Deep Neural Networks-based Classification Methodologies of Speech, Audio and Music, and its Integration for Audio Metadata Tagging

Hosung Park, Yoonseo Chung and Ji-Hwan Kim*

Sogang University, Seoul, South Korea
E-mail: hosungpark@sogang.ac.kr; ys1231@sogang.ac.kr; kimjihwan@sogang.ac.kr
**Corresponding Author*

Received 24 November 2022; Accepted 11 January 2023;
Publication 14 April 2023

Abstract

Videos contain visual and auditory information. Visual information in a video can include images of people, objects, and the landscape, whereas auditory information includes voices, sound effects, background music, and the soundscape. The audio content can provide detailed information on the story by conducting a voice and atmosphere analysis of the sound effects and soundscape. Metadata tags represent the results of a media analysis as text. The tags can classify video content on social networking services, like YouTube. This paper presents the methodologies of speech, audio, and music processing. Also, we propose integrating these audio tagging methods and applying them in an audio metadata generation system for video storytelling. The proposed system automatically creates metadata tags based on speech, sound effects, and background music information from the audio input. The proposed system comprises five subsystems: (1) automatic speech recognition, which generates text from the linguistic sounds in the audio, (2) audio event classification for the type of sound effect, (3) audio scene

Journal of Web Engineering, Vol. 22_1, 1–26.
doi: 10.13052/jwe1540-9589.2211

classification for the type of place from the soundscape, (4) music detection for the background music, and (5) keyword extraction from the automatic speech recognition results. First, the audio signal is converted into a suitable form, which is subsequently combined from each subsystem to create metadata for the audio content. We evaluated the proposed system using video logs (vlogs) on YouTube. The proposed system exhibits a similar accuracy to handcrafted metadata for the audio content, and for a total of 104 YouTube vlogs, achieves an accuracy of 65.83%.

Keywords: Content retrieval, speech recognition, music detection, audio event classification, audio scene classification.

1 Introduction

Video understanding is an essential characteristic of human intelligence, and video dramas are the best medium for its stimulation. The video Turing test (VTT) has been proposed for measuring video understanding intelligence. Figure 1 describes the process of VTT, which is as follows: (i) players, including AI, and the jury watch a video. (ii) A question on the video is given to both the players and the jury. (iii) Each player submits an answer to the question. (iv) The submitted responses are presented to the jury. (v) After checking all the answers, the jury predicts who the AI agent is and votes for their prediction.

A video for VTT represents a multimodal dataset, where continuous images and sounds are mixed with implicit, complex, and spatiotemporal

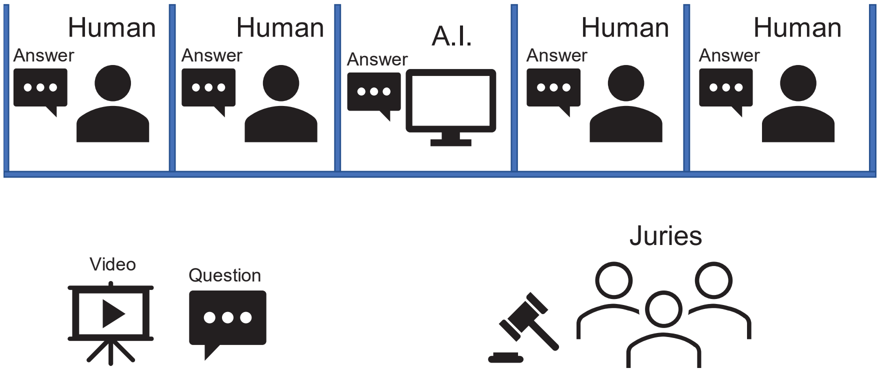


Figure 1 An illustration of the video Turing test (VTT).

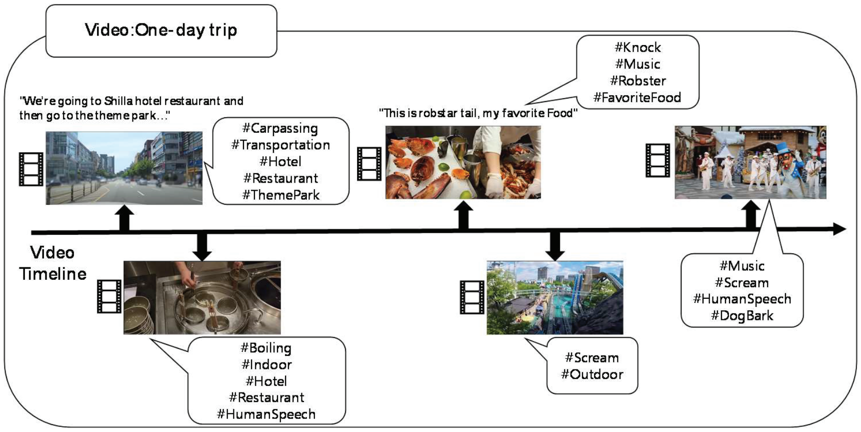


Figure 2 An example of extracting auditory information by metadata tags.

causal relationships. For this reason, one cannot obtain information using only one modality. Video information can be represented using textual metadata. The textual metadata are the means of implicitly representing information, called data-containing information. There are four levels classifying the difficulty of VTT for metadata extraction [1]. Difficulty 1 represents the ability to analyse one fact for a still picture, and difficulty 2 represents the ability to analyse several facts for a still picture. Difficulty 3 represents a fact about the video, and Difficulty 4 includes a causal relationship that does not appear in the video. Visual and auditory information is required to solve the problems of difficulties 3 and 4 [2].

Auditory information in particular plays an important part in solving difficulties 3 and 4. For example, Figure 2 presents a one-day trip video clip. This video is about a one-day trip to Seoul and includes visiting a famous hotel, restaurant, and theme park. To describe this video, it must be possible to transcribe the voice description of the narrator and extract the location information through the audio as metadata. The first scene is where the car travels to the restaurant, where the narrator explains the plan for the trip. Through the narrator's explanation, we can see that he is visiting the hotel's restaurant and theme park today. The second and third scenes are in the hotel restaurant, where the narrator explains ready-to-cook food and food. Also, you can know the atmosphere of the video through the background music. The fourth and fifth are information about the theme park. You can get information about people screaming while riding the rides, and you can find out information about the sound of a dog being included along with the music during a parade. The purpose of this study is to represent information that can be obtained through hearing with metadata tags.

This paper focus on the metadata extraction of auditory information. Unlike visual information which contains metadata on the attributes and locations of people and objects, auditory information includes metadata about the atmosphere and story through a person's voice, sound effects, and background music. Audio information in a video includes speech, sound effects, background music, and the soundscape [1]. The speech provides information on the characters and the story told in the video. The sound effects include information about events and indirect details of the visual information. The background music provides information on the mood of the video. The soundscape provides information regarding the location of the video [1, 3].

With the continuous rise of broadcasting content, the importance of metadata increases accordingly. For example, a recommendation system using metadata for a video is used in the media commerce market for product placement and second screens (smartphone, tablet, or PC). Recently, the possibility of automating metadata generation has attracted significant interest. To date, metadata was created manually by human resources. However, with the increase in media production, the demand for automation has grown [4, 5].

The word "vlog" combines the words "video" and "log", representing a form of social media post, mainly referring to videos acquired by amateur camera operators. For modern media posts, vlogs have the most need for metadata analysis [6]. Because a broadcasting station does not manage vlogs, the domain boundary of a video is ambiguous. Moreover, it is not easy to extract the metadata manually due to variations in the characters, filming location, and filming equipment [7, 8]. This study proposes automatic metadata generation using auditory information from videos. The proposed method consists of four parts: (1) an audio speech recognizer (ASR) for voice analysis, (2) an audio event classifier (AEC) for classifying sound effects, (3) an acoustic scene classification (ASC) for sound landscape analysis, and (4) music detection for background music analysis.

We evaluate our proposed system using YouTube vlogs. In this paper, the accuracy of the proposed system was measured by comparing it with manually tagged metadata. The metadata output from ASR, AED, ASC, music detection, and keyword extraction were compared with the handcrafted metadata for vlogs collected from YouTube.

The article is organized as follows. Section 2 presents previous ASR, ASC, AED, and music detection studies. Section 3 provides a brief description of these models and describes our proposed system in detail. Section 4 describes the experiments. Finally, we present our experimental results and conclusions in Section 5.

2 Related Works

Research on visual content analysis has been actively conducted and has shown comparable performance [1]. An integrated system for audio content has not been proposed to date. The information obtained from the audio includes the character's dialogue, soundscape, sound effects, and background music, enabling the precise analysis of the type of video content. Our proposed system consists of five subsystems: ASR, AED, AEC, music detection, and keyword extraction. In this section, we describe related studies on each system in detail.

Automatic speech recognition (ASR) accepts voice input and outputs the word sequence with the highest probability in the model [9–12]. ASR is implemented through sound and language models to create a defined pronunciation dictionary [13, 14]. The acoustic model outputs the probability of a speech feature vector for a phoneme defined in the pronunciation dictionary [9, 10]. The language model outputs the probability of the sequence of words defined in the pronunciation dictionary [14]. Speech recognition implements the pronunciation dictionary, language and sound models as a search network using a weighted finite-state transducer (WFST) [15]. Recent studies in speech recognition research made remarkable advances in the development of acoustic models. The traditional recognition process based on an acoustic model transforms the training data through forced alignment of a hidden Markov model (HMM) [16, 17], and outputs the probability distribution of the HMM state showing high probability using a deep neural network (DNN) [9]. However, owing to the development of DNNs, such as convolutional neural networks (CNNs), recurrent neural networks, and attention mechanisms, a model with similar performance to the existing acoustic model without forced alignment of the HMM has been proposed [18–20].

Kaldi, the most common speech recognition toolkit is used to implement automatic speech recognition for extracting metadata. It contains almost any algorithm currently used in ASR systems. It also contains recipes for training acoustic models on commonly used speech corpora such as LibriSpeech, Wall

Street Journal Corpus, TIMIT, and more. These recipes can also serve as a template for training acoustic models on speech data.

Audio event classification identifies a defined event from the audio generated from video content. Audio scene classification identifies places through soundscapes in the video content [21–24]. Audio scene classification (ASC) classifies places through soundscapes that occur in the video content [25]. These two audio classification models are implemented using CNNs suitable for image classification problems. They convert audio signals into two-dimensional images called spectrograms, which feature vectors that express an audio signal on the time and frequency axes [26]. The input spectrogram passes through a convolutional layer. It is converted into a plurality of filter images, and the transformed filter image outputs the probability of occurrence of classes learned through a fully connected neural network.

Audio event classification and audio scene classification is a regular task in the detection and classification of acoustic scenes and events (DCASE) challenges. In this paper, MobileNet v2 is used for AEC and ASC tasks. The Mobilnetv2 is a convolutional neural network based on an inverted residual structure where the residual connections are between the bottleneck layers. This model showed remarkable performance in DCASE 2021 task 1. This model achieved an accuracy of 72.6% within the 128 KB model size.

Music detection aims to identify the section of an audio stream containing music [27]. In this study, background music is detected with one-class classification, which is used in image classification to identify musical and non-musical sounds. In one-class classification, samples of the same concept are used only to recognize instances of the concept. The Gaussian mixture model (GMM) yields comparable performance with one-class classification [28].

Keyword extraction receives the text from speech recognition as input and extracts keywords corresponding to metadata using the TextRank algorithm [29]. TextRank is a method employed for extracting critical keywords for a given text using the PageRank algorithm after constructing a co-occurrence graph of the input sentence [30]. An advantage of TextRank is that it does not require domain-specific training data for supervised learning. A disadvantage is that the possible representation is limited compared to a DNN-based model, because it is based only on a given input.

3 Proposed System

This section describes the architecture of the proposed metadata generation system. We discuss each subsystem: speech recognition with keyword extraction, audio event classification, audio scene classification, and music detection.

3.1 System Architecture

Figure 3 shows the structure of the proposed system – the audio signal from a video pass through the music detection, which extracts information about the background music. The audio signal given to each subsystem is segmented; for ASR, the audio signal is divided into segments that have an only speech by the voice activity detector (VAD); for AEC, AED, and music detection, the audio signal is split into segments of equal length. For AEC, these have a length of 0.1 s, whereas, for ASC, they have a length of 1.0 s. For music detection, the length is 30 s.

Each subsystem can recognize several classes. To recognize speech, ASR has a pronunciation dictionary that contains 200,000 words. AEC can identify 11 classes of audio events: a baby crying, car horn, bicycle horn, car passing, scream, chattering, dog barking, door knocking, water boiling, whispering, and jackhammer. ASC identifies just three classes: indoors, outdoors, and on transportation.

3.2 Automatic Speech Recognition

Figure 4 depicts the structure of the ASR and keyword extraction system. The proposed ASR system consists of several modules: VAD, feature extractor, acoustic model, decoding network including a pronunciation model, and a language model.

The VAD is a pre-processing module for speech recognition tasks and is used to improve their performance. The VAD is mainly used in speech recognition and speech synthesis. When the signal is inputted, the audio signal is spilled into voice segments by VAD. Numerous vlog videos usually contain non-speech audio signals, such as audio events and soundscapes. These non-speech signals are considered noise in speech recognition. The VAD based on short-time average energy is widely used to erase these noises. The method depends on a threshold to distinguish between speech and non-speech and works efficiently [31]. Furthermore, contextual information (CI) is significant for the VAD. The CI helps improve the performance of the VAD in low SNR. The energy-based method and CI are employed in the proposed system of this study. Feature extraction transforms a speech signal into a feature vector containing compressed information for speech recognition. This method

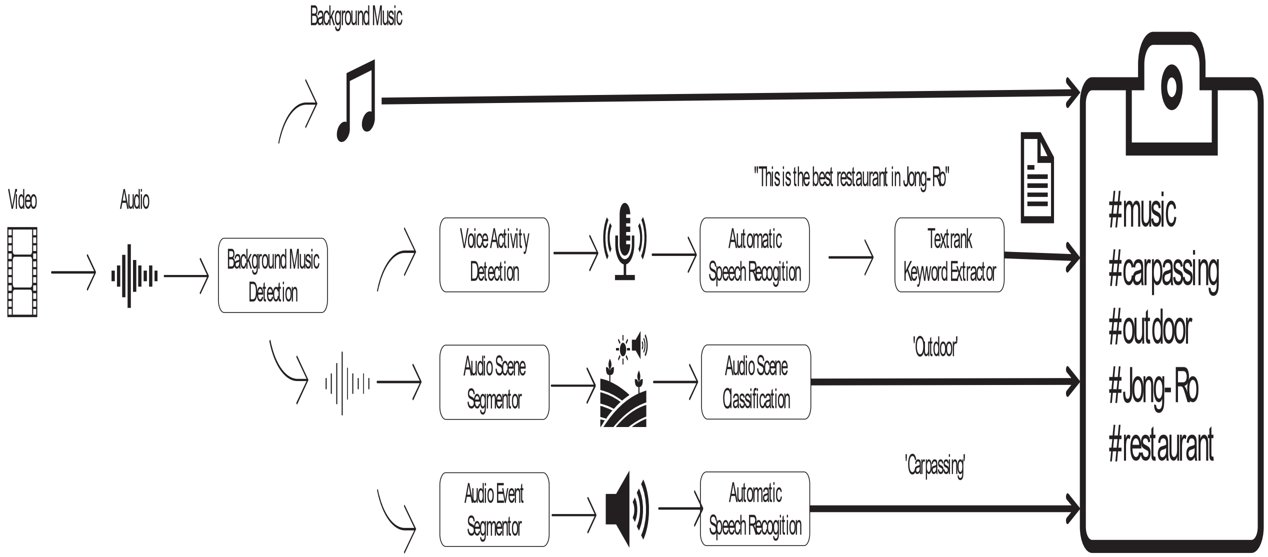

Background Music
Video
Audio
Background Music Detection
Voice Activity Detection
Automatic Speech Recognition
"This is the best restaurant in Jong-Ro"
Textrank Keyword Extractor
Audio Scene Segmentor
Audio Scene Classification
'Outdoor'
Audio Event Segmentor
Automatic Speech Recognition
'Carpassing'
#music
#carpassing
#outdoor
#Jong-Ro
#restaurant

Figure 3 Simplified diagram of the proposed system. There are four subsystems: ASR, AEC, ASC, and music detection. Each subsystem functions independently. ASR generates text from an audio signal, and the text is converted to keywords by the keyword extraction subsystem. AEC generates audio events from the audio signal, and ASC generates places from the audio signal. The result from each subsystem is concatenated to form the metadata.

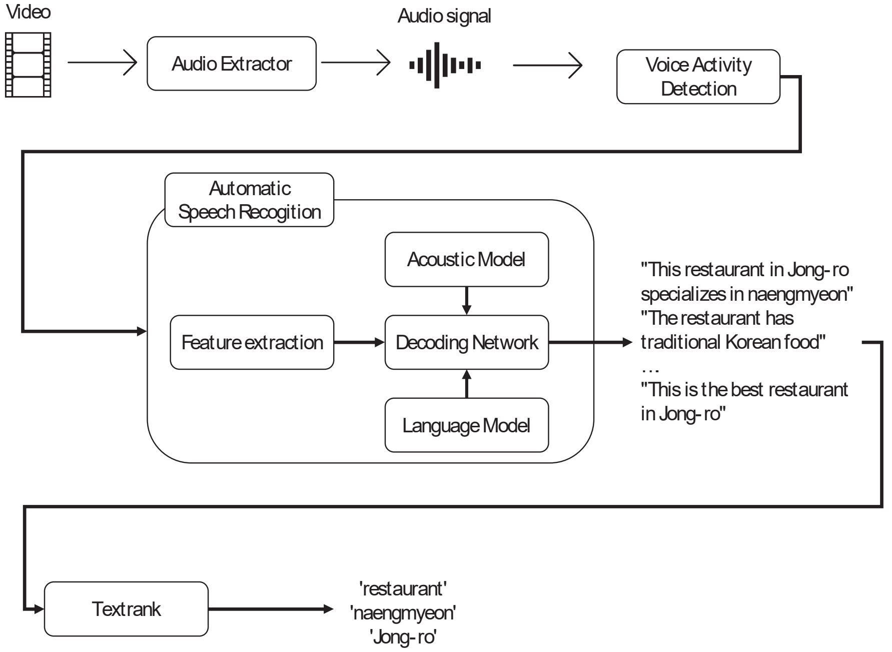

Video
Audio Extractor
Audio signal
Voice Activity Detection
Automatic Speech Recogition
Acoustic Model
Feature extraction
Decoding Network
Language Model
"This restaurant in Jong-ro specializes in naengmyeon"
"The restaurant has traditional Korean food"
...
"This is the best restaurant in Jong-ro"
Textrank
'restaurant'
'naengmyeon'
'Jong-ro'

Figure 4 Detailed diagram of ASR and keyword extraction in the proposed system.

significantly reduces the complexity of speech recognition. In this study, Mel-frequency cepstral coefficients are used for feature extraction. This method represents the short-time power spectrum of speech signals using a Fourier transform and Mel-scale frequencies.

The acoustic model determines the probability of a feature vector representing a phoneme sequence. A phoneme is a unit of speech sound. A word is pronounced based on the phonemes it contains. An HMM is used in the acoustic model to train a phoneme's length for a feature vector. Afterward, a DNN is used to classify phonemes from feature vectors as input. The relationship between feature vectors and phonemes is calculated by forced alignment in HMM. Forced alignment is a pre-processing process for model learning, which automatically extracts the data required for learning by HMM by identifying the position where a specific word is uttered in the entire learning data. The Viterbi algorithm is used to perform the forced alignment. It is possible to generate learning materials for each recognition unit for data given as a word string.

HMM determines the optimal sequence of states through the Viterbi algorithm. The acoustic model of HMM shows high performance for a given learning dataset. However, because it learns only on the fixed feature parameter dimension of the learning data, it offers a problem that the performance is low for feature vectors with noise. A deep neural network (DNN) method can efficiently change the feature parameter dimension and improve performance. However, it still depends on HMM because DNN cannot align feature vectors and phonemes. Because each state of DNN does not indicate a specific phoneme, unlike HMM, it is impossible to distinguish the recognition unit for speech. Therefore, a general acoustic modelling method uses a hybrid DNN-HMM model that employs a fusion of DNN and HMM.

The acoustic model in ASR is constructed from a time-delay neural network (TDNN) and a recurrent neural network with long short-term memory (RNN-LSTM). The TDNN, similar to one-dimensional convolution, transforms the input feature vector into a vector with a context correlation between vectors to learn the relationship between the feature vector and a phoneme. RNN-LSTM is a kind of sequence-to-sequence model. The input vector is a combination of the previous vector and the present vector. These two DNN architectures are widely used to build acoustic models in speech recognition. In this study, TDNN-LSTM is used to build the acoustic model. Two TDNN layers and one LSTM layer are stacked as one block of TDNN-LSTM, and then three blocks of TDNN-LSTM are stacked to build the acoustic model. Each TDNN layer and each LSTM layer have 520-dimensional vectors.

The TDNN structure is shown in Figure 5. A general FFNN learns entire input features for processing contexts. However, the TDNN architecture is learned in narrow contexts, where the upper layers of the networks process broader contexts of the input features. Each layer in a TDNN is updated by a different resolution that increases in higher network layers.

A decoding network, a search network for speech recognition, contains four graphs: HMM, context, lexicon, and grammar. The HMM and the context graph represent the relationship between the feature vector and phonemes. These graphs are built from the acoustic model. A feature vector is converted into the HMM states of a phoneme by an HMM graph. The context graph represents the left and right contexts of the phonemes in speech. The lexicon graph is built from a pronunciation model. The pronunciation model describes the relationship between phonemes and words. The grammar graph contains information about the sequence of words from the *n*-gram-based language model. The *n*-gram model is a kind of probability model for a contiguous sequence of *n* symbols from a given text. The WFST is widely used for decoding networks in speech recognition, and it can compose

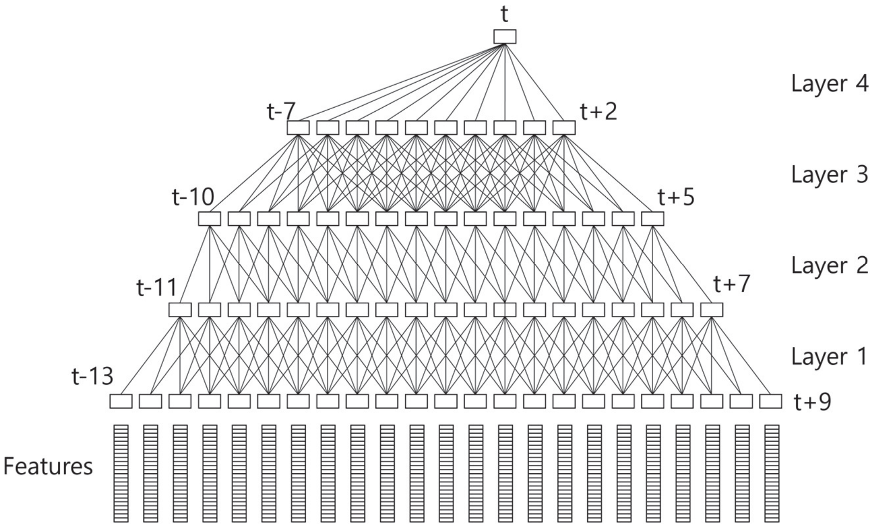


Figure 5 TDNN architecture.

and optimize two complexity graphs as a transducer by using composition, determination, and minimization [15]. Composition, determination, and minimization combine all ASR components into an integrated transducer using a convenient, efficient, and general method.

3.3 Keyword Extraction

Keyword extraction retrieves essential words from the text found by speech recognition. The keyword extraction can be learned by supervised or unsupervised. Supervised keyword extraction, usually based on a DNN, extracts keywords based on a sizeable in-domain training dataset. It works efficiently for text in the same domain but requires a large amount of training data. Unsupervised keyword extraction, usually implemented by a statistical model [31], does not require much training data and has a relatively faster processing speed than the supervised model. The TextRank algorithm is a graph-based ranking algorithm using PageRank to extract keywords [29, 30]. Google's website search engine uses the PageRank algorithm to measure the importance of web pages by assigning them a score or rank. TextRank represents words as a vertex in a graph and the relationship between the words as an edge. This co-occurrence graph finds keywords using the PageRank algorithm.

3.4 Audio Event and Audio Scene Classification

In audio event and scene classification, audio events and their locations are determined from soundscapes. Audio events and soundscapes are used to predict the atmosphere, shot changes, and occurrence of events in the video. Figure 6 shows the structure of the AEC and ASC. These systems have three components: the segmenter, feature extractor, and classifier.

The audio signal from a video is split into chunks by both the audio event and soundscape segmenter. The audio event segmenter divides the audio signal into segments of 400 ms length, whereas the soundscape segmenter divides the audio signal into segments of length 1000 ms. These lengths were chosen based on the properties of the sound types. An audio event, such as a scream or door knock, has a shorter duration than a soundscape, such as the background sound of transportation.

Both the AEC and ASC models are built with a CNN. CNNs are widely used to model audio events from audio feature vectors. This type of neural network was mainly used to transform acoustic feature vectors into spectrograms and then to train them [32]. The correlations between local information and feature vectors are learned by the CNN. Recently, residual learning by a CNN has significantly contributed to the improvement of AEC and ASC [33]. In this study, MobileNet v2 was used to implement the AEC and ASC classifiers [34]. MobileNet v2 is used for the low-complexity classification of a fixed-length feature vector. This model uses a residual CNN, which has low complexity and high accuracy.

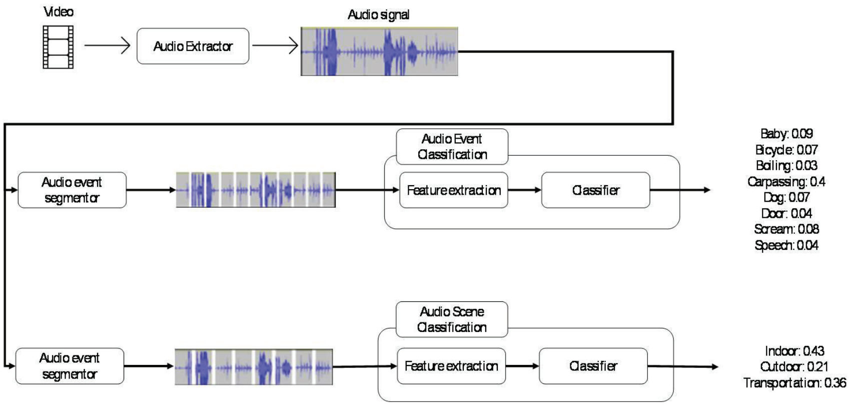


Figure 6 Diagram of AEC and ASC in the proposed system.

3.5 Music Detection

Background music refers to music played in the background of a video. In directing a video, the background music creates the atmosphere and the composition of emotions. Background music detection uses the emotional information in a video to build metadata. Algorithms can discriminate between musical and non-musical sounds. In addition, various feature extraction algorithms are used to detect the score and pitch information in the music. This study uses the zero-crossing rate, Mel-frequency cepstral coefficients, and chroma features in music detection [35].

One-class classification is used to distinguish between music and non-music. The classification is a kind of unary classification that differs from classification methods that distinguish between two or more classes. This study uses GMM and isotonic regression to implement a one-class classification. Isotonic regression is used to train the probability distribution sequence from a Gaussian mixture of music training data. In Gaussian assumption, a mixture of a finite number of Gaussian distributions generates all data samples. This model generalizes k-means clustering based on the centre and covariance of the Gaussian distribution. Isotonic regression predicts the sequence of observations and must be non-decreasing [28]. It is suitable for audio signals, depending on the time sequence.

4 Experiments

This section evaluates the proposed automatic metadata generation system using YouTube vlogs. Furthermore, each subsystem was assessed using speech data, the audio event audio scene, and music data. Labelling of all datasets was hand-crafted, and the accuracy was measured by comparing the labels and hypotheses of the model. Metrics for each system were described in the experimental setup.

4.1 Speech Recognition Task

4.1.1 Experimental setup

The Kaldi, the most common speech recognition toolkit, was employed to build the acoustic model. This toolkit shows The YouTube corpus was used as training and test data. Altogether, 917,346 Korean YouTube videos were collected as part of the training data, and 5147 videos were used as test data. KsponSpeech, SiTEC DICT, and ETRI Korean reading were also used as training data. Table 1 shows a description of the training data.

Table 1 Training data of ASR

Training Dataset	Hours	Number of Utterances
ETRI Korean reading	277.78	100,000
SiTEC DICT	57.79	20,806
Korean mobile assistant	100.00	92,874
KsponSpeech	965.2	620,000

The language model in ASR is constructed from the n-gram model, a probability model for a contiguous sequence of n symbols from a given text. The symbols include words, characters, syllables, morphemes, and sub-words. In particular, sub-words are pieces of words identified by the statistical tokenization model to avoid the out-of-vocabulary problem. In this experiment, the SRILM toolkit was employed to build our language model, and Sentencepiece was employed to tokenize sub-words [36, 37]. Text data from various Korean websites can be used as training data, such as YouTube transcriptions from the Google speech recognition API and YouTube video subtitles.

OpenFST was used to build WFST as a decoding network in our ASR. WFST combines the acoustic model, pronunciation model, and language model. This study employed the Kaldi decoding network with OpenFST libraries [38].

4.1.2 Experimental results

Model A is an acoustic model trained on corpora recorded in a quiet place: KsponSpeech, SiTEC DICT, and ETRI Korean reading. Model B is an acoustic model trained on 1053 hours of audio data from videos only on YouTube. Model C was trained with all the datasets used for models A and B.

The performance of ASR in the proposed system was assessed with four test datasets.

1. The Korean voice assistant commands (VAC) dataset has 871 Korean utterances recorded by a voice assistant, such as “오늘 날씨 어때?” (How is the weather today?). It is recorded in a quiet office environment.
2. The spontaneous speech dataset has 170 Korean utterances in various environments. It includes examples of complex spontaneous speech, such as filler words, pauses, repeated words, and word fragments.
3. The YouTube VLOG dataset contains 104 videos, with a total duration of approximately 17 hours. Various vlog creators record these in clean environments. In this dataset, only one person speaks simultaneously, and the videos are recorded in a relatively quiet place.

Table 2 Evaluation of ASR for different models for different test datasets

Test Dataset	Model A (CER, %)	Model B (CER, %)	Model C (CER, %)
VAC	2.06	6.01	3.01
Spontaneous speech	12.54	11.87	9.91
VLOG	23.05	23.33	17.31
VLOG-noisy	40.14	42.30	29.52

4. The YouTube VLOG-noisy dataset is a noisy version of YouTube VLOG. In this dataset, several people can speak simultaneously, and the videos are recorded in a relatively noisy environment, such as outdoors or in a concert hall.

Table 2 shows the evaluation of ASR in our proposed system. The model's error rate was measured with the character error rate (CER). While the word error rate is a commonly used metric for assessing speech recognition performance, for Korean text, the CER is used, as the space rules are flexible in Korean [39]. In the experiment with the VAC dataset, model A had the lowest CER, but in all other cases, model C performed superiorly.

4.2 Audio Event and Audio Scene Classification Task

4.2.1 Experimental setup

Various datasets were used for training: UrbanSound8K and BBC Sound FX for AEC and DCASE2016, TAU Urban Acoustic Scenes 2020, and FREESOUND for ASC [24]. The New York University distributed UrbanSound8K for audio event detection. It has a total of 8732 files with ten audio events. A subset was used in the evaluation. The BBC distributed the BBC Sound FX dataset for audio event detection [24]. DCASE2016 and TAU Urban Acoustic Scenes 2020 were distributed by the DCASE challenge [24]. FREESOUND, a cloud database with specific keywords, was uploaded by numerous users. Table 3 shows a description of training data for AEC and ASC.

Both the AEC and ASC models were built using MobileNet v2, which is based on a CNN. To improve the performance, SpecAugment was used to augment the data during training [40, 41].

4.2.2 Experimental results

The performance of AEC and ASC in the proposed system was measured with audio data from the YouTube VLOG dataset.

There were five versions of the AEC model:

1. Vanilla MobileNet v2
2. MobileNet v2 plus time augmentation
3. MobileNet v2 plus frequency augmentation
4. MobileNet v2 plus frequency augmentation and parameter tuning
5. MobileNet v2 plus time augmentation, frequency augmentation, and parameter tuning.

The results were compared with those for VGG-Resnet, which is an ensemble version of a CNN-based VGG network (VGGnet) [32] and a residual network [33].

Table 4 lists the accuracies of AEC in our proposed system. In this experiment, vanilla MobileNet v2 was less accurate than VGG-Resnet, although the MobileNet v2 model performed better after data augmentation. VGG-Resnet had a comparable performance in the audio event classification domain [32]. MobileNet v2 with time and frequency masking was considerably more accurate than the VGG-Resnet model. Its average accuracy increased by 0.57.

The ASC experiment employed coordinate attention and the early fusion method. Coordinate attention uses bi-directional average pooling to decompose channel attention into two-feature encoding [41]. This attention mechanism trains the long-range dependencies in feature maps [42]. Early fusion uses strides from two convolution layers with different strides in the first layer of the model [43]. The two convolutional layers generate two output features and concatenate them.

There were four versions of the ASC model:

1. Vanilla MobileNet v2
2. MobileNet v2 plus coordinate attention

Table 3 Evaluation of ASR for different models for different test datasets

Training Dataset	Number of Files	Collection
Baby crying	54	Urbansound8k
Bicycle bell	91	Google audio set
Boiling	51	YouTube, Freesound
Car passing	42	YouTube, Freesound
Dog barking	362	Urbansound8k
Door knocking	69	Urbansound8k
Human speech	196	YouTube
Scream	106	Google audio set
Indoor	4800	TAU urban acoustic scenes 2020
Outdoor	4800	TAU urban acoustic scenes 2020
Transcription	4800	TAU urban acoustic scenes 2020

Table 4 Evaluation of different AEC models for different audio events. (Aug., augmentation; Freq., frequency; Acc., accuracy)

Class	VGG-Resnet (Acc.)	MobileNet v2 (Acc.)	MobileNet v2 + time Aug. (Acc.)	MobileNet v2 + Freq. Aug. (Acc.)	MobileNet v2 + Freq. Aug. + tuning (Acc.)	MobileNet v2 + time Aug. + Freq. Aug. + tuning (Acc.)
Baby crying	0.67	0.00	0.20	1.00	0.90	1.00
Bicycle bell	1.00	0.60	1.00	1.00	1.00	1.00
Boiling	0.90	0.10	0.00	1.00	1.00	1.00
Car passing	0.36	0.00	0.00	0.00	0.80	0.80
Dog	0.00	1.00	1.00	1.00	1.00	1.00
Door	0.00	0.00	1.00	0.90	1.00	1.00
Speech	0.00	0.00	0.70	0.70	0.80	0.90
Scream	0.00	0.60	1.00	0.60	0.90	0.80
Average accuracy	0.37	0.29	0.61	0.77	0.93	**0.94**

Table 5 Evaluation of different ASC models. (Coor. Att., coordinate attention; Acc., accuracy)

Classes/Model	MobileNet v2 (Acc., %)	MobileNet v2 + Coor. Att. (Acc., %)	MobileNet v2 + Coor. Att. + Fusion (Acc., %)	MobileNet v2 + Coor. Att. + Fusion + tuning (Acc., %)
Indoor	0.61	0.71	0.70	0.68
Outdoor	0.63	0.70	0.69	0.73
Transportation	0.76	0.78	0.78	0.77
Average Acc.	0.67	**0.73**	0.72	**0.73**

3. MobileNet v2 plus coordinate attention and early fusion
4. MobileNet v2 plus coordinate attention, early fusion, and parameter tuning.

Table 5 lists the accuracies of ASC in our proposed system. The performance of MobileNet v2 with coordinate attention is higher than that of the vanilla MobileNet v2. Adding the early fusion method improves the performance of the outdoor class. However, the accuracy decreases for the indoor class when the model parameters are tuned.

4.3 Music Detection

YouTube background music and the TED-LIUM v1 corpus were used to train and evaluate music detection. Altogether, 100 files from this dataset were used as test data. The YouTube background music dataset was distributed by YouTube Studio (available at https://studio.youtube.com). There were 1271 music samples from YouTube creators. TED-LIUM v1, distributed by openSLR (available at https://www.openslr.org/resource), contained 118 h of English TED talks. This corpus was used as a counterexample for the music. GMM and isotonic regression were implemented with the scikit-learn library in Python (available at https://scikit-learn.org/). The accuracy of music detection was 73% for the test data.

4.4 Automatic Audio Metadata Generation System

4.4.1 Experimental setup

The automatic audio metadata generation system was composed of ASR with keyword extraction, AEC, ASC, and music detection. The proposed system was evaluated with the YouTube VLOG dataset, which was used in the ASR task as test data. Humans tagged all audio files. Nine audio events, three locations, whether music was detected, and keywords resulting from speech recognition were tagged in the test data.

Audio segmenters for each subsystem were developed using the FFMPEG library on Linux and the WebRTCVAD tool [44]. The acoustic model was implemented in Kaldi, the decoding network in OpenFST, and the language model in SRILM [45]. Our system used a 3.40 GHz Intel Xeon E5-2643 v4 CPU and two Nvidia Quadro RTX 5000 GPUs. To reduce the processing time, the AEC and ASC models were converted by TensorFlow-Lite, an optimization tool for DNN models. All subsystems were operated simultaneously and generated metadata, as depicted in Figure 1.

4.4.2 Experimental results

Experimental results yielded a mean accuracy of 65.83% in the test data. The average number of tags is 3.53. The videos that exhibit the low performance from the proposed system contain loud background music, so other sounds are difficult to detect in a noisy environment. In contrast, the videos that achieved 100% accuracy contain back-recorded voices in relatively quiet places.

5 Conclusions

We propose an automatic metadata generation system for audio content analysis using speech recognition, keyword extraction, audio event classification, audio scene classification, and music detection. The speech recognizer combines a TDNN-based acoustic model and an *n*-gram language model with a WFST decoding network, which learns from vlog data collected from the Korean reading corpus and YouTube, and self-collected YouTube vlogs. In the experiments, the syllable recognition rate was 17.31%. After training a MobileNet v2 model with UrbanSound8K and BBC Sound FX data, the AEC exhibits an average accuracy of 94% for YouTube vlog data. For the ASC, a MobileNet v2 model was trained with DCASE2016, TAU Urban Acoustic Scenes 2020, and FREESOUND data. It had an average accuracy of 73% for YouTube vlog data. Music detection yields an accuracy of 73% after learning a GMM and isotonic regression model trained on YouTube studio music and TED-LIUM data. For a total of 104 YouTube vlogs, the automatic audio metadata generation system combining all these models yields an average accuracy of 65.83%.

Acknowledgement

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2022-0-00621, Development of artificial intelligence technology that provides dialog-based multi-modal explainability).

References

[1] J. Jeon, H. Jo, 2020, "Effects of audio-visual interactions on soundscape and landscape perception and their influence on satisfaction with the urban environment," *Building and Environment*, vol. 169, p. 106544.

[2] S. Choi, K. On, Y. Heo, A. Seo, Y. Jang, et al., 2021, "DramaQA: character-centered video story understanding with hierarchical QA," in *Proc. Association for the Advancement of Artificial Intelligence*, pp. 1166–1174.

[3] T. Izumitani, R. Mukai, K. Kashino, 2008, "A background music detection method based on robust feature extraction," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Las Vegas, Nevada, USA, pp. 13–16.

[4] J. Wang, A. Yessenalina, A. Roshan-Ghias. 2021, "Exploring heterogeneous metadata for video recommendation with two-tower model," in *Proc. Recsys 2021 Workshop on Context-Aware Recommender Systems*, Amsterdam, Netherlands, pp. 1–8.

[5] Y. Ou, Z. Chen, F. Wu, 2021, "Multimodal local-global attention network for affective video content analysis," *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 31, no. 5, pp. 1901–1914.

[6] J. Kanna, S. Raj, M. Meena, S. Meghana, S. Roomi, 2020, "Deep learning based video analytics for person tracking," in *Proc. International Conference on Emerging Trends in Information Technology and Engineering*, VIT University Vellore Campus, India, pp. 1–6.

[7] M. Yoon, J. Lee, I. Jo, 2021, "Video learning analytics: Investigating behavioral patterns and learner clusters in video-based online learning," *The Internet and Higher Educations*, vol. 50, pp. 1–10.

[8] K. Yordanova, F. Kruger, T. Kirste, 2018, "Providing semantic annotation for the CMU grand challenge dataset," in *Procs. International Workshop on Annotation of useR Data for UbiquitOUs Systems*, Athens, Greece, pp. 579–584.

[9] J. Li, D. Yu, J. Huang, Y. Gong, 2012, "Improving wideband speech recognition using mixed-bandwidth training data in CD-DNN-HMM," in *Proc. IEEE Spoken Language Technology Workshop*, Miami, Florida, USA, pp. 131–136.

[10] M. Seltzer, D. Yu, Y. Wang, 2013, "An investigation of deep neural networks for noise robust speech recognition," in *Proc IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, BC, Canada, pp. 7398–7402.

[11] X. Chen, A. Ragni, X. Liu, M. Gales, 2017, "Investigating bidirectional recurrent neural network language models for speech recognition," in *Proc. the 18th Annual Conference of the International Speech Communication Association*, Stockholm, Sweden, pp. 269–273.

[12] G. Dahl, D. Yu, L. Deng, A. Acero, 2012, "Context-dependent pretrained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, pp. 30–42.

[13] M. Suzuki, N. Itoh, T. Nagano, G. Kurata, S. Thomas, 2019, "Improvements to n-gram language model using text generated from neural language model," in *Proc. the 44th International Conference on Acoustics, Speech and Signal Processing*, Brighton, UK, pp. 7245–7249.

[14] Y. Bengio, R. Ducharme, P. Vincent, C. Jauvin, 2003, "A neural probabilistic language model," *Journal of Machine Learning Research*, vol. 3, pp. 1137–1155.

[15] M. Mohri, F. Pereira, M. Riley, 2002, "Weighted finite-state transducers in speech recognition," *Computer Speech & Language*, vol. 16, pp. 69–88.

[16] S. Eddy, 2004, "What is a hidden Markov model?," *Nature Biotechnology*, vol. 22, pp. 1315–1316.

[17] L. Rabiner, 1989, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, pp. 257–286.

[18] E. Arisoy, A. Sethy, B. Ramabhadran, S. Chen, 2015, "Bidirectional recurrent neural network language models for automatic speech recognition," in *Proc. the 40th IEEE International Conference on Acoustics, Speech and Signal Processing*, South Brisbane, QLD, Australia, pp. 5421–5425.

[19] W. Chan, N. Jaitly, Q. Le, O. Vinyals, 2016, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Shanghai, China, pp. 4960–4964.

[20] A. Graves, S. Fernandez, F. Gomez, J. Schmidhuber, 2006, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proc. the international conference on Machine learning*, pp. 369–376.

[21] L. Lu, H. Jiang, H. Zhang, 2001, "A robust audio classification and segmentation method," in *Proc. ACM International Conference on Multimedia*, Ottawa, Canada, pp. 203–211.

[22] H. Lee, P. Pham, Y. Largman, Y. Ng, 2009, "Unsupervised feature learning for audio classification using convolutional deep belief networks," in *Proc. of Advances in Neural Information Processing Systems*, Vancouver, BC, Canada, pp. 1096–1104.

[23] J. Portelo, M. Bugalho, I. Trancoso, J. Neto, A. Abad, A. Serralheiro, 2009, "Non-speech audio event detection," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, Shanghai, China, pp. 1973–1976.

[24] M. Lim, D. Lee, H. Park, J. Oh, J. Kim, et al., 2018, "Convolutional neural network based audio event classification," *KSII Transactions on Internet and Information Systems*, vol. 12, pp. 2748–2760.

[25] A. Mesaros, T. Heittola, T. Virtanen, 2018, "A multi-device dataset for urban acoustic scene classification," in *Proc. of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop*, Surrey, UK, pp. 09–13.

[26] D. Jurafsky,, J. Martin, 2020, "Phonetics," in *Speech and Language Processing*, 3rd edn., London: Pearson, pp. 526–547.

[27] Y. Hao, Y. Chen, W. Zhang, G. Chen, L. Ruan, 2021, "A real-time music detection method based on convolutional neural network using Mel-spectrogram and spectral flux," in *Proc. INTER-NOISE and NOISE-CON Congress and Conference*, Washington, DC, USA, pp. 4919–5918.

[28] M. Zhang, J. Wu, H. Lin, P. Yuan, Y. Song, 2017, "The application of one-class classifier based on CNN in image defect detection," *Procedia Computer Science*, vol. 114, pp. 341–348.

[29] R. Mihalcea, P. Tarau, 2004, "Textrank: Bringing order into text," in *Proc. the 2004 Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain. pp. 404–411.

[30] P. Berkhin, 2011, "A survey on PageRank computing," *Internet Mathematics*, vol. 2, pp. 73–120.

[31] J. Chang, N. Kim, S. Mitra, 2006, "Voice activity detection based on multiple statistical models," *IEEE Transactions on Signal Processing*, vol. 54, pp. 1965–1976.

[32] S. Seo, C. Kim, J. Kim, 2022, "Convolutional neural networks using log mel-spectrogram separation for audio event classification with unknown devices, *Journal of Web Engineering*, vol. 21, pp. 497–522.

[33] K. He, X. Zhang, S. Ren, J. Sun, 2016, "Deep residual learning for image recognition," in *Proc. CVPR 2016 – 29th IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA, pp. 770–778.

[34] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L. Chen, 2016, "Mobilenetv2: inverted residuals and linear bottlenecks," in *Proc. CVPR 2018 – 31st IEEE Conference On Computer Vision and Pattern Recognition*, Salt Lake City, USA, pp. 4510–4520.

[35] Q. Zhang, F. Xu, J. Bai, 2021, "Audio fingerprint retrieval method based on feature dimension reduction and feature combination," *KSII Transactions on Internet and Information Systems*, vol. 15, pp. 522–539.

[36] A. Stolcke, 2002, "SRILM – an extensible language modeling toolkit," in *Proc. International Conference on Spoken Language Processing.*

[37] A. Kumar, R. Baruah, R. Mundotiya, A. Singh, 2020, "Transformer-based neural machine translation system for Hindi–Marathi: WMT20

shared task," in *Proc. Fifth Conference on Machine Translation*, pp. 393–395.

[38] C. Allauzen, M. Riley, J. Schalkwyk, W. Skut, M. Mohri, 2007, "OpenFst: A general and efficient weighted finite-state transducer library," in *Proc. International Conference on Implementation and Application of Automata*, Berlin, Heidelberg: Springer, pp. 11–23.

[39] J. Bang, S. Yun, S. Kim, M. Choi, M. Lee, et al., 2020, "KsponSpeech: Korean spontaneous speech corpus for automatic speech recognition, *Applied Sciences*, vol. 10, pp. 6936–6953.

[40] D. S. Park, W. Chan, Y. Zhang, C. Chiu, B. Zoph, E. D. Cubuk, Q. V. Le, 2019, "SpecAugment: a simple data augmentation method for automatic speech recognition," in *Proc. INTERSPEECH 2019 – 20th Annual Conference of the International Speech Communication Association*, Graz, Austria, pp. 2019–2680.

[41] S. You, C. Liu, J. Li, 2021, "Improvement of vocal detection accuracy using convolutional neural networks," *KSII Transactions on Internet and Information Systems*, vol. 15, pp. 729–748.

[42] Q. Hou, D. Zhou, J. Feng, 2021, "Coordinate attention for efficient mobile network design," in *Proc. the IEEE/CVF Conference on Computer Vision, Pattern Recognition*, pp. 13713–13722.

[43] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, *et al.*, 2017, "Attention is all you need," in *Proc. the 31st International Conference on Neural Information Processing Systems*, Long Beach, CA, USA, pp. 6000–6010.

[44] C. Snoek, M. Worring, A. Smeulders, 2005, "Early versus late fusion in semantic video analysis," in *Proc. the 13th annual ACM international conference on Multimedia*, Singapore, pp. 399-402.

[45] D. Zhu, S. Liu, C. Liu, 2020, Design and Construction of Intelligent Voice Control System, in *Proc. Artificial Intelligence in China*, pp. 319–325.

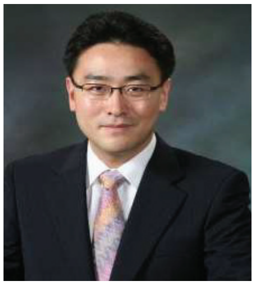
[46] B. Ojokoh, E. Adebisi, 2018, "A review of question answering systems, *Journal of Web Engineering*, vol. 17, pp. 717–758.

[47] U. Yadav, 2021, "Efficient retrieval of data using semantic search engine based on NLP and RDF," *Journal of Web Engineering*, vol. 20, pp. 717–758.

Biographies

Hosung Park received his B.E. degree in Computer Science and Engineering from Handong Global University in 2016. He also received his M.E. degree in Computer Science and Engineering from Sogang University in 2018. He is currently pursuing a Ph.D. degree in Computer Science and Engineering at Sogang University. His research interests include speech recognition and spoken multimedia content.

Yoonseo Chung received his B.E. degree in Computer Science and Engineering from Sogang University in 2022. He is currently pursuing an M.E. degree in Computer Science and Engineering at Sogang University. His research interests include speech recognition and audio event classification.



Ji-Hwan Kim received his B.E. and M.E. degrees in Computer Science from KAIST (Korea Advanced Institute of Science and Technology) in 1996 and 1998, respectively, and his Ph.D. degree in Engineering from the University of Cambridge in 2001. From 2001 to 2007, he was a chief research engineer and a senior research engineer for LG Electronics Institute of Technology, where he was engaged in the development of speech recognizers for mobile devices. In 2004, he was a visiting scientist at the MIT Media Lab. Since 2007, he has been a faculty member in the Department of Computer Science and Engineering, Sogang University. Currently, he is a full professor. His research interests include spoken multimedia content search, speech recognition for embedded systems and dialogue understanding.