mWhisper-Flamingo for Multilingual Audio-Visual Noise-Robust Speech Recognition

Andrew Rouditchenko, Samuel Thomas, *Senior Member, IEEE*, Hilde Kuehne, *Member, IEEE*, Rogerio Feris, *Senior Member, IEEE*, and James Glass, *Fellow, IEEE*

***Abstract*—Audio-Visual Speech Recognition (AVSR) combines lip-based video with audio and can improve performance in noise, but most methods are trained only on English data. One limitation is the lack of large-scale multilingual video data, which makes it hard to train models from scratch. In this work, we propose mWhisper-Flamingo for multilingual AVSR which combines the strengths of a pre-trained audio model (Whisper) and video model (AV-HuBERT). To enable better multi-modal integration and improve the noisy multilingual performance, we introduce decoder modality dropout where the model is trained both on paired audio-visual inputs and separate audio/visual inputs. mWhisper-Flamingo achieves state-of-the-art WER on MuAViC, an AVSR dataset of 9 languages. Audio-visual mWhisper-Flamingo consistently outperforms audio-only Whisper on all languages in noisy conditions.**

***Index Terms*—Audio-visual speech recognition (AVSR), multilingual.**

I. Introduction

AUTOMATIC Speech Recognition (ASR) has seen great progress thanks to large-scale training [1], [2], [3], [4], but models still struggle with background noise [5], [6]. To improve performance, Audio-Visual Speech Recognition (AVSR) models combine lip-based video with audio inputs [7], [8], [9], [10], [11], [12], [13], [14], [15], [16]. In particular, Whisper-Flamingo [17] proposed an audio-visual adaptation of Whisper [1], a pre-trained ASR model, and showed significant improvements in noise robustness compared to the original audio-only model. Whisper-Flamingo was motivated by the limitation of previous AVSR systems which often lack large-scale transcribed videos and face difficulty training models from scratch on only a few hundred hours of data. To overcome this, Whisper-Flamingo combines the strength of Whisper's audio encoder and text decoder trained on 680 k hours with AV-HuBERT [18], a pre-trained lip-reading visual encoder. The model integrates lip-based visual features from AV-HuBERT into Whisper's decoder and achieves State-of-the-Art (SOTA) performance on English AVSR.

Received 12 February 2025; accepted 3 May 2025. Date of publication 12 May 2025; date of current version 28 May 2025. This work was supported in part by the MIT-IBM Watson AI Lab. The work of Andrew Rouditchenko was supported in part by the NDSEG Fellowship. The associate editor coordinating the review of this article and approving it for publication was Dr. Anurag Kumar. *(Corresponding author: Andrew Rouditchenko.)*

Andrew Rouditchenko and James Glass are with MIT, Cambridge, MA 02139 USA (e-mail: roudi@mit.edu; glass@mit.edu).

Samuel Thomas and Rogerio Feris are with MIT-IBM Watson AI Lab, Cambridge, MA 02142 USA.

Hilde Kuehne is with the University of Tuebingen, 72074 Tuebingen, Germany.

Code at https://github.com/roudimit/whisper-flamingo

Digital Object Identifier 10.1109/LSP.2025.3569210

In this work, we propose mWhisper-Flamingo: a novel multilingual extension of Whisper-Flamingo which achieves SOTA performance on multilingual AVSR. mWhisper-Flamingo combines Whisper's strong multilingual audio encoder and text decoder with a new AV-HuBERT visual encoder pre-trained on multilingual videos [19]. Unlike the previous Whisper-Flamingo model which could only take English videos as input, the new mWhisper-Flamingo model can handle videos in 9 different languages (including English). However, we show that Whisper-Flamingo's default training process applied to multilingual videos yields poor noisy multilingual AVSR performance, despite achieving good English performance. To address this, we propose a novel decoder modality dropout technique by training the model both on paired audio-visual inputs and separate audio/video inputs. We show this to be key for good noisy multilingual AVSR performance with a thorough analysis and ablation study.

We test our method across 9 languages in clean and noisy conditions on the MuAViC dataset [20]. In clean audio conditions, mWhisper-Flamingo outperforms previous audio-visual methods and achieves SOTA across multilingual languages. In noisy conditions, mWhisper-Flamingo consistently outperforms the audio-only Whisper model on 6 different noise types and 5 levels of noise. We release our code and models.

II. Method

Our method builds upon Whisper-Flamingo [17], an audio-visual extension of Whisper [1]. Whisper-Flamingo adds new gated cross-attention layers (originally proposed for the Flamingo vision-language model [21]) into each of Whisper's decoder blocks which attend to the visual features from the AV-HuBERT visual encoder [18]. The layers are initialized as identity functions and initially ignore the visual input, but the weights are adjusted to integrate the visual features for AVSR during training on audio-visual inputs.

Whisper-Flamingo uses two-stages of training. First, all of Whisper's parameters are fine-tuned on noisy audio inputs, enhancing its domain-specific performance and noise robustness. Second, the gated cross-attention layers are initialized and trained on audio-visual data. During this stage, all of Whisper's and AV-HuBERT's parameters are frozen to preserve the pre-trained knowledge and to facilitate multi-modal integration.

mWhisper-Flamingo inherits Whisper-Flamingo's architecture, except the English AV-HuBERT [18] is replaced with a version pre-trained on multilingual videos [19]. We propose to train mWhisper-Flamingo on multilingual videos jointly with English, similar to Whisper's training process. However, using Whisper-Flamingo's default training process on multilingual

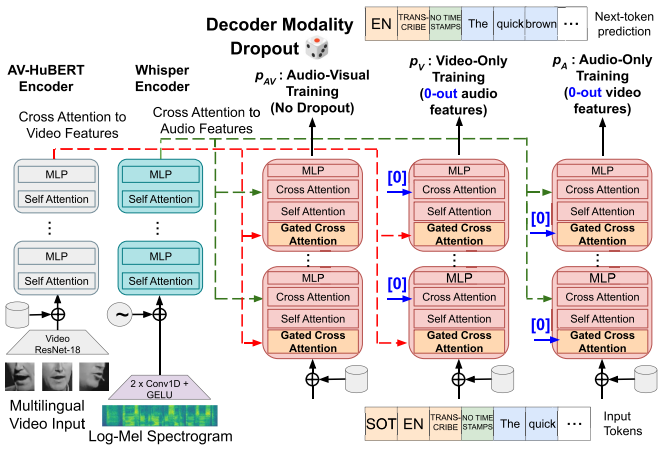

Fig. 1. In mWhisper-Flamingo, the AV-HuBERT and Whisper encoders extract visual and audio features from multilingual videos. Separate cross attention layers in Whisper's decoder attend to the visual and audio features. Decoder modality dropout randomly replaces the audio or video features by 0, forcing the decoder to train on video-only and audio-only inputs.

videos yielded only minor improvements in noisy multilingual AVSR performance, despite achieving significant improvements for English (see Section III-D). We suspect this is due to English having on average 13.6x more data than other languages in our dataset. To address this, we introduce decoder modality dropout by training the model both on paired audio-visual inputs and separate audio/video inputs.

Dropout, originally proposed to prevent overfitting in neural networks [22], was extended to multi-modal learning as modality dropout [23]. Modality dropout randomly drops input modalities during training to better capture cross-modality correlations while preserving unique contributions of specific modalities. Modality dropout has been used in AVSR [18], [24], [25], [26], [27] to prevent models from over-relying on the audio modality, which is typically easier to transcribe than the lip-based visual modality. However, current methods mainly use early-fusion where audio and visual features are combined and used as input into a single Transformer [28] encoder. During modality dropout, the Transformer encoder must handle the incomplete input stream and output an embedding sequence which is used as input to the decoder's cross-attention layers. In contrast, Whisper-Flamingo uses late-fusion where separate Transformer encoders process audio and visual features independently. Each encoder outputs an embedding sequence used as input to separate audio and video cross attention layers in the decoder, where the modality fusion occurs. During modality dropout, one of the audio and video embedding sequences is replaced by a 0-vector, forcing the decoder to handle the incomplete input stream in each of the cross-attention layers corresponding to the missing modality. This enables the encoders to remain specialized on their specific modalities while the decoder learns better multi-modal integration. This process is shown in Fig. 1. Note that cross-attention with a 0-vector results in a 0-vector, but the output from the layer incorporates the bias in the linear transformations.

Our method is conceptually similar to LayerDrop [29], which randomly drops Transformer layers for model pruning on text-based tasks. However, our focus lies in enhancing the decoder's ability to learn better multi-modal integration and to handle inputs where one modality may be unreliable.

We define the probabilities of using audio-visual (p_{AV}), audio-only (p_A), and visual-only (p_V) inputs during training as follows. If audio-visual inputs are selected, both modalities are used. If audio-only inputs are selected, visual features are zeroed out at the decoder. If video-only inputs are selected, audio features are zeroed out at the decoder. In Section III-D, we show results using different probabilities. The setting $p_{AV} = 0.5$, $p_A = 0$, $p_V = 0.5$ performs the best on noisy multilingual AVSR, so we use it for training our main models.

III. Experiments

A. Experimental Setup

We use the MuAViC [20] dataset of 1,141 h of videos in 9 languages. The dataset is based on the LRS3 English video dataset [30] and mTedX dataset [31]. The hours of video per language are: English (En): 433, Arabic (Ar): 16, German (De): 10, Greek (El): 25, Spanish (Es): 178, French (Fr): 176, Italian (It): 101, Portuguese (Pt): 153 and Russian (Ru): 49.

We use Whisper [1] small, medium, and large-v2 with 244 M, 769 M, and 1.55B parameters. We fine-tuned Whisper small and medium on 4 A6000 GPUs with 48 GB memory, but could not fine-tune Whisper large due to the GPU memory limits. We use AV-HuBERT pre-trained on unlabeled multilingual videos [19] with 325 M parameters. We selected this model instead of other English-only lip-reading models [32], [33], [34] due to its superiority on multilingual videos [35], [36], [37], [38]. Different to Whisper-Flamingo, we also fine-tune its parameters. The gated cross attention layers add 82 M and 296 M for the small and medium models, bringing the total to 651 M and 1.39B parameters for mWhisper-Flamingo small / medium. The other dataloading details and hyperparameters closely follow Whisper-Flamingo [27]. Spectrogram frames are used as input to Whisper at 100 Hz while grayscale videos are used as input to AV-HuBERT at 25 fps. The videos are cropped on the lips using Dlib [39] and are aligned to a reference mean face [40]. Models were trained using PyTorch [41] and PyTorch Lightning [42] with the AdamW optimizer [43].

During training, we randomly add noise to the audio with a Signal-to-Noise Ratio (SNR) of 0 dB. Based on prior work [13], [20], "natural", "music" and "babble" noise are sampled from the MUSAN dataset [44], and overlapping "speech" is sampled from LRS3 [30]. We monitor the token prediction accuracy on the noisy validation set every 1 k steps to select the best checkpoints. We normalize the text by lower-casing and removing punctuation except single apostrophes. Our goal is to improve the multilingual Word Error Rate (WER), so we compute average WER on all languages except English. Given the data imbalance, we also separately compute the average on the "higher resource" languages with over 100 h of data (Es, Fr, It, Pt), and the other "lower resource" languages (Ar, De, El, Ru) with as low as 10 h of data.

B. Clean Results

In Table I, we show the results on MuAViC using the original, clean audio. For previous SOTA audio-visual methods, initial work fine-tuned pre-trained English AV-HuBERT [18] on multilingual videos in MuAViC [20], [46]. Some methods trained individual models for specific non-En languages [48], [49], [50], however, they are outperformed by multilingual models. The current SOTA models are XLAVS-R [45] and Fast Conformer [47]. The former adapted XLS-R [51], a multilingual self-supervised audio model, and the latter trains a model from scratch. Moreover, Fast Conformer achieved even better results

TABLE I
MULTILINGUAL WER ON MUAVIC (CLEAN)

Model	Total Params	FT Vid. Hours	Mod.	En	Ar	De	El	Es	Fr	It	Pt	Ru	Avg non-En	Avg H.R.	Avg L.R.
Multilingual Audio-Visual Baselines															
AV-HuBERT [18], [45]	477M	1,141	AV	**1.7**	89.4	52.0	46.2	17.4	20.3	20.8	22.1	44.7	39.1	20.2	**58.1**
Intuitive Multilingual [46]	477M	1,141	AV	2.5	90.9	48.1	42.4	18.0	21.5	19.7	20.8	54.1	39.5	20.0	58.9
XLAVS-R 300M [45]	452M	1,141	AV	2.4	81.7	44.7	24.3	10.9	14.4	12.8	13.2	32.7	29.3	12.8	45.9
XLAVS-R 2B [45]	2.15B	1,141	AV	**1.7**	**79.3**	**44.4**	**19.0**	9.1	12.3	**10.6**	**11.2**	**25.0**	**26.4**	**10.8**	**41.9**
Fast Conformer [47]	197M	1,141	AV	-	-	46.8	-	**9.0**	**11.4**	11.6	11.8	-	-	11.0	-
Multilingual Audio-Visual Baselines (Using Extra Multilingual Training Videos)															
Fast Conformer [47]	197M	4,957	AV	-	-	**23.6**	**-**	**8.2**	**10.3**	**10.4**	**9.9**	-	-	**9.7**	**-**
Our Audio-Only Whisper Zero-Shot Baselines															
Whisper Small Zero-Shot	244M	-	A	2.6	78.9	23.9	23.5	11.4	17.7	20.0	17.1	23.7	27.0	16.6	37.5
Whisper Medium Zero-Shot	769M	-	A	2.3	75.6	21.3	15.7	9.5	15.6	11.6	13.1	20.7	22.9	12.5	33.3
Whisper Large Zero-Shot	1.5B	-	A	**2.1**	**73.2**	**20.1**	**12.7**	**8.9**	**14.5**	**10.2**	**11.7**	**19.0**	**21.3**	**11.3**	**31.3**
Our Models															
Whisper Small Fine-Tuned	244M	-	A	1.1	73.3	26.4	18.4	9.5	13.8	12.8	12.8	21.2	23.5	12.2	34.8
mWhisper-Flamingo Small	651M	1,141	AV	1.2	74.9	26.6	18.9	9.6	13.8	12.7	12.9	21.2	23.8	12.3	35.4
Whisper Medium Fine-Tuned	769M	-	A	**0.74**	**68.9**	**21.5**	**13.6**	**7.9**	**11.3**	**10.1**	**10.7**	**16.7**	**20.1**	**10.0**	**30.2**
mWhisper-Flamingo Medium	1.39B	1,141	AV	0.88	70.2	22.0	14.0	8.1	**11.3**	10.2	10.8	**16.7**	20.4	10.1	30.7

↓ is better. Hours of video used to fine-tune each model are shown. Mod. = Modality. Avg. non-En: average WER without english. H.R. = High Resource (Es, Fr, It, Pt). L.R. = Low Resource (Ar, De, El, Ru).

TABLE II
MULTILINGUAL WER ON MUAVIC (BABBLE NOISE ADDED AT 0-SNR)

Model	Total Params	FT Vid. Hours	Mod.	En	Ar	De	El	Es	Fr	It	Pt	Ru	Avg non-En	Avg H.R.	Avg L.R.
Small Models															
Whisper Small Zero-Shot	244M	-	A	27.9	102	61.2	76.4	63.3	57.8	76.2	73.4	57.6	71.0	67.7	74.3
Whisper Small Fine-Tuned	244M	-	A	16.1	**101**	59.9	56.8	41.7	35.9	50.6	50.2	46.7	55.3	44.6	66.0
mWhisper-Flamingo Small	651M	1,141	AV	**8.3**	**101**	**56.7**	**51.1**	**33.6**	**31.9**	**41.2**	**42.8**	**44.6**	**50.4**	**37.4**	**63.5**
Relative Improvement		-		*48.4*	*-0.8*	*5.3*	*10.0*	*19.4*	*11.1*	*18.6*	*14.7*	*4.5*	*10.4*	*16.0*	*4.8*
Medium Models															
Whisper Medium Zero-Shot	769M	-	A	22.5	105	53.1	61.3	49.1	47.7	60.3	60.9	47.5	60.7	54.5	66.8
Whisper Medium Fine-Tuned	769M	-	A	12.3	96.4	51.8	45.7	36.1	30.4	43.5	42.2	37.9	48.0	38.1	58.0
mWhisper-Flamingo Medium	1.39B	1,141	AV	**7.4**	**95.3**	**49.4**	**41.8**	**28**	**27.5**	**35.2**	**36**	**36.1**	**43.7**	**31.7**	**55.7**
Relative Improvement		-		*39.8*	*1.1*	*4.6*	*8.5*	*22.4*	*9.5*	*19.1*	*14.7*	*4.7*	*10.6*	*16.4*	*4.8*
Large Models															
Whisper Large Zero-Shot	1.5B	-	A	**21.3**	**96.5**	**51.5**	**54.7**	**46.9**	**41.7**	**56.6**	**58.7**	**41.3**	**56.0**	**51.0**	**61.0**

Hours of video used to fine-tune each model are shown. Mod. = modality. Avg. non-En: average WER without english. H.R. = High Resource (Es, Fr, It, Pt). L.R. = Low Resource (Ar, De, El, Ru).

on the higher-resource languages by using extra multilingual training videos from other datasets. While these methods worked well on the higher-resource languages, their performance on lower-resource languages is less satisfactory.

Next, we establish baselines using Whisper. Compared to XLAVS-R 2B, Whisper small zero-shot (without fine-tuning) achieves better performance on the lower resource languages (37.5% vs 41.9%), while Whisper medium zero-shot achieves better overall average non-En WER (22.9% vs 26.4%). This shows Whisper's strong multilingual capabilities and motivates its selection as the foundation of our proposed models.

Fine-tuning Whisper on MuAViC leads to further gains. Fine-tuned Whisper medium achieves a new **SOTA non-En average WER of 20.1%,** surpassing all previous audio-visual methods fine-tuning solely on MuAViC. Notably, the English performance is also improved from 2.3% (zero-shot) to 0.74%, approaching the current SOTA of 0.68% reported in Whisper-Flamingo [17]. This shows that fine-tuning improves multilingual performance without compromising English accuracy.

mWhisper-Flamingo performs similarly to audio-only fine-tuned Whisper (20.4% vs. 20.1% WER for the medium models). This small difference suggests that video provides limited benefit for clean audio. However, mWhisper-Flamingo still significantly outperforms all previous AVSR models fine-tuned solely on the 1,141 h MuAViC videos, achieving a new SOTA on all languages with this setup. It even surpasses Fast Conformer trained with 4,957 h of videos on all languages except Fr and Pt. This shows the strength of using Whisper as initialization for AVSR. Finally, mWhisper-Flamingo medium consistently outperforms mWhisper-Flamingo small, showing the benefit of increased model size.

C. Noisy Results

Table II shows the performance on MuAViC with babble noise at 0-SNR. The babble noise is from Whisper-Flamingo [17] and was constructed from 30 LRS3 speakers. We compare Whisper zero-shot with Whisper fine-tuned and mWhisper-Flamingo. The noisy results show a significant performance degradation compared to the clean results (Table I). For example, the average non-En WER for Whisper small zero-shot increases from 27.0% in the clean setting to 71.0% under 0-SNR babble noise. However, fine-tuning significantly improves WER. For the small model, fine-tuning improves the average non-En WER from 71.0% (zero-shot) to 55.3%. A similar trend is observed for the medium model, with WER decreasing from 60.7% to 48.0%.

The integration of visual features in mWhisper-Flamingo provides further gains. mWhisper-Flamingo small achieves an average non-En WER of 50.4%, a **10.4% relative improvement** over fine-tuned audio-only Whisper small (55.3%). Similarly, mWhisper-Flamingo medium achieves an average non-En WER of 43.7%, a **10.6% relative improvement** compared to fine-tuned audio-only Whisper medium (48.0%). The relative improvements are better for the higher-resource languages (**16.0% and 16.4%** for small and medium models), indicating that more video training data is helpful. The relative improvements for English are much better at 48.4% and 39.8% for the small and medium models, which we attribute to having more English training data.

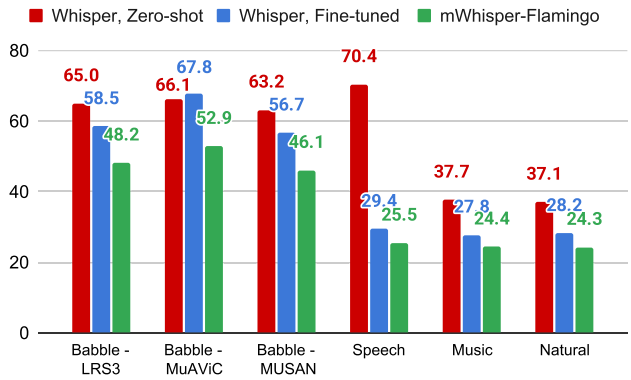

Fig. 2. Multilingual WER (↓ is better) for different noise types averaged over 4 languages (Es, Fr, It, Pt) and 5 SNR levels $\{-10, -5, 0, 5, 10\}$.

TABLE III
ABLATION STUDY AND ANALYSIS ON MUAVIC (0DB SNR BABBLE NOISE)

Setup	En	Es	Fr	It	Pt	Avg non-En
Whisper-Small Fine-Tuned						
Whisper-Small	16.2	41.7	35.8	50.0	50.2	44.4
Decoder Modality Dropout and Fine-Tunable Visual Encoder						
Mod. Drop. ✓, FT vis. encoder ✓	**8.0**	**32.9**	**31.9**	**40.5**	**41.2**	**36.6**
Mod. Drop. ✗, FT vis. encoder ✓	16.0	41.6	36.2	50.2	50.4	44.6
Mod. Drop. ✗, FT vis. encoder ✗	11.5	39.1	35.6	47.8	47.8	42.6
Mod. Drop. ✓, FT vis. encoder ✗	9.7	36.7	34.1	45.1	46.5	40.6
Decoder Modality Dropout Probabilities						
AV: 0.5, A: 0, V: 0.5	**8.0**	**32.9**	**31.9**	40.5	**41.2**	**36.6**
AV: 1.0, A: 0, V: 0	16.0	41.6	36.2	50.2	50.4	44.6
AV: 0.5, A: 0.5, V: 0	16.0	41.6	36.1	50.3	50.5	44.6
AV: 0.5, A: 0.25, V: 0.25	8.4	33.7	32.5	42.1	42.4	37.7
AV: 0.25, A: 0.25, V: 0.5	8.5	34.5	32.2	41.5	42.6	37.7
AV: 0.75, A: 0, V: 0.25	8.6	33.9	33.0	41.3	42.5	37.7
AV: 0.25, A: 0, V: 0.75	8.3	33.5	32.7	**40.0**	41.5	36.9
Multilingual vs English AV-HuBERT						
Multilingual AV-HuBERT	8.0	**32.9**	**31.9**	**40.5**	**41.2**	**36.6**
English AV-HuBERT	**7.5**	34.8	32.4	40.8	41.6	37.4

Performance is WER%, (↓ is better).

mWhisper-Flamingo medium achieves the best multilingual WER (43.7%), even outperforming Whisper large zero-shot (56.0%) despite having fewer parameters (1.39B vs 1.5B). This shows that the performance gains are not only due to more parameters but also due to the visual modality. As additional evidence, mWhisper-Flamingo small (651 M parameters) outperforms audio-only fine-tuned Whisper medium (769 M parameters) on the higher-resource languages (37.4% vs. 38.1%). This highlights that smaller audio-visual models can achieve better performance than larger audio-only models.

Finally, we tested the models on 6 different noise types, 5 SNR levels $\{-10, -5, 0, 5, 10\}$, and 4 languages (Es, Fr, It, Pt). The noise setups follow Whisper-Flamingo [17] and AV-HuBERT [13]. Fig. 2 shows the WER for Es, Fr, It, Pt averaged over all SNR values for each noise type. There is a clear trend of mWhisper-Flamingo outperforming the audio-only models, especially for the more challenging noises (3 types of babble and side speech). Overall, these results confirm the advantage of mWhisper-Flamingo on diverse noise types.

D. Ablation Study and Analysis

In Table III, we present an ablation study and model analysis. Due to computational constraints, we trained and evaluated the small model on 5 of the 9 languages. We compare the average non-En WER between models. We show the noisy results since the performance differences between models in the clean results were minor. The baseline audio-only fine-tuned Whisper small achieves 44.4%.

First, we show that the combination of decoder modality dropout and a fine-tunable visual encoder leads to the best performance (36.6%). The performance is significantly worse if decoder modality dropout is disabled using either a fine-tunable visual encoder (44.6%) or a frozen visual encoder encoder (42.6%). It shows that our proposed decoder modality dropout is *crucial* for obtaining the best multilingual performance. Note that the latter setup represents the default training configuration from Whisper-Flamingo [17]. Considering that the latter setup improved the English noisy WER from 16.2% to 11.5%, it is reasonable that modality dropout was not necessary in the original, English-only Whisper-Flamingo. Finally, using decoder modality dropout with a fine-tunable visual encoder is better than using it with a frozen visual encoder (36.6% vs 40.6%).

mWhisper-Flamingo uses decoder modality dropout probabilities of $p_{AV} = 0.5$, $p_A = 0$, $p_V = 0.5$, meaning that the model trains with audio-visual inputs 50% of the time and visual-only inputs 50% of the time. Without dropout, the modal trains only on audio-visual inputs ($p_{AV} = 1$, $p_A = 0$, $p_V = 0$), and the performance is much worse (44.6% vs 36.6%). Using dropout with audio-only inputs 50% of the time instead of video-only inputs ($p_{AV} = 0.5$, $p_A = 0.5$, $p_V = 0$) does not improve performance (44.6%). It shows that it is necessary for the model to train with video-only inputs. Trying different probabilities between AV, A, and V such as $p_{AV} = 0.5$, $p_A = 0.25$, $p_V = 0.25$ and $p_{AV} = 0.25$, $p_A = 0.25$, $p_V = 0.5$, the performance is improved to 37.7%, however the best performance is achieved without audio-only inputs ($p_{AV} = 0.5$, $p_A = 0$, $p_V = 0.5$). Our explanation is that Whisper was fine-tuned in the first stage of training, so the model can already handle audio-only inputs well. Training on video-only inputs allows the model to better integrate the visual modality. We also tried other proportions of audio-visual and video-only training such as $p_{AV} = 0.75$, $p_A = 0$, $p_V = 0.25$ and $p_{AV} = 0.25$, $p_A = 0$, $p_V = 0.75$ which performed slightly worse.

Finally, we compare multilingual AV-HuBERT [19] with AV-HuBERT pre-trained only on English videos [18]. The model using the multilingual AV-HuBERT encoder achieves better multilingual performance (36.6% vs 37.4%) but worse English performance (8.0% vs 7.5%), which is reasonable.

IV. CONCLUSION

We introduce mWhisper-Flamingo, a novel multilingual model for AVSR. mWhisper-Flamingo outperforms all audio-visual methods trained on MuAViC, and even surpasses a model trained with substantially more videos on most languages. mWhisper-Flamingo outperforms audio-only Whisper in diverse noise settings. Our proposed decoder modality dropout significantly improved the noisy multilingual WER.

ACKNOWLEDGMENT

We thank Videet Mehta for help with the demo and Tatiana Likhomanenko and Saurabhchand Bhati for helpful discussions.

REFERENCES

[1] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 28492–28518.

[2] K. C. Puvvada et al., "Less is more: Accurate speech recognition & translation without web-scale data," in *Proc. Interspeech*, 2024, pp. 3964–3968.

[3] A. Rouditchenko et al., "Comparison of multilingual self-supervised and weakly-supervised speech pre-training for adaptation to unseen languages," in *Proc. Interspeech*, 2023, pp. 2268–2272.

[4] J. Shi et al., "Ml-superb: Multilingual speech universal performance benchmark," in *Proc. Interspeech*, 2023, pp. 884–888.

[5] Y. Gong, S. Khurana, L. Karlinsky, and J. Glass, "Whisper-AT: Noise-robust automatic speech recognizers are also strong general audio event taggers," in *Proc. Interspeech*, 2023, pp. 2798–2802.

[6] Y. Hu et al., "Large language models are efficient learners of noise-robust speech recognition," in *Proc. Int. Conf. Learn. Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=ceATjGPTUD

[7] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 8717–8727, Dec. 2022.

[8] S. Petridis, T. Stafylakis, P. Ma, G. Tzimiropoulos, and M. Pantic, "Audio-visual speech recognition with a hybrid CTC/attention architecture," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2018, pp. 513–520.

[9] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic, "End-to-end audiovisual speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 6548–6552.

[10] B. Xu, C. Lu, Y. Guo, and J. Wang, "Discriminative multi-modality speech recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 14433–14442.

[11] P. Ma, S. Petridis, and M. Pantic, "End-to-end audio-visual speech recognition with conformers," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 7613–7617.

[12] D. Serdyuk, O. Braga, and O. Siohan, "Transformer-based video front-ends for audio-visual speech recognition for single and muti-person video," in *Proc. Interspeech*, 2022, pp. 2833–2837.

[13] B. Shi, W. -N. Hsu, and A. Mohamed, "Robust self-supervised audio-visual speech recognition," in *Proc. Interspeech*, 2022, pp. 2118–2122.

[14] P. Ma, A. Haliassos, A. Fernandez-Lopez, H. Chen, S. Petridis, and M. Pantic, "Auto-AVSR: Audio-visual speech recognition with automatic labels," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2023, pp. 1–5.

[15] M. Burchi and R. Timofte, "Audio-visual efficient conformer for robust speech recognition," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2023, pp. 2258–2267.

[16] U. Cappellazzo et al., "Large language models are strong audio-visual speech recognition learners," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2025.

[17] A. Rouditchenko et al., "Whisper-Flamingo: Integrating visual features into whisper for audio-visual speech recognition and translation," in *Proc. Interspeech*, 2024, pp. 2420–2424.

[18] B. Shi, W.-N. Hsu, K. Lakhotia, and A. Mohamed, "Learning audio-visual speech representation by masked multimodal cluster prediction," in *Proc. Int. Conf. Learn. Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=Z1Qlm11uOM

[19] M. Kim, J. Yeo, S. J. Park, H. Rha, and Y. M. Ro, "Efficient training for multilingual visual speech recognition: Pre-training with discretized visual speech representation," in *Proc. 32nd ACM Int. Conf. Multimedia*, 2024, pp. 1311–1320.

[20] 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. Interspeech*, 2023, pp. 4064–4068.

[21] J.-B. Alayrac et al., "Flamingo: A visual language model for few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 23716–23736.

[22] G. Hinton, "Improving neural networks by preventing co-adaptation of feature detectors," 2012, *arXiv:1207.0580*.

[23] N. Neverova, C. Wolf, G. Taylor, and F. Nebout, "ModDrop: Adaptive multi-modal gesture recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1692–1706, Aug. 2016.

[24] T. Makino et al., "Recurrent neural network transducer for audio-visual speech recognition," in *Proc. Autom. Speech Recognit. Understanding*, 2019, pp. 905–912.

[25] W.-N. Hsu and B. Shi, "u-HuBERT: Unified mixed-modal speech pretraining and zero-shot transfer to unlabeled modality," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 21157–21170.

[26] J. Lian, A. Baevski, W.-N. Hsu, and M. Auli, "Av-data2vec: Self-supervised learning of audio-visual speech representations with contextualized target representations," in *Proc. IEEE Autom. Speech Recognit. Understanding*, 2023, pp. 1–8.

[27] A. Rouditchenko, R. Collobert, and T. Likhomanenko, "AV-CPL: Continuous pseudo-labeling for audio-visual speech recognition," 2023, *arXiv:2309.17395*.

[28] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[29] A. Fan, E. Grave, and A. Joulin, "Reducing transformer depth on demand with structured dropout," in *Proc. Int. Conf. Learn. Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=SylO2yStDr

[30] T. Afouras, J. S. Chung, and A. Zisserman, "Lrs3-ted: A large-scale dataset for visual speech recognition," 2018, *arXiv:1809.00496*.

[31] E. Salesky et al., "The multilingual TEDx corpus for speech recognition and translation," in *Proc. Interspeech*, 2021, pp. 3655–3659.

[32] A. Haliassos, P. Ma, R. Mira, S. Petridis, and M. Pantic, "Jointly learning visual and auditory speech representations from raw data," in *Proc. Int. Conf. Learn. Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=BPwIgvf5iQ

[33] A. Haliassos, A. Zinonos, R. Mira, S. Petridis, and M. Pantic, "Braven: Improving self-supervised pre-training for visual and auditory speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2024, pp. 11431–11435.

[34] A. Haliassos, R. Mira, H. Chen, Z. Landgraf, S. Petridis, and M. Pantic, "Unified speech recognition: A single model for auditory, visual, and audiovisual inputs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2024. [Online]. Available: https://openreview.net/forum?id=vWSll6M9pj&referrer=%5Bthe%20profile%20of%20Stavros%20Petridis%5D(%2Fprofile%3Fid%3D~Stavros_Petridis1)

[35] P. Ma, S. Petridis, and M. Pantic, "Visual speech recognition for multiple languages in the wild," *Nature Mach. Intell.*, vol. 4, no. 11, pp. 930–939, 2022.

[36] A. Zinonos, A. Haliassos, P. Ma, S. Petridis, and M. Pantic, "Learning cross-lingual visual speech representations," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2023, pp. 1–5.

[37] M. Kim, J. H. Yeo, J. Choi, and Y. M. Ro, "Lip reading for low-resource languages by learning and combining general speech knowledge and language-specific knowledge," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 15359–15371.

[38] J. H. Yeo, M. Kim, S. Watanabe, and Y. M. Ro, "Visual speech recognition for languages with limited labeled data using automatic labels from whisper," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2024, pp. 10471–10475.

[39] D. E. King, "Dlib-ml: A machine learning toolkit," *J. Mach. Learn. Res.*, vol. 10, pp. 1755–1758, 2009.

[40] B. Martinez, P. Ma, S. Petridis, and M. Pantic, "Lipreading using temporal convolutional networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 6319–6323.

[41] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019. [Online]. Available: https://papers.nips.cc/paper_files/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html

[42] W. Falcon and The PyTorch Lightning team, "PyTorch Lightning," 2023. [Online]. Available: https://github.com/Lightning-AI/pytorch

[43] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7

[44] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," 2015, *arXiv:1510.08484*.

[45] H. Han et al., "XLAVS-R: Cross-lingual audio-visual speech representation learning for noise-robust speech perception," in *Proc. 62nd Annu. Meeting Assoc. Comput. Linguistics*, 2024, pp. 12896–12911.

[46] J. Hong, S. Park, and Y. Ro, "Intuitive multilingual audio-visual speech recognition with a single-trained model," in *Proc. Findings EMNLP*, 2023, pp. 4886–4890.

[47] M. Burchi et al., "Multilingual audio-visual speech recognition with hybrid CTC/RNN-T fast conformer," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2024, pp. 10211–10215.

[48] 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. IEEE Autom. Speech Recognit. Understanding*, 2023, pp. 1–8.

[49] D. Gimeno-Gómez and C.-D. Martínez-Hinarejos, "Tailored design of audio-visual speech recognition models using branchformers," *Comput. Speech Lang.*, vol. 94, 2025, Art. no. 101811. [Online]. Available: https://doi.org/10.1016/j.csl.2025.101811

[50] Z. Li et al., "Interleaved audio/audiovisual transfer learning for AV-ASR in low-resourced languages," in *Proc. Interspeech*, 2024, pp. 2524–2528.

[51] A. Babu et al., "XLS-R: Self-supervised cross-lingual speech representation learning at scale," 2021, *arXiv:2111.09296*.