Interspeech 2024
1-5 September 2024, Kos, Greece

Whisper-Flamingo: Integrating Visual Features into Whisper for Audio-Visual Speech Recognition and Translation

Andrew Rouditchenko[1], *Yuan Gong*[1], *Samuel Thomas*[2,3], *Leonid Karlinsky*[2,3], *Hilde Kuehne*[3,4], *Rogerio Feris*[2,3], *James Glass*[1]

[1]MIT, USA [2]IBM Research AI, USA [3]MIT-IBM Watson AI Lab, USA
[4]University of Bonn, Germany

roudi@mit.edu

Abstract

Audio-Visual Speech Recognition (AVSR) uses lip-based video to improve performance in noise. Since videos are harder to obtain than audio, the video training data of AVSR models is usually limited to a few thousand hours. In contrast, speech models such as Whisper are trained with hundreds of thousands of hours of data, and thus learn a better speech-to-text decoder. The huge training data difference motivates us to adapt Whisper to handle video inputs. Inspired by Flamingo which injects visual features into language models, we propose Whisper-Flamingo which integrates visual features into the Whisper speech recognition and translation model with gated cross attention. Our audio-visual Whisper-Flamingo outperforms audio-only Whisper on English speech recognition and En-X translation for 6 languages in noisy conditions. Moreover, Whisper-Flamingo is a versatile model and conducts all of these tasks using one set of parameters, while prior methods are trained separately on each language.

Index Terms: audio-visual speech recognition, noise-robust

1. Introduction

In recent years, major improvements in Automatic Speech Recognition (ASR) performance have been achieved by models trained on large-scale data [1, 2], but performance still declines in noise [3]. To enhance performance in noise, Audio-Visual Speech Recognition (AVSR) uses lip-based video in addition to audio inputs. Self-Supervised Learning (SSL) methods such as AV-HuBERT [4] pre-train on large-scale datasets of unlabeled videos and fine-tune on a few hundred hours of labeled videos to perform noise-robust AVSR. However, due to the difficulty in collecting publicly accessible videos, these models are usually trained on only a few thousand hours of data.

To overcome the lack of video data, recent methods fine-tune audio-only models pre-trained on hundreds of thousands of hours of audio for audio-visual speech recognition [5–7]. The results show that such audio models combined with audio-visual fine-tuning on a few hundred hours of videos can approach the performance of video models pre-trained on thousands of hours of video [6]. However, these methods often train video models and text decoders from scratch on only a few hundred hours of data, which is suboptimal compared to training on large-scale data. Furthermore, only English data has been used.

In this work, we propose to integrate visual features from AV-HuBERT into Whisper [1], an audio-only model trained on 680k hours of speech with a strong *multilingual* decoder. Compared to the prior audio-visual adaptation methods, our video model and text decoder are pre-trained with large-scale data. This allows our method to perform well on audio-visual speech translation, a task not explored by prior methods [5–7].

How to fuse modalities effectively in multi-modal models is an ongoing research question. One recent work, Flamingo [8], fuses visual features into text-only language models using gated cross attention and fine-tuning on a paired text-image dataset. The gated cross attention layers are initialized to the identity function and learn to attend to the visual features during fine-tuning. These layers have been shown to generalize to different modality pairs; Audio Flamingo [9] recently applied them for text-audio reasoning. Inspired by this method, we propose Whisper-Flamingo which inserts gated cross attention layers into Whisper's decoder and enables Whisper to use lip-based features for speech recognition.

Our experiments on the English (En) LRS3 video dataset [10] show that our novel audio-visual Whisper-Flamingo significantly outperforms the audio-only Whisper baseline in noise. Moreover, Whisper-Flamingo achieves competitive noise-robust results compared to prior audio-visual models. Next, we demonstrate Whisper's multilingual capabilities by extending Whisper-Flamingo for En-X translation on the MuAViC dataset [11]. Our model performs En transcription and En-X translation into 6 other languages, while the previous audio-visual SOTA requires fine-tuning on each language separately. Once again, Whisper-Flamingo significantly outperforms audio-only Whisper in noise for both En transcription and En-X translation. Code and models at https://github.com/roudimit/whisper-flamingo

2. Method

In this section, we review audio-visual fusion methods for AVSR, and then explain our method. Two common fusion methods are early and late fusion. In early fusion, both modalities are first separately processed by light-weight encoders and then combined with feature addition or concatenation and used as input to an audio-visual Transformer [12, 13]. Both SSL models [4, 14, 15] and fully-supervised models [16–18] use this design. In late fusion, audio and video are processed separately by Transformer encoders, and afterwards features are fused with an MLP. The audio-visual features are then passed to a linear-layer or Transformer decoder. This approach is common for fully-supervised models [19–21]. Both early and late fusion need identical audio and visual feature rates so that they can be fused at each time step; a common design is to match the video's frame rate by downsampling the audio features to 25 Hz.

Most methods which adapt pre-trained audio-only models for AVSR through audio-visual fine-tuning use early fusion. FAVA [6] adapts BEST-RQ [22], an audio self-supervised model, through early fusion with a video model trained from scratch. Adaptive AV [7] prepends Whisper with an audio-visual Transformer to output a de-noised spectrogram, but does

10.21437/Interspeech.2024-322

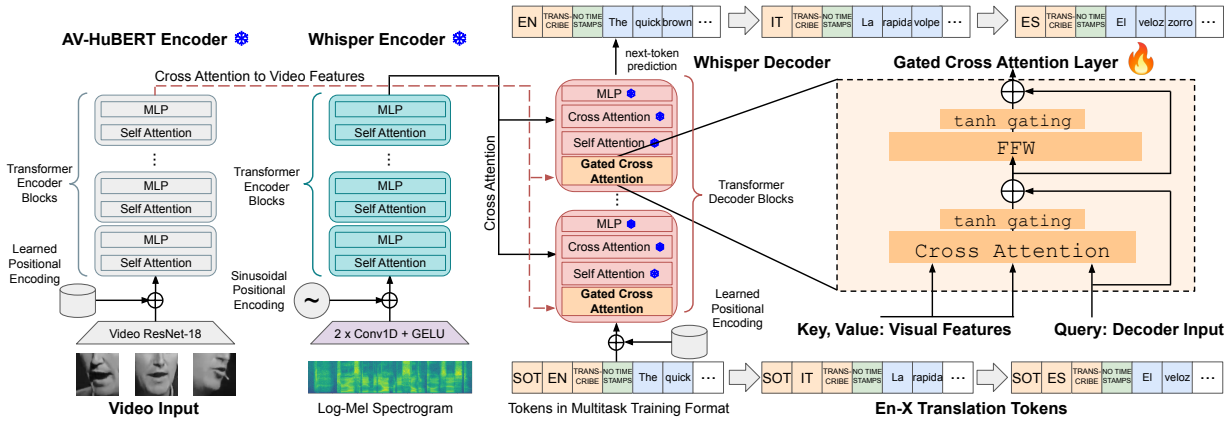

Figure 1: *Diagram of Whisper-Flamingo based on Whisper [1] and Flamingo [8]. We first fine-tune all of Whisper's parameters using English audio for English transcription and En-X translation. To train Whisper-Flamingo, we freeze the audio model, add gated cross attention layers into Whisper's decoder attending to visual features from AV-HuBERT, and train the model on audio-visual inputs.*

Table 1: *Hyperparameter summary. We first train Whisper-Large FT (Fine-tune) with audio-only, then use it to initialize Whisper-Flamingo. A=audio, AV=audio-visual. †=per sample.*

	Whisper-Large FT	Whisper-Large FT	Whisper-Flamingo	Whisper-Flamingo
Test Modalities	A	A	AV	AV
En Recognition	✓	✓	✓	✓
En-X Translation	✗	✓	✗	✓
GPUs	1	4	1	4
Total Params.	1.55B	1.55B	2.5B	2.5B
AV-HuBERT Params.	-	-	325M	325M
Gated X-Attn Params.	-	-	630M	630M
Trainable Params.	1.55B	1.55B	631M	631M
Warmup Steps	1k	1k	5k	5k
Total Steps	90k	225k	20k	40k
Learning Rate	5×10^{-6}	5×10^{-6}	1×10^{-4}	1×10^{-4}
Batch per GPU (s)	80	40	160	30
Max Length (s) †	10	10	15	15
Max Characters †	350	300	350	250

not use visual features directly in Whisper. However, we found gated cross attention with features from pre-trained AV-HuBERT worked better than early fusion. Note that separate research focuses on using visual features from images or instructional videos for AVSR, where the visuals provide context and are only loosely synchronized with the audio [23, 24].

We propose to adapt Whisper's decoder with visual features from AV-HuBERT using gated cross attention, as shown in Figure 1. Each of Whisper's decoder blocks consists of a self-attention layer, cross attention layer attending to the audio features, and a Multi-Layer Perceptron (MLP). Based on Flamingo [8], the gated cross attention layer is defined as follows, where $\mathbf{x}$ is the input to the decoder block, $\mathbf{v}$ are the visual features, Attn is multi-head cross attention, LN is Layernorm [25], and FFW is an MLP:

$$\mathbf{x}' = \mathbf{x} + \tanh(\alpha_{\text{xattn}}) \times \text{Attn}(\text{LN}(\mathbf{x}), \mathbf{v}) \quad (1)$$

$$\mathbf{y} = \mathbf{x}' + \tanh(\alpha_{\text{mlp}}) \times \text{FFW}(\text{LN}(\mathbf{x}')) \quad (2)$$

The learnable parameters α_{xattn} and α_{mlp} are initialized to 0 so that the layers initially function as the identity since $\tanh(0) = 0$. Through audio-visual fine-tuning, the model adjusts the weights of α_{xattn} and α_{mlp} and learns to attend to the visual features. We insert the gated cross attention layers in Whisper's decoder in the beginning of each block, before the self-attention layer. We tried to insert them in other orders within the decoder blocks, but the performance was slightly worse. We also tried inserting them into the encoder, but the performance was significantly worse. Note that since the gated cross attention separately attends to the video features, the audio and video features can have different feature rates (for example, 50 Hz and 25 Hz).

Table 2: *Fusion ablation with Whisper-Medium on LRS3. We report results on the original test set (Clean) and with babble noise injected at 0-SNR (Noisy). A=audio, AV=audio-visual.*

Model	Test Modalities	Clean WER↓	Noisy WER↓
Whisper, Zero-shot	A	2.3	22.2
Whisper, Fine-tuned	A	**1.9**	**12.6**
Whisper-Early-Fusion	AV	1.7	10.0
Whisper-Late-Fusion	AV	2.1	16.5
Whisper-Flamingo	AV	**1.5**	**7.0**

Training pipeline. Before adding gated cross attention, we first fine-tune all layers of the audio-only Whisper model to adapt it to the domain of interest (denoted as Whisper Fine-tuned). We also add noise during fine-tuning to increase the noise-robustness. We use the standard cross-entropy loss between the model's predicted transcript and the ground-truth tokens. To train Whisper-Flamingo, we *freeze* the fine-tuned Whisper, insert the gated cross attention layers, and fine-tune the model with audio-visual inputs. The gated cross attention layers and a linear layer on top of the visual features are trained from scratch, while all other parameters are frozen. The new layers can therefore be seen as a (large) set of adaptors [27]: removing them results in the audio-only Whisper weights.

From English to Multilingual. Whisper was trained for multilingual transcription and X-En translation (multilingual audio to En text). We tried Whisper-Flamingo on multilingual speech recognition and X-En translation using the videos in the MuAViC dataset [11] but found several issues. Most languages in the dataset have less than a third of the hours of English data available, which makes training new layers from scratch difficult. Also, the multilingual videos are longer on average than the English videos. This causes increased GPU memory pressure and requires a reduced batch size, which also makes training difficult. Therefore we focused on *En-X* translation (English audio to multilingual text) and propose to handle multilingual recognition and translation in future work [28–32].

Prior research shows that Whisper can be prompted for En-X translation, but it requires language-specific logit filtering and the performance can still be unsatisfactory [24]. Since fine-

Table 3: *Results for English transcription on LRS3. We report results on the original test set (Clean) and with babble noise added at 0-SNR (Noisy). A= Audio, AV= Audio-Visual. Noise dataset= dataset used to make babble noise. Hours of unlabeled / labeled data used to train each model are shown. Note† that u-HuBERT [14] and AV-HuBERT [26] use a different noise file than us.*

Model	Test Modalities	Noise Dataset	Unlabeled Hrs A	Unlabeled Hrs AV	Labeled Hrs A	Labeled Hrs AV	WER↓ Clean	WER↓ Noisy
Audio-Visual SSL Methods								
AV-BEST-RQ [6]	AV	NoiseX	-	1759	-	433	2.1	6.8
AV2vec [15]	AV	MUSAN	-	433	-	433	2.5	6.7
AV-HuBERT [26]	AV	LRS3†	-	1759	-	433	1.4	5.8
u-HuBERT [14]	AV	LRS3†	-	1759	-	433	**1.3**	**4.6**
Audio-Only Pre-train + Audio-Visual Fine-Tune Methods								
Adaptive AV [7]	AV	MUSAN	-	400	680k	30	2.3	16.3
FAVA [6]	AV	NoiseX	1759	-	-	433	1.7	6.6
FAVA-USM [6]	AV	NoiseX	12M	-	5000	433	**1.3**	**6.2**
Our Audio-Only Whisper Baselines								
Whisper-Large, Zero-shot (No Fine-Tuning)	A	LRS3	-	-	680k	-	**2.1**	20.8
Whisper-Large, Fine-tuned on LRS3	A	LRS3	-	-	680k	-	2.3	**11.7**
Proposed Audio-Visual Fine-tuning Method								
Whisper-Flamingo (**Ours**)	AV	LRS3	-	1759	680k	433	**1.5**	**5.6**

tuning Whisper has been shown to enable transcription of unseen languages [33], we propose to fine-tune Whisper for En-X translation. We fine-tune the audio model in a multi-task style to transcribe English audio and translate it to the other languages. To train Whisper-Flamingo, we freeze the fine-tuned audio model, add the gated cross attention layers and the linear layer on top of the visual features, and train the model on audio-visual inputs.

3. Experiments

3.1. Experimental Setup

To train our models, we use LRS3 [10] - the largest, publicly-available AVSR dataset in English (En), sourced from TED talks. We followed AV-HuBERT [4] to create a 433h training set, 1h validation set, and 1h test set. For En-X translation, we use the MuAViC [11] dataset which has translations of LRS3's English text into 6 languages: Greek (El), Spanish (Es), French (Fr), Italian (It), Portuguese (Pt), and Russian (Ru).

We use Whisper Small, Medium, and Large-v2 with 244M, 769M, and 1.55B parameters [1]. We extract 80-bin log-Mel spectrograms with a stride of 10ms and window size of 25ms from audio sampled at 16kHz. We extract video features from the AV-HuBERT Large [4] encoder fine-tuned on LRS3 with 325M parameters. For Whisper Large, the gated cross attention layers add 630M parameters, bringing the total number of parameters to 2.5B (including AV-HuBERT). We freeze AV-HuBERT but enable dropout and batch normalization updating during Whisper-Flamingo training. The videos have a frame rate of 25fps and are converted to grayscale. Dlib [35] is used to extract 96x96 crops centered on the lips which are aligned to a reference mean face [36]. During training, a random 88x88 crop is used and the video is flipped horizontally with probability 0.5. For testing, the center 88x88 crop is used.

Table 1 summarizes the hyperparameters for the main experiments. We used A6000 GPUs with 48GB memory. Audio/video samples with similar lengths are batched together, and short samples are 0-padded. AdamW was used as the optimizer [37]. Following [1], we used SpecAugment [38] (Librispeech-Basic) with Whisper-Large and did not use it with Whisper-Medium. Training was done with PyTorch [39] and PyTorch Lightning [40]. We used the SpecAugment and batch sorter implementations from ESPnet [41].

During training, we randomly add noise to the audio with a signal-to-noise ratio (SNR) of 0. Following prior work [11, 26], the "natural", "music" and "babble" noise are sampled from the MUSAN dataset [42], and overlapping "speech" noise is sampled from LRS3 [10]. To select the best checkpoints, we monitor the highest token prediction accuracy on the noisy validation set every 1k steps. We report beam search decoding results with beam size 15. Following prior work [11], we use the Fairseq normalizer [43] to remove punctuation and lower-case text before calculating WER. For translation, we use SacreBLEU [44] with the default 13-a tokenizer to calculate BLEU [45].

3.2. Modality Fusion Ablation with Whisper-Medium

We first compared gated cross attention to early and late fusion using Whisper Medium. For early fusion, we duplicate AV-HuBERT's 25 Hz video features to temporally align them with Whisper's 50 Hz audio features (after the CNN layers) and use addition to fuse them before Whisper's Transformer encoder. For late fusion, we use an MLP to fuse the video features with Whisper's audio features after its Transformer encoder. In both cases, all of Whisper's parameters are fine-tuned. For audio-only baselines, we use Whisper zero-shot (no fine-tuning) and fine-tuned on LRS3. We test models in both the clean and noisy conditions with babble-noise injected at 0-SNR. The results are shown in Table 2. Fine-tuning audio-only Whisper decreases the noisy WER of the zero-shot model from 22.2% to 12.6%. We then use the fine-tuned model as initialization to train the models with audio-visual fusion. Early-fusion obtained a small improvement in both the clean and noisy WERs. Late-fusion could not fuse the modalities well and performance became worse in both clean and noisy conditions. Finally, Whisper-Flamingo with gated cross attention obtained the best noisy WER, significantly improving the audio-only Whisper fine-tuned baseline from 12.6% to 7.0%, while the clean WER was slightly improved from 1.9% to 1.5%. Freezing Whisper helps retain its strong audio skills while new cross attention layers enable it to integrate the visual modality more effectively.

3.3. Whisper-Flamingo English Speech Recognition

For our main experiments, we scale up to Whisper-Large. In the 3rd section of Table 3, we compare audio-only Whisper-Large zero-shot and fine-tuned on LRS3; the fine-tuned model outperforms the zero-shot model in noise (11.7% vs 20.8%). Fine-tuned Whisper-Large performs slightly worse than Whisper-Medium on clean audio (Table 2), but Whisper-Large performs better on noisy audio. The bottom part of Table 3 shows our audio-visual Whisper-Flamingo initialized from fine-tuned Whisper-Large. Compared to the audio-only baseline, Whisper-

Table 4: *Results for English transcription on LRS3 and En-X Translation on MuAViC. Babble noise is added at 0-SNR (Noisy). One Model= the model translates to all languages with one set of parameters. Test Mod.= inference modalities (Text: T, audio: A, audio-visual: AV). Note† that Bilingual AV-HuBERT [11] use a different noise file than us that was not publicly available.*

	Test	One	Noise	En	El	Es	Fr	It	Pt	Ru	Avg
Model	Mod.	Model	Dataset	WER↓	BLEU↑						w/o En
Text-to-Text Translation											
Bilingual Transformer [11]	T	✗	-	-	**25.8**	**29.5**	**27.0**	**22.6**	**23.9**	**17.2**	**24.3**
M2M-100 [11, 34]	T	✓	-	-	24.5	28.7	25.6	21.8	22.2	15.8	23.1
Speech-to-Text Translation (Clean Audio)											
Bilingual AV-HuBERT [11]	A	✗	-	-	23.0	27.5	25.1	20.7	20.1	14.7	21.9
Whisper-Small, Fine-tuned	A	✓	-	2.0	22.4	27.1	24.9	20.9	**21.6**	15.6	22.1
Whisper-Medium, Fine-tuned	A	✓	-	2.1	22.9	27.5	**26.1**	**21.9**	21.4	15.1	22.5
Whisper-Large, Fine-tuned	A	✓	-	**1.5**	**23.7**	**27.9**	26.0	21.8	21.4	**15.7**	**22.7**
Bilingual AV-HuBERT [11]	AV	✗	-	-	23.4	26.6	25.3	20.7	20.5	14.6	21.9
(**Ours**) Whisper-Flamingo (Small)	AV	✓	-	2.0	22.6	27.0	24.7	20.7	21.3	15.5	22.0
(**Ours**) Whisper-Flamingo (Medium)	AV	✓	-	1.6	23.8	**28.0**	**26.1**	**22.5**	21.3	**16.0**	**23.0**
(**Ours**) Whisper-Flamingo (Large)	AV	✓	-	**1.3**	**24.4**	27.9	25.9	22.1	**21.8**	15.7	22.9
Speech-to-Text Translation (Noisy Audio from MuAViC)											
Bilingual AV-HuBERT [11]	A	✗	MuAViC†	-	15.9	19.2	17.1	12.9	14.4	10.3	15.0
Whisper-Small, Fine-tuned	A	✓	MuAViC	17.3	17.5	20.1	19.4	15.3	16.3	11.8	16.7
Whisper-Medium, Fine-tuned	A	✓	MuAViC	14.8	18.1	22.1	19.8	16.2	17.3	12.1	17.6
Whisper-Large, Fine-tuned	A	✓	MuAViC	**13.8**	**19.7**	**23.4**	**20.4**	**17.4**	**17.7**	**13.3**	**18.6**
Bilingual AV-HuBERT [11]	AV	✗	MuAViC†	-	**22.7**	24.8	**23.8**	**20.0**	**20.0**	13.7	**20.8**
(**Ours**) Whisper-Flamingo (Small)	AV	✓	MuAViC	10.7	19.0	22.1	21.1	17.1	18.3	13.2	18.5
(**Ours**) Whisper-Flamingo (Medium)	AV	✓	MuAViC	8.3	20.7	24.5	21.6	18.8	18.6	13.7	19.6
(**Ours**) Whisper-Flamingo (Large)	AV	✓	MuAViC	**7.2**	21.1	**25.4**	22.4	19.3	19.9	**14.7**	20.5

Flamingo significantly improves the noisy performance from **11.7% to 5.6% (52.1% relative improvement).** It also improves the clean WER from 2.3% to 1.5%.

Table 3 also shows a comparison with prior audio-visual SSL methods and audio-visual fine-tuning methods on LRS3. Direct comparison in noisy conditions is challenging since different noise datasets were used to generate the babble noise. SSL methods AV-HuBERT [26] and u-HuBERT [14] used LRS3 to generate babble noise, but the noise file they generated was not publicly available. We followed their procedure to generate the noise, so our noisy conditions are similar but not identical. Compared with AV-HuBERT, Whisper-Flamingo achieves comparable clean performance (1.5% vs 1.4%) and slightly better noisy results (5.6% vs 5.8%), which shows that Whisper-Flamingo is effective at adapting Whisper to the visual features from AV-HuBERT. Moreover, a major advantage of Whisper-Flamingo over AV-HuBERT is improved translation performance (Section 3.4). The best noisy performance (4.6%) is reported by u-HuBERT; we would like to try it as a visual encoder for Whisper-Flamingo, but the weights are not publicly available. Finally, Whisper-Flamingo outperforms other methods in noise which adapt audio-only models through audio-visual fine-tuning [6, 7], including FAVA-USM [6] which was pre-trained on 12M hours of unlabeled audio [2]. However, the babble noise was generated from different datasets making results not strictly comparable.

3.4. Whisper-Flamingo En-X Speech Translation

Audio Results. In Table 4, we show the result of fine-tuning audio-only Whisper-Large for En-X translation using the 6 languages in the MuAViC dataset ("Whisper-Large, Fine-tuned"). Although Whisper was not originally trained for En-X translation, it adapts well to the new task. Testing with clean audio, we achieve an **average BLEU score of 22.7**, which outperforms the previous SOTA of 21.9 from Bilingual AV-HuBERT. Moreover, our model transcribes En audio (WER of 1.5%) and translates to 6 languages with a single set of parameters, while Bilingual AV-HuBERT fine-tunes separately for each language pair and trains language-specific decoders from scratch. Our model nearly reaches the text-to-text performance from machine translation models using the ground-truth English text; those models achieve average BLEU scores of 23.1 from a multilingual model and 24.3 from bilingual models.

Audio-Visual Results. Once we fine-tune audio-only Whisper for En-X translation, we use it to train Whisper-Flamingo by freezing the weights and adding gated cross attention layers. Testing with clean audio, Whisper-Flamingo slightly outperforms the audio-only model with an average BLEU score of 22.9 and En WER of 1.3%. In noisy conditions, we use multilingual babble noise constructed following MuAViC [11] by adding audio in 9 different languages from 30 speakers. Note that their noise file was not publicly available, so our noisy conditions are similar but not identical. With multilingual babble noise, Whisper-Flamingo significantly outperforms the audio-only Whisper model in average BLEU score (**20.5 vs 18.6**) and En WER (**7.2% vs 13.8%**). Compared with the previous SOTA bilingual AV-HuBERT, our audio-only average BLEU is much better (18.6 vs 15.0), but our audio-visual performance is slightly worse (20.5 vs 20.8). However, our models perform **both En-X translation and En transcription with a single model**, while their models fine-tune separately for each language pair. Finally, we show the results using Whisper-Medium and Whisper-Small: Whisper-Flamingo always does better in noise compared to the audio-only baselines, and performance tends to improve as the model size increases.

4. Conclusion

We introduced Whisper-Flamingo, a novel audio-visual model that combines the strengths of AV-HuBERT and Whisper using gated cross attention. Our audio-visual Whisper-Flamingo significantly outperforms audio-only Whisper in noise. We showed that Whisper can be fine-tuned for the new task of X-En translation. Our model performs both En speech recognition and En-X speech translation using one set of parameters while previous methods fine-tune separately on each language. Our method is a generic way of fusing a visual encoder into the decoder of an ASR model to enable AVSR, and it could work with other models trained on more data in the future.

5. Acknowledgments

We thank Alex H. Liu, Mohamed Anwar, and the reviewers for helpful discussion. This research was supported by the MIT-IBM Watson AI Lab and an NDSEG Fellowship to A.R.

6. References

[1] A. Radford *et al.*, "Robust speech recognition via large-scale weak supervision," in *ICML*, 2023.

[2] Y. Zhang *et al.*, "Google usm: Scaling automatic speech recognition beyond 100 languages," *arXiv preprint*, 2023.

[3] Y. Gong, S. Khurana, L. Karlinsky, and J. Glass, "Whisper-AT: Noise-Robust Automatic Speech Recognizers are Also Strong General Audio Event Taggers," in *Interspeech*, 2023.

[4] B. Shi, W.-N. Hsu, K. Lakhotia, and A. Mohamed, "Learning audio-visual speech representation by masked multimodal cluster prediction," in *ICLR*, 2022.

[5] X. Pan, P. Chen, Y. Gong, H. Zhou, X. Wang, and Z. Lin, "Leveraging unimodal self-supervised learning for multimodal audio-visual speech recognition," in *ACL*, 2022.

[6] A. May, D. Serdyuk, A. P. Shah, O. Braga, and O. Siohan, "Audio-visual fine-tuning of audio-only asr models," *ASRU*, 2023.

[7] C. Simic and T. Bocklet, "Self-supervised adaptive av fusion module for pre-trained asr models," in *ICASSP*, 2024.

[8] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds *et al.*, "Flamingo: a visual language model for few-shot learning," *NeurIPS*, 2022.

[9] Z. Kong *et al.*, "Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities," *arXiv preprint*, 2024.

[10] T. Afouras, J. S. Chung, and A. Zisserman, "Lrs3-ted: a large-scale dataset for visual speech recognition," *arXiv preprint*, 2018.

[11] M. A. *et al.*, "MuAViC: A Multilingual Audio-Visual Corpus for Robust Speech Recognition and Robust Speech-to-Text Translation," in *Interspeech*, 2023.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *NeurIPS*, 2017.

[13] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE TPAMI*, 2018.

[14] W.-N. Hsu and B. Shi, "u-hubert: Unified mixed-modal speech pretraining and zero-shot transfer to unlabeled modality," *NeurIPS*, 2022.

[15] J.-X. Zhang, G. Wan, Z.-H. Ling, J. Pan, J. Gao, and C. Liu, "Self-supervised audio-visual speech representations learning by multimodal self-distillation," in *ICASSP*, 2023.

[16] T. Makino, H. Liao, Y. Assael, B. Shillingford, B. Garcia, O. Braga, and O. Siohan, "Recurrent neural network transducer for audio-visual speech recognition," in *ASRU*, 2019.

[17] D. Serdyuk, O. Braga, and O. Siohan, "Transformer-Based Video Front-Ends for Audio-Visual Speech Recognition for Single and Muti-Person Video," in *Interspeech*, 2022.

[18] A. Rouditchenko, R. Collobert, and T. Likhomanenko, "Av-cpl: Continuous pseudo-labeling for audio-visual speech recognition," *arXiv preprint*, 2023.

[19] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic, "End-to-end audiovisual speech recognition," in *ICASSP*, 2018.

[20] P. Ma, S. Petridis, and M. Pantic, "End-to-end audio-visual speech recognition with conformers," in *ICASSP*, 2021.

[21] P. Ma, A. Haliassos, A. Fernandez-Lopez, H. Chen, S. Petridis, and M. Pantic, "Auto-avsr: Audio-visual speech recognition with automatic labels," in *ICASSP*, 2023.

[22] C.-C. Chiu, J. Qin, Y. Zhang, J. Yu, and Y. Wu, "Self-supervised learning with random-projection quantizer for speech recognition," in *ICML*, 2022.

[23] P. H. Seo, A. Nagrani, and C. Schmid, "Avformer: Injecting vision into frozen speech models for zero-shot av-asr," in *CVPR*, 2023.

[24] P. Peng, B. Yan, S. Watanabe, and D. Harwath, "Prompting the Hidden Talent of Web-Scale Speech Models for Zero-Shot Task Generalization," in *Interspeech*, 2023.

[25] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint*, 2016.

[26] B. Shi, W.-N. Hsu, and A. Mohamed, "Robust Self-Supervised Audio-Visual Speech Recognition," in *Interspeech*, 2022.

[27] 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 *ICML*, 2019.

[28] X. Cheng *et al.*, "Mixspeech: Cross-modality self-learning with audio-visual stream mixup for visual speech translation and recognition," in *ICCV*, 2023.

[29] J. Hong, S. Park, and Y. Ro, "Intuitive multilingual audio-visual speech recognition with a single-trained model," in *Findings of EMNLP*, 2023.

[30] Z. Li *et al.*, "Parameter-efficient cross-language transfer learning for a language-modular audiovisual speech recognition," in *ASRU*, 2023.

[31] M. Burchi *et al.*, "Multilingual audio-visual speech recognition with hybrid ctc/rnn-t fast conformer," in *ICASSP*, 2024.

[32] H. Han, M. Anwar, J. Pino, W.-N. Hsu, M. Carpuat, B. Shi, and C. Wang, "Xlavs-r: Cross-lingual audio-visual speech representation learning for noise-robust speech perception," *ACL*, 2024.

[33] A. Rouditchenko, S. Khurana, S. Thomas, R. Feris, L. Karlinsky, H. Kuehne, D. Harwath, B. Kingsbury, and J. Glass, "Comparison of Multilingual Self-Supervised and Weakly-Supervised Speech Pre-Training for Adaptation to Unseen Languages," in *Interspeech*, 2023.

[34] A. Fan, S. Bhosale, H. Schwenk, Z. Ma, A. El-Kishky, S. Goyal, M. Baines, O. Celebi, G. Wenzek, V. Chaudhary *et al.*, "Beyond english-centric multilingual machine translation," *JMLR*, 2021.

[35] D. E. King, "Dlib-ml: A machine learning toolkit," *The Journal of Machine Learning Research*, 2009.

[36] B. Martinez, P. Ma, S. Petridis, and M. Pantic, "Lipreading using temporal convolutional networks," in *ICASSP*, 2020.

[37] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *ICLR*, 2019.

[38] D. S. P. *et al.*, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Interspeech*, 2019.

[39] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *NeurIPS*, 2019.

[40] W. Falcon and The PyTorch Lightning team, "PyTorch Lightning," 2023.

[41] S. W. *et al.*, "ESPnet: End-to-End Speech Processing Toolkit," in *Interspeech*, 2018.

[42] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint*, 2015.

[43] C. Wang, Y. Tang, X. Ma, A. Wu, D. Okhonko, and J. Pino, "Fairseq S2T: Fast speech-to-text modeling with fairseq," in *AACL: System Demonstrations*, 2020.

[44] M. Post, "A call for clarity in reporting BLEU scores," in *WMT*, 2018.

[45] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002.