ETRI Journal WILEY

KMSAV: Korean multi-speaker spontaneous audiovisual dataset

Kiyoung Park[1,2] | Changhan Oh[1,2] | Sunghee Dong[1]

[1]Superintelligence Creative Research Laboratory, Electronics and Telecommunications Research Institute, Daejeon, Republic of Korea

[2]Integrated Intelligence Research Laboratory, University of Science and Technology, Daejeon, Republic of Korea

Correspondence
Kiyoung Park, Superintelligence Creative Research Laboratory, Electronics and Telecommunications Research Institute, Daejeon, Republic of Korea.
Email: pkyoung@etri.re.kr

Abstract

Recent advances in deep learning for speech and visual recognition have accelerated the development of multimodal speech recognition, yielding many innovative results. We introduce a Korean audiovisual speech recognition corpus. This dataset comprises approximately 150 h of manually transcribed and annotated audiovisual data supplemented with additional 2000 h of untranscribed videos collected from YouTube under the Creative Commons License. The dataset is intended to be freely accessible for unrestricted research purposes. Along with the corpus, we propose an open-source framework for automatic speech recognition (ASR) and audiovisual speech recognition (AVSR). We validate the effectiveness of the corpus with evaluations using state-of-the-art ASR and AVSR techniques, capitalizing on both pretrained models and fine-tuning processes. After fine-tuning, ASR and AVSR achieve character error rates of 11.1% and 18.9%, respectively. This error difference highlights the need for improvement in AVSR techniques. We expect that our corpus will be an instrumental resource to support improvements in AVSR.

KEYWORDS
audiovisual data, dataset, multimodal data, multi-speaker spontaneous data, speech recognition

1 | INTRODUCTION

For speech recognition, humans depend on both auditory and visual cues. Such multimodal speech recognition is useful when a modality is impaired owing to factors such as noise interference. Therefore, speech recognition using multimodal inputs is among the longstanding topics in this field. Moreover, rapid advancements in deep learning technology have led to remarkable progress in both the speech and visual recognition domains. This progress has naturally led to the development of multimodal speech recognition methods that integrate conventional speech and visual recognition techniques or audiovisual speech recognition (AVSR), resulting in numerous studies and innovative outcomes [1, 2].

With the development of deep learning techniques, many different solutions for speech recognition have been developed in recent years. They include automatic speech recognition (ASR) using only audio information, visual speech recognition or lip reading using only video information, and AVSR using audiovisual information. As ASR relies solely on speech, its performance drops in noisy environments [3]. Visual speech recognition uses facial videos and extracts features from lip movements, thus being insensitive to noise in audio. Compared with ASR, however, visual speech recognition is very challenging,

even for humans, owing to the lack of information and ambiguities in the input data [4]. AVSR leverages both audio and visual information to improve the recognition performance and increase robustness in noisy environments. This research topic has been widely explored, and various representative studies are available [5, 6].

In previous research on AVSR, lip reading emerged as a predominant area of study, leading to the development of numerous datasets specifically tailored for this purpose. Initial research efforts were primarily concentrated on the recognition of words, short phrases, and simple sentences articulated in a deliberate manner. These datasets were collected in highly controlled environments and employing predefined scripts as reference points [7, 8]. With the advent and development of deep learning techniques, the size of such datasets has notably expanded [9, 10], aiming to accommodate the increased complexity of deep learning models and include a variety of languages, such as Korean [11].

The styles of speech in datasets should be diversified beyond reading of predefined scripts. To capture fully natural and unstructured audiovisual data, numerous datasets have begun sourcing material from existing media platforms such as broadcast shows and movies. However, these sources are predominantly produced in English. Moreover, most of these datasets have focused extensively on monologue-style videos, such as news presentations or lectures.

In this paper, we present a real-world dataset for Korean AVSR focused on the spontaneous conversations between multiple participants. Furthermore, we release a complete system from data preparation to performance evaluation of ASR and AVSR using publicly available toolkits and models. We aim to assist researchers in this field to easily set up a baseline system and use the constructed corpus. We expect that this corpus and the associated methods will help accelerate the development of AVSR and related techniques.

1.1 | Related work

Recently, various remarkable studies on AVSR have been conducted. Originating from Google Bidirectional Encoder Representations from Transformers (BERT), the adoption of self-supervision for pretraining expansive deep learning models has ushered in considerable advancements across various research domains, encompassing both speech and vision. In speech recognition, pretrained models such as wav2vec 2.0 [12] and Hu-BERT [13] with self-supervised learning techniques have demonstrated comparable ASR performance even with small amounts of labeled speech data. AVSR also necessitates pretrained models learned

from large-scale unlabeled data owing to the shortage of synchronized transcription data from audio and video recordings. AV-HuBERT is one of the most successful approach to provide a large pretrained model for AVSR [14]. The pretrained model with 1759 h of unlabeled data combined with only 30 h of labeled data showed a high AVSR performance.

In [5], the lack of data to train a large model was addressed with a different approach. Transcription was generated for large unlabeled audiovisual datasets using the publicly available pretrained ASR models. With pseudo-labeled data and an end-to-end conformer architecture [2], state-of-the-art results were achieved on the LRS2 [4] and LRS3-TED [15] corpora. For the unlabeled datasets, 1323 h of unlabeled data from the AVSpeech corpus [16] and 1307 h from the VoxCeleb2 corpus [17] were used as an additional training set.

Labeled data are easier to obtain for ASR than for AVSR. Hence, training large ASR models using massive datasets has become common [18–20]. Recently released by OpenAI, the Whisper model employed a large amount of web speech data, approximately 680 000 h, and pretrained the transformer encoder–decoder structure under weak supervision [19]. This model exhibits robust recognition and translation capabilities, not only for English but also for over 100 other languages, including Korean.

1.2 | Existing datasets

We briefly describe currently available audiovisual datasets. Lip reading, a topic of significant interest, is supported by datasets such as AVLetters [7], CUAVE [8], and GRID [21]. These datasets primarily feature read speeches of isolated alphabets and both isolated and connected digits. For more complex tasks, such as recognizing phrases and sentences with larger vocabularies, datasets like AV-TIMIT [10], AVI-CAR [22], OuluVS [23], and OuluVS2 [9] have been developed. Notably, the AVICAR and OuluVS2 datasets include multiple viewing angles, such as side and frontal views. Nevertheless, these datasets were collected in well-controlled environments with predefined scripts.

The recent surge in AVSR applications has created the demand for datasets collected in real-world environments. This need led to the collection and labeling of existing audiovisual data. The LRW [24] and LRS2 [4] datasets, which focus on words and sentences, respectively, were among the first datasets created for real-world applications. To further expand the scope, the LRS3-TED dataset [1] was introduced. Regarding Korean audiovisual datasets, the OLKAVS dataset [11] features a large scale of 1150 h from over 1000 speakers reading sentences from multiple views. However, a

comprehensive real-world audiovisual dataset is yet to be developed in Korean. Table 1 lists characteristics of existing datasets divided into three categories according to the degree of freedom in speech.

The LRS3-TED dataset is the most widely used dataset for AVSR. It comprises 151819 videoclips from TED Talks, totaling 439 h of data. The dataset comprises 9506 videos and includes English subtitles from manual transcription. Using ASR systems, word-level alignment was produced, and sentences were extracted from this alignment based on the punctuation marks in the transcript. As a result, 118 516 (408 h), 31 982 (30 h), and 1321 clips (0.9 h) were obtained for the pretraining, training–validation, and test sets, respectively. Additionally, the dataset provides facial region data, extracted using a face detector implemented using a convolutional neural network and based on the single shot multibox detector [26].

1.3 | Contributions

We introduce a novel dataset, the Korean multi-speaker spontaneous audiovisual (KMSAV) speech corpus. This corpus is designed for use in various fields of audiovisual research including AVSR, active speaker detection (ASD), lip reading, and audiovisual speech enhancements. This is the first audiovisual corpus of multi-speaker spontaneous conversations in the Korean language. With its size, encompassing 150 h of transcribed content and an additional 2000 h of unlabeled data, it is sufficiently large to be suitable for both supervised and unsupervised training of audiovisual data processing algorithms.

Distinctively, this corpus emphasizes multiple speakers conversing naturally, in contrast to existing datasets, which primarily feature lecture-style, single-person speech. Similar to the LRS3-TED dataset, the constructed KMSAV dataset also provides frame-wise coordinates of facial regions. This enables researchers to concentrate on the development of core algorithms by mitigating the effort required for data preprocessing. We constructed this corpus using YouTube videos available under the Creative Commons License, such that the data can be used freely. As with the LRS3-TED dataset, we extracted sentences from the video and the facial region of the active speaker per sentence.

Using the KMSAV dataset, we conducted ASR and AVSR experiments to assess its quality on open-source toolkits and models. For ASR, we employed Whisper and ESPnet [27]. With pre-released models and an inference tool, we observed a character error rate (CER) ranging from 15.3% to 32.2% depending on the model size. By fine-tuning with the training portion of the KMSAV dataset, the CER decreased to between 11.1% and 23.5%. For the AVSR experiments, we employed the AV-HuBERT toolkit and its pretrained models. Notably, although these pretrained models were trained on English data, they still showed a reasonable performance on the KMSAV dataset, achieving a CER of 21.7% with audio-only inputs and 18.9% with audiovisual inputs. The same evaluation metrics and dataset were used in all the experiments to ensure a fair comparison.

We have ascertained that ASR trained with a considerable volume of labeled data surpasses AVSR, even though supplementary visual information is integrated

TABLE 1 Summary of existing audiovisual datasets.

Dataset	Environment/source	Language	No. of utterances	No. of subjects	Utterances
AVLetters [7]	Lab	English	780	10	Isolated English alphabets
CUAVE [8]	Sound booth	English	7.9k	36	Isolated and connected digits
GRID [21]	Sound booth	English	34k	34	Simple command sentences
AV-TIMIT [10]	Lab	English	4.4k	223	Phonetically balanced short sentences
OuluVS [23]	Lab	English	817	20	Short phrases
OuluVS2 [9]	Studio	English	1.5k	53	Phrases and sentences
AVICAR [22]	Car	English	59k	86	Phrases and sentences
OLKAVS [11]	Studio	Korean	2.5M	1107	Read sentences
LRW [24]	Broadcast news	English	539k	>1000	Isolated words in spontaneous sentences
LRS2 [1]	Broadcast news	English	144k	–	Spontaneous sentences
LRS3-TED [15]	Online lecture	English	152k	9545	Spontaneous sentences
VoxCeleb1 [25]	YouTube video	English	154k	1251	Spontaneous sentences without transcription
VoxCeleb2 [17]	YouTube video	Multilingual	1.1M	6112	Spontaneous sentences without transcription
KMSAV (Ours)	YouTube video	Korean	>43.8k	–	Spontaneous sentence

into AVSR. This emphasizes the prevailing developmental disparity between ASR and AVSR. We expect substantial advancements in AVSR techniques in the near future, for which our work may play a pivotal role. All the resources required to replicate our results, including the datasets, code, and models, are publicly available online (https://github.com/etri/kmsav).

The contributions of this study are summarized as follows:*The first real-world Korean audiovisual dataset*. To the best of our knowledge, this is the first large-scale, real-world audiovisual dataset with high-quality Korean transcription. The dataset not only includes transcriptions but also tags for speech overlap and filler words. In addition, manually verified frame-wise facial regions are provided to standardize data preprocessing.

Open-source framework for audiovisual research. We utilized open-source frameworks to conduct ASR and AVSR experiments on the KMSAV dataset. All the resources needed to replicate our results, including datasets, codes, and models, will be made available online.

Baseline ASR and AVSR system. We present a baseline ASR and AVSR system using the KMSAV dataset. The system encompasses all processes, from data preparation to performance evaluation. We expect that the proposed standardized procedure will facilitate comparisons with results from other studies and support further research.

2 | KMSAV DATASET

The KMSAV dataset is the first Korean dataset designed for AVSR of natural conversations between multiple speakers. To construct a real-world corpus, we performed the following steps:

1. source videos from YouTube,
2. manually transcribe them at the utterance level using audio,
3. extract facial regions from video frames aligned with the respective utterances, and
4. manually verify the extracted audiovisual utterances.

In this section, we detail the KMSAV dataset construction procedure and its characteristics. The pipeline to collect, segment, and refine the audiovisual utterances is shown in Figure 1.

2.1 | Data collection

To construct the real-world corpus, we sourced YouTube videos based on the following criteria:

1. The video must be licensed for public use.
2. The video should feature natural conversations in Korean involving multiple speakers.
3. At least half of the video duration should consist of speaking voices from the participants.
4. Voices not emanating from individuals present in the video (e.g., commentator narration and translated speech) are not considered valid speaking samples.
5. A diverse range of subjects and conversation formats should be gathered.

In line with the selection criteria, we curated a collection of 5214 videos, accounting for approximately 2162 h of data. We then grouped these videos into 12 distinct categories considering their content and format, as detailed in Table 2. Because the KMSAV corpus emphasizes multi-speaker conversations, we tallied the number of speakers actively participating within each video. Not every speaker may be visible on the screen simultaneously, and the count represents the total number of speakers appearing throughout a video. While transcribing their speech, we also assigned a speaker index for each video, which can be useful for applications such as speaker diarization. Moreover, with numerous instances of speaker overlap, we carefully annotated the intervals of overlap to facilitate research in applications such as overlap detection and speech separation. The detailed histogram of the number of people is shown in Figure 2. Approximately half of the videos (52.5%) consisted of clips with two speakers. Videos containing from three to
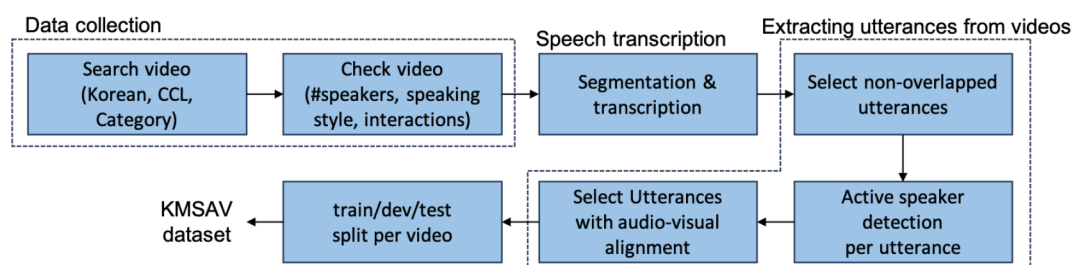

FIGURE 1 Pipeline to collect, segment, and refine audiovisual utterances from video.

TABLE 2 Categories and description of videos selected for KMSAV corpus.

| Category	Description	No. of videos		Total length (h)		Avg. no. of speakers
		All	Trans.	All	Trans.	
beau	Videos related to makeup or beauty	313	45	80.0	5.9	2.5
docu	Documentaries on various topics	42	9	17.7	2.2	6.0
dram	TV or web dramas	65	33	22.1	7.2	7.2
ente	Entertainment programs or variety shows	368	14	208.4	3.3	4.0
game	Reviews and conversations on games	177	3	56.6	0.6	3.1
inte	Interviews on various topics	1748	61	778.5	23.2	2.6
issu	Discussion programs on current events	1891	271	756.4	81.5	3.5
movi	TV or web movies	17	0	7.1	0	6.4
mukb	"Mukbang" or eating show	279	12	111.6	3.5	3.7
news	News broadcasts	25	3	14.0	0.9	3.5
scie	Lectures on science	270	61	104.2	19.7	4.5
vlog	Video blog	19	0	5.8	0	3.5
Total	-	5214	512	2162.1	147.9	3.3
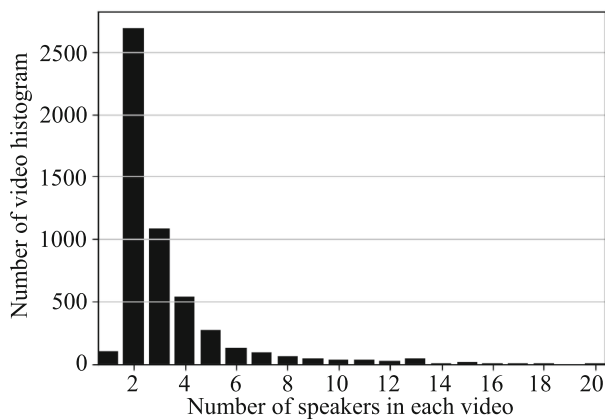
FIGURE 2 Histogram of number of speakers in a video, who are actively participating in the conversation.

five speakers constituted 37.1%, and those containing more than five speakers constituted 10.2% of the collected data.

2.2 | Speech transcription

Approximately 10% of the collected videos were randomly selected and transcribed by human annotators. While subtitles provided with the videos were available and referenced, all transcriptions were essentially created anew by the annotators. During transcription, the audio data were divided into utterances. Typically, an utterance corresponded to a sentence, but it might encompass two or more sentences if splitting was impractical due to brief pauses or other reasons. The script and timestamp of each utterance were noted. If there was an overlap between two or more speakers, the utterances were separately annotated to ensure they could be distinguished for use. In addition, a speaker index was assigned to each utterance. The speaker index was distinct only within each video but not across different videos.

Because the corpus was primarily intended for Korean speech recognition, transcription rules were specifically elaborated for this purpose. First, common filler words or hesitations words in Korean were actively tagged for subsequent use. Depending on the application, such as translation, these filler words could be excluded from the system output. Second, all the transcriptions were represented in the numerical and original forms for digits and the English alphabet, rather than in the textual or transliterated form. In Korean, numerals can be written and pronounced in multiple ways. In addition, English words in the Roman alphabet are commonly used in everyday life. To enhance the readability of the output of the trained speech recognition system, we aimed to perform consistent and uniform transcription. Lastly, to enhance readability of the machine output, punctuation marks such as periods, question marks, and exclamation marks were scrupulously added.

2.3 | Utterance extraction from videos

During transcription, utterances were extracted solely based on audio information from each video. We then identified the utterances that aligned well with the

corresponding video content for a speaker. To this end, we applied ASD to identify the facial region of the current speaker per utterance. For ASD, we used TalkNet-ASD [28], where an audiovisual cross-attention mechanism is utilized to capture long-term speaking evidence. To reliably extract data from individual speakers in videos featuring multiple free-flowing conversations, we implemented the following procedure:

1. Initially, all the videos were processed using ASD to identify the tracks of every active speaker. A track was a temporal progression of a rectangular region where the active speaker was presumed to be located. During track extraction, brief interruptions without any active speaker lasting up to 24 ms were permitted.
2. The track identified in the previous step was not synchronized with the audio utterances. Hence, for each utterance segmented during transcription, we determined the corresponding video track. Only tracks that overlapped by more than 50% of the utterance duration were retained.
3. In the final step, we further refined the segments by discarding audiovisual utterances where the starting and ending points between audio and video differed by more than 1 s.

After applying the abovementioned procedure, we retained 53 375 utterances, which accounted for approximately 89.0 h out of the total 147.9 h of transcribed audiovisual content.

For every audiovisual utterance, we conducted a manual verification, with human workers meticulously checking the transcriptions and extracted videos. This review task was delegated to workers who had not been previously involved with the project. Consequently, any utterances with potential issues were excluded from the final compilation. The excluded utterances included falsely detected facial regions (5.12%), overlapping utterances from multiple speakers (4.47%), and inaccurate utterance segmentation (0.22%).

Ultimately, from the collected videos, a total of 43 805 utterances spanning 84.4 h were extracted. The transcribed audiovisual data were divided into training, validation, and test sets at the video level. This division was made to ensure a consistent distribution of video categories across sets. Details of the data quantity for each split are presented in Table 3.

Figure 3 shows the histogram of the utterance lengths for the training set and the validation and test sets, respectively.

Most utterances (79.5%) were shorter than 10 s, while utterances ranging from 10 to 20 s accounted for 18.3%, those between 20 and 30 s comprised 1.5%, and those longer than 30 s accounted for 0.6% of the data. The longest utterance spanned 104 s.

2.4 | Characteristics of KMSAV dataset

The LRS3-TED dataset is a widely used audiovisual corpus for AVSR, lip reading, and other tasks. We collected the KMSAV dataset following a similar procedure to collect the LRS3-TED dataset but considering the Korean language. The KMSAV dataset, as a multi-speaker database, exhibits characteristics distinct from LRS3-TED and other datasets. First, the videos are selected to include dialogues with multiple speakers, resulting in frequent turn changes and overlapping speech intervals. Table 4 presents the frequency of turn changes in conversations, given by the average number of turn changes over 10 s for each domain.
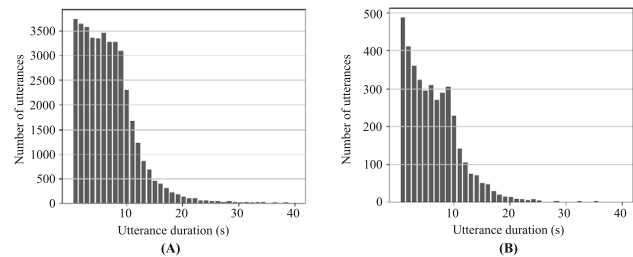

FIGURE 3 Histograms of extracted utterance lengths for (A) training and (B) validation and test sets.

TABLE 3 Dataset after segmenting utterances from video.

Split	No. of videos	Length of videos (h)	No. of utterances	Length of utterances (h)
Train	474	135.4	39 926	77.5
Valid.	16	5.8	2032	3.6
Test	22	6.7	1847	3.4
All	512	147.9	43 805	84.4

Additionally, the extent of overlapping speech is given by the ratio of its duration. In domains such as `beau` and `ente`, turn changes and overlaps are more common than in domains such as `news` and `scie`.

Second, the KMSAV data are sourced from natural conversations, as opposed to script readings in controlled laboratory settings or lecture environments. This real-world data presents challenges, such as frequent changes in the orientation of speakers' faces and instances where the lips are partially or fully obscured. It is common for speakers to be positioned sideways relative to the camera and microphones partially occluding facial features, thus hindering lip reading.

To assess the dataset applicability to lip reading, facial landmark detection was performed using dlib [29], a conventional machine learning library. The face direction angles were calculated by finding the centroid of the outer face landmarks and the centroid of the eye, nose, and mouth landmarks. The outcomes of this detection process and corresponding angles are detailed in Table 5. Although the KMSAV dataset provides a higher failure rate in facial landmark detection than the LRS3-TED dataset, the angles estimated for successfully detected frames are comparable, suggesting the suitability of the KMSAV dataset for lip-reading studies.

The speaking style in the KMSAV dataset is spontaneous rather than dictatorial, adding to the complexity of speech recognition. Consequently, the error rate is expected to be considerably higher than that obtained from the LRS3-TED dataset, which has a word error rate below 2.0% when used in state-of-the-art AVSR methods [5].

3 | ASR

To demonstrate the effectiveness of the constructed KMSAV dataset for ASR, we conducted ASR experiments using the OpenAI Whisper system. Whisper models were trained using approximately 640 000 h of speech data, including 438 000 h of English, and 800 h Korean speech recognition data [19].

Initially, without fine-tuning, we deployed the Whisper models to decode the test data in the KMSAV dataset. For these experiments, we used the Whisper tools and models as explained at https://github.com/openai/whisper. During decoding, we set the beam size to 3 without conditioning the current output on previous text. The second column of Table 6 lists the CER for various Whisper model sizes. Notably, even without fine-tuning, the results were acceptable, especially given the spontaneous nature of the utterances.

The performance was quantified in terms of the CER owing to the inherent ambiguity in Korean word spacing.

T A B L E 6 Speech recognition performance with OpenAI Whisper models.

Whisper model	Zero-shot	Fine-tuned	ERR
`tiny`	32.22	23.54	26.9
`base`	24.30	18.40	24.3
`small`	18.83	14.04	25.4
`medium`	16.02	11.90	25.7
`large-v2`	15.31	11.08	27.6

Note: CERs (%) are shown for the zero-shot and fine-tuned models for each size of the models. The error reduction ratio (ERR, %) is also shown.

T A B L E 4 Multi-speaker characteristics of KMSAV dataset.

Category	Turn change	%Overlap	Category	Turn change	%Overlap
beau	3.77	18.48	inte	1.58	10.57
docu	0.68	6.12	issu	0.87	8.78
dram	2.52	11.11	mukb	2.78	7.62
ente	3.66	24.25	news	0.59	5.80
game	2.12	6.64	scie	0.26	2.14

Note: Turn change is the average number of speaker changes in a videoclip over 10 s, and %Overlap is the duration of overlapping speech divided by the entire speech duration.

T A B L E 5 Results of facial landmark detection and orientation estimation for each frame of transcribed utterances in KMSAV and LRS3-TED datasets.

	%Failed frames	Horizontal orientation	No. of frames
KMSAV	11.7	8.99 ± 7.26	7.6M
LRS3-TED	8.83	10.1 ± 7.66	2.8M

Note: The horizontal orientation is averaged for all the frames, and the mean and standard deviation are shown.

All punctuation marks were included in the error calculations, and no text cleanup or postprocessing was applied to the recognizer output. Throughout the experiments in this study, we used the same error metrics for consistency.

We used the ESPnet toolkit (https://github.com/espnet/espnet) for fine-tuning the Whisper model. This fine-tuning utilized the training split of the KMSAV dataset. Over 10 training epochs, without employing early stopping, we averaged the three models that demonstrated the best accuracy on the validation split of the KMSAV dataset. Despite having only 77.5 h of fine-tuning data, we observed a substantial decrease in the CER across all models, resulting in a relative error reduction of approximately 25%.

4 | AVSR

To evaluate the effectiveness of the KMSAV dataset in audiovisual contexts, we conducted AVSR experiments using the same setup as for ASR. To highlight the benefits of the visual modality, we also tested speech recognition under noisy conditions. For these experiments, we employed the AV-HuBERT framework.

AV-HuBERT is an AVSR framework that uses audio and visual information by expanding the HuBERT input to multiple modalities. Conventional AVSR relies on supervised learning, necessitating a large amount of labeled data. On the other hand, AV-HuBERT employs self-supervised learning, thereby enhancing the performance while using only 10% of the conventionally required labeled data.

4.1 | Evaluation of pretrained model and data modality

We fine-tuned the pretrained model using the KMSAV dataset as described in [6]. Multiple pretrained models were available with different sizes and training methods.

We selected the base and large models that had been pretrained with LRS3-TED and English portion of VoxCeleb2 data, totaling 1759 h of unlabeled audiovisual data. In addition, we evaluated both the clean-pretrained and noisy-pretrained variants per model size. The noisy-pretrained models, during their pretraining, incorporated a noise-augmentation technique to enhance their noise robustness. During fine-tuning, we used 77.5 h from the KMSAV training set. A randomly initialized transformer decoder was coupled with a pretrained transformer encoder. For the first 4000 updates, the encoder remained frozen, allowing only the decoder to train. In our configuration, the 4000 updates roughly translated to 35 epochs. Throughout the fine-tuning experiments, we ran 100 epochs and selected the model that demonstrated the best performance on the validation set as the final model.

Table 7 lists the recognition performance in terms of CER after fine-tuning the AV-HuBERT model using the KMSAV dataset. The large model consistently outperformed the base model but by a small margin. However, when the pretrained model was not used and the model weights were randomly initialized, the CER considerably deteriorated. This result indicated the benefits of using a pretrained model. Note that the pretrained models were trained exclusively on English data and no Korean data were included.

To investigate the impact of the data modality on the AVSR performance, we adopted two approaches for fine-tuning and recognition: using solely audio and combining both audio and video. Integrating visual information with audio led to a reduction in CER across all the configurations, showing 9.8% to 16.6% of error reduction rate.

Notably, the lowest AVSR CER was 18.86%. However, upon fine-tuning and evaluation using the configuration detailed in Section 4, we obtained a CER of 11.08% for ASR. This discrepancy may be attributed to the larger dataset used for pretraining the ASR model, namely, 640 000 h for ASR compared with 1759 h for AVSR.

T A B L E 7 CER (%) of AV-HuBERT models fine-tuned with KMSAV data in different pretraining configurations and input modalities.

Model size	Pretrain type	Audio only	Audiovisual	ERR
Base	Clean	23.02	19.80	14.0
	Noisy	22.94	20.28	11.6
	None	66.68	55.62	16.6
Large	Clean	21.69	18.86	13.0
	Noisy	21.66	19.53	9.8
	None	74.83	66.05	11.7

Note: Clean, clean pretraining; noisy, noise-augmented pretraining; none, pretrained model is not used by randomly initializing the weights. Audio only and audiovisual indicate the data modalities used during fine-tuning and inference. ERR (%) indicates the relative error reduction rate between audio-only and audiovisual cases.

TABLE 8 CER (%) under audio interference at 0 dB, 5 dB, and 10 dB SNRs.

Pretrained model	Fine-tuning type	SNR (dB)	Modality: audio only				Modality: audiovisual			
			All	Speech	Music	Noise	All	Speech	Music	Noise
`large-clean`	*Clean*	0	73.21	95.59	60.46	54.71	56.26	72.70	46.27	40.90
		5	53.55	68.47	44.05	39.71	38.43	47.75	33.44	30.81
		10	36.95	43.41	32.05	31.28	28.27	31.07	25.48	25.52
		Clean	21.69				18.86			
`large-noisy`	*Clean*	0	64.99	82.40	53.19	48.25	47.58	61.62	39.91	37.70
		5	46.62	56.11	38.51	36.70	33.93	40.36	29.69	28.87
		10	33.18	37.70	29.69	29.39	26.42	28.83	24.42	24.30
		Clean	21.66				19.53			
`large-noisy`	*Noisy*	0	54.38	68.84	46.87	43.93	32.98	40.88	30.19	29.22
		5	39.94	47.04	35.49	35.38	26.20	30.09	25.00	24.75
		10	32.19	34.62	29.91	30.36	22.72	24.00	21.95	22.25
		Clean	24.42				19.38			

Note: Two Whisper pretrained models, `large-clean` and `large-noisy`, were fine-tuned using the KMSAV training data. The "fine-tuning type" indicates whether noise augmentation was applied (*noisy*) or not (*clean*). The modality refers to the types of input data used during both fine-tuning and inference.

4.2 | Evaluation of audio interference

To determine the robustness of the AVSR models to audio interference, we mixed noise to the test data and evaluated the recognition results. Table 8 lists the CER for various pretrained models and fine-tuning methods when exposed to such noise. For these noise robustness experiments, only the `large` pretrained models were employed.

As the noise source, we used the MUSAN dataset [30], as described in [6]. This dataset comprises the following signal categories: `speech`, `music`, and `noise`. We added each signal in signal-to-noise ratios (SNRs) of 0, 5, and 10 dB to the input audio signal. `all`-type indicates that all categories of signals were selected randomly. As shown in the first and second rows of Table 8, where fine-tuning type is *Clean*, the CER values worsened rapidly with increasing noise level. In the second row, where the pretrained model is `large-noisy` and fine-tuning type is *Clean*, the noise-augmented pretrained model was used. While the CER still worsened for a lower SNR, a consistent improvement in performance was observed compared with the use of the `large-clean` pretrained model.

To further enhance the AVSR performance under noisy conditions, we adopted noise-augmented fine-tuning, wherein noise was added to the input audio during fine-tuning. Adhering to the procedure detailed in [6], we infused random noise from the MUSAN dataset into our training data at a 0 dB SNR with a 25% probability of occurrence during fine-tuning. The AVSR performance results are listed in the third row of Table 8 for fine-tuning type *noisy*. The CER under noisy conditions was substantially and consistently enhanced by applying noise-augmentation fine-tuning. For example, the CER was reduced by 30.7% for the audiovisual case (from 47.58% to 32.98%) at 0 dB SNR under interference type `all`.

5 | CONCLUSION

To support a variety of research initiatives, such as speech recognition and speaker diarization, we meticulously transcribed 150 h of content and annotated each utterance with speaker labels. Particularly for AVSR, we enriched our dataset with 84.4 h of high-quality data, which was refined through automated detection and rigorous manual review.

We used Whisper and AV-HuBERT as state-of-the-art pretrained models for ASR and AVSR, respectively, to conduct speech recognition experiments and thus evaluate the applicability of the constructed KMSAV dataset. The experiments showed up to 11.8% of CER for ASR experiments, and 18.9% for AVSR when fine-tuned using a publicly available large-scale pretrained model. In addition, we found that the AVSR performance in noisy environments was less influenced by audio interference when using the audio and video modalities than when using only audio.

The discrepancy between the best performances of ASR and AVSR underscores the need for further research on utilizing multimodal data and fully harnessing the potential of each modality. All the resources required to replicate our results, including the datasets, code, and models, will be made publicly available online.

CONFLICT OF INTEREST STATEMENT

The authors declare that there are no conflicts of interest.

ORCID

Kiyoung Park https://orcid.org/0009-0004-8673-7608
Changhan Oh https://orcid.org/0009-0003-3444-0901
Sunghee Dong https://orcid.org/0000-0003-4092-506X

REFERENCES

1. T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, *Deep audio-visual speech recognition*, IEEE Trans. Pattern Anal. Machine Intellig. **44** (2022), no. 12, 8717–8727.

2. S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic, *End-to-end audiovisual speech recognition*, (IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP), Calgary, Canada), 2018, pp. 6548–6552.

3. J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, *An overview of noise-robust automatic speech recognition*, IEEE/ACM Trans. Audio Speech Lang. Process. **22** (2014), no. 4, 745–777.

4. J. Chung, A. Senior, O. Vinyals, and A. Zisserman, *Lip reading sentences in the wild*, (IEEE Conf. Comput. Vision Pattern Recognit. (CVPR), Honolulu, HI, USA), 2017, pp. 3444–3453.

5. P. Ma, A. Haliassos, A. Fernandez-Lopez, H. Chen, S. Petridis, and M. Pantic, *Auto-AVSR: audio-visual speech recognition with automatic labels*, (IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP), Rhodes Island, Greece), 2023, pp. 1–5.

6. B. Shi, W.-N. Hsu, and A. Mohamed, *Robust self-supervised audio-visual speech recognition*, arXiv preprint, 2022, DOI 10.48550/arXiv.2201.01763

7. I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, and R. Harvey, *Extraction of visual features for lipreading*, IEEE Trans. Pattern Anal. Machine Intell. **24** (2002), no. 2, 198–213.

8. E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy, *CUAVE: a new audio-visual database for multimodal human-computer interface research*, (IEEE Int. Conf. Acoust. Speech Signal Process., Orlando, FL, USA), 2002, DOI 10.1109/ICASSP.2002.5745028.

9. I. Anina, Z. Zhou, G. Zhao, and M. Pietikäinen, *OuluVS2: a multi-view audiovisual database for non-rigid mouth motion analysis*, (11th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG), Ljubljana, Slovenia), 2015, DOI 10.1109/FG.2015.7163155

10. T. J. Hazen, K. Saenko, C.-H. La, and J. R. Glass, *A segment-based audio-visual speech recognizer: data collection, development, and initial experiments*, (Proc. 6th Int. Conf. Multimodal Interfaces, ICMI '04, Association for Computing Machinery, New York, NY, USA), 2004, pp. 235–242.

11. J. Park, J.-W. Hwang, K. Choi, S.-H. Lee, J. H. Ahn, R.-H. Park, and H.-M. Park, *OLKAVS: an open large-scale Korean audio-visual speech dataset*, arXiv preprint, 2023, DOI 10.48550/arXiv.2301.06375.

12. A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, *wav2vec 2.0: a framework for self-supervised learning of speech representations*, (34th Conference Neural Information Processing Systems, Vancouver, Canada), 2020, pp. 12449–12460.

13. Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, *Multimodal transformer for unaligned multimodal language sequences*, (Proc. 57th Annu. Meet. Assoc. Comput. Ling., Florence, Italy), 2019, pp. 6558–6569.

14. B. Shi, W.-N. Hsu, K. Lakhotia, and A. Mohamed, *Learning audio-visual speech representation by masked multimodal cluster prediction*, arXiv preprint, 2022, DOI 10.48550/arXiv.2201.02184

15. T. Afouras, J. Son Chung, and A. Zisserman, *LRS3-TED: a large-scale dataset for visual speech recognition*, arXiv preprint, 2018, DOI 10.48550/arXiv.1809.00496

16. A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. Freeman, and M. Rubinstein, *Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation*, ACM Trans. Graph. **37** (2018), no. 4, 1–11.

17. J. S. Chung, A. Nagrani, and A. Zisserman, *VoxCeleb2: deep speaker recognition*, (Proc. INTERSPEECH, Hyderabad, India), 2018, pp. 1086–1090. DOI 10.21437/Interspeech.2018-1929

18. T. Likhomanenko, Q. Xu, V. Pratap, P. Tomasello, J. Kahn, G. Avidov, R. Collobert, and G. Synnaeve, *Rethinking evaluation in ASR: are our models robust enough?* (INTERSPEECH, Brno, Czechia), 2021, pp. 311–315.

19. A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, *Robust speech recognition via large-scale weak supervision*, (Int. Conf. Mach. Learn., Honolulu, HI, USA), 2023, pp. 28492–28518.

20. Y. Zhang, D. Park, W. Han, J. Qin, A. Gulati, J. Shor, A. Jansen, Y. Xu, Y. Huang, S. Wang, Z. Zhou, B. Li, M. Ma, W. Chan, J. Yu, Y. Wang, L. Cao, K. Sim, B. Ramabhadran, and Y. Wu, *BigSSL: exploring the frontier of large-scale semi-supervised learning for automatic speech recognition*, IEEE J. Sel. Top. Signal Process. **16** (2022), 1–14.

21. M. Cooke, J. Barker, S. P. Cunningham, and X. Shao, *An audio-visual corpus for speech perception and automatic speech recognition*, J. Acoust. Soc. Am. **120** (2006), no. 5, 2421–2424, DOI 10.1121/1.2229005

22. B. Lee, M. Hasegawa-Johnson, C. Goudeseune, S. Kamdar, S. Borys, M. Liu, and T. Huang, AVICAR: audio-visual speech corpus in a car environment, (Proc. INTERSPEECH, Jeju, Rep. of Korea), 2004, pp. 2489–2492.

23. G. Zhao, M. Barnard, and M. Pietikainen, *Lipreading with local spatiotemporal descriptors*, IEEE Trans. Multimed. **11** (2009), no. 7, 1254–1265.

24. J. S. Chung and A. Zisserman, *Lip reading in the wild*, (Proc. Asian Conf. Comput. Vision, Taipei, Taiwan), 2016, pp. 87–103.

25. A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, *Voxceleb: large-scale speaker verification in the wild*, Comput. Speech Lang. **60** (2020), 101027, DOI 10.1016/j.csl.2019.101027

26. W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C.-Y. Fu, and A. C. Berg, *SSD: single shot multibox detector*, Computer Vision—ECCV 2016, B. Leibe, J. Matas, N. Sebe, and M.

Welling, (eds.), Lecture Notes in Computer Science, Vol. **9905**, Springer, Cham, 2016, pp. 21–37.

27. S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, *ESPnet: end-to-end speech processing toolkit*, (Proc. INTERSPEECH, Hyderabad, India), 2018, pp. 2207–2211.

28. R. Tao, Z. Pan, R. K. Das, X. Qian, M. Z. Shou, and H. Li, *Is someone speaking? Exploring long-term temporal features for audio-visual active speaker detection*, (Proc. 29th ACM Int. Conf. Multimedia, Association for Computing Machinery, New York, NY, USA), 2021, pp. 3927–3935.

29. D. E. King, *Dlib-ml: a machine learning toolkit*, J. Mach. Learn. Res. **10** (2009), 1755–1758.

30. D. Snyder, G. Chen, and D. Povey, *MUSAN: a music, speech, and noise corpus*, arXiv preprint, 2015, DOI 10.48550/arXiv.1510.08484

AUTHOR BIOGRAPHIES

Kiyoung Park received the MS and PhD degrees in Electrical Engineering from the Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea, in 1999 and 2003, respectively. From 2003 to 2005, he worked with the Samsung Advanced Institute of Technology, Yongin, Republic of Korea, where he contributed to the research and development of human–machine interaction systems. Since 2005, he has been with the Electronics and Telecommunication Research Institute, Daejeon, Republic of Korea, where he is now a principal researcher. Since 2022, he also has been with the University of Science and Technology. His main research interests include speech recognition, signal processing, and machine learning.

Changhan Oh received the BS degree from the Department of Computer Engineering, Sejong University, Seoul, Republic of Korea, in 2023. He is currently pursuing the MS degree at the University of Science and Technology, Daejeon, Republic of Korea, and is a student researcher at the Electronics and Telecommunication Research Institute, Daejeon, Republic of Korea. His main research interests include speech recognition and machine learning.
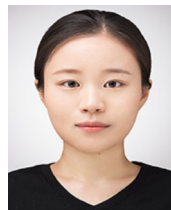
Sunghee Dong received the BS degree from the Department of Physics, Korea University, Seoul, Republic of Korea, in 2014, and the PhD degree from the Department of Brain and Cognitive Engineering, Korea University, Seoul, Republic of Korea, in 2019. Since 2019, she has been with the Electronics and Telecommunication Research Institute, Daejeon, Republic of Korea, where she is now a senior researcher. Her main research interests include speech separation, speech enhancement, speech recognition, and machine learning.