

OpenSR: Open-Modality Speech Recognition via Maintaining Multi-Modality Alignment

Xize Cheng^{1*}, Tao Jin^{1*}, Linjun Li^{1*}, Wang Lin¹, Xinyu Duan², Zhou Zhao^{1 †}

¹Zhejiang University, ²Huawei Cloud

{chengxize, jint_zju, zhaozhou}@zju.edu.cn duanxinyu@huawei.com

Abstract

Speech Recognition builds a bridge between the multimedia streaming (audio-only, visual-only or audio-visual) and the corresponding text transcription. However, when training the specific model of new domain, it often gets stuck in the lack of new-domain utterances, especially the labeled visual utterances. To break through this restriction, we attempt to achieve zero-shot modality transfer by maintaining the multi-modality alignment in phoneme space learned with unlabeled multimedia utterances in the high resource domain during the pre-training (Shi et al., 2022), and propose a training system Open-modality Speech Recognition (OpenSR) that enables the models trained on a single modality (e.g., audio-only) applicable to more modalities (e.g., visual-only and audio-visual). Furthermore, we employ a cluster-based prompt tuning strategy to handle the domain shift for the scenarios with only common words in the new domain utterances. We demonstrate that OpenSR enables modality transfer from one to any in three different settings (zero-, few- and full-shot), and achieves highly competitive zero-shot performance compared to the existing few-shot and full-shot lip-reading methods. To the best of our knowledge, OpenSR achieves the state-of-the-art performance of word error rate in LRS2 on audio-visual speech recognition and lip-reading with 2.7% and 25.0%, respectively. The code and demo are available at <https://github.com/Exgc/OpenSR>.

1 Introduction

Speech Recognition (Afouras et al., 2018a; Ren et al., 2021; Zhao et al., 2020) (e.g., Audio-Visual Speech Recognition known as AVSR) transcribes visual and audio data into text form, building a bridge between multi-media speech (Cheng et al., 2023; Huang et al., 2023a,b; Cui et al., 2022; Aoxiong et al., 2023) and natural language (Yin et al.,

*Equal contribution.

†Corresponding author.

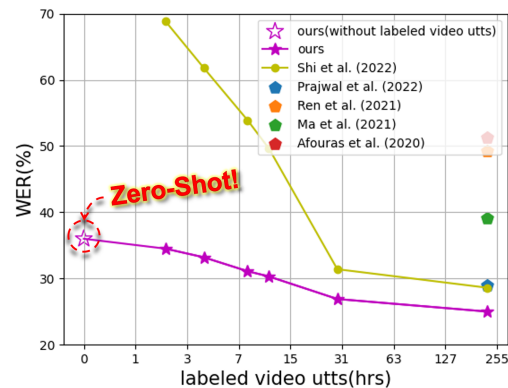


Figure 1: Comparison between OpenSR and previous methods under different sizes of labeled visual utterance. As the results highlighted by the red circle show, OpenSR with zero-shot even outperforms most previous methods with full-shot.

2022, 2021; Jin et al., 2022a,b; Jin, 2021). Among them, ASR (Automatic Speech Recognition) and lip-reading (VSR, Visual Speech Recognition) are twin tasks transcribed using only audio and only vision, respectively. Audio utterances with clear pronunciation are sufficient for ASR training, most of which can be easily collected from recordings of telephone conversations and audiobooks (Korvas et al., 2014). While current lip-reading training systems require mostly-frontal and high-resolution videos with a sufficiently high frame rate, such that motions around the lip area are clearly captured (Prajwal et al., 2022). The significant difficulty of collecting labeled visual utterances hinders the training of lip-reading models suitable for new domains or low-resource domains, resulting in the relatively low-speed development of domain-specific lip-reading models, compared with ASR models.

Since audio speech is easier to obtain, can we use audio utterance alone to train lip-reading models for target domains? Humans, once they have mastered a language, can immediately predict the lip movements of a phoneme (Meltzoff and Moore, 1977). Similarly, if we align the phoneme represen-

tations of acoustic and image fragments with each other, we can then apply the term distribution and syntax specification of the target domain learned from the audio utterance to lip-reading. Building on this novel idea, we employ the audio-visual aligned encoder, such as AV-Hubert, co-trained on a large number of multi-modality high resource domain utterances, to align the multi-modality utterances in the same phoneme space, and train a domain-specific decoder from phoneme space to text using the labeled audio-only utterances. Since the audio-visual encoder is trained to embed different modalities of the same phoneme near one another (for instance, the visual phoneme feature of /tu:/ is aligned with its audio phoneme feature), we can flexibly transfer the knowledge of the target domain (e.g., the mapping from phoneme /tu:/ to homophones `two` and `too`, and the syntax used to distinguish them) from a single modality (i.e., audio) to more modalities (i.e., visual-only and audio-visual). As shown in Figure 1, the zero-shot lip-reading performance of our method is even better than most full-shot methods.

Furthermore, we propose to enhance the discrimination of visually similar words (e.g., `pet` and `bet`, which are audio distinguishable but visually similar (Kim et al., 2022)) with more readily available utterances containing only common words. However, the uneven word distribution causes the sequence modeling to be easily corrupted by domain shift during fine-tuning (Huang et al., 2022). We propose a cluster-based prompt tuning strategy, `Cluster Prompt`, only tuning the limited parameters of the fine-grained prompt-embedding of each visual phoneme cluster, to prevent excessive impact on sequence modeling.

The main contributions are as follows:

- To the best of our knowledge, OpenSR is the first to achieve zero-shot lip-reading, which fully considers the modality transfer.
- OpenSR is the first low-resource lip-reading method that attempts to leverage common words, adopting the `Cluster Prompt` strategy to overcome the ensuing domain shift problem and improve the accuracy of lip-reading by 19.1% to 36.1%.
- OpenSR achieves modality transfer and the state-of-the-art performance in all three settings (zero-shot, few-shot and full-shot). In particular, it achieves 2.7% and 25% word error rate on LRS2 in AVSR and lip-reading.

2 Related Work

2.1 Lip Reading

The lip-reading task has attracted many researchers (Cooke et al., 2006; Afouras et al., 2018a), aiming to recognize spoken sentences according to the given video of lip movements without relying on the audio stream. With the support of a large amount of visual speech utterances (Afouras et al., 2018a,b), Assael et al. (2016) first proposes to use neural networks for lip reading. (Ma et al., 2021) and (Shi et al., 2022) adopt different pre-training strategies, attempting to obtain fine-grained lip representation by pre-training on a large amount of additional unlabeled audio-visual utterances (Chung et al., 2018). Some works (Makino et al., 2019; Serdyuk et al., 2021) adopt massive labeled visual utterances (even more than 90,000h) for training and promote the generalization of the model. However, none of these methods can train lip-reading models without large amounts of labeled visual utterances, making them unusable in low-resource domains where labeled visual utterances are scarce or unavailable. This is also the original intention of our work, for which we propose a training system that can employ only labeled audio utterances to train the lip-reading model.

2.2 Transfer Learning from Audio to Video

Although the two parallel speech modalities, audio and video, remain aligned at the temporal level, ASR is still far more accurate than lip reading, benefiting from its easy access to labeled audio utterance and fine-grained phoneme audio representation. Ma et al. (2021) hopes to take advantage of this natural temporal alignment and use audio to assist with lip-reading training. Ren et al. (2021) proposes different distillation strategies (from ASR to lip-reading) which enables lip-reading to learn complementary and discriminant clues from ASR. Shi et al. (2022) has adopted audio assisted pre-training methods, that regard audio as the auxiliary supervision for visual utterances in order to obtain fine-grained phoneme visual representations. However, in previous methods (Ren et al., 2021; Huang et al., 2021), audio only played the role of auxiliary supervision, and the lip-reading models could not be trained using non-visual utterances (i.e., audio) alone. In our work, we attempt to maintain the alignment of audio and video in phoneme space, and employ only labeled audio utterances instead of visual utterances to train the lip-reading model.

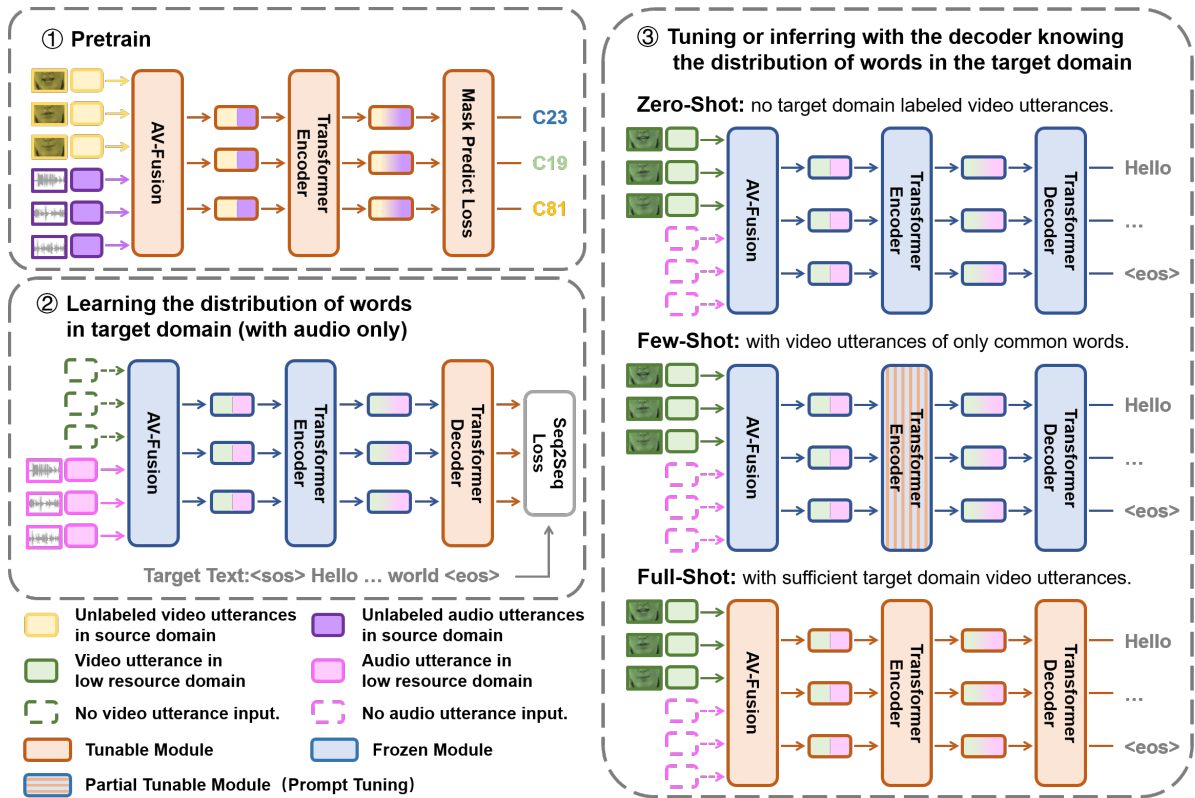


Figure 2: Illustration of OpenSR training system. In the second stage, parameters other than those in the transformer decoder are frozen to maintain the cross modality alignment state achieved in the first stage. We propose three different tuning and inference strategies in the third stage for scenarios with different scales of labeled visual utterances in the target domain.

3 Method

3.1 Audio-Visual Alignment Learning

AV-Hubert (Shi et al., 2022) is a self-supervised representation learning method for audio-visual speech, alternating between feature clustering and mask prediction, as shown in the first stage of Figure 2. During the feature clustering, the audio-visual speech is labeled as a sequence of frame-level assignments $z = \{z_t^a\}_{t=1}^T$ based on the sequence of image sequences $V = \{V_t\}_{t=1}^T$ and audio acoustic frames $A = \{A_t\}_{t=1}^T$ (MFCC or audio-visual features from the previous encoder) with a discrete latent variable model (e.g., k-means). Subsequently, with the paired data (A, V, z) , the model learns a better audio-visual representation in phoneme space $f^p = \{f_t^p\}_{t=1}^T \in \mathbb{R}^{T \times D}$, where T is the length of the sequence and D is the dimension of the embedding, by reducing the mask prediction loss, just like the mask language modeling in BERT (Devlin et al., 2018). The above two steps are repeated during training to improve the quality of audio-visual speech clustering and representation.

Furthermore, the random dropout of modalities

in the framework maps the speech features of different modalities into the same phoneme space, which not only optimizes the representation quality of the uni-modality, but also achieves cross-modality representation alignment with a large amount of source domain (i.e., high resource domain) unlabeled audio-visual utterances. OpenSR is the first attempt to take full advantage of this cross-modality alignment of speech representations.

3.2 Decoder Training with Audio Only

Now that the audio and visual representations have been mapped into the same phoneme space, we can use audio of the target domain as an alternative to video when labeled visual utterance is scarce.

As shown in the second stage in Figure 2, we adopt the AV-Fusion and the pre-trained Transformer Encoder in the first stage to obtain the features of phoneme space $f^p = \{f_t^p\}_{t=1}^T \in \mathbb{R}^{T \times D}$. With only labeled audio utterance of the target domain $f^a = \{f_t^a\}_{t=1}^T \in \mathbb{R}^{T \times D}$ as input, the audio-visual feature fed into AV-Fusion can be formally expressed as: $f_t^{av} = \text{concat}(f_t^a, \mathbf{0}_D) \in \mathbb{R}^{T \times 2D}$, similar to the modality dropout mechanism in pre-

training stage. With the parameters of AV-Fusion and the Transformer Encoder being frozen, we obtain the fusion features $f^m = \{f_t^m\}_{t=1}^T \in \mathbb{R}^{T \times D}$:

$$f^m = \text{AV-Fusion}_{(\text{frozen})}(f^{av}) \quad (1)$$

and then encode them into a common phoneme space,

$$f^p = \text{encoder}_{(\text{frozen})}(f^m) \quad (2)$$

The freezing parameter allows speech recognition training to focus on the decoder (from the phoneme space to the text space), while avoiding the missing visual modality utterances from destroying the cross-modality alignment relationship in the phoneme space. A tunable Transformer Decoder is appended to autoregressively decode the phoneme feature f^p into the target probabilities:

$$p(w_t | \{w_i\}_{i=1}^{t-1}, f^p) = \text{decoder}_{(\text{tunable})}(f^p) \quad (3)$$

, where $\{w_i\}_{i=1}^s$ is the ground-truth transcription. In the second and third stages of training, the overall model is trained with cross-entropy loss $L_{s2s} = -\sum_{t=1}^s \log p(w_t | \{w_i\}_{i=1}^{t-1}, f^p)$.

3.3 Tuning of Target-domain Decoder in Lip-reading

The decoder trained with the target domain audio utterances fully learns the word distribution and syntactic characteristics of the target domain. Furthermore, the OpenSR training system can further tune the lip-reading model with the labeled visual utterances. Depending on the amount of visual utterances, it can be divided into three types: zero-shot, few-shot and full-shot.

Zero-Shot In OpenSR, the target domain decoder trained in the ASR can be directly transferred to the target domain lip-reading. Both audio and video are aligned in the same phoneme space, and the co-usable decoder for multi-modality speech recognition only needs to map from phoneme space to text. In inferring on lip-reading, only the visual utterances $f^v = \{f_t^v\}_{t=1}^T \in \mathbb{R}^{T \times D}$ inputs into AV-Fusion, and the fusion feature can be formally expressed as: $f_t^{av} = \text{concat}(\mathbf{0}_D, f_t^v)$.

Full-Shot When sufficient labeled visual utterance is available for training, the model parameters can be further fine-tuned. With the word distribution and specific syntax of the target domain, the decoder can steadily boost the discrimination and representation of the pre-trained encoder for visual utterance.

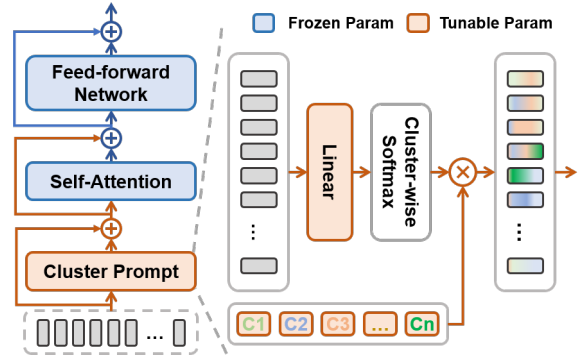


Figure 3: Illustration of Cluster Prompt. The cluster-wise softmax enables the model to assign appropriate prompt embeddings to features based on clustering.

Few-Shot More commonly, when we can only use visual utterances containing only common words for training, the training of the model is likely to be affected by the data distribution, and the model will be more inclined to recognize common words. We adopt a prompt tuning strategy called `Cluster Prompt` to make the model pay more attention to local phoneme features, as shown in the Figure 3.

In the first stage, the pre-training process uses k -means to put cluster labels on audio-visual features during training. We further explore this cluster-based approach by tuning the learnable clustering embeddings $\mathbf{c}_j = \{\mathbf{c}_j^i\}_{i=1}^N \in \mathbb{R}^{N \times D}$ for each cluster in each layer layer_j of the encoder, where N is the number of the clusters. Cluster embedding is the cluster-specific fine-grained bias from audio-based to visual-based phoneme features, which is used to further enhance the visual discriminability between visually similar clusters. The input $\mathbf{x}_j = \{\mathbf{x}_j^t\}_{t=1}^T \in \mathbb{R}^{T \times D}$ to each layer is first fed into a cluster network `Meta`, which is consisted of a linear projection layer and a cluster-wise softmax layer, to extract the cluster weights $\mathbf{u}_j \in \mathbb{R}^{T \times N}$ for each phoneme features \mathbf{x}_j :

$$\mathbf{u}_j = \text{Meta}_{(\text{tunable})}(\mathbf{x}_j) \quad (4)$$

and then combine the cluster embedding vector to update each audio-visual fusion feature:

$$\mathbf{x}'_j = \mathbf{x}_j + \mathbf{u}_j \times \mathbf{c}_j(\text{tunable}) \quad (5)$$

At the same time, we freeze the parameters of the encoder and the decoder to maintain the learned syntactic knowledge and reduce the tuning computational resources:

$$\mathbf{x}_{j+1} = \text{layer}_j(\text{frozen})(\mathbf{x}'_j) \quad (6)$$

Table 1: Comparison of LRS2-COMMON with LRS2: TF stands for word frequency threshold, red numbers indicate the number of utterances containing words that do not appear in the corresponding vocab.

Split-Name	Train	Test	Val	Vocab-Size	Hours	TF
LRS2-224H	142 157	1 082	1 243	41 370	224h	-
LRS2-29H	45 839	1 082	1 243	17 660	29h	-
LRS2-COMMON _{TF>10}	22 669	1 082 ₍₅₁₄₊₅₆₈₎	1 243 ₍₇₅₁₊₄₉₂₎	2 385	11h	>10
LRS2-COMMON _{TF>20}	17 247	1 082 ₍₃₈₉₊₆₉₃₎	1 243 ₍₆₃₁₊₆₁₂₎	1 413	8h	>20
LRS2-COMMON _{TF>50}	10 122	1 082 ₍₂₃₁₊₈₅₁₎	1 243 ₍₄₁₆₊₈₂₇₎	626	4h	>50
LRS2-COMMON _{TF>100}	5 885	1 082 ₍₁₃₅₊₉₄₇₎	1 243 ₍₂₅₃₊₉₉₀₎	344	2h	>100

4 Experiment

4.1 Datasets

LRS2 Afouras et al. (2018a) is one of the most commonly used publicly available English wild lip-reading datasets, including 224 hours of video extracted from shows on BBC television. In the original dataset, the training data is divided into two partitions: *Pretrain* (195H) and *Train* (29H), both of which are transcribed from videos to text at the sentence level. The only difference is that the video clips in the *Pretrain* partition is not strictly trimmed and sometimes longer than the corresponding text. We conducted experiments on LRS2 with different training data amounts (i.e., *Pretrain+Train*(224h) and *Train*(29h)). Note that since the video of the LRS3 dataset has already been used in the pre-training process of AV-Hubert, we do not conduct experiments on it.

LRS2-COMMON Based on the LRS2 dataset, we further propose the LRS2-COMMON to verify the lip-reading performance of the few-shot model trained with labeled visual utterances containing only common words. We counted the word frequency of each word in the *Train* partition of LRS2, and extracted new training sets with only common words according to the word frequency. Note that during the inference, we use the complete *Test* and *Validation* that contain not only common words.

4.2 Evaluation and Implementation Details

For all experiments on LRS2, we use the *word error rate* (WER) as the evaluation index of speech recognition (both lip-reading and AVSR). WER can be defined as $WER=(S+D+I)/M$, where S, D, I, M represent the number of words replaced, deleted, inserted and referenced respectively. During validation, the inference is only performed when all of the validation utterances are of the same modality as the training utterances. For ex-

Table 2: Comparison of full-shot, few-shot and zero-shot methods on LRS2.¹ The experiments highlighted by the underline are in the zero-shot setting.

Type	Method	Labeled Utt(hrs)		WER(%)
		Video	Audio	
Full-Shot	Son Chung et al. (2017)	224	-	70.4
	Afouras et al. (2018a)	698	-	49.8
	Zhao et al. (2020)	698	698	65.3
	Zhang et al. (2019)	698	-	51.7
	Afouras et al. (2020)	224	808	51.3
	Ren et al. (2021)	698	698	49.2
	Prajwal et al. (2022)	698	-	28.9
	Shi et al. (2022)	224	-	28.6
	OpenSR(ours)	224	224	25.0
Few-Shot	Afouras et al. (2020)	224	1032	54.2
	Ma et al. (2021)	224	-	39.1
Zero-Shot	OpenSR(ours)	×	29	<u>39.2</u>
	OpenSR(ours)	×	224	36.0

ample, zero-shot lip-reading trained on labeled audio utterances should also be validated with audio utterances (the inference is performed on the visual utterances, during the testing). In Section A, we present more implement details.

4.3 Main Result

As shown in Table 2, we compare our method with the previous methods in LRS2 to highlight the effect of our proposed training system OpenSR. As the first training system that achieves zero-shot lip-reading, OpenSR not only achieves the state-of-the-art zero-shot and few-shot performance, but even outperforms most full-shot methods. This demonstrates that our training system can effectively train lip-reading models suitable for domains lacking labeled visual utterances. Furthermore, we demonstrate that OpenSR can improve the lip-reading capability of the full-shot lip-reading model. Since the features of audio and video are projected in the

¹The results presented are all trained using publicly available datasets.

Table 3: Comparison of cross-domain method and cross-modality method in training lip-reading model for LRS2-BBC. 🔥 and ❄️ indicate that the parameters in the Transformer Encoder are tunable and frozen, respectively. The experiments highlighted by the underline are in the zero-shot setting.

Pretrain Model	Training Split	Training		Encoder Frozen	WER(%)	ID	
		Utt(hrs)	A V				
Transformer-BASE	Train	LRS3(30h)	✓	🔥	54.9	(1)	
		LRS2(29h)	✓	✓	🔥	51.2	(2)
			✓		🔥	<u>98.3</u>	(3)
			✓		❄️	46.0	(4)
			✓	✓	❄️(A)+🔥(V)	35.7	(5)
	Train + Pretrain	LRS3(433h)		✓	🔥	45.3	(6)
		LRS2(224h)		✓	🔥	39.7	(7)
			✓		🔥	<u>98.5</u>	(8)
			✓		❄️	<u>42.3</u>	(9)
			✓	✓	❄️(A)+🔥(V)	31.7	(10)
Transformer-LARGE	Train	LRS3(30h)		✓	🔥	43.0	(11)
		LRS2(29h)		✓	🔥	31.4	(12)
			✓		🔥	<u>98.2</u>	(13)
			✓		❄️	<u>39.2</u>	(14)
			✓	✓	❄️(A)+🔥(V)	29.5	(15)
	Train + Pretrain	LRS3(433h)		✓	🔥	38.8	(16)
		LRS2(224h)		✓	🔥	28.6	(17)
			✓		🔥	<u>97.4</u>	(18)
			✓		❄️	<u>36.0</u>	(19)
			✓	✓	❄️(A)+🔥(V)	25.0	(20)

same phoneme space after pre-training, a decoder suitable for both ASR and VSR can be trained using only labeled audio utterances. Benefiting from this well-trained decoder, the performance is improved by 2.1% compared to (Shi et al., 2022) using a similar framework.

4.4 Cross Domain VS Cross Modality

For the domains without labeled visual utterances, there are two ways to train lip-reading models using knowledge transfer: cross-domain and cross-modality. The experiments in Table 3 provide answers to the following questions about knowledge transfer:

- How much does the domain shift affect the lip-reading model?
- Is OpenSR training with cross-modality better than cross-domain transferring?

When there is no labeled visual utterances for lip-reading model training in the target domain (Here with LRS2), most of the current methods train on labeled visual utterances from other high resource domains (Here with LRS3). Compared with the model trained with in-domain utterances (ID: 2,7,12,17), the performance of models trained

with other domains (ID: 1,6,11,16) utterances decreased by 3.7% - 11.6%, mainly because there may be terminological and syntactic differences among domains. In Section B, we thoroughly discuss the domain differences between two datasets LRS2&3 of the similar domain from the perspective of term distribution.

In contrast, OpenSR training system learns the word distribution and syntax of the target domain from audio, which effectively avoids domain shift that affects the lip-reading model transferring between domains. As shown in Table 3, even training without labeled visual utterances, the performance of OpenSR training with only audio utterances (ID: 4,9,14,19) can achieve considerable improvement (2.8% - 8.9%) in varying degrees comparing with cross-domain methods (ID: 1,6,11,16). In Section C, we further present qualitative results on cross-domain and cross-modality.

4.5 Pre-training alone cannot achieve zero-shot modality transfer.

With the extraordinary performance of AV-Hubert on the uni-modality speech recognition, there might be some doubt that this zero-shot modality

transfer benefits entirely from AV-Hubert. Indeed, pre-training does align different modalities into the phoneme space and enhance the feature representation, but as shown in the experiment (ID: 3,8,13,18) in Table 3, the model with simple fine-tuning cannot achieve zero-shot modality transfer yet. The idea of parameter freezing in the second stage of OpenSR makes the model better maintain the alignment state between audio and visual features in phoneme space. Maintaining the multi-modality alignment is the key to zero-shot modality transfer, and OpenSR consistently achieves performance commensurate with the scale of the training utterances and model parameters with this property.

Table 4: Comparison of different strategies using only common words for training. ‘>N’ indicates that the training visual utterances only contain words with TF>N in LRS2. #Param is the tunable parameters during tuning.

Training Strategy	# Param(MB)	WER in LRS2-COMMON			
		>100	>50	>20	>10
Shi et al. (2022)	477.33($\times 1.00$)	68.8	61.8	53.9	49.6
OpenSR					
+Finetune	477.33($\times 1.00$)	34.5	33.2	32.1	30.3
+Cluster prompt	9.84($\times 0.02$)	32.7	32.1	30.8	30.5

4.6 Model Tuning with Common Word Videos

Table 4 compares the performance and tuning parameter scale with the state-of-the-art method in LRS2-COMMONs with different word frequencies. The domain shift seriously affects the training of the decoder in AV-Hubert, when the word frequency threshold of the tuning dataset (from LRS2-224H to LRS2-COMMON_{TF>100}) gradually increased, the WER of lip-reading increases sharply, from 28.6% to 68.8%. By learning the word distribution and specific syntax in the target domain from audio, OpenSR ensures that the model will not overfit the common words even if only the video utterances containing a small number of common words is used, as shown in training strategy ‘OpenSR + Finetune’. Furthermore, our proposed Cluster Prompt training strategy in the few-shot scenario shows a further improvement compared to training strategy ‘OpenSR + Finetune’ in terms of common words with a high word frequency threshold (‘>100’, ‘>50’ and ‘>20’). Also, note that the amount of the tuning parameters ($\times 0.02$) of the Cluster Prompt is significantly smaller than the amount ($\times 1.00$) in the other strategies. In particular, with the decrease of the frequency threshold

Table 5: Ablation experiments of different tuning layers in the model. The encoder and decoder have 24 and 9 layers respectively.

Method	WER(%)
OpenSR(zero-shot)	35.995
<i>OpenSR+Tuning the Encoder Layers</i>	
+ w/ encoder.layer.[18,24]	28.273
+ w/ encoder.layer.[12,18]	26.996
+ w/ encoder.layer.[6,12]	26.711
+ w/ encoder.layer.[0,6]	26.426
<i>OpenSR+Tuning the Decoder Layers</i>	
+ w/ decoder.layer.[0,9]	32.522
<i>OpenSR+Tuning the Encoder and Decoder Layers</i>	
+ w/ encoder.layer.[0,24] & decoder.layer.[0,9]	24.954

of common words, the amount of tuning utterances keeps increasing, and the influence of domain shift gradually disappears. Compared to Cluster Prompt, the strategy of fine-tuning learns more about the distribution of the target domain that is applicable to lip-reading from the common word utterances, such as sequence modeling in the decoder specifically for the visual utterances.

4.7 Is the audio trained decoder suitable for lip reading?

We conducted ablation experiments on modules participating in fine-tuning to explore why OpenSR could optimize the upper accuracy limit of the full-shot lip-reading, as shown in Table 5. Only 3.473% (from 35.995% to 32.522%) improvement comes from tuning the decoder, while fine-tuning the pre-trained encoder can achieve additional 9.569% improvement (from 35.995% to 26.426%). The limited improvement achieved by continuing to tune the decoder demonstrates that the word distribution and syntax learned in the audio modality can be zero-shot transferred to the visual modality. Meanwhile, this is why OpenSR can increase the accuracy ceiling, as the encoder pre-trained on large amounts of utterances and the decoder with knowledge of the target domain can steadily boost each other. Furthermore, we also attempt to determine the parameters that need to be tuned most during modality transfer. When tuning the last few layers in the encoder, the performance boost is the most pronounced, increasing the performance by 7.722% (from 35.995% to 28.273%), while additional tuning of all the other layers (encoder.layer.[0,18]) only provides a limited increase of 1.847% (from 28.273% to 26.426%). This demonstrates that what

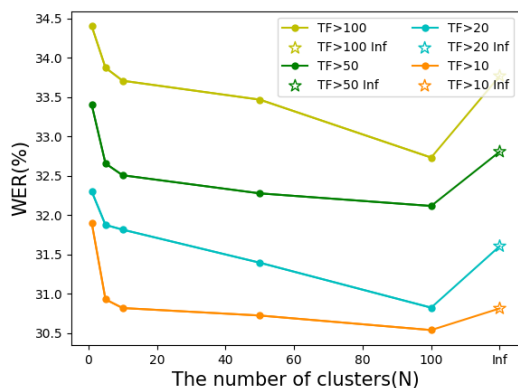


Figure 4: Comparison of lip-reading performance with different number of clusters(N). For comparison, the performance of inf in the figure is simply represented by N=500.

needs to be further tuned while modality transfer is the target-modality representation of the encoder, so that it can distinguish visually similar words (e.g., PET and BET). By maintaining multi-modality feature alignment, the decoder trained with a single modality (i.e., audio) utterance can directly apply the knowledge of the target domain to lip-reading models.

4.8 Modality Prompt VS Cluster Prompt

The Cluster Prompt enables the model to provide prompt embeddings according to the clusters of different phoneme features. In Figure 4, we show the influence of the number of clusters on the performance of accuracy. Obviously, when there is only one cluster, it can be considered as providing the prompt embedding only for the visual modality, which can also be called Modality Prompt. With the increase of the number of clusters (N), the cluster division of the model becomes more refined, so that the prompt embedding also becomes more consistent with each representation. In particular, we noticed that the accuracy decreases when the number of clusters increases to inf (the performance of inf in the figure is simply represented by n=500). This is because there is not enough data to support the training of over-differentiated cluster specific prompt embedding.

4.9 Performance of AV Speech Recognition.

OpenSR can also achieve zero-shot modality transfer from a single modality to multi-modality, by maintaining both audio and visual modality features aligned at the same time. As shown in Table

6, OpenSR outperforms other methods in audio-visual speech recognition. Note that, even in the zero-shot setting, OpenSR performs better than the full-shot performance previous of most previous methods. Furthermore, only using 29h visual utterance to participate in training is enough to comprehensively surpass the previous methods. Comparing with AV-Hubert of the same backbone, the performance of OpenSR is 0.4% better, demonstrating the significance of the term distribution and syntax specification learned during a single modality training for full-shot speech recognition in other-modality.

Table 6: Comparison of audio-visual speech recognition performance on LRS2. The experiment highlighted by the underline is in the zero-shot setting.

Method	Labeled Utt(hrs)		WER(%)
	Video	Audio	
Afouras et al. (2018a)	1428	1428	8.5
Petridis et al. (2018)	381	381	7.0
Yu et al. (2020)	224	224	5.9
Ma et al. (2021)	224	224	3.7
Shi et al. (2022)	224	224	3.1
OpenSR(ours)	-	224	<u>3.3</u>
	29	224	2.8
	224	224	<u>2.7</u>

5 Conclusion

The bottleneck of domain-specific models is the lack of target domain data, such as the challenge of visual utterance collection in lip-reading. We propose OpenSR, a training system that can train target domain lip-reading models without using labeled visual utterance. The other modality models can directly utilize the target domain knowledge obtained from the single modality (e.g., audio) utterance, via maintaining the multi-modality feature alignment state learned during pre-training. This zero-shot modality transfer idea can alleviate the training problem caused by the severely scarce labeled data of some modalities. For example, despite there is no labeled visual utterance of the target domain, we can still use OpenSR to train the lip-reading model with the labeled audio utterance for the target-domain specifically. Furthermore, our training system achieves zero-shot modality transfer in a tuning-based manner, bringing a new perspective utilizing the pre-trained models, which can be transferred to other multi-modality pre-training models such as CLIP (Radford et al., 2021).

6 Ethical Discussion

Lip-reading has many applications, including instruction dictation in public areas or a noisy environment and information accessibility for the hearing impaired. OpenSR makes it possible to quickly build domain-specific lip-reading models for low-resource domains (lack of labeled visual utterances), which greatly enhances the fairness of lip-reading technology across the domains and languages.

For speech recognition, there may be concerns about the risk of information leakage. But in fact, as mentioned above, the lip-reading model has a relatively high requirement on the visual utterances used for lip-reading, requiring mostly-frontal and high-resolution videos with a sufficiently high frame rate, such that motions around the lip area are clearly captured. In general, only close-range cameras or online meetings have similar video conditions, which ensures that the lip-reading model will not be abused in potentially privacy revealing scenarios such as surveillance videos.

Acknowledgements

This work was supported in part by the National Key R&D Program of China under Grant No.2022ZD0162000, National Natural Science Foundation of China under Grant No. 62222211, Grant No.61836002 and Grant No.62072397, and Yiwise.

References

- Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Senior. 2018a. Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence*.
- Triantafyllos Afouras, Joon Son Chung, and Andrew Senior. 2018b. Lrs3-ted: a large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496*.
- Triantafyllos Afouras, Joon Son Chung, and Andrew Senior. 2020. Asr is all you need: Cross-modal distillation for lip reading. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2143–2147. IEEE.
- Yin Aoxiong, Zhong Tianyun, Tang Li, Jin Weike, Jin Tao, and Zhao Zhou. 2023. Gloss attention for gloss-free sign language translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE.
- Yannis M Assael, Brendan Shillingford, Shimon Whiteson, and Nando De Freitas. 2016. Lipnet: End-to-end sentence-level lipreading. *arXiv preprint arXiv:1611.01599*.
- Xize Cheng, Linjun Li, Tao Jin, Rongjie Huang, Wang Lin, Zehan Wang, Huangdai Liu, Ye Wang, Aoxiong Yin, and Zhou Zhao. 2023. Mixspeech: Cross-modality self-learning with audio-visual stream mixup for visual speech translation and recognition. *arXiv preprint arXiv:2303.05309*.
- Joon Son Chung, Arsha Nagrani, and Andrew Senior. 2018. Voxceleb2: Deep speaker recognition. In *INTERSPEECH*.
- Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. 2006. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120(5):2421–2424.
- Chenye Cui, Yi Ren, Jinglin Liu, Rongjie Huang, and Zhou Zhao. 2022. Varietysound: Timbre-controllable video to sound generation via unsupervised information disentanglement. *arXiv preprint arXiv:2211.10666*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Rongjie Huang, Feiyang Chen, Yi Ren, Jinglin Liu, Chenye Cui, and Zhou Zhao. 2021. Multi-singer: Fast multi-singer singing voice vocoder with a large-scale corpus. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3945–3954.
- Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. 2023a. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. *arXiv preprint arXiv:2301.12661*.
- Rongjie Huang, Mingze Li, Dongchao Yang, Jiaotong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, et al. 2023b. Audioopt: Understanding and generating speech, music, sound, and talking head. *arXiv preprint arXiv:2304.12995*.
- Rongjie Huang, Yi Ren, Jinglin Liu, Chenye Cui, and Zhou Zhao. 2022. Genspeech: Towards style transfer for generalizable out-of-domain text-to-speech synthesis. *arXiv preprint arXiv:2205.07211*.
- Tao Jin. 2021. Contrastive disentangled meta-learning for signer-independent sign language translation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5065–5073.
- Tao Jin, Zhou Zhao, Meng Zhang, and Xingshan Zeng. 2022a. Mc-slt: Towards low-resource signer-adaptive sign language translation. In *Proceedings of*

- the 30th ACM International Conference on Multimedia*, pages 4939–4947.
- Tao Jin, Zhou Zhao, Meng Zhang, and Xingshan Zeng. 2022b. Prior knowledge and memory enriched transformer for sign language translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3766–3775.
- Minsu Kim, Jeong Hun Yeo, and Yong Man Ro. 2022. Distinguishing homophenes using multi-head visual-audio memory for lip reading. In *AAAI*.
- Davis E King. 2009. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10:1755–1758.
- Matěj Korvas, Ondřej Plátek, Ondřej Dušek, Lukáš Žilka, and Filip Jurčiček. 2014. Free English and Czech telephone speech corpus shared under the CC-BY-SA 3.0 license. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2014)*, page To Appear.
- Pingchuan Ma, Rodrigo Mira, Stavros Petridis, Björn W Schuller, and Maja Pantic. 2021. Lira: Learning visual speech representations from audio through self-supervision. *arXiv preprint arXiv:2106.09171*.
- Pingchuan Ma, Stavros Petridis, and Maja Pantic. 2022. Visual speech recognition for multiple languages in the wild. *ArXiv*, abs/2202.13084.
- Takaki Makino, Hank Liao, Yannis Assael, Brendan Shillingford, Basi García, Otavio Braga, and Olivier Siohan. 2019. Recurrent neural network transducer for audio-visual speech recognition. *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 905–912.
- Ander Martinez. 2021. [The fujitsu DMATH submissions for WMT21 news translation and biomedical translation tasks](#). In *Proceedings of the Sixth Conference on Machine Translation*, pages 162–166, Online. Association for Computational Linguistics.
- Andrew N Meltzoff and M Keith Moore. 1977. Imitation of facial and manual gestures by human neonates. *Science*, 198(4312):75–78.
- Stavros Petridis, Themis Stafylakis, Pingchuan Ma, Georgios Tzimiropoulos, and Maja Pantic. 2018. Audio-visual speech recognition with a hybrid ctc/attention architecture. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 513–520. IEEE.
- KR Prajwal, Triantafyllos Afouras, and Andrew Zisserman. 2022. Sub-word level lip reading with visual attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5162–5172.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *ICML*.
- Sucheng Ren, Yong Du, Jianming Lv, Guoqiang Han, and Shengfeng He. 2021. Learning from the master: Distilling cross-modal advanced knowledge for lip reading. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13325–13333.
- Dmitriy Serdyuk, Otavio Braga, and Olivier Siohan. 2021. Audio-visual speech recognition is worth 32x32x8 voxels. *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 796–802.
- Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. 2022. Learning audio-visual speech representation by masked multimodal cluster prediction. *arXiv preprint arXiv:2201.02184*.
- David Snyder, Guoguo Chen, and Daniel Povey. 2015. Musan: A music, speech, and noise corpus. *ArXiv*, abs/1510.08484.
- Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. 2017. Lip reading sentences in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6447–6456.
- Aoxiong Yin, Zhou Zhao, Weike Jin, Meng Zhang, Xingshan Zeng, and Xiaofei He. 2022. Mlslt: Towards multilingual sign language translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5109–5119.
- Aoxiong Yin, Zhou Zhao, Jinglin Liu, Weike Jin, Meng Zhang, Xingshan Zeng, and Xiaofei He. 2021. Simulslt: End-to-end simultaneous sign language translation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4118–4127.
- Jianwei Yu, Shi-Xiong Zhang, Jian Wu, Shahram Ghorbani, Bo Wu, Shiyin Kang, Shansong Liu, Xunying Liu, Helen Meng, and Dong Yu. 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)*, pages 6984–6988. IEEE.
- Xingxuan Zhang, Feng Cheng, and Shilin Wang. 2019. Spatio-temporal fusion based convolutional sequence learning for lip reading. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 713–722.
- Ya Zhao, Rui Xu, Xinchao Wang, Peng Hou, Haihong Tang, and Mingli Song. 2020. Hearing lips: Improving lip reading by distilling speech recognizers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6917–6924.

A Implementation Details

Audio and Visual Utterance Preprocessing.

For the visual utterance, we only intercept the lip region for lip-reading. As the previous methods (Shi et al., 2022; Afouras et al., 2018a,b), we adopt dlib (King, 2009) to detect the 68 facial keypoints and align each face with its neighbors. We crop a 96×96 region-of-interest (ROI) talking head video centered on the mouth from each visual utterance. For the audio utterance, we also remain same preprocessing steps as the prior works (Ma et al., 2021; Shi et al., 2022). We extract the 26-dimensional log filterbank energy feature from the raw waveform and stack the 4 neighboring acoustic frames for synchronization. During training, for data enhancement, we randomly crop 88×88 from the whole ROI and flipped it horizontally with 0.5 probability. To improve noise robustness, we apply noise with a probability of 0.25 to each audio utterance from (Snyder et al., 2015) as steps in (Afouras et al., 2018a).

Pre-training Setup. OpenSR builds on pretraining process of AV-Hubert (Shi et al., 2022), directly utilizing its checkpoint for the subsequent stages. During pre-training, a modified ResNet-18 used in prior works (Ma et al., 2021; Martinez, 2021) and a linear projection layer are adopted as visual and audio encoders, respectively. It considers two models with different configurations: Transformer-BASE and Transformer-LARGE have 12/24 Transformer layers with the embedding dimension/feed-forward dimension/attention heads of 768/3072/12 and 1024/4096/16. We simply adopted the pre-trained model obtained by training on LRS3 (Afouras et al., 2018b) and VoxCeleb2 (Chung et al., 2018).

OpenSR Tuning Setup in a Single Modality. In the second stage of OpenSR, we fine-tune the decoder with the labeled audio utterance and the absent modality (visual) feature is replaced by a zero-vector. For comparison, we adopt the same decoder configuration as (Shi et al., 2022), with 6 and 9 Transformer layers in Transformer-BASE and Transformer-LARGE, respectively. With the encoder parameters frozen, we fine-tune the decoder on a single 3090 GPU for 45K/120K steps in the 29h/224h setting. Note that, also only audio can be used while tuning the hyperparamters on the validation set during the second stage. Conversely, during the inference or further tuning with the visual utterance (the third stage), we only adopt

the visual utterance as input and replace the audio feature with the zero-vector $\mathbf{0}_D$. Each stage of OpenSR is trained with Adam, with the learning rate being warmed up for the first 50% of updates to 0.0005.

B Word distribution differences between domains

In this section, starting with LRS2 and LRS3, we explore the differences in word distribution in different domains. Although both datasets are extracted from television shows (BBC and TED respectively), we still find a huge difference in word distribution between them, perhaps due to the different content of the shows.

Different frequencies of the same word in different domains.

In Figure 5, we visually show the word frequency of each word in LRS2&3, each domain has more than 4,000 words that do not appear in the other’s dictionary, as shown by the red dots. There are also a number of words that vary widely in word frequency from domain to domain, as shown by the points off the diagonal. In general, words that are far off the diagonal tend to be terms specific to the domain. Except for a few words with high frequencies that are common to all domains, most words have different word frequencies in different domains. From the perspective of common terms in LRS2, we further quantitatively reveal the differences of the common terms distribution in Figure 6. There are 834 words (2.02% of the LRS2 vocabulary size) with word frequency differences (greater than and less than 10, respectively), demonstrating that a large part of the terms is domain-specific. At the same time,

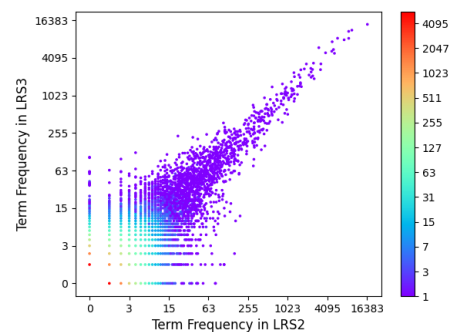


Figure 5: Different frequencies of the same word in LRS2&3. The color shows the number of words at that position.

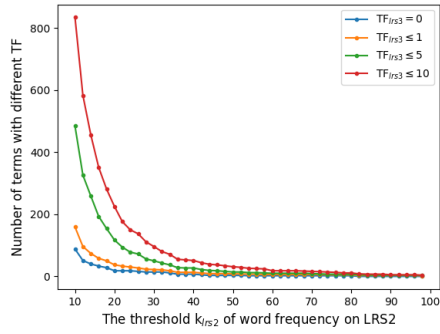


Figure 6: The number of terms whose $TF_{lrs2} \geq K_{lrs2}$ and $TF_{lrs3} \leq K_{lrs3}$. $TF_{lrs2\&3}$ and $K_{lrs2\&3}$ are the word frequency of each word and the counting threshold on LRS2&3, respectively.

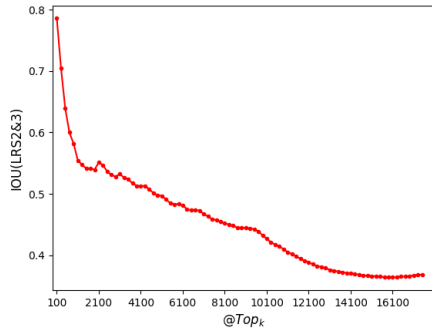


Figure 7: The IOU of the words with the highest $@Top_k$ word frequency between LRS2 and LRS3.

there are a number of terms that varied greatly between domains: 88 words with frequency over 10 (e.g., *bargain*, *crafts* and *saxon*) and 2 words with frequency over 80 (*cos* and *antiques*) never appears in LRS3.

The IoU of words in LRS2 and LRS3 In Figure 7, we show the IoU (Intersection over Union) of the words with the highest $@Top_k$ word frequency in two datasets. Among the top 100 words in the two datasets, there are still 21 different words in addition to non-domain-specific generic words such as ‘The’, ‘A’ and ‘I’ et al. From the perspective of the whole dataset, the IoU of the dictionaries in the two datasets is only 36.925%, which means that there are a large number of words that occur only in their respective domains. Even between near-domain datasets LRS2 and LRS3, there are differences both in the most commonly used words and in the whole dictionary. In certain domains, such as biomedicine and electronic information, the greater difference in word distribution between

Table 7: Video clips of visually confusable words.

Reference: in touch	confusable answer: searched
Reference: rectify	confusable answer: ratify
Reference: britain's	confusable answer: prints & print's
Reference: fingers	confusable answer: fears & feelers

domains makes the lip-reading model unable to transfer across domains.

C Qualitative Results

Performance on visually confusing words. Zero-shot OpenSR trained using only audio performed worse on similar-sounding words. We show several video clips of visually confusable words in Table 7, and show the performance comparison of the model trained with different scale visual utterances in Table 8. With the gradual introduction of visual utterances, the lip-reading performance of the model for visually confusable words is significantly enhanced, demonstrating the significance of our proposed further training using the utterances of common words.

Performance with different term distributions

The distribution of terms will seriously affect the recognition results, while the training speech recognition models. If a term appears never or only a few times, it is essentially unrecognized, as shown in the Table 9. We notice that the models are more likely to come up more common answers (words with high frequency in the in-domain dictionary), probably because the model needs to be trained to fit the word distribution in the training utterances.

Table 8: Qualitative comparison on visually confusing words. Red words highlights misidentified words, (strikeouts) in parentheses highlight corresponding visually similar words and the (red words) in parentheses highlight the absent words.

	
Ground Truth:	people getting in touch and wanting to rectify wrongs
OpenSR _{Zero-shot} (0h):	people getting searched (in touch) and wanting to ratify (rectify) what was
OpenSR _{Few-shot} (2h):	people getting in touch and wanting to rectify what is
OpenSR _{Full-shot} (433h):	people getting in touch and wanting to rectify what's
<hr/>	
	
Ground Truth:	regarding one of britain's most noted
OpenSR _{Zero-shot} (0h):	regarding one of prints (britain's) most noting
OpenSR _{Few-shot} (2h):	regarding one of print's (britain's) most noting
OpenSR _{Full-shot} (433h):	regarding one of britain's most noting
<hr/>	
	
Ground Truth:	all my fingers were hanging off
OpenSR _{Zero-shot} (0h):	all my fears (fingers) were hanging (off)
OpenSR _{Few-shot} (2h):	all my feelers (fingers) were hanging off
OpenSR _{Full-shot} (433h):	all my fingers were hanging off

Table 9: Qualitative performance comparison of models trained with utterances of different word distribution. The misidentified words are highlighted in red and the absent words are highlighted with (red) in parentheses. The table on the right shows the word frequency shift of the misidentified words in different domains.




		Word Frequency	
Ground Truth:	he absolutely insisted on writing his own intro	Term	Dataset
AV-Hubert(LRS3):	(he) absolutely assisted on writing his own intro		
OpenSR _{zero-shot} (LRS2):	he absolutely insisted on writing his own entro	insisted	2 0
OpenSR _{full-shot} (LRS2):	he absolutely insisted on writing his own injury	intro	0 1
<hr/>		<hr/>	
		Word Frequency	
Ground Truth:	an opportunity to pay tribute to all members of the armed forces	Term	Dataset
AV-Hubert(LRS3):	an opportunity to pay attribute to all members of the armed forces		
OpenSR _{zero-shot} (LRS2):	an opportunity to pay tribute to all members of the armed forces	tribute	5 0
OpenSR _{full-shot} (LRS2):	an opportunity to pay tribute to all members of the armed forces	attribute	0 3
<hr/>		<hr/>	
		Word Frequency	
Ground Truth:	let's take a look behind the scenes at how it all came	Term	Dataset
AV-Hubert(LRS3):	let's take a look behind the seeds (at) how it all came		
OpenSR _{zero-shot} (LRS2):	let's take a look behind the scenes (at) how it all came	scenes	14 2
OpenSR _{full-shot} (LRS2):	let's take a look behind the scenes (at) how it all game	seeds	3 7

Table 10: Comparison of Lip-Reading performance on LRS3. VC2-EN stands for the English utterances of VoxCeleb2. For unlabelled utterances, only audio-visual speech can be employed for training (no corresponding transcription). Experiments labeled with † used non publicly available dataset.

Mode	Method	Unlabeled Utts	Labeled Utts Video(hrs)	WER(%)
Full-Shot	Lira (Ma et al., 2021)	-	590	43.3
	VisualSR (Ma et al., 2022)	-	1 459	31.5
	† Sub (Prajwal et al., 2022)	-	2 676	30.7
	† RecurrentNN (Makino et al., 2019)	-	31 000	33.6
	† AV-VIT (Serdyuk et al., 2021)	-	90 000	25.9
	AV-Hubert (Shi et al., 2022)	LRS3+VC2-EN	433	28.6
	OpenSR(ours)	LRS3+VC2-EN	433	28.5
Zero-Shot	OpenSR(ours)	LRS3+VC2-EN	X	30.6

D Zero-Shot Lip-Reading on LRS3.

We further present the performance on the LRS3 dataset in Table 10 (although it has been used in pre-training). During pre-training, the encoder has fully mastered the domain-specific knowledge (the word distribution and syntax) in LRS3, resulting in the performance of 30.6% under the zero-shot setting that is very close to the performance of 28.5% under full-shot (+2.1%). Note that, on the LRS2 dataset, which is not used during the pretraining, the performance of zero-shot is 11% worse than that of full-shot (36.0%→25.0%). In fact, the difference between the further tuning effects on LRS2 and LRS3 (11% and 2.1%) effectively can also effectively demonstrate the domain shift between the LRS2 and LRS3 datasets.

ACL 2023 Responsible NLP Checklist

A For every submission:

- A1. Did you describe the limitations of your work?
Section 6.Ethical Discussion
- A2. Did you discuss any potential risks of your work?
Section 6.Ethical Discussion
- A3. Do the abstract and introduction summarize the paper’s main claims?
In abstract and introduction
- A4. Have you used AI writing assistants when working on this paper?
Left blank.

B Did you use or create scientific artifacts?

Left blank.

- B1. Did you cite the creators of artifacts you used?
In reference
- B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
In reference
- B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.
- B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.
- B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.
- B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
In section4.1

C Did you run computational experiments?

Left blank.

- C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
In appendix.A

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

- C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?

No response.

- C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?

Not applicable. Left blank.

- C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?

In appendix.A

D Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

- D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?

Not applicable. Left blank.

- D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?

Not applicable. Left blank.

- D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?

Not applicable. Left blank.

- D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?

Not applicable. Left blank.

- D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?

Not applicable. Left blank.