ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10446901

OLKAVS: AN OPEN LARGE-SCALE KOREAN AUDIO-VISUAL SPEECH DATASET

Jeongkyun Park[1] *Jung-Wook Hwang*[2] *Kwanghee Choi*[3] *Seung-Hyeon Lee*[2]
Jun Hwan Ahn[4] *Rae-Hong Park*[2,5] *Hyung-Min Park*[1,2,5,†]

[1] Department of Artificial Intelligence, Sogang University, Seoul 04107, Republic of Korea
[2] Department of Electronic Engineering, Sogang University, Seoul 04107, Republic of Korea
[3] Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA
[4] Mindslab Inc., Gyeonggi-do 13493, Republic of Korea
[5] ICT Convergence Disaster/Safety Research Institute, Sogang University, Seoul 04107, Republic of Korea

ABSTRACT

Inspired by humans comprehending speech in a multi-modal manner, various audio-visual datasets have been constructed. However, most existing datasets focus on English, developed from pre-existing videos using various prediction models, and have only a small number of multi-view videos. To mitigate the limitations, we constructed the Open Large-scale Korean Audio-Visual Speech (OLKAVS) dataset, which is the largest among publicly available audio-visual speech datasets. The dataset contains 1,150 hours of transcribed audio from 1,107 Korean speakers in a studio setup with nine different viewpoints and various noise situations. We also provide the pre-trained baseline models for two tasks: audio-visual speech recognition and lip reading. We conducted experiments based on the models to verify the effectiveness of multi-modal and multi-view training over uni-modal and frontal-view-only training. We expect the OLKAVS dataset to facilitate multi-modal research in broader areas.

***Index Terms*—** Audio-visual speech datasets, multi-view datasets, lip reading, audio-visual speech recognition, deep learning

1. INTRODUCTION

People comprehend speech in a multi-modal manner in real-world communication, not only relying on listening to utterances but also reading speakers' faces and lips to improve perception. Inspired by this, many audio-visual speech datasets have been introduced, capturing both utterances and talking faces in a synchronized manner. These datasets enabled the research on various speech-related applications, such as noise-robust speech recognition [11], lip reading [12, 13], mouth motion analysis [1], and speaker recognition [6]. To further embolden future research, it is essential to develop datasets with more diverse situations, such as various speakers [4, 13, 14], languages [6, 7, 14], sentences [4, 13], and multiple viewpoints of the head [1, 5].

For the research on lip reading, multiple datasets [15, 16, 17, 18] were constructed, albeit the limited vocabulary and sample size of classical datasets, containing less than 50 speakers. OuluVS2 [1] and AVICAR [2] considered full sentences with more various vocabularies, more than a thousand words. Interestingly, they also provided multi-view videos from four and five different angles, respectively.

To develop larger datasets for word utterances and reflect real-world speech with unrestricted vocabulary, LRW proposed a data collection pipeline to label TV news footage automatically [12]. The pipeline-based approach has influenced numerous follow-up sentence-based audio-visual speech datasets: LRS2-BBC [3, 13], LRS3-TED [4], and MV-LRS [5]. MV-LRS primarily focused on providing multi-view video, dividing the data into five categories based on the face angle to facilitate the multi-view evaluation. Leveraging the development of the collection pipeline from [12], some non-English datasets were introduced such as LRW-1000 [14], GLips [7], and MISP2021 [8]. VoxCeleb2 [6] also contains multilingual speech in more than 100 languages, but there are no transcriptions available in the dataset.

Despite the development in audio-visual datasets, only a few attempts have been made to collect Korean speech. [9] collected predefined syllables of a single speaker with 7 views. Also, [10] and [19] collected predefined word utterances, such as digits and city names, of 56 and 9 speakers, respectively. Unfortunately, the size of all the datasets is too minuscule to support deep learning-driven models; moreover, some are not publicly available.

Despite the increasing attention to the audio-visual speech domain, existing datasets have a number of limitations.

† Corresponding author: hpark@sogang.ac.kr.

This work was partly supported by National Information Society Agency (NIA) grant (AI HUB Voice and Motion Data - Lip Voice) and Institute of Information & communications Technology Planning & evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2022-0-00989, Development of Artificial Intelligence Technology for Multi-speaker Dialog Modeling). The audio-visual dataset collection project was proposed based on the National Information Society Agency (NIA)'s request for proposal (AI HUB Voice and Motion Data - Lip Voice), funded by the Korea government (MSIT). After obtaining consent from the participants, the dataset was collected and published on https://aihub.or.kr, an official portal that discloses various AI data. All of this was done under the supervision of the NIA.

Table 1: Summarized statistics of various audio-visual speech datasets. The largest values are written in bold. Some datasets containing various head poses without explicitly fixed views are listed under 'unfixed'. Hours, #Utt., and #Subj. of LRS2-BBC and LRS3-TED are approximations that can be slightly larger due to potential overlaps in the dataset.

Database	Hours	#Utt.	#Subj.	Views	Language	Content
OuluVS2 [1]	-	1.5K	53	5 views	English	digits, phrases, TIMIT sentences
AVICAR [2]	-	59K	86	4 views	English	digits, letters, TIMIT sentences
LRS2-BBC [3]	224.5	144K	-	unfixed	English	sentences
LRS3-TED [4]	438	152K	**9,545**	unfixed	English	sentences
MV-LRS [5]	777.2	504K	-	5 views	English	sentences
VoxCeleb2 [6]	2,442	1.1M	6,112	unfixed	Multilingual	speaker info. only
GLips [7]	81	250K	-	unfixed	German	words
MISP2021 [8]	141.2	-	263	3 views	Mandarin	sentences
Jo *et al.* [9]	-	168	1	7 views	Korean	syllables, consonants, vowels
Lee *et al.* [10]	-	4.5K	56	frontal	Korean	digits, words
OLKAVS	Video: **5,750** Audio: 1,150	**2.5M**	1,107	**9 views**	Korean	sentences

1) Foremost, most datasets, especially large-scale datasets, only focus on English alone, with a few exceptions. 2) Even though many large-scale datasets have been introduced, they often rely on publicly available videos on the TV and web, introducing an inherent dependency on various detection models in collection pipelines and sometimes encountering copyright restrictions. 3) Many speech datasets with multi-view videos are too small for modern deep learning, with less than 100 hours. The MV-LRS [5] stands out with its 777 hours of face rotations, but it offers only one view at a time. Refer to Table 1 for more details.

To mitigate the limitations, we developed a new audio-visual speech dataset, **Open Large-scale Korean Audio-Visual Speech (OLKAVS)** dataset. We gathered 1,150 hours of audio from 1,107 speakers in a studio setup with corresponding Korean transcriptions. We recorded a speaker from 9 viewpoints: frontal, cross (left, right, up, and down), and diagonal (upper left, upper right, lower left, and lower right). The OLKAVS' advantages can be summarized as follows.

1. It is the only Korean audio-visual speech dataset with a significant amount of more than hundreds of hours while many datasets focus on English.
2. Among publicly available datasets, it is the largest audio-visual speech dataset. In particular, it provides complete manual transcriptions for all utterances.
3. It is a large-scale dataset providing synchronously recorded multi-view videos.

OLKAVS can advance research in multi-modal tasks like audio-visual speech recognition (AVSR) and lip reading, as well as in speech-only tasks such as speech and speaker recognition. For AVSR and lip-reading, we have publicly released the pre-trained baseline models to underscore the dataset's value and assist future research.[1]

In this paper, we aim to provide a complete guideline for the dataset. The details of how the dataset was constructed are enumerated in Section 2. Also, we demonstrated our dataset works on two downstream tasks, AVSR and multi-view lip reading in Section 3.

2. OLKAVS DATASET

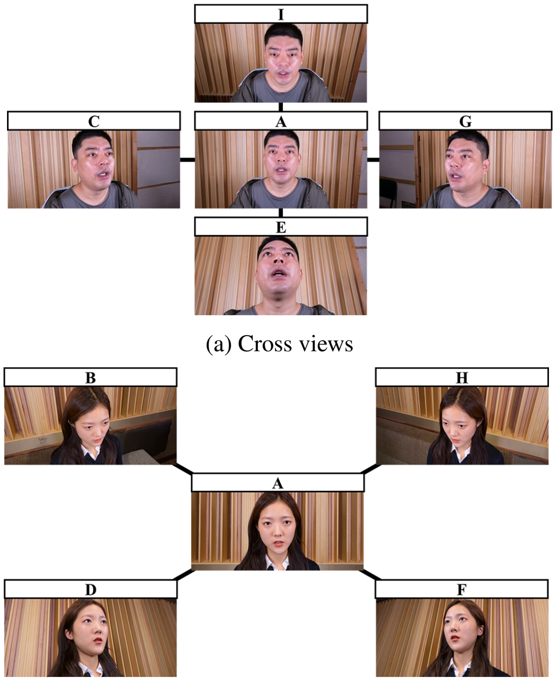
In this section, we describe our dataset in detail. OLKAVS contains 1,150 and 5,750 hours of audio and video in total, respectively, where we recorded five videos per single audio. To the best of our knowledge, this is the largest audio-visual speech dataset, especially for multi-view datasets.

2.1. Scripts and Speakers

The OLKAVS dataset contains 14 different topics, where we constructed a set of sentences based on their corresponding keywords. Each topic has at least 2,000 sentences, so the total sentences exceed 28,000. Each sentence is written in the Korean language. Then, randomly chosen sentences from each topic are combined to build a script with an average of 2,750 characters (5 minutes long on average). Generated scripts are provided to the speakers when recording.

A total of 1,107 speakers participated in constructing the OLKAVS dataset, and all the speakers formally consented to using portrait rights and disclosing personal information. They were recruited while considering the demography of biological sex and age. The numbers of males and females are balanced with 555 males and 552 females. The age of speakers ranges from 10 to 60 and above, following its distribution within Korea. Note that the teenagers were specifically asked to speak on six topics chosen from the total set of 14, as these topics were deemed more understandable for younger people (e.g., food, education/school), while the others spoke on all 14 topics.

[1] The dataset and the pre-trained models are available at `https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=538` and `https://github.com/IIP-Sogang/olkavs-avspeech`, respectively.



(a) Cross views

(b) Diagonal views

Fig. 1: Video samples from the OLKAVS dataset. Five different views were synchronously recorded for each recording session. (a) and (b) show two types of views and denote their labels.

30% of the speakers are speech experts, such as announcers, actors, would-be announcers, and would-be actors. Also, 50% of the experts' speeches were done spontaneously about the provided news article, while it was not collected from the other speakers due to the difficulty of spontaneous speech.

2.2. Recording Environment

Each utterance was recorded from five different views simultaneously by GoPro Hero7 cameras. As shown in Fig. 1, the captured views comprise two types of recorded views: cross views (frontal, up, down, left, and right) and diagonal views (frontal, upper left, upper right, lower left, and lower right). Notably, the frontal view is present in both categories. Unlike prior datasets [1, 5, 9] that only consider horizontal views, i.e., left, frontal, and right, OLKAVS dataset also considers the vertical variants, i.e., upper left, upper center, upper right, lower left, lower center, and lower right.

A condenser microphone was employed for recording speech in stereo, ensuring a consistent distance of about 1 m between the speaker and microphone, as well as the frontal-view camera. A soundproof booth measuring about 1.5 m × 1.5 m × 2.0 m was utilized to capture the speech. To simulate noisy conditions inside the studio environment, we played the pre-recorded noise inside the studio. We constructed six noise scenarios: no noise, indoor ambiance, indoor noise (laughing and clapping), traffic noise, construction site noise, and natural outdoor noise (flowing water and animal sounds). Noise-free scenarios account for 29% of total hours, each of the others making up 14%. All the noise audios were collected from the corresponding real-world situations.

2.3. Data Cleansing and Post-processing

After recording, we transcribed all the scripted and spontaneous speech. Even though the scripts were given to the speakers, some of the speakers mispronounced a few words or spoke pause fillers such as 'um' or 'ah,' where everything was transcribed afterward. All transcriptions are the ground truth for speech, being provided only in Korean language.

We manually filtered out the problematic recordings by the following failure criteria: the entire lip or facial region is not recorded, the filming environment is too dark, the noise level is too high, or videos and the audio are not in sync. We also trimmed every audio-visual pair to have the same length.

The OLKAVS dataset provides predefined splits of train, validation, and evaluation with a ratio of 8:1:1. The dataset is split based on speaker ID so that no speaker co-occurs on different splits. Only the train and validation sets are made publicly available. For ease of use, we also provide the bounding box coordinates of the lip and the face. We used the face alignment estimation module from [20]. Facial landmarks were extracted every 3 seconds and linearly interpolated for the frames in between.

Videos are stored in MP4 format with the FHD (1,920×1,080) resolution with 30 fps, while audios are encoded in WAV format with stereo channels and 48 kHz sampling rates. Label data include transcriptions, the utterances' start and end times, speaker information on ID, age, gender, and whether the speaker is an expert.

3. EXPERIMENTS

In this section, we outline OLKAVS dataset baselines for AVSR and multi-view lip reading. The former examines multi-modal versus uni-modal training, while the latter assesses the advantage of multi-view data over frontal-view-only training.

In both experiments, videos were cropped with the lip-bounding boxes we provided. Then, we split the speech sentence-wise, based on the timestamps in the transcripts. We tokenized the transcripts into grapheme units of 21 vowels, 19 initial consonants, and 27 final consonants, removed punctuations, and converted numbers into corresponding words.

3.1. AVSR

We employed the conformer sequence-to-sequence (CM-Seq2seq) architecture [11] for the audio-visual model (AV-model), using 8-head attention and 256 feature dimensions for

Table 2: Baseline performances of A-model, V-model, and AV-model on the OLKAVS validation set. 'Val.' refers to the full validation dataset published on the official portal. 'Clean' indicates a noise-free validation subset while '(number) dB' represents the same subset, but corrupted by adding babble noise from NOISEX-92 [21] to get the designated input SNR

Model	CER (%)							WER (%)							sWER (%)						
	Val.	Clean	15dB	10dB	5dB	0dB	−5dB	Val.	Clean	15dB	10dB	5dB	0dB	−5dB	Val.	Clean	15dB	10dB	5dB	0dB	−5dB
A	**3.57**	2.00	2.29	2.89	4.79	11.15	30.76	**10.61**	**7.27**	7.96	9.24	12.97	24.33	54.03	**8.11**	**4.36**	5.13	6.47	10.73	23.82	58.72
V	26.64	25.08	25.08	25.08	25.08	25.08	25.08	47.89	45.20	45.20	45.20	45.20	45.20	45.20	50.00	46.41	46.41	46.41	46.41	46.41	46.41
AV	3.64	**1.98**	**2.24**	**2.70**	**4.09**	**8.42**	**19.75**	10.82	7.29	**7.83**	**8.88**	**11.87**	**20.34**	**39.97**	8.18	4.37	**5.01**	**6.08**	**9.32**	**18.54**	**39.57**

Table 3: CERs (%) of lip reading on a subset of validation data with no noise in the OLKAVS dataset

View / Model	Upper			Medium-high			Lower		
	left	center	right	left	center (frontal)	right	left	center	right
F-model (Frontal only)	73.58	75.99	77.88	63.19	41.24	71.78	59.22	53.99	61.99
All-model (All views)	**41.02**	**46.14**	**42.43**	**34.93**	**32.16**	**34.60**	**31.91**	**33.55**	**32.21**

the 12-layered conformer encoder and 6-layered transformer decoder. For visual processing, we utilized a pre-trained front-end from [3] trained on several English lip reading datasets [3, 5, 12] and froze them during training.

During training, we introduced additional noise from DEMAND [22] and NOISEX-92 [21] to simulate more challenging environments than those encountered in our recordings, including scenarios with superimposed noise. They varied six different input signal-to-noise ratios (SNRs): −5 dB, 0 dB, 5 dB, 10 dB, 15 dB, and 20 dB. We used the Adam optimizer [23] with a 0.0001 learning rate and the Noam learning rate scheduler [24]. Furthermore, for the audio-only (A-model) and visual-only (V-model) baselines, we simply removed the corresponding encoder and the fusion module, following [11].

For the baseline evaluation (Table 2), we evaluated all three models, A-model, V-model, and AV-model, on the publicly available validation set containing various noises, as mentioned in Subsection 2.2. To evaluate the noise robustness of the models, we varied the noise levels explicitly by mixing the noise-free validation subset with babble noise from NOISEX-92 [21]. The performance was measured in terms of character error rate (CER), word error rate (WER), and space-normalized word error rate (sWER) [25], which were computed for both Korean characters and words.

We provided speech recognition baselines for the OLKAVS dataset in Table 2, which contains the first published outcomes for the Korean AVSR and lip reading at the sentence level. The AV-model demonstrated a notable performance improvement over the A-model under noisy conditions while exhibiting comparable performance with the A-model for the clean data, indicating the effectiveness of utilizing audio-visual modalities in speech recognition tasks.

3.2. Multi-view Lip Reading

We adopted the V-model architecture and the optimization settings from Subsection 3.1 but trained it from scratch to thoroughly examine the visual representations.

To assess the benefit of using all the views in enhancing the lip-reading performance [26], we compared how the lip-reading models of the same architecture perform when trained only with the frontal view (F-model) versus using all the views (All-model). For evaluation, we used a noise-free subset of the validation data.

Table 3 demonstrates that the frontal lip reading performance of the All-model (32.16%) was superior to that of the F-model (41.24%) with an 22% relative reduction in CER, which reaffirmed the results of [26] that utilizing multi-view training data was superior to that of the frontal-view-only. We also analyzed the difficulty of lip reading from each view by comparing the performance on various views. Confirming [26], centered views underperformed right and left views in lip reading, except at the Medium-high level. This may result from model bias towards the frontal view, which has double the training data of the others. It is noteworthy that its performance was better than that of the upper and medium-high angles since the lower angle showed lip movement better. These results underscore the promise of using non-frontal views, guiding our future works to leverage all synchronized views for enhanced lip representations.

4. CONCLUSION

In this paper, we introduced a novel audio-visual speech dataset, the Open Large-scale Korean Audio-Visual Speech (OLKAVS) dataset, which is the largest publicly available audio-visual speech dataset, to the best of our knowledge. We also provided Korean baselines for AVSR and multi-view lip reading, demonstrating that using both multi-modal and multi-view data during training improved upon the uni-modal and frontal-view-only data. Finally, we emphasize that the OLKAVS dataset can be utilized for broader research areas outside of AVSR and lip reading, such as Korean speech recognition, speaker recognition, and mouth motion analysis.

5. REFERENCES

[1] Iryna Anina, Ziheng Zhou, Guoying Zhao, and Matti Pietikäinen, "OuluVS2: A multi-view audiovisual database for non-rigid mouth motion analysis," in *Proc. IEEE Int. Conf. and Workshops on Automatic Face and Gesture Recognition (FG)*, 2015, vol. 1, pp. 1–5.

[2] Bowon Lee, Mark Hasegawa-Johnson, Camille Goudeseune, Suketu Kamdar, Sarah Borys, Ming Liu, and Thomas Huang, "AVICAR: Audio-visual speech corpus in a car environment," in *Proc. Int. Conf. on Spoken Language Processing*, 2004.

[3] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman, "Deep audio-visual speech recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 8717–8727, 2022.

[4] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman, "LRS3-TED: A large-scale dataset for visual speech recognition," *arXiv preprint arXiv:1809.00496*, 2018.

[5] Joon Son Chung and Andrew Zisserman, "Lip reading in profile," in *Proc. British Machine Vision Conf.*, 2017.

[6] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "VoxCeleb2: Deep speaker recognition," in *Proc. Interspeech*, 2018, pp. 1086–1090.

[7] Gerald Schwiebert, Cornelius Weber, Leyuan Qu, Henrique Siqueira, and Stefan Wermter, "A multimodal German dataset for automatic lip reading systems and transfer learning," in *Proc. Language Resources and Evaluation Conf.*, 2022, pp. 6829–6836.

[8] Hang Chen, Jun Du, Yusheng Dai, Chin Hui Lee, Sabato Marco Siniscalchi, Shinji Watanabe, Odette Scharenborg, Jingdong Chen, Bao Cai Yin, and Jia Pan, "Audio-visual speech recognition in MISP2021 challenge: Dataset release and deep analysis," in *Proc. Interspeech*, 2022, pp. 1766–1770.

[9] Cheol-Woo Jo, Roland Goecke, and Bruce Millar, "Collection of Korean audio-video speech data," *Speech Sciences*, vol. 7, no. 1, pp. 5–15, 2000.

[10] Jong-Seok Lee and Cheol Hoon Park, "Robust audio-visual speech recognition based on late integration," *IEEE Transactions on Multimedia*, vol. 10, no. 5, pp. 767–779, 2008.

[11] Pingchuan Ma, Stavros Petridis, and Maja Pantic, "End-to-end audio-visual speech recognition with conformers," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, 2021, pp. 7613–7617.

[12] Joon Son Chung and Andrew Zisserman, "Lip reading in the wild," in *Proc. Asian Conf. on Computer Vision*, 2016, pp. 87–103.

[13] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman, "Lip reading sentences in the wild," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2017, pp. 6447–6456.

[14] Shuang Yang, Yuanhang Zhang, Dalu Feng, Mingmin Yang, Chenhao Wang, Jingyun Xiao, Keyu Long, Shiguang Shan, and Xilin Chen, "LRW-1000: A naturally-distributed large-scale benchmark for lip reading in the wild," in *Proc. IEEE Int. Conf. and Workshops on Automatic Face and Gesture Recognition (FG)*, 2019, pp. 1–8.

[15] Iain Matthews, Timothy Cootes, Andrew Bangham, Stephen Cox, and Richard Harvey, "Extraction of visual features for lipreading," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 2, pp. 198–213, 2002.

[16] Eric Patterson, Sabri Gurbuz, Zekeriya Tufekci, and John Gowdy, "CUAVE: A new audio-visual database for multimodal human-computer interface research," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, 2002, vol. 2, pp. 2017–2020.

[17] Guoying Zhao, Mark Barnard, and Matti Pietikainen, "Lipreading with local spatiotemporal descriptors," *IEEE Transactions on Multimedia*, vol. 11, no. 7, pp. 1254–1265, 2009.

[18] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[19] Sung-Won Lee, Je-Hun Yu, Seung Min Park, and Kwee-Bo Sim, "Visual speech recognition of Korean words using convolutional neural network," *Int. Journal of Fuzzy Logic and Intelligent Systems*, vol. 19, no. 1, pp. 1–9, 2019.

[20] Adrian Bulat and Georgios Tzimiropoulos, "How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks)," in *Proc. Int. Conf. on Computer Vision*, 2017, pp. 1021–1030.

[21] Andrew Varga and Herman JM Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[22] Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent, "The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings," *Proceedings of Meetings on Acoustics*, vol. 19, no. 1, pp. 035081, 2013.

[23] Diederik Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *Proc. International Conference on Learning Representations*, 2015.

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, vol. 30.

[25] Jeong-Uk Bang, Seung Yun, Seung-Hi Kim, Mu-Yeol Choi, Min-Kyu Lee, Yeo-Jeong Kim, Dong-Hyun Kim, Jun Park, Young-Jik Lee, and Sang-Hun Kim, "Ksponspeech: Korean spontaneous speech corpus for automatic speech recognition," *Applied Sciences*, vol. 10, no. 19, pp. 6936, 2020.

[26] Daehyun Lee, Jongmin Lee, and Kee-Eung Kim, "Multi-view automatic lip-reading using neural network," in *Proc. Asian Conf. on Computer Vision*, 2017, pp. 290–302.