Interleaved Audio/Audiovisual Transfer Learning for AV-ASR in Low-Resourced Languages

Zhengyang Li[1], *Patrick Blumenberg*[1], *Jing Liu*[2], *Thomas Graave*[1], *Timo Lohrenz*[1], *Siegfried Kunzmann*[2], *Tim Fingscheidt*[1]

[1]Technische Universität Braunschweig, Institute for Communications Technology, 38106 Braunschweig, Germany [2]Amazon AGI, 15203 Pittsburgh, PA, USA

{zhengyang.li,p.blumenberg,t.fingscheidt}@tu-bs.de, {jlmk,kunzman}@amazon.com

Abstract

Cross-language transfer learning from English to a target language has shown effectiveness in low-resourced audiovisual speech recognition (AV-ASR). We first investigate a 2-stage protocol, which performs fine-tuning of the English pre-trained AV encoder on a large audio corpus in the target language (1st stage), and then carries out cross-modality transfer learning from audio to AV in the target language for AV-ASR (2nd stage). Second, we propose an alternative interleaved audio/audiovisual transfer learning to avoid catastrophic forgetting of the video modality and to overcome 2nd stage overfitting to the small AV corpus. We use only 10h AV training data in either German or French target language. Our proposed interleaved method outperforms the 2-stage method in all low-resource conditions and both languages. It also excels the former state of the art both in the noisy benchmark (babble 0dB, 53.9% vs. 65.9%) and in clean condition (34.9% vs. 48.1%) on the German MuAVIC test set.

Index Terms: audiovisual speech recognition, interleaved training, cross-language transfer learning, low-resourced speech recognition

1. Introduction

The development of audiovisual speech recognition (AV-ASR) systems, which utilize a speaker's lip movement as compensation to auditory speech, has been inspired by the psychological finding that speech perception is inherently multimodal [1, 2]. AV-ASR systems have shown their superior performance compared to ASR based on acoustics especially in noisy conditions [3, 4, 5, 6]. The noise robustness of AV-ASR systems promote the deployment in smart home devices [7] and automobiles [8].

Building upon advancements of the all-attention-based transformer architecture in neural machine translation [9] and speech recognition [10, 11, 12, 13], transformer models have been applied to AV-ASR as well [4, 14]. For English AV-ASR, the rapid improvement also benefits from the availability of large public audiovisual datasets, e.g., 224 hours Lip Reading Sentences 2 (LRS2) [15], 433 hours Lip Reading Sentences 3 (LRS3) [14], and 1326 hours English subset of the VoxCeleb2 dataset [16]. In recent years, researchers have made efforts to release a few public multi-lingual audiovisual datasets [17, 18]. However, most languages are under-represented and still face a low-resource problem among public labeled audiovisual data, e.g., the German split in the MuAViC dataset [18] comprises only 10 hours of AV data. To solve this issue, cross-language transfer learning from English as the source language to a resource-constrained target language has been applied in speech tasks [19, 20, 21, 22] including AV-ASR [18, 23]. Anwar et al. [18] fine-tune the `AV-HuBERT` encoder pre-trained on English datasets [3] on a target language. Li et al. [23] build a language-modular AV-ASR system by applying parameter-efficient adapters [24] in `AV-HuBERT`. These studies show that AV-ASR systems in low-resourced languages can benefit from the visual modality in noisy conditions. However, the overall accuracy remains at a somewhat low level.

For low-resource ASR tasks, effective training protocols have been investigated [25, 19]. Chen et al. [25] trained deep neural networks (DNNs) of a hybrid ASR system on multiple tasks to improve ASR performance in low-resourced South African languages. Wang et al. [19] investigated a cross-language transfer learning for end-to-end ASR in Dutch and Mongolian by pre-training the encoder-decoder transformer on a speech translation task in a first learning phase. Both works [25, 19] show that a transfer learning protocol in two phases, which trains the model on a related speech task in the target language in the first phase, is able to improve final low-resourced ASR performance after the second phase training on the ASR task. In a deep noise suppression task, Xu et al. [26, 27] proposed a novel epoch-level alternating training protocol for two neural sub-networks, keeping the knowledge of sub-networks up-to-date during training. From education, we know that interleaved (alternating) training between topics can foster longer lasting and more generalizable learning [28].

In this work, compared to conventional cross-language transfer learning for AV-ASR (which only utilizes limited AV data in the target language), we first introduce rich target language knowledge by training an AV-ASR system on limited audiovisual data *as well as* on more easily accessible audio-only data of the target language. Second, we investigate effective training protocols of applying AV and audio-only data for AV-ASR. Specifically, we investigate a 2-stage training protocol which—in a first fine-tuning stage—leverages vast amounts of audio-only training data from the target language, and in a second stage, fine-tunes on the small target AV dataset. Through an empirical ablation study, we found that this 2-stage transfer learning protocol encounters catastrophic forgetting of the video modality in the first stage and overfitting on low-resourced AV data in the target language in the second stage. In resource-constrained settings, we then demonstrate that our newly proposed minibatch-level interleaved audio/audiovisual training in the target language improves recognition accuracy and strengthens noise robustness simultaneously in a single training (i.e., fine-tuning) stage, achieving a longer lasting and more generalizable learning.

The paper is structured as follows. In Section 2, we introduce the baseline and our investigated transfer learning protocols. The experimental setup is described in Section 3. Section 4 comprises experimental results and discussion on the multi-lingual MuAViC audiovisual speech recognition task [18]. The paper is concluded in Section 5.

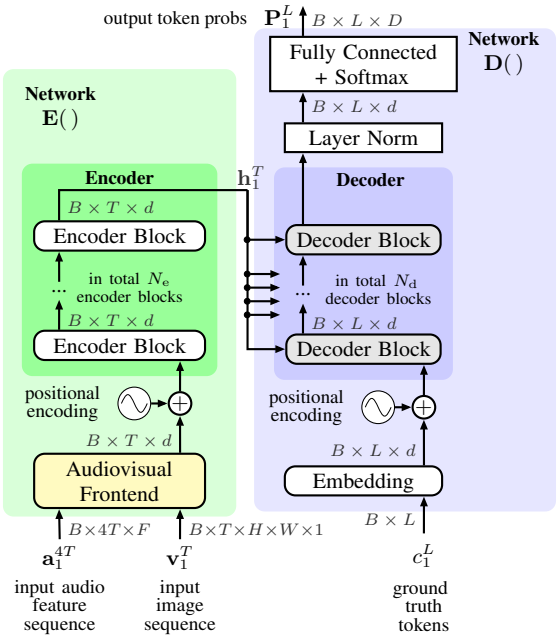

Figure 1: ***Transformer encoder-decoder model*** *used for audiovisual speech recognition during training. The audiovisual encoder network* $\mathbf{E}(\,)$ *is pre-trained on English data [3].*

2. Methods

2.1. Baseline Model and Training Protocol

Baseline model: As shown in Figure 1, we utilize a transformer encoder-decoder model for AV-ASR, which consists of an encoder network $\mathbf{E}(\,)$ (green block) and a decoder network $\mathbf{D}(\,)$ (blue block). The encoder network $\mathbf{E}(\,)$ comprises an audiovisual frontend followed by positional encoding and N_e serial transformer encoder blocks. Inputs are the image sequence $\mathbf{v}_1^T{=}(\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_T)$ and the audio feature sequence $\mathbf{a}_1^{4T}{=}(\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_{4T})$. Note that in our case, the frame rate is 25 Hz (video) and 100 Hz (audio), causing the fourfold length $4T$ of the audio feature sequence. During training, the decoder network $\mathbf{D}(\,)$ leverages the encoded audiovisual representation $\mathbf{h}_1^T = \mathbf{E}(\mathbf{x}_1^T) = (\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_T)$ and the ground truth tokens $c_1^L{=}(c_1, c_2, ..., c_\ell, ..., c_L)$ to predict the probability distributions of output tokens $\mathbf{P}_1^L = \mathbf{D}(\mathbf{h}_1^T, c_1^L) = (\mathbf{P}_1, \mathbf{P}_2, ..., \mathbf{P}_L)$, with probability distribution $\mathbf{P}_\ell \in \mathbb{R}^D$, and D being the size of the token vocabulary. All entries in each probability distribution $\mathbf{P}_\ell$ sum up to one.

Baseline training protocol: Our baseline follows the same training protocol as in [18]. The encoder network $\mathbf{E}(\,)$ is initialized by the `large AV-HuBERT` model [3], which is pre-trained on large unlabeled audiovisual data in English as the source language $\mathcal{D}_{\mathrm{AV}}^{\mathrm{S}}$, and thus contains English language knowledge and models audiovisual interaction with transformer encoder blocks. In the baseline training protocol, the decoder network $\mathbf{D}(\,)$ is fine-tuned from scratch jointly with the pre-trained encoder network $\mathbf{E}(\,)$ on audiovisual data in a target language $\mathcal{D}_{\mathrm{AV}}^{\mathrm{T}}$, thereby transferring the English AV encoder from English to the target language.

2.2. Investigated Transfer Learning Protocols

2-stage protocol: Cross-language transfer learning relies on acoustic similarities between languages. Recent studies [29, 30] show that the AV encoder $\mathbf{E}(\,)$ pre-trained by self-supervised learning (SSL) captures acoustic features primarily by the initial acoustic frontend and early transformer encoder blocks, while language-specific linguistic features are learned mainly through late transformer encoder blocks close to the decoder. In addition, the decoder network $\mathbf{D}(\,)$, which leverages ground truth tokens c_1^L as input during training, learns an internal language model [31, 12, 32]. However, fine-tuning the English AV encoder on a small AV corpus in the target language cannot incorporate enough target language knowledge and can result in low recognition accuracy. Compared to AV data, audio-only data is much easier to obtain [33], and therefore has a much larger data size $|\mathcal{D}_{\mathrm{A}}^{\mathrm{T}}|{>}|\mathcal{D}_{\mathrm{AV}}^{\mathrm{T}}|$. In the 2-stage method, we fine-tune the transformer encoder-decoder model on a large audio-only corpus $\mathcal{D}_{\mathrm{A}}^{\mathrm{T}}$ in the target language to enhance target language knowledge (1st stage). Then, we perform the second stage fine-tuning on the AV corpus in the target language $\mathcal{D}_{\mathrm{AV}}^{\mathrm{T}}$ for AV-ASR.

Interleaved protocol (proposed): In contrast to the 2-stage protocol, our proposed interleaved training protocol utilizes a large-scale audio-only dataset $\mathcal{D}_{\mathrm{A}}^{\mathrm{T}}$ and a resource-constrained audiovisual dataset $\mathcal{D}_{\mathrm{AV}}^{\mathrm{T}}$ in a single target-language fine-tuning stage. During training, for each minibatch, with a probability $p \in [0, 1]$, we select only audiovisual samples from the AV dataset in the target language $\mathcal{D}_{\mathrm{AV}}^{\mathrm{T}}$, while the probability of a minibatch with audio-only samples from the audio dataset $\mathcal{D}_{\mathrm{A}}^{\mathrm{T}}$ is then $(1-p)$. In the minibatch with audio-only samples, the input image sequence $\mathbf{v}_1^T$ is replaced by $\mathbf{0}$ vectors with dimensionality of $B{\times}T{\times}H{\times}W{\times}1$. In this way, the model is trained on audio and AV data in a minibatch-level interleaved fashion. The motivation is to solve two issues of the 2-stage protocol: First, the audiovisual interaction, which is learned during pre-training on the English AV data $\mathcal{D}_{\mathrm{AV}}^{\mathrm{S}}$, can be forgotten in the 1st stage training on audio-only data in the target language $\mathcal{D}_{\mathrm{A}}^{\mathrm{T}}$. Second, as some target languages are resource-constrained [18], the AV-ASR based on a transformer encoder-decoder architecture overfits easily on the small amount of AV data in the target language $\mathcal{D}_{\mathrm{AV}}^{\mathrm{T}}$. With the interleaved protocol, we can simultaneously maintain the learned audiovisual interaction and enhance target language knowledge by fine-tuning the model on an interleaved minibatch of audiovisual samples with a probability of p and on audio-only samples with a probability of $(1-p)$.

3. Experimental Setup

Datasets and pre-processing: We leverage the `large AV-HuBERT` encoder [3] pre-trained on unlabeled English AV datasets $\mathcal{D}_{\mathrm{AV}}^{\mathrm{S}}$, i.e., Lip Reading Sentences 3 (LRS3) [34] and the English subset of the VoxCeleb2 dataset [16], to initialize the AV encoder. Cross-language transfer learning experiments are performed from English as the source language to a target language. We utilize German (DE), a Germanic language, and French (FR), a Romance language, as our target languages. The German or French `train` set of the MuAViC dataset is used for fine-tuning, which comprises 10 hours and 176 hours data, respectively. In addition, for comparison of investigated training protocols in low-resourced settings, we randomly sample 10 hours data from the MuAViC French `train` set as a low-resourced `train` set. The audio-only data used for 2-stage and interleaved protocols are the German and French subsets of the CommonVoice v13.0 dataset [33]. We remove audios longer than 20 seconds in the training sets, building an 854h German and a 732h French `train` set. The video frame rate is 25 Hz and the speech signal sample rate is 16 kHz. In accordance with

the pre-processing pipeline of the MuAViC dataset [18], we use 26-dimensional log-filterbank outputs as input audio features, which are extracted with a 25 ms window and a frame shift of 10 ms, resulting in 100 audio frames per second. Regarding video frames, we convert them to grayscale and crop them to a region of interest measuring 96×96 based on face alignment.

Transfer learning: For a fair comparison, we leverage the same encoder-decoder architecture as the baseline [18] for all experiments, which comprises 477M parameters in total. The encoder network $\mathbf{E}(\,)$ contains 24 transformer encoder blocks, and the decoder network $\mathbf{D}(\,)$ has 6 transformer decoder blocks. The outputs of the encoder-decoder architecture are subword tokens with a vocabulary size of 1000 generated by `SentencePiece` [35]. The fine-tuning process is done using the `PyTorch`-based `fairseq` toolkit. We fine-tune the encoder-decoder model for 30k updates in the baseline protocol, in each stage of the 2-stage protocol, and in the interleaved protocol. The fine-tuning process uses minibatches of up to 1000 image frames. The learning rate is linearly increased to 0.001 for the first 10k updates, then linearly decreased to 0. We apply the same data augmentation as in the baseline method [18], where 25% of the training data is augmented with an SNR of 0dB noise chosen from the babble, music, natural noise, and second interfering talker conditions. There is no speaker overlap in babble noise and second interfering talker condition among different splits.

Evaluation in noisy environments: We evaluate models on the `dev` and `test` sets of the CommonVoice v13.0 and the MuAViC dataset. To add noise to our speech data, we follow the exact same procedure as detailed in [18, 3]. We generate babble noise by mixing utterances from 30 different speakers from the MUSAN dataset [36] where each speaker is used exclusively for either the `train`, `dev`, or the `test` split. We also evaluate our approaches for speech with a *second interfering talker* from LRS3 data, and speech with *music* noise from the MUSAN dataset [36], following [3] for comparability reasons.

4. Results and Discussion

To compare and analyze different transfer learning protocols, we perform an ablation study as shown in Table 1. The word error rates (WERs) are evaluated on the German (DE) `dev` and `test` splits of the CommonVoice [33] (audio-only) and the AV MuAViC [18] datasets. For comparison of the noise robustness, on the AV MuAViC dataset, the inference results with audio-only (A) input and audiovisual input (AV) are presented on clean speech and on speech at a signal-to-noise ratio (SNR) of 0dB babble noise.

The top table segment shows the results of cited and resimulated baselines [18] as well as a recent benchmark [37]. The first two rows are cited numbers from [18], where the WERs are only reported on the `test` split of the MuAViC dataset. Compared to inference with audio-only (A) input in the first row, the inference with audiovisual (AV) input in the second row reduces the WER (lower is better) in both clean and noisy conditions. We retrained the baseline [18] and observed WERs above 80% on the CommonVoice `dev` and `test` splits. Reasons for the poor performance can be either *limited target language knowledge* from the small AV dataset $\mathcal{D}_{\mathrm{AV}}^{\mathrm{T}}$ in the target language, or *domain mismatch* of the CommonVoice dataset [33] with read speech and the MuAViC dataset [18] with more challenging talking speech collected from TED and TEDx talks. On the MuAViC dataset, our retrained baseline can basically reproduce the results reported in [18]. As expected, the baseline model has a larger WER at an SNR of 0dB babble noise than with clean speech.

Table 1: ***Ablation study:*** *WER (%) of AV-ASR on the* ***German*** *(DE)* `dev` *and* `test` *splits of the CommonVoice (audio-only) and AV MuAViC datasets. Inference results with audio-only input (A) and audiovisual input (AV) are presented. Best results of* `dev` *and* `test` *splits are in* ***bold*** *font, second best are* <u>*underlined*</u>*.*

Training Protocol	Inference Modalities	CommonVoice		MuAViC			
		Clean		Clean		0dB Babble	
		`dev`	`test`	`dev`	`test`	`dev`	`test`
Baseline [18]	A	-	-	-	61.1	-	83.5
	AV	-	-	-	52.4	-	74.6
Baseline [18], retrained	A	82.4	84.5	54.0	57.2	81.4	81.3
	AV	-	-	51.6	54.5	72.3	71.1
Hong et al. [37]	A	-	-	-	-	-	-
	AV	-	-	-	48.1	-	65.9
2-Stage after Stage 1	A	**11.1**	**12.7**	46.5	48.2	68.9	68.8
	AV	-	-	46.6	48.6	69.9	67.0
2-Stage after Stage 2	A	32.2	35.6	36.3	39.2	68.8	67.4
	AV	-	-	35.6	38.8	63.1	62.4
Interleaved A:20%,AV:80%	A	19.5	21.7	36.9	41.2	68.3	67.6
	AV	-	-	35.7	40.2	58.9	59.7
Interleaved A:50%,AV:50%	A	14.7	16.4	31.7	36.2	63.2	63.1
	AV	-	-	30.6	35.5	<u>53.8</u>	<u>54.2</u>
Interleaved A:80%,AV:20%	A	12.9	14.7	31.8	35.6	63.1	62.3
	AV	-	-	**29.9**	<u>34.9</u>	**52.4**	**53.9**
Interleaved A:90%,AV:10%	A	<u>12.5</u>	<u>14.3</u>	33.3	36.2	64.4	63.2
	AV	-	-	<u>30.6</u>	**34.5**	55.7	54.6
A:100%	A	**11.1**	**12.7**	46.5	48.2	68.9	68.8
	AV	-	-	46.6	48.6	69.9	67.0

In addition, compared to audio-only input (A), we observe a larger WER reduction with AV input in the noisy condition than in the clean condition, showing the noise robustness nature of AV-ASR. Hong et al. [37] designed a language identifier in the `AV-HuBERT`-based multi-lingual AV-ASR, reporting a recent state of the art with 48.1% WER in clean condition and 65.9% WER at an SNR of 0dB babble noise.

The center table segment exhibits details of the 2-stage protocol. First, results after the first fine-tuning stage on the German audio-only data $\mathcal{D}_{\mathrm{A}}^{\mathrm{T}}$ are presented, which are equivalent to 100% audio-only input in the lower table segment. After the first stage, on the CommonVoice dataset, the model achieves the best results of 11.1% WER and 12.7% WER on the `dev` and `test` splits, respectively. Compared to the baseline, the model after the first fine-tuning stage already performs better on the MuAViC dataset, showing the benefits of rich target language knowledge from a large audio corpus $\mathcal{D}_{\mathrm{A}}^{\mathrm{T}}$ in the target language. However, the WERs with audiovisual inputs are close or even worse than with audio-only inputs on the MuAViC dataset, *revealing the catastrophic forgetting of the visual modality in the first stage*. Second, results after the second fine-tuning stage on the German AV data $\mathcal{D}_{\mathrm{AV}}^{\mathrm{T}}$ are presented. Expectedly, the performance on the in-domain MuAViC `dev` and `test` splits is improved. However, on the CommonVoice dataset, the performance drops significantly to 32.2% (vs. 11.1%) WER and 35.6% (vs. 12.7%) WER on the `dev` and `test` splits, respectively, *showing an overfitting on the small AV MuAViC dataset and degraded generalization ability*.

In the lower table segment, we perform an ablation study in our proposed interleaved protocol to find a suitable hyperparameter p for the probability of the audiovisual minibatch. Expectedly, the more probable the audio-only minibatch during interleaved training is, the better is the performance on the CommonVoice `dev` and `test` splits. On the AV MuAViC dataset, the interleaved protocol with a 20% probability for the AV mini-

Table 2: *WER (%) on the **German** (DE) and **French** (FR)* `test` *splits of the CommonVoice ($\mathcal{D}_{\mathrm{A}}^{\mathrm{T}}$) and the MuAViC ($\mathcal{D}_{\mathrm{AV}}^{\mathrm{T}}$) datasets. Models are evaluated with clean speech, and at various SNRs of music noise, second interfering speaker, and babble noise. Best results of each segment in the table are in **bold** font. * Numbers are taken from respective paper.*

Training AV Data	Training Protocol	WER (%) on CommonVoice (test on audio-only)	WER (%) on MuAViC (test on AV)										
			Music			2nd Speaker			Babble			Clean	Mean
			-5dB	0dB	5dB	-5dB	0dB	5dB	-5dB	0dB	5dB		
FR 176h = full	Baseline [18]*	-	-	32.7	-	-	32.5	-	-	48.1	-	23.7	-
	Baseline [18], retrained	47.7	**34.4**	24.5	19.5	32.3	23.6	19.5	**56.9**	**32.5**	22.4	15.3	28.1
	Hong et al. [37]*	-	-	-	-	-	-	-	-	37.4	-	21.5	-
	2-Stage	39.7	34.7	**23.4**	**18.7**	**32.0**	**22.7**	**18.7**	58.8	32.6	**21.4**	**14.8**	**27.8**
	Interleaved (AV:50%)	**19.4**	36.8	25.3	20.1	33.5	24.4	20.4	61.2	34.9	23.1	15.5	29.5
FR 10h = reduced	Baseline [18], retrained	92.6	73.3	63.9	58.8	71.9	65.3	60.2	85.2	71.0	61.8	52.5	66.4
	2-Stage	27.3	58.2	43.3	34.6	58.3	44.5	36.5	85.0	57.0	40.5	27.3	48.5
	Interleaved (AV:50%)	20.4	48.3	35.6	29.2	44.6	35.2	29.1	70.7	46.5	32.3	22.8	39.4
	Interleaved (AV:20%)	18.5	**45.8**	**32.8**	**26.6**	**43.4**	**32.9**	**27.0**	**70.3**	**44.1**	**30.5**	**21.0**	**37.4**
	Interleaved (AV:10%)	**18.0**	47.6	34.5	27.4	45.7	33.6	27.7	74.4	46.5	31.0	21.1	39.0
DE 10h = full	Baseline [18]*	-	-	-	-	-	-	-	-	74.6	-	52.4	-
	Baseline [18], retrained	84.5	73.3	65.8	61.0	73.3	67.1	62.5	84.4	71.1	63.3	54.5	67.6
	Hong et al. [37]*	-	-	-	-	-	-	-	-	65.9	-	48.1	-
	2-Stage	35.6	64.2	53.0	46.6	65.5	55.3	48.5	84.1	62.4	50.5	38.8	56.9
	Interleaved (AV:50%)	16.4	**54.4**	45.8	40.5	53.5	45.3	41.6	**72.5**	54.2	43.4	35.5	48.7
	Interleaved (AV:20%)	14.7	54.7	**45.7**	**40.0**	**53.4**	**44.6**	**40.0**	74.2	**53.9**	**42.9**	34.9	**48.4**
	Interleaved (AV:10%)	**14.3**	56.7	45.9	40.2	54.7	46.9	41.2	76.8	54.6	43.8	**34.5**	49.5

batch achieves the strongest performance on the `dev` splits in both clean and noisy conditions. We thus select p=0.2 (i.e., 20% AV minibatch) in our proposed interleaved protocol for further low-resourced experiments. With the best 52.4% and 53.9% WERs at an SNR of 0dB babble noise, the interleaved protocol (AV:20%) with AV input outperforms audio-only input with a 10.7% (vs. 63.1%) and an 8.4% (vs. 62.3%) absolute WER reduction on the `dev` and `test` splits, respectively. The 2-stage protocol achieves only about half of this improvement. Therefore, *our proposed interleaved protocol preserves more audiovisual interaction and empowers the noise robustness of the AV-ASR system*.

In Table 2, we evaluate the training protocols at various SNRs of music noise, second interfering talker, and babble noise. The mean WER over clean and all 9 noisy conditions is reported as well. In addition, we apply the training protocols to French with the full 176h AV training data, as shown in the top table segment, and with the reduced 10h AV training data in the center segment to simulate a low-resourced condition. In the lower table segment, we present results with the full 10h German (DE) training AV data. We observe that in the top table segment with enough French AV data, the 2-stage protocol shows the best performance in 7 out of 10 test conditions on the MuAViC dataset. However, the 2-stage protocol is much worse than the interleaved training on the CommonVoice `test` split.

In the center table segment with 10h French AV data, the interleaved protocol with 20% AV minibatches performs best in clean and all noisy conditions on the MuAViC dataset, validating the efficacy of the chosen hyperparameter in our ablation study in Table 1. Our interleaved protocol (AV:20%) surpasses the 2-stage protocol with a WER of 18.5% (vs. 27.3%) on the CommonVoice `test` split and with a 37.4% (vs. 48.5%) mean WER on the MuAViC `test` split. This corresponds to an 8.8% and an 11.1% absolute WER reduction, respectively, *proving a longer lasting and more generalizable learning of the interleaved protocol*.

In the lower table segment with the full 10h German AV data, our proposed interleaved protocol with 20% minibatches achieves the best performance in 7 out of 10 test conditions on the MuAViC dataset, achieving the best 48.4% mean WER on the MuAViC `test` split. In the interleaved protocol with varying probablities of AV minibatches, we observe that a higher AV minibatch probability (AV:50%) improves the performance in extremely noisy conditions, e.g., music and babble noise with an SNR of -5dB, showing the benefits of the visual modality in noisy conditions. Our interleaved protocol (AV:20%) excels the 2-stage protocol with a WER of 14.7% (vs. 35.6%) on the CommonVoice `test` split and with a 48.4% (vs. 56.9%) mean WER on the MuAViC `test` split, which means a 20.9% and an 8.5% absolute WER reduction, respectively. Even more, it outperforms the former state of the art by Hong et al. [37] both in noisy (babble 0dB, 53.9% vs. 65.9%) and in clean condition (34.9% vs. 48.1%) on the German MuAVIC `test` set.

5. Conclusions

In this work, we improve the performance of low-resourced AV-ASR systems by leveraging additional audio-only data in the target language for cross-language transfer learning. We also investigated effective training protocols to further enhance the AV-ASR system both in clean and noisy conditions. Our proposed minibatch-level interleaved training protocol is able to learn target language knowledge from the audio-only data effectively. At the same time, it preserves the audiovisual interaction learned in the pre-trained encoder, empowering the noise-robustness nature of AV-ASR systems. In the 10h resource-constrained setting, we significantly outperformed a 2-stage protocol which first fine-tunes towards target language audio data, then to the limited target AV data. In this setting, it also excels the former state of the art both in noisy (babble 0dB, 53.9% vs. 65.9%) and in clean condition (34.9% vs. 48.1%) on the German MuAVIC `test` set.

6. Acknowledgments

The research leading to these results has received funding from the Bundesministerium für Bildung und Forschung (BMBF) under funding code 03VP10991 (BesserLesen project).

7. References

[1] H. McGurk and J. MacDonald, "Hearing Lips and Seeing Voices," *Nature*, vol. 264, no. 5588, pp. 746–748, 1976.

[2] L. D. Rosenblum, "Speech Perception as a Multimodal Phenomenon," *Current Directions in Psychological Science*, vol. 17, no. 6, pp. 405–409, 2008.

[3] B. Shi, W.-N. Hsu, K. Lakhotia, and A. Mohamed, "Robust Self-Supervised Audio-Visual Speech Recognition," *arXiv:2201.02184*, Jul. 2022.

[4] P. Ma, S. Petridis, and M. Pantic, "End-To-End Audio-Visual Speech Recognition With Conformers," in *Proc. of ICASSP*, Toronto, ON, Canada, Jun. 2021, pp. 7613–7617.

[5] S. Receveur, R. Weiss, and T. Fingscheidt, "Turbo Automatic Speech Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 5, pp. 846–862, May 2016.

[6] Z. Li, C. Liang, T. Lohrenz, M. Sach, B. Möller, and T. Fingscheidt, "An Efficient and Noise-Robust Audiovisual Encoder for Audiovisual Speech Recognition," in *Proc. of Interspeech*, Dublin, Ireland, Aug. 2023, pp. 1583–1587.

[7] H. Chen, H. Zhou, J. Du, C.-H. Lee, J. Chen, S. Watanabe, S. M. Siniscalchi, O. Scharenborg, D.-Y. Liu, B.-C. Yin *et al.*, "The First Multimodal Information Based Speech Processing (MISP) Challenge: Data, Tasks, Baselines and Results," in *Proc. of ICASSP*, Singapore, May 2022, pp. 9266–9270.

[8] H. Wang, P. Guo, P. Zhou, and L. Xie, "MLCA-AVSR: Multi-Layer Cross Attention Fusion based Audio-Visual Speech Recognition," *arXiv:2401.03424*, Jan. 2024.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," in *Proc. of NIPS*, Long Beach, CA, USA, Dec. 2017, pp. 1–11.

[10] L. Dong, S. Xu, and B. Xu, "Speech-Transformer: A No-Recurrence Sequence-to-Sequence Model for Speech Recognition," in *Proc. of ICASSP*, Calgary, AB, Canada, Apr. 2018, pp. 5884–5888.

[11] T. Lohrenz, Z. Li, and T. Fingscheidt, "Multi-Encoder Learning and Stream Fusion for Transformer-Based End-to-End Automatic Speech Recognition," in *Proc. of Interspeech*, Brno, Czech Republic, Sep. 2021, pp. 2846–2850.

[12] T. Lohrenz, P. Schwarz, Z. Li, and T. Fingscheidt, "Relaxed Attention: A Simple Method to Boost Performance of End-to-End Automatic Speech Recognition," in *Proc. of ASRU*, Cartagena, Colombia, Dec. 2021, pp. 177–184.

[13] Z. Li, T. Lohrenz, M. Dunkelberg, and T. Fingscheidt, "Transformer-Based Lip-Reading with Regularized Dropout and Relaxed Attention," in *Proc. of SLT*, Doha, Qatar, Jan. 2023, pp. 723–730.

[14] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep Audio-Visual Speech Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–11, Dec. 2018, (early access).

[15] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip Reading Sentences in the Wild," in *Proc. of CVPR*, Honolulu, HI, USA, Jul. 2017, pp. 3444–3453.

[16] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep Speaker Recognition," in *Proc. of Interspeech*, Hyderabad, India, Sep. 2018, pp. 1086–1090.

[17] A. Zadeh, Y. S. Cao, S. Hessner, P. P. Liang, S. Poria, and L.-P. Morency, "CMU-MOSEAS: A Multimodal Language Dataset for Spanish, Portuguese, German and French," in *Proc. of EMNLP*, vol. 2017, virtual, Nov. 2020, pp. 1801–1812.

[18] M. Anwar, B. Shi, V. Goswami, W.-N. Hsu, J. Pino, and C. Wang, "MuAViC: A Multilingual Audio-Visual Corpus for Robust Speech Recognition and Robust Speech-to-Text Translation," in *Proc. of Interspeech*, Dublin, Ireland, Aug. 2023, pp. 4064–4068.

[19] C. Wang, J. Pino, and J. Gu, "Improving Cross-Lingual Transfer Learning for End-to-End Speech Recognition with Speech Translation," in *Proc. of Interspeech*, Shanghai, China, Oct. 2020, pp. 4731–4735.

[20] P. Ma, S. Petridis, and M. Pantic, "Visual Speech Recognition for Multiple Languages in the Wild," *arXiv:2202.13084*, Feb. 2022.

[21] C.-H. H. Yang, B. Li, Y. Zhang, N. Chen, R. Prabhavalkar, T. N. Sainath, and T. Strohman, "From English to More Languages: Parameter-Efficient Model Reprogramming for Cross-Lingual Speech Recognition," in *Proc. of ICASSP*, Rhodes, Greece, Jun. 2023, pp. 1–5.

[22] J. Peng, T. Stafylakis, R. Gu, O. Plchot, L. Mošner, L. Burget, and J. Černocký, "Parameter-Efficient Transfer Learning of Pre-Trained Transformer Models for Speaker Verification Using Adapters," in *Proc. of ICASSP*, Rhodes, Greece, Jun. 2023, pp. 1–5.

[23] Z. Li, T. Graave, J. Liu, T. Lohrenz, S. Kunzmann, and T. Fingscheidt, "Parameter-Efficient Cross-Language Transfer Learning for a Language-Modular Audiovisual Speech Recognition," in *Proc. of ASRU*, Taipei, Taiwan, Dec. 2023, pp. 1–8.

[24] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-Efficient Transfer Learning for NLP," in *Proc. of ICML*, Long Beach, CA, USA, Jun. 2019, pp. 2790–2799.

[25] D. Chen and B. K.-W. Mak, "Multitask Learning of Deep Neural Networks for Low-Resource Speech Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 7, pp. 1172–1183, 2015.

[26] Z. Xu, M. Strake, and T. Fingscheidt, "Deep Noise Suppression Maximizing Non-Differentiable PESQ Mediated by A Non-Intrusive PESQNet," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1572–1585, 2022.

[27] ——, "Does a PESQNet (Loss) Require a Clean Reference Input? The Original PESQ Does, But ACR Listening Tests Don't," in *Proc. of IWAENC*, Bamberg, Germany, Sep. 2022, pp. 1–5.

[28] J. Samani and S. C. Pan, "Interleaved Practice Enhances Memory And Problem-Solving Ability in Undergraduate Physics," *npj Science of Learning*, vol. 6, no. 1, p. 32, 2021.

[29] A. Pasad, J.-C. Chou, and K. Livescu, "Layer-Wise Analysis of a Self-Supervised Speech Representation Model," in *Proc. of ASRU*, Cartagena, Colombia, Dec. 2021, pp. 914–921.

[30] A. Pasad, B. Shi, and K. Livescu, "Comparative Layer-Wise Analysis of Self-Supervised Speech Models," in *Proc. of ICASSP*, Rhodes, Greece, Jun. 2023, pp. 1–5.

[31] M. Zeineldeen, A. Glushko, W. Michel, A. Zeyer, R. Schlüter, and H. Ney, "Investigating Methods to Improve Language Model Integration for Attention-Based Encoder-Decoder ASR Models," in *Proc. of Interspeech*, Brno, Czech Republic, Sep. 2021, pp. 2856–2860.

[32] T. Lohrenz, B. Möller, Z. Li, and T. Fingscheidt, "Relaxed Attention for Transformer Models," in *Proc. of IJCNN*, Gold Coast, QLD, Australia, Jun. 2023, pp. 1–10.

[33] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common Voice: A Massively-Multilingual Speech Corpus," *arXiv:1912.06670*, Mar. 2019.

[34] T. Afouras, J. S. Chung, and A. Zisserman, "LRS3-TED: A Large-Scale Dataset for Visual Speech Recognition," *arXiv:1809.00496*, Oct. 2018.

[35] T. Kudo and J. Richardson, "SentencePiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing," *arXiv:1808.06226*, Aug. 2018.

[36] D. Snyder, G. Chen, and D. Povey, "MUSAN: A Music, Speech, and Noise Corpus," *arXiv:1510.0848v1*, pp. 1–4, Oct. 2015.

[37] J. Hong, S. Park, and Y. Ro, "Intuitive Multilingual Audio-Visual Speech Recognition with a Single-Trained Model," in *Findings of the EMNLP*, Singapore, Dec. 2023, pp. 4886–4890.