Zero-AVSR: Zero-Shot Audio-Visual Speech Recognition with LLMs by Learning Language-Agnostic Speech Representations

Supplementary Material

7. Detailed Information of MARC

The MARC dataset is driven by combining the existing audio-visual speech datasets. Specifically, the labeled audio-visual speech datasets, LRS3 [33] and MuAViC [26], and the unlabeled audio-visual speech datasets, VoxCeleb2 [58] and AVSpeech [59], are combined. The information of each dataset is as follows:

Lip Reading Sentences 3 (LRS3) [33] is a dataset designed for AVSR and is one of the most widely used resources. It contains 433 hours of audio-visual English data with human-annotated transcriptions, sourced from TED and TEDx talks.

Multilingual Audio-Visual Corpus (MuAViC) [26] is a dataset for multilingual audio-visual speech recognition and translation, collected from TED and TEDx talks across nine languages and comprising 1,200 hours of data with human-annotated transcriptions. Since its English portion overlaps with LRS3 and the preprocessing differs due to a different landmark detector, we exclusively use the English portion of LRS3 following the process in [9].

VoxCeleb2 [58] is a dataset for speaker recognition containing 2,442 hours of multilingual audio-visual data. Although it includes speaker ID information, it lacks human-annotated text transcriptions and language labels.

AVSpeech [59] is a dataset aimed at isolating a target speaker's voice from mixed audio, sourced from YouTube videos and comprising 4,700 hours of multilingual audio-visual data. Like VoxCeleb2, it does not provide human-annotated text transcriptions or language labels.

The unlabeled audio-visual datasets, VoxCeleb2 and AVSpeech, are labeled using language identification and ASR. For language identification, we use the MMS-LID-1024 model[2] with a threshold of 0.95 for the confidence score. To generate language-specific graphemes, we utilize the pre-trained MMS-1B-ALL[3] ASR model in conjunction with a language adapter, which is selected based on the identified language. Furthermore, during the decoding stage, we leverage language-specific language models[4]. The resulting MARC dataset consists of 82 languages and approximately 2,916 hours of audio-visual data. The languages and their respective families [65] in the MARC dataset are listed in Tables 8 and 9.

[2] https://huggingface.co/facebook/mms-lid-1024
[3] https://huggingface.co/facebook/mms-1b-all
[4] https://huggingface.co/facebook/mms-cclms

Method	Unseen Lang.	Target Language (CER(%)↓)								Avg (w/o Eng)
		Ara	Deu	Ell	Spa	Fra	Ita	Por	Rus	
Cascaded Zero-AVSR (Llama3.2-3B)	Ara	86.6	25.2	56.3	13.5	33.4	15.5	18.5	55.3	37.9
	Deu	76.8	67.0	55.0	12.4	21.5	13.6	14.8	53.1	44.3
	Ell	68.5	25.5	75.2	13.9	29.2	15.3	14.8	58.8	39.3
	Spa	73.6	25.0	53.8	34.8	20.3	18.8	16.1	53.9	37.1
	Fra	76.8	24.7	56.7	15.5	82.2	14.7	17.6	54.3	40.0
	Ita	73.7	25.2	54.2	12.7	26.4	36.7	14.8	55.9	37.6
	Por	74.5	25.7	54.1	17.1	21.4	14.4	63.0	55.2	41.4
	Rus	68.6	24.6	55.8	14.1	22.2	16.8	14.6	79.8	38.1
Zero-AVSR (Llama3.2-3B)	Ara	76.5	16.2	21.6	6.8	7.4	7.0	7.7	18.5	19.7
	Deu	56.9	52.9	22.4	7.4	8.2	6.9	7.7	18.7	26.4
	Ell	53.9	16.1	62.1	6.8	8.1	6.8	7.8	18.4	24.3
	Spa	57.8	16.7	24.4	19.7	8.5	7.5	8.6	20.2	19.0
	Fra	55.3	16.0	21.2	6.9	54.6	7.2	7.9	18.1	20.7
	Ita	61.6	17.2	22.6	7.1	8.1	25.1	7.7	18.6	20.7
	Por	58.0	16.3	24.7	7.5	7.7	7.4	44.0	18.8	22.3
	Rus	57.7	16.0	22.7	7.1	7.7	6.9	7.6	45.4	21.5

Table 7. The zero-shot language AVSR performances of Cascaded Zero-AVSR and Zero-AVSR using the same LLM (Llama3.2-3B) on MuAViC dataset. We train 8 AV-Romanizers, setting each language as an unseen language, and evaluate their zero-shot performance, which are shown in blue-colored cells.

8. The Effectiveness of Zero-AVSR

In order to confirm the effectiveness of the Zero-AVSR compared to the Cascaded Zero-AVSR, we also report the zero-shot language AVSR performance of Cascaded Zero-AVSR using Llama3.2-3B, in Table 7. Therefore, since Cascaded Zero-AVSR and Zero-AVSR utilize the same LLM here, we can evaluate the effectiveness of finetuning the LLM by employing the proposed multi-task framework. By comparing the zero-shot speech recognition performances (*i.e.*, shown in blue-colored cells), we can confirm that Zero-AVSR improves performance over all 8 languages, demonstrating the effectiveness of incorporating speech features directly into the LLM instead of employing a text formula. Furthermore, when comparing the average CER, taking into account both seen and unseen languages, Zero-AVSR outperforms the Cascaded Zero-AVSR model. These results show the promise of Zero-AVSR that if a better LLM is employed and finetuned, performance can be improved even more.

9. Noise-robustness Experiments

By employing audio-visual speech inputs, we can achieve more robust noise performance in speech recognition compared to when we employ audio-only speech inputs. In

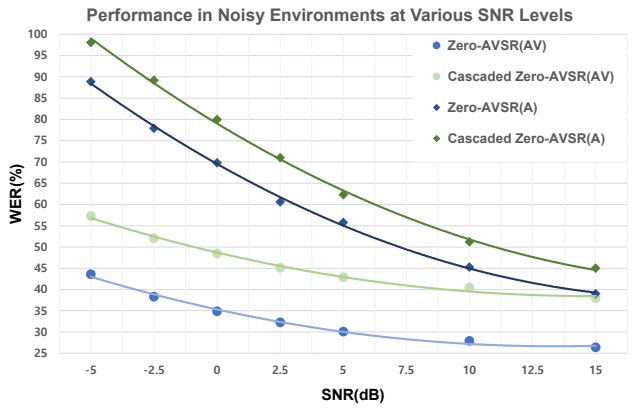

Figure 3. The performances of Cascaded Zero-AVSR and Zero-AVSR using audio-only (A) and audio-visual (AV) inputs under different SNR noise levels.

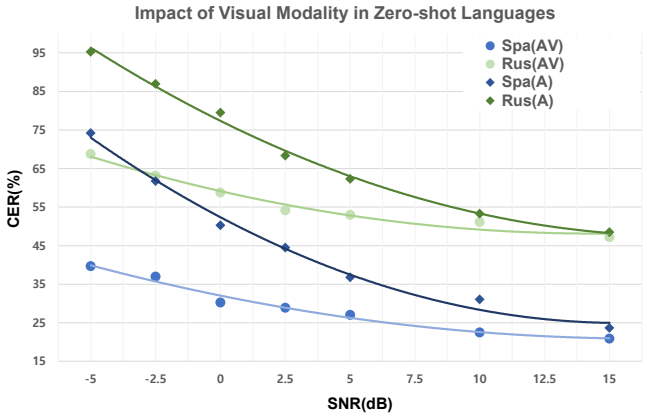

Figure 4. Performance of Zero-AVSR on zero-shot languages, Spanish (Spa) and Russian (Rus), using audio-only (A) and audio-visual (AV) inputs across various SNR levels.

this section, we analyze the performance of both Cascaded Zero-AVSR and Zero-AVSR by differing the acoustic noise levels from -5 dB SNR to 15 dB SNR. The acoustic noise is uniformly sampled among natural, babble, music, speech partitions from MUSAN [64]. The analysis results are shown in Fig. 3. It shows the average WER for all 9 languages. In both Cascaded Zero-AVSR and Zero-AVSR, the audio-only (A) models' performances are significantly degraded according to the noise become strong. However, we can confirm that the audio-visual (AV) models show robust performances over the different noise levels.

We have seen that by using audio-visual speech inputs, we can achieve more robust speech recognition performances. Here, we also analyze this in zero-shot language settings. To this end, we measure CERs of Zero-AVSR model under various SNR levels on two unseen languages, Spanish and Russian. The results in Fig. 4 show that, similar to the seen language scenarios, the audio-visual (AV) model outperforms the audio-only (A) framework across all SNR levels. Especially, in the zero-shot language setting,

Ground Truth :	このくつしたは じょうぶでやすいです
Prediction :	このくたしゃたばこぶであすいです
Roman Prediction :	konu kutcashitava cobudeasuides
Filename :	Common_voice_ja_20184324.wav
Ground Truth :	といれからでるとき てをあらいます
Prediction :	といれがてるついておあらいます
Roman Prediction :	to irega terutui teo alaimas
Filename :	Common_voice_ja_20853323.wav
Ground Truth :	だれかわたしのぼ るぺんをもっていませんか
Prediction :	だれかばだしのぽろぺんおもちあせんか
Roman Prediction :	dareka badasinoporopen omoti asenka
Filename :	Common_voice_ja_22727312.wav
Ground Truth :	あそこにきむらさんがたっています
Prediction :	あそこにきたらさんがたたます
Roman Prediction :	asokoni kimera sanggatata mas
Filename :	Common_voice_ja_24592150.wav

Figure 5. Examples of prediction results from the Cascaded Zero-AVSR on an unseen language, Japanese, on out-of-domain data.

we can observe that while both models exhibit comparable performance under clean environment (*i.e.*, higher SNR), the performance gain by using audio-visual inputs over the audio-only inputs increases as the SNR decreases.

10. Zero-Shot Performance on Out-of-Domain

We evaluate the extent to which the proposed Zero-AVSR framework can perform on out-of-domain data. To this end, we evaluate the zero-shot speech recognition performance on Japanese, whose language family is not presented in the training set of MARC. We measure the Japanese performance on the test set of CommonVoice [66] by using audio-only inputs. The Cascaded Zero-AVSR achieves 60.9% CER and the Zero-AVSR achieves 64.9% CER. These results demonstrate that the proposed Zero-AVSR framework can be employed for languages even when no data from the same language family is used, showing its scalability to more languages. The examples of prediction using the Cascaded Zero-AVSR are shown in Fig. 5. For example, as shown in the last row, the AV-Romanizer predicts the Roman text as 'asokoni kimera sanggatata mas', and the LLM de-romanizes this text into Japanese as 'あそこにきたらさんがたたます'. Notably, despite not being perfect, the prediction was made without Japanese data being employed during the training of the proposed AV-Romanizer.

11. Qualitative Error Analysis

While our proposed Zero-AVSR framework enables speech recognition in zero-shot languages, its performance still lags behind that on seen languages. In this section, we conduct a qualitative error analysis to better understand the types and root causes of failures. Specifically, we investigate two error categories: mis-romanization and

Failure case : Mis-romanization.	
Ground Truth :	я не знаю почему так происходит
Roman Ground Truth :	ya ne znayu pochemu tak proiskhodit
Roman Prediction :	ila kzau pochemu tak prosvod
Grapheme Prediction :	Ила кзау почему так просвод
Language:	Russian
Ground Truth :	não havia esta confusão
Prediction :	nao havia esta confusao
Roman Prediction :	via essa confusa
Grapheme Prediction :	via essa confusão
Language :	Portuguese
Failure case : LLM de-romanization error	
Ground Truth :	اختيار منك حتى تجاوب على بعض الأسئلة،
Roman Ground Truth :	taban, ana musta'id. eltahab as'ilatak.
Grapheme Prediction :	تابع، أنا مستعد. التحق بأسئلتك
Language :	Arabic
Ground Truth :	أن أبطال هذه القصص سيكونون أحد الحاضرين في هذا المؤتمر
Roman Ground Truth :	n abtal hadhih al-qisas sayakunun ahad al-hadirin fi hadha al-mu'tamar
Grapheme Prediction :	أنت بخير؟
Language :	Arabic

Figure 6. Qualitative examples of mis-romanization and LLM-deromanization errors.

LLM-deromanization. Mis-romanization errors occur when the AV-romanizer's output differs from the ground-truth romanization. LLM-deromanization errors arise when, after feeding the ground-truth romanization into the LLM, the model mispredicts the original graphemes. Examples of both error types are illustrated in Fig. 6. In our analysis, we found that the vast majority of errors stem from mis-romanization stage. This suggests that there is still room for improving the overall zero-shot language recognition performance by refining the romanization stage (*i.e.*, the AV-Romanizer).

Limitation

Despite demonstrating strong zero-shot performance, our framework exhibits two key limitations: 1) Language coverage depends on LLM support. Our AVSR pipeline relies on an LLM for each target language. When the LLM underperforms on a given language, particularly low-resource ones, pipeline accuracy degrades accordingly. 2) Prosodic information is lost through romanization. We convert all inputs to unaccented Roman characters, which inherently discards tone, stress, and vowel-length distinctions. This omission is especially critical in tonal languages, where pitch shifts can change word meaning entirely. To address these issues, future zero-shot AVSR systems may consider 1) incorporating prosodic features, tone, stress, and rhythm into language-agnostic encodings, and 2) leveraging next-generation LLMs trained on low-resource languages. These enhancements will broaden robust support across diverse linguistic contexts.

Language Family	Subgroup	Branch	Number	Language Name	Language Code	Video Hours
Indo-European	Germanic	Western	1	English	eng	435.2
			2	German	deu	327.5
			3	Dutch	nld	72.5
			4	Afrikaans	afr	1.7
			5	Luxembourgish	ltz	1.9
		Northern	6	Swedish	swe	19.3
			7	Danish	dan	18.1
			8	Norwegian	nob	2.8
			9	Icelandic	isl	0.4
	Romance	-	10	Italian	ita	146.8
			11	French	fra	291.7
			12	Spanish	spa	216.2
			13	Portuguese	por	408.0
			14	Romanian	ron	16.0
			15	Catalan	cat	7.0
			16	Galician	glg	8.7
			17	Asturian	ast	0.1
			18	Occitan	oci	1.5
	Celtic	Brythonic	19	Welsh	cym	97.1
		Goidelic	20	Irish	gle	0.1
	Hellenic	-	21	Greek	ell	21.5
	Slavic	Eastern	22	Russian	rus	123.8
			23	Ukrainian	ukr	3.8
			24	Belarusian	bel	5.9
		Western	25	Polish	pol	50.5
			26	Czech	ces	20.6
			27	Slovak	slk	4.9
		Southern	28	Bulgarian	bul	3.9
			29	Slovene	slv	4.6
			30	Macedonian	mkd	0.3
			31	Bosnian	bos	0.8
			32	Croatian	hrv	2.6
			33	Serbian	srp	0.9
	Indo-Iranian	Iranian	34	Persian	fas	7.6
			35	Kurdish	ckb	0.05
			36	Tajik	tgk	0.1
			37	Pushto	pus	0.3
		Indic	38	Hindi	hin	99.0
			39	Urdu	urd	8.6
			40	Bengali	ben	8.8
			41	Punjabi	pan	3.0
			42	Marathi	mar	8.2
			43	Gujarati	guj	1.6
			44	Assamese	asm	0.2
			45	Nepali	npi	4.5
			46	Sindhi	snd	0.5
			47	Odia	ory	0.1

Table 8. The Data Statistics of the MARC dataset 1.

Language Family	Subgroup	Branch	Number	Language Name	Language Code	Video Hours
Indo-European	**Baltic**	-	48	Lithuanian	lit	2.9
		-	49	Latvian	lav	1.3
	-	-	50	Armenian	hye	0.7
Uralic	**Finno-Ugric**	**Finnic**	51	Finnish	fin	9.7
			52	Estonian	est	2.1
		Ugric	53	Hungarian	hun	11.0
Altaic	**Turkic**	**Southwestern**	54	Turkish	tur	50.6
			55	Azerbaijani	aze	1.9
		Northwestern	56	Kazakh	kaz	3.8
			57	Kyrgyz	kir	0.1
		Southeastern	58	Uzbek	uzb	0.3
	Mongolian	-	59	Mongolian	mon	1.1
Caucasian	**Southern**	-	60	Georgian	kat	1.5
Dravidian	-	-	61	Telugu	tel	18.8
	-	-	62	Tamil	tam	18.3
	-	-	63	Kannada	kan	3.6
	-	-	64	Malayalam	mal	15.5
Independent	-	-	65	Korean	kor	128.6
Mon-Khmer	-	-	66	Vietnamese	vie	32.9
Austronesian	**Western**	-	67	Indonesian	ind	15.0
		-	68	Javanese	jav	0.1
		-	69	Tagalog	tgl	6.1
	Polynesian	-	70	Maori	mri	4.9
Niger-Congo	**Atlantic**	-	71	Wolof	wol	1.1
	Benue-Congo	-	72	Swahili	swh	1.2
		-	73	Lingala	lin	0.9
		-	74	Ganda	lug	0.04
		-	75	Shona	sna	3.8
Afro-Asiatic	**Semitic**	**North Arabic**	76	Arabic	ara	96.7
			77	Maltese	mlt	1.5
		Canaanitic	78	Hebrew	heb	14.3
		Ethiopic	79	Amharic	amh	1.2
	Cushitic	-	80	Somali	som	4.6
		-	81	Oromo	orm	0.1
	Chadic	-	82	Hausa	hau	0.7

Table 9. The Data Statistics of the MARC dataset 2.