Received: 11 July 2023 | Revised: 10 November 2023 | Accepted: 20 December 2023

DOI: 10.4218/etrij.2023-0266

ORIGINAL ARTICLE

ETRI Journal WILEY

Multimodal audiovisual speech recognition architecture using a three-feature multi-fusion method for noise-robust systems

Sanghun Jeon[1,2] | Jieun Lee[2] | Dohyeon Yeo[2] | Yong-Ju Lee[1] | SeungJun Kim[2]

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

[2]Gwangju Institute of Science and Technology, School of Integrated Technology, Gwangju, Republic of Korea

Correspondence
SeungJun Kim, Gwangju Institute of Science and Technology, School of Integrated Technology, Gwangju, Republic of Korea.
Email: seungjun@gist.ac.kr

Funding information
This study was supported by the following grants: an Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (Ministry of Science and ICT, MSIT) (No. 2022-0-00871, Development of AI Autonomy and Knowledge Enhancement for AI Agent Collaboration, 60%); a National Research Foundation of Korea (NRF) grant funded by the Ministry of Science and ICT (2021R1A4A1030075, 20%); and a GIST-MIT Research Collaboration grant funded by the GIST in 2023 (20%).

Abstract

Exposure to varied noisy environments impairs the recognition performance of artificial intelligence-based speech recognition technologies. Degraded-performance services can be utilized as limited systems that assure good performance in certain environments, but impair the general quality of speech recognition services. This study introduces an audiovisual speech recognition (AVSR) model robust to various noise settings, mimicking human dialogue recognition elements. The model converts word embeddings and log-Mel spectrograms into feature vectors for audio recognition. A dense spatial–temporal convolutional neural network model extracts features from log-Mel spectrograms, transformed for visual-based recognition. This approach exhibits improved aural and visual recognition capabilities. We assess the signal-to-noise ratio in nine synthesized noise environments, with the proposed model exhibiting lower average error rates. The error rate for the AVSR model using a three-feature multi-fusion method is 1.711%, compared to the general 3.939% rate. This model is applicable in noise-affected environments owing to its enhanced stability and recognition rate.

KEYWORDS
application programming interface, audiovisual speech recognition, lip reading, multimodal interaction

1 | INTRODUCTION

Human speech is bimodal and involves audio and visual information. Audio information detects the acoustic waveform of a speaker, whereas visual information detects lip movements [1]. Despite the challenges such as auditory recognition in noisy environments, audiovisual speech recognition (AVSR) is widely investigated and is reported to exhibit excellent recognition capabilities [2–6]. AVSR is used in technologies such as Microsoft Azure, Google Assistant, and Amazon Alexa, which convert analog signals into digital formats by acoustically analyzing speech and automatically transcribing it into

Sanghun Jeon, Jieun Lee, and Dohyeon Yeo equally contributed to this work.

text [7–9]. However, commercial speech recognition application programming interface (API) services are primarily used indoors or in specific environments because of low recognition rates in outdoor settings, where background noise can degrade the quality of acoustic waveforms. This has led to the emergence of noise robustness as a critical factor for the realization of large-scale real-world applications because open cloud-based speech recognition application programming interface (OCSR API) services must demonstrate improved capabilities in challenging acoustic scenarios. [10, 11] Nevertheless, AVSR technologies that combine auditory and visual information can mitigate the influence of ambient noise, resulting in high recognition rates, even in the presence of background noise. Accordingly, few studies have reported that superior performance is achieved when hearing and vision are combined in several groups. [12, 13].

Humans use a bimodal interaction method based on the multisensory integration of auditory (represented in red, processed in the temporal lobe) and visual (represented in blue, processed in the occipital lobe) inputs to recognize language. Accordingly, a system that can reduce the impact of ambient noise was proposed (Figure 1) [12]. In other words, the human nervous system separately transmits information from different sensory organs, and the combination of inputs enhances the ability to react to, evaluate, and perceive external events with high accuracy [14–16]. Therefore, bimodality is preferred over unimodality to minimize the effect of ambient noise.

Speech and visual recognition present distinct challenges. In the case of speech recognition, because optimal performance cannot be guaranteed in all circumstances, exposure to varied noise environments degrades recognition performance. A system with decreased performance can be used only in a limited particular context before being applied to a real-world environment, resulting in low-quality service and reduced expectations for consumers. In the case of visual recognition, when audio information is not used, recognition issues exist in homophony words with identical mouth shapes. As a result, we must overcome each issue with speech and visual recognition. As a result, in this study, we developed a novel approach to address this issue.

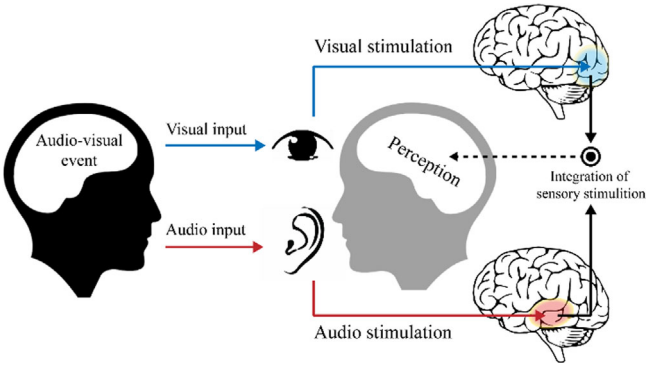

FIGURE 1 Multisensory integrated cognitive process according to external events.

We propose a noise-resistant system that combines an OCSR API, log-Mel spectrogram, and lip-movement data using multiple convergence methods. The bimodal system outperforms the single-modality system. For audio-based speech recognition, the recognized word lists from the OCSR API were represented as vectors using a pre-trained model to generate a one-dimensional (1D) vector. In contrast, for vision-based speech recognition, the log-Mel spectrogram image, in which the frequency unit of the standard spectrogram is converted to a log-Mel unit, and the lip-movement image are input to a newly proposed deep neural network comprising an end-to-end neural network to visualize audio speech data. A dense spatial–temporal three-dimensional (3D) convolutional neural network (CNN) was used for specific sequence image extraction to reduce the number of parameters, training time, and overfitting. Furthermore, an attention-mechanism-based spatial attention approach was used to extract features from the input sequence images, intensify the feature representation of the region of interest, focus on the information section locations, and supplement channel attention. Bidirectional gated recurrent units (GRUs) with linear layers were connected to prevent the overfitting of small-scale data, overcome the lack of visual information due to sequence image data, and obtain specific image features. Subsequently, a new vector matrix was created by combining the 1D word vectors generated by the audio and visual speech recognition models, log-Mel spectrogram, and feature vector generated by the lip movement image. The newly created vector matrix was decoded using a connectionist temporal classification loss function based on the beam search approach to obtain a predicted word, which was used to train the newly created vector matrix.

We evaluated the precision and effectiveness of our architecture compared with existing visual feature extraction algorithms that have exhibited excellent performance on a collected dataset to assess the performance of speech, visual, and spectrograms using LipNet [17] as the baseline model. Numerous evaluation findings demonstrate that the proposed architecture outperforms existing deep-learning techniques in terms of state-of-the-art performance and efficiency.

The remainder of this paper is organized as follows: Section 2 provides detailed information on the individual components of the proposed architecture, Section 3

describes our data collection environment and experimental process, and Section 4 presents the quantitative evaluation results. Finally, Section 5 concludes the paper.

2 | RELATED WORKS

Google [18] introduced Google Assistant, an artificial intelligence (AI)-powered speech recognition assistant with an open API that can be applied to various fields such as automobiles and home appliances. Microsoft's Azure Cognitive Services [19] is a collection of preconstructed, customizable AI models that can be used to add AI capabilities to applications across multiple domains such as speech, reading, language, and vision, including speech recognition technology. IBM Watson [20] is a cloud-native solution that utilizes deep-learning AI algorithms to create tailored speech recognition for optimal text transcription by leveraging knowledge of grammar, language structure, and audio/speech signal composition. Amazon Transcribe [21] is a service that converts speech to text and allows easy integration with any application. This technology has added a new privacy option for commercial services that removes personally identifiable information like name, credit card or bank number, and social security number automatically [22]. In addition to speech recognition, companies in Korea such as ETRI [23], NAVER [24], and KAKAO [25] offer various speech recognition services.

Deep learning technology has recently exhibited significant performance improvements in sentence-level speech recognition, visual speech recognition, and speech-visual recognition studies using traditional prediction techniques [26–29]. When comparing the word recognition rate performance for phrases on the same benchmark dataset, the performance of deep learning-based visual recognition studies [17] improved by 36.4% over conventional visual recognition studies [30]. However, with increasing training data complexity, traditional prediction techniques are faced with challenges such as a large number of speakers, posture changes, lighting conditions, and background environmental changes, among other issues. In response to these challenges, Google DeepMind introduced LipNet [17] in 2016, which is an end-to-end model that utilizes visual speech recognition technology to process sentence-level images of user lips and predicts them as character sequences. LipNet comprises an encoder that processes inputs combined with continuous two-dimensional (2D) images using a combination of space–time CNN and GRU, and a decoder that uses the connectionist temporal classification (CTC) loss function. The benchmark dataset GRID corpus [R] was used to evaluate the model performance. While lip-reading experts achieved a word error rate (WER) of 47.7% on the GRID corpus, LipNet demonstrated an impressive performance with a WER of 11.4%.

In 2022, Jeon and others proposed a new architecture that combined a commercial speech recognition API with a visual speech recognition model [30]. They proposed a learning method in which feature vector outputs from a visual speech recognition model are combined with word vectors generated by a pretrained word-embedding model for words recognized in a commercial speech recognition API and evaluated the performance in eight noisy environments. The results demonstrated that combining visual and audio information improved the performance across all environments compared to using only visual information. In conclusion, deep-learning-based prediction techniques have shown the ability to learn more deeply and extract more comprehensive features from complex data compared to traditional prediction techniques. Further, they have demonstrated superior suitability to big data and tackle visual ambiguity.

Therefore, the previous studies mainly proposed models that combine cloud-based speech recognition services with visual information, and log-Mel spectrogram information with visual information. However, the model proposed in this study differs from previous studies as it fuses three modalities using cloud-based speech recognition results in audio information, log-Mel spectrograms in visual information, and lip-movement information. Previous studies confirm that deep-learning-based visual recognition techniques must be utilized to address visual ambiguity in speech recognition using visual information. Additionally, previous research [31–33] primarily focused on evaluating and comparing the performance of OCSR APIs in noise-free or artificially generated noise environments using benchmark datasets. However, performance evaluation in various real-world noise environments is essential to using speech recognition technology in real-world applications, and to this end, we conducted a performance evaluation of the proposed model in nine real-world noise environments. Furthermore, we attempted to merge two types of visual information (lip-movement information and log-Mel spectrograms) into the OCSR API for noise-robust and superior performance in diverse noisy conditions, and the proposed model exhibited greater recognition rates in noisy environments when compared to the existing OCSR API system. Unlike the audiovisual speech recognizer model, which uses only the user's lip information, this performance gain was accomplished by combining lip movement information, log-Mel spectrograms, and auditory information.

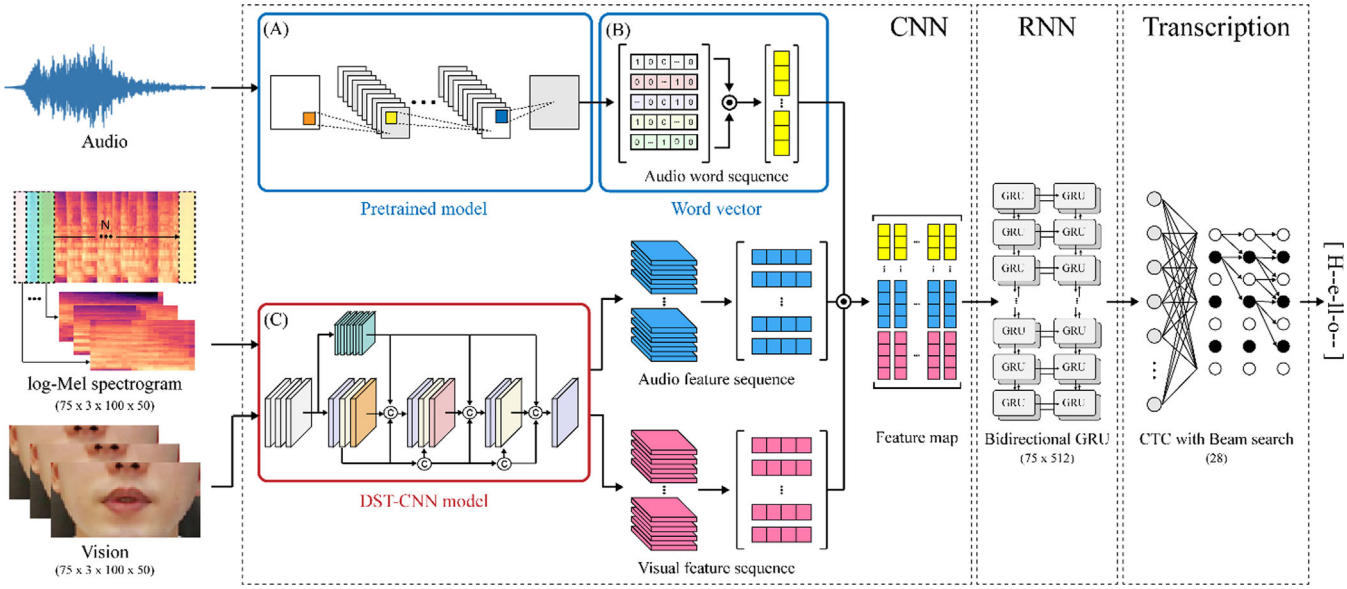

FIGURE 2 Block diagram of the proposed multimodal AVSR architecture. (A) Pretrained Word2Vec embedding model; (B) word vector process; (C) dense spatial–temporal CNN model structure.

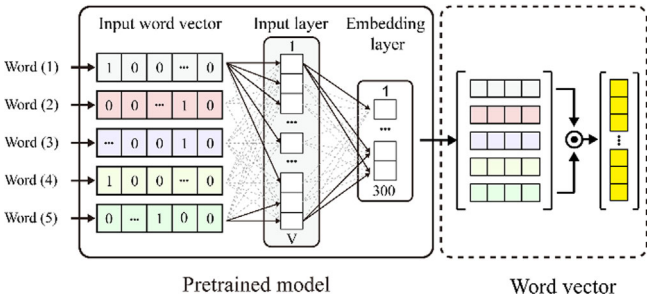

FIGURE 3 Process of the proposed audio module from input audio to word vector.

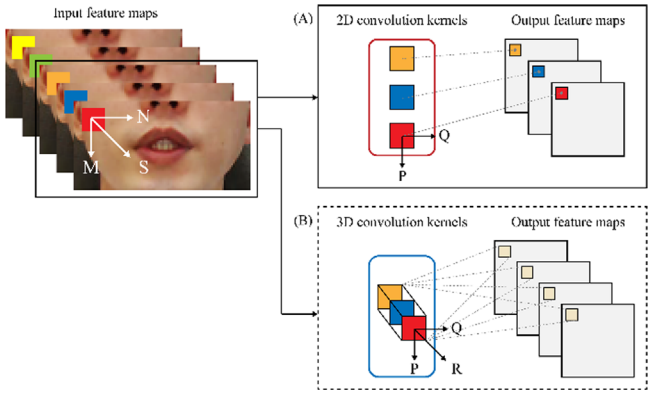

FIGURE 4 Comparison of convolutions in (A) 2D and (B) 3D. The width, length, sequence, height, weight, and 3D kernel of the kernel are denoted by N, M, S, P, Q, and R, respectively.

3 | ARCHITECTURE

The proposed audio recognition module is illustrated in Figures 2A,B and 3. It comprises an open cloud-based speech-recognition API system, a log-Mel spectrogram generated from input audio information, and user lip-movement data.

3.1 | Audio module

The proposed audio recognition module is illustrated in Figure 3. In previous studies [34, 35], we used an open cloud-based speech recognition API using Microsoft's Azure Cognitive Services API [19], which had approximately 5%–10% better word recognition rates than Google Assistant and Amazon Transcribe. To mitigate the impact of performance changes over time, a new API that surpasses the current API is provided whenever available. Only recognizable words were utilized, allowing for easy replacement if necessary. Audio data from a local device with a microphone were input into the audio module in real time and transmitted to Microsoft Azure's open cloud-based server for recognition. A list of recognized words was generated, and using the Word2Vec embedding model pretrained on 100 billion words from the Google News corpus, the produced word list was transformed into a 300-dimensional word vector, which was then integrated into a 1D vector (Figure 2A). Word2Vec [36, 37] embedding model is a distributed word representation method that expresses words with similar contexts in similar vectors.

3.2 | Visual module

As illustrated in Figure 2, the visual module comprises four components: a 3D CNN, dense spatial–temporal CNN, spatial attention module, and bidirectional-GRU. In contrast to the 2D CNN (1) in Figure 4A, the first element, the 3D CNN, is effective in extracting features for various lip-reading tasks, such as lip, tongue, and teeth movements, as it encodes motion information in multiple consecutive frames (Figure 4B) [17, 38]. A 3D CNN (2) adds a 3D kernel to the 2D CNN to convert sequence frames into a single-frame structure. In this structure, the feature map is connected to the sequence frames of the previous layer to capture the lip-movement information in the image. Equation (2) of the 3D CNN adds 3D kernel information to (1) of the 2D CNN, allowing feature maps in consecutive frames to be linked to consecutive frames of the previous layer and to be converted into a single frame, thereby capturing the object motion information. In (2), (x, y, z) denotes the coordinates of the feature map/volume, $(p, q)^{\text{th}}$ denotes the spatial dimension index of the kernel, and r denotes the temporal dimension index. j^{th} and m denote the feature map/volume, and i^{th} denotes the convolution layer. B_{ij} denotes the bias of the feature map/volume, and tanh(·) denotes the hyperbolic tangent function [37].

$$v_{ij}^{xy} = \tanh\left(b_{ij} + \sum_{m}\sum_{p=0}^{P_i-1}\sum_{q=0}^{Q_i-1} w_{ijm}^{pq} v_{(i-1)m}^{(x+p)(y+q)}\right), \quad (1)$$

$$v_{ij}^{xyz} = \tanh\left(b_{ij} + \sum_{m}\sum_{p=0}^{P_i-1}\sum_{q=0}^{Q_i-1}\sum_{r=0}^{R_i-1} w_{ijm}^{pqr} v_{(i-1)m}^{(x+p)(y+q)(z+r)}\right). \quad (2)$$

We constructed a 3D CNN to extract lip-movement features from consecutive frames. Additionally, to minimize the transformation of internal variables, accelerate the training process, and reduce the spatial size of the 3D feature map, we incorporated batch normalization (BN), rectified linear unit (ReLU), and max-pooling 3D layers sequentially, followed by another 3D CNN (Table 1).

After the 3D CNN, Figure **S1**A(b–e) illustrates the use of a proposed dense spatial–temporal CNN model to reduce training time and save resources owing to smaller parameters. Additionally, by establishing short and dense connections between various layers, the network depth increased, the gradient vanishing problem was mitigated, and training efficiency improved. Parameter reduction, bottleneck layers, transition layers, and slow growth rates were implemented to minimize computation while avoiding common overfitting issues that can arise during training with limited data.

In Figure S1A(b–e, purple square), the "Dense block" structure and diagram of the BN layer, ReLU layer, 3D convolution layer, BN, ReLU, and 3D convolution layers are sequentially connected. Figure S1A(b–d, yellow square) shows a "transition" structure in which BN, ReLU, 3D convolutional, and average 3D pooling layers are sequentially connected. Additionally, following the "transition" structure shown in Figure **S1**A(b) (Table S1), a standard dropout layer was connected, and pixels were randomly dropped to prevent strong correlation in feature maps between successive frames [39]. In addition, the spatial dropout layer connected to the "transition" structure, shown in Figure S2C, was effectively used to extract fine movement features such as lips, teeth, and tongue with strong spatial correlation [31–43]. Therefore, the proposed dense spatial–temporal CNN network comprises one layer that represents a nonlinear transformation Hl, and the output of the layer can be expressed as x_l (3), where x_0, x_1, ⋯, and $x_{(l-1)}$ denote the volume of the 3D feature created in the previous layer and [⋯] denotes a concatenation operation.

$$x_l = H_l([x_0, x_1, \cdots, x_{l-1}]). \quad (3)$$

In the proposed model, a spatial attention module that focuses on the position of the information section and complements the attention channel is combined to effectively extract the features of fine movements (lips, teeth, and tongue) with a strong spatial correlation (Figure S1A(a) and Table S1) [44]. The spatial attention module focuses on utilizing inter-space interaction to accurately detect the most identifiable and useful part of the input continuous image frame, that is, the lip movement [44]. The spatial attention module focuses on utilizing inter-space interaction to accurately detect the most identifiable and useful part of the input continuous image frame, that is, the lip movement [44]. The spatial attention map ($M_S(F) \in R^{(H\times W)}$) aggregates the channel information of the feature maps by concatenating the convolutional, average pooling, and max pooling layers sequentially, followed by the concatenated feature descriptors, and generates two 3D maps ($F_{\text{avg}}^{s} \in R^{(H\times W)}$ and $F_{\max}^{s} \in R^{(H\times W)}$). The 3D spatial attention map generated by connecting and convolving with the existing convolution layer is shown in (4) and (5), where σ denotes the sigmoid function, and $f^{(7\times 7)}$ is the filter size of 7 × 7 representing the convolution operation.

$$M_S(F) = \sigma\left(f^{7\times 7}([\text{AvgFool}(F); \text{MaxPool}(F)])\right), \quad (4)$$

$$M_S(F) = \sigma\left(f^{7\times7}\left(\left[F^s_{\text{avg}}; F^s_{\text{max}}\right]\right)\right). \quad (5)$$

3.3 | Concatenation module

The proposed model combines the word vector output from the audio module, the log-Mel spectrogram output from the visual module, and the feature vector of the lip-motion image to create a single vector (Figure 5). Recurrent neural network (RNN) models, such as RNN, LSTM, and GRU algorithms, should be used to classify time-series input data, such as log-Mel spectrograms and lip movements. In the proposed model, we trained the propagation and control of the flow of time-series data using a two-layer bidirectional GRU structure with update and reset gates (Figure 5D) (Table S1) [44]. By utilizing update and reset gates, it is possible to address the vanishing gradient problem that exists in traditional RNN algorithms and capture rich information between two networks owing to the bidirectional structure.

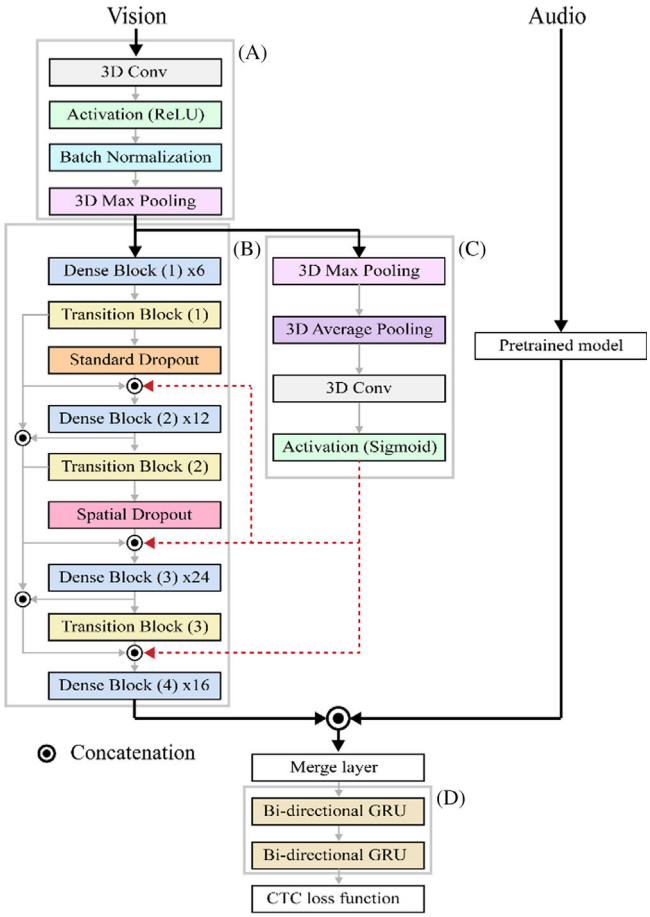

FIGURE 5 Detailed schematic of the proposed multimodal AVSR architecture. (A) 3D CNN. (B) Dense spatial–temporal CNN. (C) Spatial attention module. (D) Bi-GRU.

Consequently, the input sequence that passed through the bidirectional GRU was fed through the merge layer and output as a tensor. Following the Bi-GRU module, learning was performed using the CTC loss function (Figures 2 and 5) [45]. The CTC loss function produces a sequence of random vectors to parameterize the distribution of a sequence of label tokens without explicit alignment with continuous input data. The output probability vector sequence is conditionally independent of the marginal distribution generated at each time step. When the language model is ambiguous in terms of probability, it is decoded using a beam search approach to restore the label temporal dependencies.

4 | EXPERIMENT

4.1 | Dataset and data preprocessing

To construct a word table, we referred to the speech command dataset introduced by Google [46, 47] and collected data for Version 3 following a previous study (Table S2) [31, 35]. A total of 40 participants (20 males, 20 females; average age 29.14 years) were recruited based on their familiarity with speech recognition devices as they used speech recognition technology at least five times a day in noise environments. To balance gender and English language familiarity, we included both native speakers (10 males and 10 females) and bilingual, advanced English-speaking participants (10 males and 10 females) (Figure 6A). Participants wore a head-mounted device (Figure 6B) that was equipped with a webcam and a built-in microphone, sat in front of a display, and read and repeated 70-word lists displayed 100 times at 3-s intervals (25 fps) (Figure 6C). To collect generalized data from participants, we selected a within-subject design

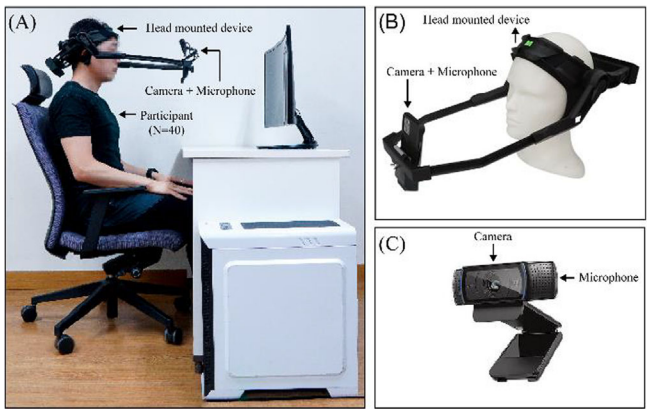

FIGURE 6 Data collection with an audio–video recording device: (A) experiment setup; (B) head-mounted device; and (C) camera for collecting lip-movement data.

that can minimize variability due to individual differences between participants. In addition, the within-subject design used a randomized partial counterbalancing, one of the partial counterbalancing, to prevent order and carryover effects problems. Audio and video of the participants were acquired via a Logitech webcam (C920 HD PRO WEBCAM) with stereo microphone specifications at a resolution of 1920 × 1080 (FHD) and a frame rate of 30 fps. In total, 280000 video clips including full audio (stereo, sample rate of 44100 Hz) and video (resolution of 640 × 480 pixels at 30 fps) information were collected.

The proposed model was trained using training and validation datasets obtained by dividing the data collected 100 times per class in a 7:3 (training: validation) ratio. Because of concerns about participants' concentration and deteriorating pronunciation over the course of 100 repetitions of the same experiment, three sections of the collected data, namely, the early (21–30), middle (51–60), and last (81–90) utterances, were used exclusively for the validation set. Specifically, the training dataset comprised seven subsets (1–9, 10–20, 31–40, 41–50, 61–70, 71–80, and 91–100) out of a total of 100 subsets, with a 7:3 ratio for training, and three subsets (21–30, 51–60, and 81–90) for validation purposes. This approach was adopted to prevent overfitting caused by poor pronunciation data.

The problem caused by differences in frequency-domain emphasis by various audio input devices was addressed in this study's audio preprocessing step using a log-Mel spectrogram because it can provide more accurate and detailed characteristics in the high- and low-frequency domains than the Mel-spectrogram [48]. In addition, log-Mel spectrograms can improve performance, as demonstrated by the DCASE 2020 challenge for audio scene classification [48]. At a sampling rate of 16000, each 3-s sample was processed to create a single log-Mel time-frequency spectrogram. A log-Mel spectrogram was produced using a linearly spaced triangular filter at the Mel scale. A total of 751 window frames, each with a window size of 25 ms, were created from the 3 s speech sample, of which 750 were used because the initial blank frame without speech input was disregarded as it was silent to preserve consistency.

In the visual data preprocessing step, a histogram of oriented gradients (HoG) feature-based Dlib linear classifier was utilized to extract the subject's face and lip regions [49]. The Dlib classifier can detect face and lip regions and create a bounding box with diagonal edge coordinates (x, y) along with 68 landmarks predicted using an online Kalman filter-based iBug program [50]. An affine transformation with a mean and variance of 0 was thereafter applied for Red Green Blue (RGB) channel normalization on the central lip region of the entire training dataset extracted from each frame using the two techniques. Data augmentation was performed by horizontally mirroring the sequence data during the training process [17]. All the proposed models were trained and evaluated using the same data after applying preprocessing and data augmentation techniques.

4.2 | Implementation

To evaluate the performance of the proposed model trained using the CTC loss function based on the beam search technique, Keras with a TensorFlow backend was used in a Linux Ubuntu experimental environment. The evaluation environment comprised an Intel® Core™ i7-7700K CPU, 32GB RAM, and an NVIDIA GeForce RTX 2080-Ti GPU running the Linux Ubuntu 18.04 LTS operating system. The initialization was performed on the network parameters of the proposed model, excluding the initialized orthogonal GRU matrix and hyperparameters. A mini-batch size of 8, learning rate of 0.0001, and Adam optimization [51] were used for training. The training included a baseline model trained on the collected dataset until the proposed model was overfitted. Because of the constraints of the computational requirements for training, the mini-batch size was set to be small, resulting in uneven real-value fluctuations, which were smoothed using a moving average and visualized as smooth curves.

Considering factors such as lighting that may adversely affect the visual information of the proposed model, the data collection and evaluation environments were considered to be different during the performance evaluation. The performance evaluation was conducted in a noise environment without any control over lighting and noise, which was completely different from the existing data collection environment. We intended to conduct an evaluation of the proposed model in a similar environment, assuming everyday environments where real speech recognition systems could be used. The participants of the audiovisual data collection stage were excluded from the performance evaluation stage. The audiovisual model used for performance evaluation was divided into three categories: audio-only, visual-only, and audiovisual speech.

4.3 | Performance evaluation metrics

The performance and efficiency of the proposed model were compared through an evaluation process using character error rate (CER) as the standard measure. The evaluation included a comparison of parameters, learning

time, and other relevant factors. The CER represents the percentage of incorrectly predicted characters, with a lower value indicating a better performance of the speech recognition system. The calculation of the CER was used in the evaluation.

$$\mathrm{CER}(\%) = \frac{S+D+I}{N} \times 100, \tag{6}$$

where S, D, I, and N denote the different types of errors made in speech recognition, namely, substitution, deletion, insertion, and the total number of characters in the ground truth, respectively. These errors were used to calculate the CER, which are standard measures of the performance of automatic speech recognition systems. In addition to the CER, the quantitative performance of the proposed model was evaluated using synthesized noise benchmark data to assess the signal-to-noise ratio (SNR). SNR is a measurement method used to compare the desired signal with the signal of background noise, defined as the ratio of the signal power to the noise power. Noise was generated using the multi-channel acoustic noise database (DEMAND), which includes ambient noise from eight different environments recorded with a 16-channel array microphone, such as public, transportation, nature, and streets [52]. The recorded speech data were synthesized using this noise for evaluation.

5 | RESULTS

5.1 | Convergence rate

We compared the tendency of the loss function to change according to three factors (audio only, visual only, and audiovisual). Figure S2 depicts the training and verification losses for Models B–G listed in Table S3, and Table S2 summarizes the performances of all trials using the obtained datasets. Under identical experimental settings and learning rates (learning rate = 0.0001), models (B–D), which did not employ audio, had fewer parameters and faster convergence rates than models (E–F) that employed audio. Loss reduction assumes an exponential form for rapidly converging "high learning rates," which can quickly converge to generate overfitting difficulties and learning stagnation. However, compared with the other models, the log-Mel spectrogram of the audio-based model group demonstrated more consistent convergence rates and trends, and our suggested Model G exhibited the lowest overfitting, thereby avoiding the steadiest training loss. Therefore, Model G tended to converge more steadily than other models while additionally concatenating log-Mel spectrogram information.

5.2 | Characteristic accuracy rate

Figures 7 and 8 and Table S3 present the performances of the proposed and comparative models, respectively. Figure 7 depicts the actual values as shadow portions of the image and smooth values as curves for training the proposed model. Although the proposed model increased the average epoch time by approximately 30 s compared with the baseline (Model A), the CER improved significantly by 9.157%. In addition, Model B, which did not use speech as additional information, and Model E, which used speech as additional information, exhibited a decrease in the number of parameters and learning time compared to Model A, and the accuracy performance improved by 4.226% and 7.183%, respectively. In addition, when comparing the group that did not use speech (Models B–D) to the baseline (Model A), the performance improvements were 4.931%, 4.245%, and 1.974%, respectively, without a significant change in the average epoch time, while the group that used speech (Models E–G) exhibited improvements of 7.579%, 8.033%, and 9.157%, respectively. In addition, Models C and F, using the spatial attention technique, improved the performance by 0.686% and 0.454% compared to Models B and E, respectively, without significant changes in the number of parameters. Consequently, the proposed model demonstrates superiority by exhibiting significant performance improvements without large parameters and learning time changes compared with the baseline and other models (Figure 8).

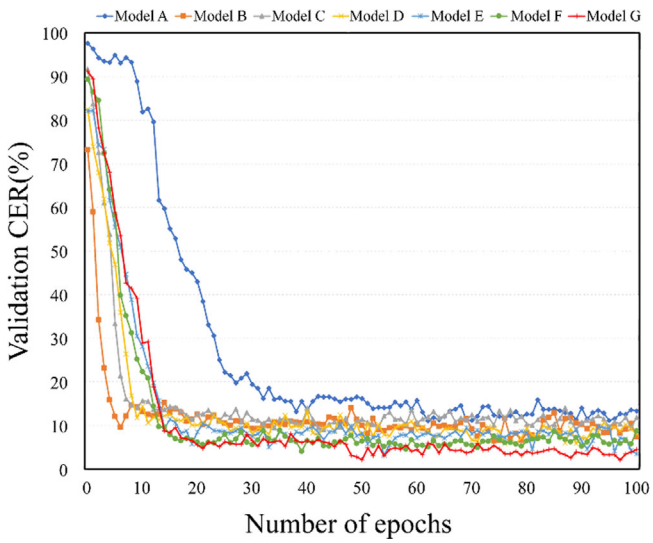

FIGURE 7 Training steps for CER comparison between the proposed model and other models: models (A) a; (B) B; (C) C; (D) D; (E) E; (F) F; and (G) G.

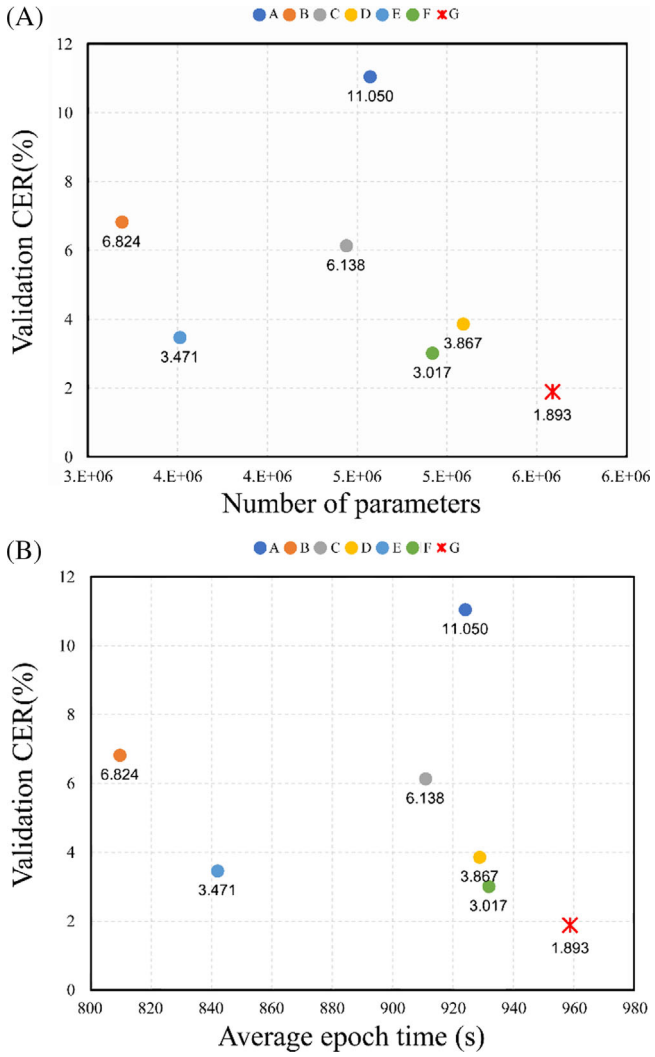

FIGURE 8 Comparison of the number of parameters and learning time between the proposed model and other models; (A) number of parameters; (B) average epoch time.

5.3 | Performance in noisy environments

As shown in Figure S4 and Table S4, the word accuracy of each model was evaluated at different SNR (dB) values by synthesizing each of the nine noisy environments in the speeches of the participants. Depending on the composition of the auditory and visual information, seven models were evaluated as A, $V(L)$ for visual information that used only lip information, $V(L + S)$ for visual information using lip and log-Mel spectrogram information, and $A + V(L)$ for auditory and lip and log-Mel spectrogram information. The use of only the auditory data (black dotted lines) of the single-modal approach exhibited approximately 3%–4% WER performance in nine noisy environments, and the OCSR API, which performed well only in certain environments, such as indoors or vehicles, was better than the quantitative performance shown in two previous studies [31, 35]. Therefore, in the clean environment listed in Table S4, the OCSR API exhibited an accuracy of $3.939 \pm 0.313\%$ (Table S5). However, when speech recognition was attempted using only visual information (yellow and blue dotted lines) with a single-modal approach, the performance was somewhat inferior compared to using only auditory information, and, on average, the WER performance was approximately 4%–6% in nine noisy environments. The average WER in the nine noisy environments was 6.138% when only lip information was used, and it improved to $4.016 \pm 0.241\%$ when log-Mel spectrogram information was used. Further, if the log-Mel spectrogram is used as additional information, the performance changes owing to noise (Figure S4, blue line). As the SNR increased, the tendency of the blue line decreased, and the performance improved. When both auditory and visual stimuli (purple and red lines) were used, performance improved in all nine noise environments. Table S5 summarizes the performance parameters when audio and visual information were combined. When log-Mel spectrogram information was added to the existing lip model, the performance improved by 2.228% and 2.305%, respectively. To perform a more accurate quantitative evaluation, those who participated in the data collection experiment were excluded, and the evaluation was performed in an environment completely different from the data collection environment. Therefore, some actual applications can contribute to the differences between the learning and evaluation results. Figure S3 shows the statistical significance of the t-test between each group in the nine noisy environments. The WER performance statistics in the nine noisy environments improved $A + V(L)$ on average, with recognition rates of 0.678% and 0.755%, respectively, compared with A and V. In addition, when comparing $A + V\,(L + S)$, A, and V, it improved by 2.228% and 2.305%, respectively, and we demonstrated that adding the log-Mel spectrogram of speech to visual information helped improve performance in noisy environments. Unlike the previous OCSR API [31, 35], which exhibited excellent performance only in certain environments, it exhibited similar performance in all nine noise environments and an evenly stable performance overall.

6 | DISCUSSION

This study proposed a new AVSR model for the three-feature (word embedding, lip-movement, and log-Mel spectrogram) multi-fusion method that can operate reliably and repeatedly in various application scenarios. To

evaluate the proposed model, three features were combined to compare the training step, error rate, and accuracy, and SNR evaluation was performed by synthesizing speeches in nine noise scenario environments that could be used in a real environment. Consequently, the proposed model exhibited the most stable convergence process and superior performance without significant changes in parameters and training time compared to other models, and it proved to be more robust to noise than previous studies [30, 34]. In particular, by adding a new specific point, the log-Mel spectrogram information, the model exhibited a performance improvement of approximately 1.6%–1.2% over the performance of the model that used only two existing features (word embedding and lip movement).

For practical applications in future scenarios, nine noise environments were synthesized into input speech information to measure the SNR of the six stages. Audio-based speech recognition systems, which exhibited excellent performance only in certain environments, such as automobiles and indoors in previous studies [31, 35], were improved on their own and exhibited stable performance in nine noisy environments. The improved performance of the OSCR API produces better recognition results. Therefore, it acts as a more reliable input to our proposed model using the output results as input to the model.

7 | CONCLUSION

We demonstrated a new approach for a multimodal-based AVSR model that combines three features to develop a robust speech recognition system in an ambient noise environment. The proposed system uses auditory information from word embedding techniques that upload input speech to an OCSR API to generate recognized words as word vectors and visual information from the log-Mel spectrogram that transforms the RAW data of speech into a Mel-scale. As additional visual information, the movement information of the lips was used together to finally concatenate the log-Mel spectrogram and lip information as input to the visual recognition model. Previous studies demonstrated that OCSR APIs, which performed well only in certain environments, improved on their own to demonstrate performance stability. We further demonstrate that as the performance of open cloud-based speech APIs improves, the performance improves reliably when we combine our proposed models. Therefore, this system can be applied to the Internet of Things (IoT) and robot fields and will also play a significant role in applications such as cinematography, automobiles, and hospitals that require speech recognition in noisy environments. In future studies, we will demonstrate the practicality of the system by applying the proposed model to specific applications. In particular, the proposed system will be a system that can be actively used in movie shooting or hospitals in noisy environments to help patients or situations where conversation is difficult owing to speech recognition problems. Future models will attempt to develop integrated systems by combining language models that can interpret syntax. In addition, we will develop a lightweight model for application in the fields of robots and IoT.

CONFLICT OF INTEREST STATEMENT

The authors declare that there are no conflicts of interest.

ORCID

Sanghun Jeon https://orcid.org/0000-0003-4705-1254

REFERENCES

1. H. McGurk and J. J. N. MacDonald, *Hearing lips and seeing voices*, Nature **264** (1976), no. 5588, 746–748.
2. A. G. Chitu and L. J. M. Rothkrantz, *Automatic visual speech recognition*, In *Speech enhancement, modeling, recognition—algorithms, and applications*, S. Ramakrishnan (ed.), IntechOpen, London, UK, 2012, 95–120.
3. P. Agarwal and S. Kumar, *Electroencephalography-based imagined speech recognition using deep long short-term memory network*, ETRI J. **44** (2022), no. 4, 672–685.
4. L. Sun, Q. Li, S. Fu, and P. Li, *Speech emotion recognition based on genetic algorithm–decision tree fusion of deep and acoustic features*, ETRI J. **44** (2022), no. 3, 462–475.
5. S. Kumar, *Real-time implementation and performance evaluation of speech classifiers in speech analysis-synthesis*, ETRI J. **43** (2021), no. 1, 82–94.
6. S. B. Alex and L. Mary, *Variational autoencoder for prosody-based speaker recognition*, ETRI J. **45** (2023), no. 4, 678–689.
7. B. Shi, W. Hsu, and A. Mohamed, *Robust self-supervised audio-visual speech recognition*, (Interspeech, Incheon, Rep. of Korea), 2022, DOI 10.21437/Interspeech.2022-99.
8. R. Shashidhar, S. Patilkulkarni, and S. B. Puneeth, *Combining audio and visual speech recognition using LSTM and deep convolutional neural network*, Int. J. Inform. Technol. **14** (2022), 3425–3436.
9. D. Serdyuk, O. Braga, and O. Siohan, *Transformer-based video front-ends for audio-visual speech recognition for single and multi-person video*, (Interspeech, Incheon, Rep. of Korea), 2022, DOI 10.21437/Interspeech.2022-10920.
10. C. Deuerlein, M. Langer, J. Seßner, P. Heß, and J. Franke, *Human-robot-interaction using cloud-based speech recognition systems*, Proc. CIRP **97** (2021), 130–135.
11. K. Takashi, T. Nose, S. Hirooka, Y. Chiba, and A. Ito, Comparison of speech recognition performance between Kaldi and Google cloud speech API, In *Recent advances in intelligent information hiding and multimedia signal processing: proceeding of the fourteenth international conference on intelligent information hiding and multimedia signal processing, November, 26–28, 2018, Sendai, Japan*, Vol. **110**, Springer International Publishing, 2019.

12. S. Qiya, B. Sun, and S. Li, *Multimodal sparse transformer network for audio-visual speech recognition*, IEEE Trans. Neural Netw. Learn. Syst. **34** (2022), no. 12, DOI 10.1109/TNNLS.2022.3163771
13. L. D. Terissi, G. D. Sad, and J. C. Gómez, *Robust front-end for audio, visual and audio–visual speech classification*, Int. J. Speech Technol. **21** (2018), 293–307.
14. G. Calvert, C. Spence, and B. E. Stein, *The handbook of multisensory processes*, MIT Press, London, UK, 2004.
15. J. Venezia, W. Matchin, and G. Hickok, *Multisensory integration and audiovisual speech perception*, Brain Mapp. Encycl. Ref. **2** (2015), 565–572.
16. R. Campbell, *The processing of audio-visual speech: empirical and neural bases*, Philos. Trans. R. Soc. Lond. B Biol. Sci. **363** (2008), no. 1493, 1001–1010.
17. Y.M. Assael, B. Shillingford, S. Whiteson, and N. De Freitas, *LipNet: End-to-end sentence-level lipreading*, arXiv preprint, 2016, DOI 10.48550/arXiv.1611.01599
18. Google assistant, Available at: https: /developers.google.com/assistant?hl=ko/ [last accessed 27 April 2022].
19. Microsoft azure cognitive services, Available at: https://azure.microsoft.com/en-us/services/cognitive-services/ [last accessed 27 April 2022].
20. Watson speech to text, Available at: https://www.ibm.com/kr-ko/cloud/watson-speech-to-text [last accessed 27 April 2022].
21. Available at: https://aws.amazon.com/transcribe/ [last accessed 27 April 2022].
22. P. Sawers, Amazon Transcribe can now automatically redact personally identifiable data, available at https://venturebeat.com/2020/02/27/amazon-transcribe-can-now-automatically-redact-personally-identifiable-data/, VentureBeat, 27th February 2020, Retrieved 3rd February 2021.
23. Available at: https://aiopen.etri.re.kr/guide/Recognition [last accessed 27 April 2022].
24. Available at: https://clova.ai/speech [last accessed 27 April 2022].
25. Available at: https://speech-api.kakao.com/ [last accessed 27 April 2022].
26. A. Fernandez-Lopez and F. M. Sukno, *Survey on automatic lip-reading in the era of deep learning*, Image vis. Comput. **78** (2018), 53–72.
27. Y. R. Oh, K. Park, and J. G. Park, *Fast offline transformer-based end-to-end automatic speech recognition for real-world applications*, ETRI J. **44** (2022), no. 3, 476–490.
28. C. Sheng, G. Kuang, L. Bai, C. Hou, Y. Guo, X. Xu, M. Pietikäinen, and L. Liu, *Deep learning for visual speech analysis: a survey*, arXiv preprint, 2022, DOI 10.48550/arXiv.2205.10839
29. Y. R. Oh, K. Park, H. B. Jeon, and J. G. Park, *Automatic proficiency assessment of Korean speech read aloud by non-natives using bidirectional LSTM-based speech recognition*, ETRI J. **42** (2020), no. 5, 761–772.
30. D. Kolossa, S. Zeiler, A. Vorwerk, and R. Orglmeister, Audiovisual speech recognition with missing or unreliable data, *(Proc. International Conference on Auditory Visual Speech Processing*, Norwich, UK), (2009), pp. 117–122.
31. V. Këpuska and G. Bohouta, *Comparing speech recognition systems (Microsoft API, Google API and CMU sphinx)*, Int. J. Eng. Res. Appl. **7** (2017), 20–24.
32. H. J. Yoo, S. Seo, S. W. Im, and G. Y. Gim, *The performance evaluation of continuous speech recognition based on Korean phonological rules of cloud-based speech recognition open API*, Int. J. Netw. Distrib. Comput. **9** (2021), no. 1, 10–18.
33. B. Alibegović, N. Prljača, M. Kimmel, and M. Schultalbers, Speech recognition system for a service robot—a performance evaluation, *(Proc. 2020 16th Intl Conf. Control Autom. Robot. Vis. (ICARCV)*, Shenzhen, China), 2020, pp. 1171–1176.
34. S. Jeon and M. S. Kim, *End-to-end lip-reading open cloud-based speech architecture*, Sensors (Basel). **22** (2022), no. 8, 2938.
35. S. Jeon and M. S. Kim, *Noise-robust multimodal audio-visual speech recognition system for speech-based interaction applications*, Sensors (Basel). **22** (2022), no. 20, 7738.
36. T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, Distributed representations of words and phrases and their compositionality, *(Proceedings of the 26th International Conference on Neural Information Processing Systems*, Lake Tahoe, NV, USA), 2013, pp. 3111–3119.
37. T. Mikolov, K. Chen, G. Corrado, and J. Dean, *Efficient estimation of word representations in vector space*, arXiv preprint, 2013, DOI 10.48550/arXiv.1301.3781
38. S. Ji, M. Yang, and K. Yu, *3D convolutional neural networks for human action recognition*, IEEE Trans. Pattern Anal. Mach. Intell. **35** (2013), no. 1, 221–231.
39. G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, *Improving neural networks by preventing co-adaptation of feature detectors*, arXiv preprint, 2012, DOI 10.48550/arXiv.1207.0580
40. J. Tompson, R. Goroshin, A. Jain, Y. Lecun, and C. Bregler, Efficient object localization using convolutional networks, *(Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA), 2015, pp. 648–656.
41. S. Lee and C. Lee, *Revisiting spatial dropout for regularizing convolutional neural networks*, Multimed. Tools Appl. **79** (2020), 34195–34207.
42. S. Jeon, A. Elsharkawy, and M. S. Kim, *Lipreading architecture based on multiple convolutional neural networks for sentence-level visual speech recognition*, Sensors (Basel). **22** (2021), no. 1, DOI 10.3390/s22010072.
43. S. Jeon and M. S. Kim, *End-to-end sentence-level multi-view lipreading architecture with spatial attention module integrated multiple CNNs and cascaded local self-attention-CTC*, Sensors (Basel). **22** (2022), no. 9, DOI 10.3390/s22093597.
44. S. Woo, J. Park, J. Y. Lee and I. S. Kweon, *Cbam: convolutional block attention module*, (Proc. Eur. Conf. Comput. Vis. (ECCV), Munich, Germany), 2018, pp. 3–19.
45. A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks, *(Proc. 23rd Intl Conf. Mach. Learn., Pittsburgh*, PA, USA), 2006, pp. 369–376.
46. P. Warden, *Speech commands: a dataset for limited-vocabulary speech recognition* arXiv preprint, 2018, DOI 10.48550/arXiv.1804.03209
47. C. H. H. Yang, J. Qi, S. Y. C. Chen, P. Y. Chen, S. M. Siniscalchi, X. Ma, and C. H. Lee, Decentralizing feature extraction with quantum convolutional neural network for automatic speech recognition, *(ICASSP 2021–2021 IEEE*

International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, Canada), 2021, pp. 6523–6527.

48. S. Seo, C. Kim, and J. H. Kim, *Convolutional neural networks using log mel-spectrogram separation for audio event classification with unknown devices*, J. Web Eng. **21** (2022), no. 2, 497–522.
49. S. Suh, S. Park, Y. Jeong and T. Lee, *Designing acoustic scene classification models with CNN variants*, Technical Report, Detection and Classification of Acoustic Scenes and Events, Challenge, 2020.
50. C. Sagonas, G. Tzimiropoulos, S. Zafeiriou and M. Pantic, 300 faces in-the-wild challenge: the first facial landmark localization challenge, *(Proceedings of the IEEE International Conference Computability Vision Workshops 2013, Sydney*, Australia), 2013, pp. 397–403.
51. D. P. Kingmaand J. Ba, *Adam: a method for stochastic optimization*, arXiv preprint, 2014, DOI 10.48550/arXiv.1412.6980
52. J. Thiemann, N. Ito, and E. Vincent, *DEMAND: diverse environments multichannel acoustic noise database*, Proc. Mtgs. Acoust. **19** (2013), DOI 10.1121/1.4799597

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: S. Jeon, J. Lee, D. Yeo, Y.-J. Lee, and S. Kim, *Multimodal audiovisual speech recognition architecture using a three-feature multi-fusion method for noise-robust systems*, ETRI Journal (2024), e12660, DOI 10.4218/etrij.2023-0266