairs in addition to freezing the parameters of the video encoder. This is in contrast to the supervised finetuning, which is likely to significantly vary the video encoder parameters to align for text decoding. Furthermore, from Table 3, we observe our Lip2Vec trained with large encoder with 1759h of pretraining to obtain the best results of 26.0. This results in gains of 2.6, 2.2 and 2.4 over AV-HuBERT, RAVen and VATLM, respectively, when self-training (*i.e.*, pseudo-labeling the data and additionally using them for finetuning) is not employed by these approaches.

4.2. Results on VoxCeleb2-en

In Table 4, we report the WER scores on three folds of the VoxCeleb2-en test set: the first fold is randomly selected 5k videos, the second and third are subsets of this

Table 3. **Performance comparison on LRS3 test set in high-resource setting.** 'Base' and 'Large' denote the size of the self-supervised video encoder employed. Performance of supervised approaches are also reported. † denotes that our Lip2Vec does not utilize labeled video-text data during finetuning, but uses unlabeled video-audio pairs for the same. Our approach achieves favorable gains across different settings with significantly higher inference speed due to CTC decoding, compared to other approaches that require an auto-regressive decoder (CE). Particularly, when using the large encoder pretrained on 1759h, our approach achieves the best score of 26.0 and is on par with 25.9 of [47] that utilizes 90k hours of labeled data in a supervised setting.

	Method	Unlabeled AV data	Labeled Data	Decoding	VSR
Supervised	Afouras *et al.* [1]	-	1519h	CE	58.9
	Shillingford *et al.* [52]	-	3886h	CTC	55.1
	Ma *et al.* [30]	-	813h	CTC+CE	34.7
	Makino *et al.* [32]	-	31000h	Transducer	33.6
	Prajwal *et al.* [40]	-	2676h	CE	30.7
	Serdyuk *et al.* [47]	-	90000h	Transducer	25.9
	Chang *et al.* [10]	-	100000h	Transducer	12.8
Self-Supervised Base	AV-HuBERT [50]	433h	433h	CTC	49.3
	AV-HuBERT [50]	433h	433h	CE	44.0
	RAVen [20]	433h	433h	CTC+CE	39.1
	Ours: Lip2Vec	433h	433h†	CTC	42.0
	AV-HuBERT [50, 51]	1759h	433h	CTC	43.0
	AV-HuBERT [50, 51]	1759h	433h	CE	34.8
	RAVen [20]	1759h	433h	CTC+CE	33.1
	VATLM [61]	1759h	433h	CE	34.2
	Ours: Lip2Vec	1759h	433h†	CTC	34.1
Self-Supervised Large	AV-HuBERT [50]	433h	433h	CTC	44.3
	AV-HuBERT [50]	433h	433h	CE	41.6
	Ours: Lip2Vec	433h	433h†	CTC	50.1
	AV-HuBERT [50, 51]	1759h	433h	CTC	38.6
	AV-HuBERT [50, 51]	1759h	433h	CE	28.6
	AV-HuBERT [50, 51] w/ self-training	1759h	433h	CE	26.9
	RAVen [20]	1759h	433h	CTC+CE	27.8
	RAVen [20] w/ self-training	1759h	433h	CTC+CE	24.4
	VATLM [61]	1759h	433h	CE	28.4
	VATLM [61] w/ self-training	1759h	433h	CE	26.2
	Ours: Lip2Vec	1759h	433h†	CTC	26.0

5k, where Wav2Vec2.0 obtains WER scores less than 30 and 20, respectively. We follow this procedure to reduce the bias and aim for a fair comparison as the labels are obtained with another ASR model (*i.e.*, Whisper [42]). First, we observe that SoTA approaches fail to generalize under this benchmark. Both the model from [30] and the VTP [40] scores are around 70 WER. It is worth mentioning that VTP was trained on a 2.7k hours of video. As expected, the Wav2Vec2.0 gets relatively reasonable results (10 to 25 WER). Interestingly enough, our Lip2Vec approach tracks the Wav2Vec2.0 scores with an upper bound proportional to the quality of the prior network. When only trained on 30h of LRS3, our Lip2Vec deviates from Wav2Vec2.0 by an average WER score of 23 across the three folds, thereby showing the generalization capability of our approach. Note that self-trained variants of RAVen and AV-HuBERT are not considered for OOD generalization since they are trained on pseudo-labelled VoxCeleb2-en train set. It can be seen that our Lip2Vec also achieves consistent gains in terms of WER, in comparison to RAVen and AV-HuBERT across different folds, demonstrating better generalization to unseen

or novel speakers. This trend also holds for 433h finetuning. **Training on VoxCeleb2-en.** Here, we further investigate the impact of training with VoxCeleb2-en [13] data. In practice, this scenario might arise if one has access to a dataset comprising unlabelled lip sequences. Our Lip2Vec framework is a suitable fit for this setting as it does not require labeled video-text pairs to learn the prior network. We take advantage of this property and train the model variants on a low-resource (30h) setting of the VoxCeleb2-en dataset. Table 5 shows the WER scores following various training sets on both LRS3 and VoxCeleb2-en test sets. As expected, combining 30h from VoxCeleb2-en with the LRS3 low-resources setting improves the performance on the LRS3 test set as compared to training on 30h of LRS3 only (30.5 *vs.* 31.2 for large and 40.1 *vs.* 40.6 for base). It is worth mentioning that training on 30h of VoxCeleb2-en achieves similar WER compared to using LRS3 low-resource dataset. This highlights the robustness of the proposed Lip2Vec approach and its considerable advantages over the supervised finetuning.

The Whisper [42] pseudo-labelled VoxCeleb2-en test set

Table 4. **Out-of-distribution generalization on VoxCeleb2-en test set in terms of WER.** The folds are selected using Wav2Vec2.0 scores. Base and Large models are denoted by * and †. All models are fine-tuned on the 30h low-resource setting of LRS3. The performance on LRS3 test set is also shown for ease of reference. The last column reports the average computational load in seconds for decoding 100 frames video (4 seconds) on a single Nvidia A100.

Model	Unlabeled	Video folds			Runtime (s)
		01	02	03	
Wav2Vec2.0 [5]	–	25.1	15.0	10.1	0.05
Ma et al. [30]	–	69.4	64.1	61.3	3.91
VTP [39]	–	71.7	69.1	66.9	0.97
RAVen*	433h	78.2	74.3	72.3	1.40
AV-HuBERT*	433h	79.1	76.3	73.2	0.24
Ours: Lip2Vec*	433h	71.2	65.7	57.3	0.07
RAVen*	1759h	61.2	56.1	53.4	1.40
AV-HuBERT*	1759h	71.1	66.2	62.9	0.24
Ours: Lip2Vec*	1759h	58.1	53.4	49.2	0.07
AV-HuBERT†	433h	78.1	75.6	71.4	0.35
Ours: Lip2Vec†	433h	77.3	70.4	61.8	0.09
RAVen†	1759h	**47.5**	42.8	39.4	1.63
AV-HuBERT†	1759h	49.7	44.6	41.2	0.35
Ours: Lip2Vec†	1759h	48.1	**39.4**	**33.2**	0.09

Table 5. **Training Lip2Vec on VoxCeleb2-en.** Comparing the effects of varying the training set on the WER scores on both LRS3 and VoxCeleb2-en test sets. We randomly select 30h from VoxCeleb2-en, and use it in different settings. We observe that our prior network trained on VoxCeleb2-en data only can still generalize to LRS3.

Encoder	Training set	Test set	
		LRS3	VoxCeleb2-en
Base	LRS3-30h	40.6	58.2
	LRS3+VoxCeleb2-en-60h	40.1	54.6
	VoxCeleb2-en-30h	41.2	57.3
Large	LRS3-30h	31.2	39.4
	LRS3+VoxCeleb2-en-60h	30.4	33.1
	VoxCeleb2-en-30h	30.5	33.8

turns out to be more challenging due to the high variety of speakers, vocabulary, *etc*. The Lip2Vec variants scores on this benchmark are still far from their performance on the LRS3 test set. Future works will focus on the generalization aspects on both LRS3 and VoxCeleb2-en test sets.

Inference speed: As model efficiency is a key factor for real-world VSR applications, Table 4 shows a GPU runtime comparison (processing time per 100 frames) of the different approaches on sample videos. Compared with other approaches, our model exhibits a remarkable improvement, being nearly $4\times$ and $18\times$ faster than AV-HuBERT and RAVen, respectively. This is explained by the fact that CTC decoding does not require any computationally expensive auto-regressive procedures, beam search, *etc*.

Table 6. **Impact of varying α.** WER comparison on LRS3 test set when varying the weight α for $\mathcal{L}_{mse}$. When α is increased beyond 0.05, the $\mathcal{L}_{mse}$ dominates over $\mathcal{L}_{cosine}$, resulting in $\mathbf{z}_{asr}^{g}$ diverging from the target $\mathbf{z}_{asr}$. While performance is slightly degraded without MSE loss for a fixed training budget, longer training (denoted by †) can reach similar optimal performance as with $\alpha=0.01$, validating that $\mathcal{L}_{mse}$ improves the convergence.

α	0.0	0.0†	0.01	0.2	0.5
WER	34.6	31.4	31.2	52.1	91.3

4.3. Ablation Study

Here, we evaluate the performance of our Lip2Vec when ablating the key components: varying the hyperparameter α for $\mathcal{L}_{mse}$ loss and the masking function $\mathcal{M}(\cdot)$. For this study, evaluation is conducted on LRS3 test set and the large video encoder (self-supervisedly pretrained on 1759 hours of LRS3 and VoxCeleb2-en) is employed.

Impact of varying α: Table 6 presents the performance of our framework when the hyperparameter weight α (Eq. 3) is varied. We observe that higher values of α degrade the performance since the similarity between predicted representations $\mathbf{z}_{asr}^{g}$ and target $\mathbf{z}_{asr}$ diverges due to the gradients from $\mathcal{L}_{mse}$ dominating over $\mathcal{L}_{cosine}$. Furthermore, when training for a fixed number of epochs, $\alpha=0$ achieves a WER score of 34.6 compared to the best results of 31.2 when $\alpha=0.01$. We also observe that training longer without MSE loss (denoted by † in Table 6) can achieve a WER score of 31.4, indicating that $\mathcal{L}_{mse}$ aids in faster training convergence.

Impact of different masking strategies: From Table 7, we observe that not masking the audio representations $\mathbf{z}_{asr}$ results in the prior network learning a shortcut from its input to output while ignoring the video representations and thereby performing poorly at test time when no audio is available. Similarly, maintaining the same masking probability p throughout the training results in the prior network expecting the masked audio to be present at test time as well for generating synthetic $\mathbf{z}_{asr}^{g}$ accurately. In contrast, initializing p to a low value of 0.3 and gradually increasing it to 1.0 (simulating no $\mathbf{z}_{asr}$ input) by the end of training enables the prior network to learn better representations $\mathbf{z}_{asr}^{g}$ from input $\mathbf{z}_{v}$. Consequently, our progressive masking achieves a WER score of 31.2, thereby validating its efficacy for training. Additional results are provided in the supplementary.

5. Discussion and Future Work

Supervised *vs*. self-supervised video encoder: As discussed in the experiments, we employed a self-supervised video encoder from AV-HuBERT [50] for training the prior network. In contrast, here, we evaluate the efficacy of a supervised video encoder in the Lip2Vec framework by utilizing the encoder from [30]. For this experiment, we train the prior network following the low resources setting.

Table 7. **Masking strategy.** WER comparison on LRS3 test with different masking strategies $\mathcal{M}(\cdot)$ for the audio representations $\mathbf{z_{asr}}$. No masking performs poorly since the prior network discards the $\mathbf{z_v}$ input. Similarly, masking with same probability (p) throughout the training shows only marginal improvement over no masking. The best results are obtained with the proposed progressive masking, where p is gradually increased during the training.

Masking	WER
No Masking	75.2
50% Masking	66.9
80% Masking	61.5
100% Masking	65.3
Progressive Masking	31.2

This achieves WER scores of 45.0 and 76 on LRS3 and VoxCeleb2-en test sets, respectively. This is likely due to the explicit text supervision, which trains the video encoder to output representations aligned with the text decoding task rather than trained towards better representing the lip movements. This shows that self-supervised encoders are highly suited for learning the latent-to-latent mappings and are better generalizable.

This is also supported by the findings in neuroscience research, which demonstrate that silent lip-reading signal first synthesizes a coarse-grained auditory speech representation in early auditory cortices. Then, the right angular gyrus excites the temporal visual speech area, extracts and possibly predicts the slower features of lip movements. Finally, the auditory cortices are fed with this signal to decode as an audio signal [35, 23]. Consequently, our approach opens a new line of research for exploring the subtle definition of visual speech recognition embedded in the human brain [8].

VSR as interpolation *vs.* extrapolation: Most perception problems are interpolative in their nature [12] and satisfy the manifold hypothesis [16]. These tasks are intuitive for humans, and are usually solved in the early layers of the visual cortex in a matter of milliseconds (*i.e.*, classification, recognition, *etc*.) [26, 56]. For such problems, deep learning is a perfect fit with its ability to perform non-linear interpolation in a complex high-dimensional manifold, enabling arbitrary complex behavior [54, 12]. However, lip-reading experts allude to a high-level step-wise and iterative reasoning to solve the task. This likely suggests that VSR has some higher level of extrapolation as compared to the common perception tasks. Thus, we hypothesize that learning the manifold transfer without exposing the lip sequences explicitly to the text labels would induce some interpolation, thereby allowing for a better generalization. We believe improving the prior network by leveraging better training procedures and architectures such as [46] would be an important future research direction for tightening up the bound with the ASR performance.

Impact of fine-tuning on learned self-supervised encoder: From our experiments above, we observed that supervised video encoders and models pretrained on LRS3 only, are not suitable for latent-to-latent learning. A potential future direction includes studying the effect of text labels on self-supervised learned weights using measures such as Center Kernel Alignment (CKA) [25] for obtaining deeper insights into the VSR task.

6. Conclusion

We introduced Lip2Vec, a simple VSR framework that makes the most of ASR and VSR models by combining knowledge acquired by an off-the-shelf VSR encoder and an ASR model. The approach exploits the latent space structure to perform inter modality mapping, and learns how to transfer the visual representations to a suitable decoding space. Results on various benchmarks demonstrated the competitiveness and robustness of the approach. We believe this is an important step towards better VSR modeling using latent-to-latent methods. In summary, the results and discussions presented in the paper along with those in the appendices demonstrate the efficacy of our Lip2Vec approach for the task of visual speech recognition.

References

[1] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence*, 44(12):8717–8727, 2018. 1, 2, 7

[2] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. Lrs3-ted: a large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496*, 2018. 2, 5

[3] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. ASR is all you need: cross-modal distillation for lip reading. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2020-May:2143–2147, 11 2019. 1, 2

[4] Alexei Baevski, Steffen Schneider, and Michael Auli. vq-wav2vec: Self-supervised learning of discrete speech representations. *arXiv preprint arXiv:1910.05453*, 2019. 1

[5] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020. 1, 3, 5, 8

[6] Lynne E Bernstein, Nicole Jordan, Edward T Auer, and Silvio P Eberhardt. Lipreading: A review of its continuing importance for speech recognition with an acquired hearing loss and possibilities for effective training. *American Journal of Audiology*, 31(2):453–469, 2022. 1

[7] Lynne E Bernstein, Paula E Tucker, and Marilyn E Demorest. Speech perception without hearing. *Perception & Psychophysics*, 62(2):233–252, 2000. 1

[8] Mathieu Bourguignon, Martijn Baart, Efthymia C Kapnoula, and Nicola Molinaro. Lip-reading enables the brain to synthesize auditory features of unknown silent speech. *Journal of Neuroscience*, 40(5):1053–1065, 2020. 1, 9

[9] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE international conference on computer vision*, pages 1021–1030, 2017. 5

[10] Oscar Chang, Hank Liao, Dmitriy Serdyuk, Ankit Shah, and Olivier Siohan. Conformers are all you need for visual speech recogntion. *arXiv preprint arXiv:2302.10915*, 2023. 7

[11] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2

[12] François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019. 9

[13] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018. 2, 5, 7

[14] Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13*, pages 87–103. Springer, 2017. 1

[15] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5203–5212, 2020. 5

[16] Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016. 9

[17] Souheil Fenghour, Daqing Chen, Kun Guo, Bo Li, and Perry Xiao. Deep Learning-Based Automated Lip-Reading: A Survey. *IEEE Access*, 9:121184–121205, 2021. 2

[18] W Tecumseh Fitch. The evolution of speech: a comparative review. *Trends in cognitive sciences*, 4(7):258–267, 2000. 2

[19] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*, 2020. 1

[20] Alexandros Haliassos, Pingchuan Ma, Rodrigo Mira, Stavros Petridis, and Maja Pantic. Jointly learning visual and auditory speech representations from raw data. *arXiv preprint arXiv:2212.06246*, 2022. 1, 2, 6, 7

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3

[22] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2, 2020. 2

[23] Anne Keitel, Joachim Gross, and Christoph Kayser. Shared and modality-specific brain regions that mediate auditory and visual word comprehension. *Elife*, 9:e56972, 2020. 9

[24] Siavash Khodadadeh, Shabnam Ghadar, Saeid Motiian, Wei An Lin, Ladislau Boloni, and Ratheesh Kalarot. Latent to Latent: A Learned Mapper for Identity Preserving Editing of Multiple Face Attributes in StyleGAN-generated Images. *Proceedings - 2022 IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022*, pages 3677–3685, 2022. 3

[25] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pages 3519–3529. PMLR, 2019. 9

[26] Hirotaka Kosaka, Masao Omori, Tetsuya Iidaka, Tetsuhito Murata, T Shimoyama, Tomohisa Okada, Norihiro Sadato, Yoshiharu Yonekura, and Yuji Wada. Neural substrates participating in acquisition of facial familiarity: an fmri study. *Neuroimage*, 20(3):1734–1742, 2003. 9

[27] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[28] Pingchuan Ma, Rodrigo Mira, Stavros Petridis, Björn W Schuller, and Maja Pantic. Lira: Learning visual speech representations from audio through self-supervision. *arxiv preprint arXiv:2106.09171*, 2021. 1, 2

[29] Pingchuan Ma, Stavros Petridis, and Maja Pantic. End-to-end audio-visual speech recognition with conformers. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7613–7617. IEEE, 2021. 2

[30] Pingchuan Ma, Stavros Petridis, and Maja Pantic. Visual speech recognition for multiple languages in the wild. *Nature Machine Intelligence*, pages 1–10, 2022. 1, 2, 5, 7, 8

[31] Shuang Ma, Zhaoyang Zeng, Daniel McDuff, and Yale Song. Contrastive Learning of Global-Local Video Representations. *Advances in Neural Information Processing Systems*, 9:7025–7040, 4 2021. 2

[32] Takaki Makino, Hank Liao, Yannis Assael, Brendan Shillingford, Basilio Garcia, Otavio Braga, and Olivier Siohan. Recurrent neural network transducer for audio-visual speech recognition. In *2019 IEEE automatic speech recognition and understanding workshop (ASRU)*, pages 905–912. IEEE, 2019. 7

[33] Dominic W Massaro and Michael M Cohen. Perception of synthesized audible and visible speech. *Psychological Science*, 1(1):55–63, 1990. 1

[34] Paarth Neekhara, Shehzeen Hussain, Jinglong Du, Shlomo Dubnov, Farinaz Koushanfar, and Julian McAuley. Cross-modal adversarial reprogramming. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2427–2435, 2022. 3

[35] Collins Opoku-Baah, Adriana M Schoenhaut, Sarah G Vassall, David A Tovar, Ramnarayan Ramachandran, and Mark T Wallace. Visual influences on auditory behavioral, neural, and perceptual processes: a review. *Journal of the Association for Research in Otolaryngology*, 22(4):365–386, 2021. 9

[36] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE, 2015. 3

[37] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *NeurIPS*, 2019. 5

[38] Stavros Petridis, Themos Stafylakis, Pingchuan Ma, Georgios Tzimiropoulos, and Maja Pantic. Audio-visual speech recognition with a hybrid ctc/attention architecture. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 513–520. IEEE, 2018. 3

[39] KR Prajwal, Triantafyllos Afouras, and Andrew Zisserman. Sub-word level lip reading with visual attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5162–5172, 2022. 2, 8

[40] K. R. Prajwal, Triantafyllos Afouras, and Andrew Zisserman. Sub-word Level Lip Reading With Visual Attention. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022-June:5152–5162, 2021. 7

[41] Delong Qi, Weijun Tan, Qi Yao, and Jingfeng Liu. Yolo5face: Why reinventing a face detector. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*, pages 228–244. Springer, 2023. 5

[42] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*, 2022. 1, 5, 7

[43] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 3, 4

[44] Mirco Ravanelli, Jianyuan Zhong, Santiago Pascual, Pawel Swietojanski, Joao Monteiro, Jan Trmal, and Yoshua Bengio. Multi-task self-supervised learning for robust speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6989–6993. IEEE, 2020. 2

[45] Vincent Roger, Jérôme Farinas, and Julien Pinquier. Deep neural networks for automatic speech processing: a survey from large corpora to limited data. *EURASIP Journal on Audio, Speech, and Music Processing*, 2022(1):19, 2022. 1

[46] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2022-June:10674–10685, 12 2021. 9

[47] Dmitriy Serdyuk, Otavio Braga, and Olivier Siohan. Audio-visual speech recognition is worth 32x32x8 voxels. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 796–802. IEEE, 2021. 7

[48] Changchong Sheng, Gangyao Kuang, Liang Bai, Chenping Hou, Yulan Guo, Xin Xu, Matti Pietikäinen, and Li Liu. Deep Learning for Visual Speech Analysis: A Survey. 5 2022. 2

[49] Changchong Sheng, Matti Pietikäinen, Qi Tian, and Li Liu. Cross-modal Self-Supervised Learning for Lip Reading: When Contrastive Learning meets Adversarial Training. *MM 2021 - Proceedings of the 29th ACM International Conference on Multimedia*, pages 2456–2464, 10 2021. 2

[50] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. *arXiv preprint arXiv:2201.02184*, 2022. 1, 2, 3, 5, 6, 7, 8

[51] Bowen Shi, Wei-Ning Hsu, and Abdelrahman Mohamed. Robust self-supervised audio-visual speech recognition. *arXiv preprint arXiv:2201.01763*, 2022. 1, 2, 6, 7

[52] Brendan Shillingford, Yannis Assael, Matthew W. Hoffman, Thomas Paine, Cían Hughes, Utsav Prabhu, Hank Liao, Hasim Sak, Kanishka Rao, Lorrayne Bennett, Marie Mulville, Misha Denil, Ben Coppin, Ben Laurie, Andrew Senior, and Nando De Freitas. Large-Scale Visual Speech Recognition. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2019-September:4135–4139, 7 2018. 7

[53] William H Sumby and Irwin Pollack. Visual contribution to speech intelligibility in noise. *The journal of the acoustical society of america*, 26(2):212–215, 1954. 1

[54] Christopher Summerfield. *Natural General Intelligence: How understanding the brain can help us build AI*. Oxford University Press, 2022. 9

[55] Gabriel Synnaeve, Qiantong Xu, Jacob Kahn, Tatiana Likhomanenko, Edouard Grave, Vineel Pratap, Anuroop Sriram, Vitaliy Liptchinsky, and Ronan Collobert. End-to-end asr: from supervised to semi-supervised learning with modern architectures. *arXiv preprint arXiv:1911.08460*, 2019. 3

[56] Alexander Todorov. The role of the amygdala in face perception and evaluation. *Motivation and Emotion*, 36:16–26, 2012. 9

[57] Nancy Tye-Murray, Mitchell S. Sommers, and Brent Spehar. Audiovisual integration and lipreading abilities of older adults with normal and impaired hearing. *Ear and hearing*, 28(5):656–668, 9 2007. 1

[58] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 3

[59] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1, 4

[60] Ziheng Zhou, Guoying Zhao, Xiaopeng Hong, and Matti Pietikäinen. A review of recent advances in visual speech decoding. *Image and Vision Computing*, 32(9):590–605, 9 2014. 2

[61] Qiushi Zhu, Long Zhou, Ziqiang Zhang, Shujie Liu, Binxing Jiao, Jie Zhang, Lirong Dai, Daxin Jiang, Jinyu Li, and

Furu Wei. Vatlm: Visual-audio-text pre-training with unified masked prediction for speech representation learning. *arXiv preprint arXiv:2211.11275*, 2022. 2, 6, 7

[62] Ye Zhu, Kyle Olszewski, Yu Wu, Panos Achlioptas, Menglei Chai, Yan Yan, and Sergey Tulyakov. Quantized GAN for Complex Music Generation from Dance Videos. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 13697 LNCS:182–199, 2022. 3