Audio-visual keyword transformer for unconstrained sentence-level keyword spotting

Received: 25 November 2022 | Revised: 23 February 2023 | Accepted: 26 February 2023

CAAI Transactions on Intelligence Technology

DOI: 10.1049/cit2.12212

ORIGINAL RESEARCH

Audio–visual keyword transformer for unconstrained sentence-level keyword spotting

Yidi Li[1] | Jiale Ren[1,2] | Yawei Wang[1] | Guoquan Wang[1] | Xia Li[3] | Hong Liu[1]

[1]Key Laboratory of Machine Perception, Peking University, Shenzhen Graduate School, Shenzhen, China

[2]College of Electronics and Information Engineering, Sichuan University, Chengdu, China

[3]Department of Computer Science, ETH Zurich, Zurich, Switzerland

Correspondence

Jiale Ren.
Email: jialeren@stu.pku.edu.cn

Funding information

Science and Technology Plan of Shenzhen, Grant/Award Number: JCYJ20200109140410340; National Natural Science Foundation of China, Grant/Award Number: 62073004

Abstract

As one of the most effective methods to improve the accuracy and robustness of speech tasks, the audio–visual fusion approach has recently been introduced into the field of Keyword Spotting (KWS). However, existing audio–visual keyword spotting models are limited to detecting isolated words, while keyword spotting for unconstrained speech is still a challenging problem. To this end, an Audio–Visual Keyword Transformer (AVKT) network is proposed to spot keywords in unconstrained video clips. The authors present a transformer classifier with learnable CLS tokens to extract distinctive keyword features from the variable-length audio and visual inputs. The outputs of audio and visual branches are combined in a decision fusion module. As humans can easily notice whether a keyword appears in a sentence or not, our AVKT network can detect whether a video clip with a spoken sentence contains a pre-specified keyword. Moreover, the position of the keyword is localised in the attention map without additional position labels. Experimental results on the LRS2-KWS dataset and our newly collected PKU-KWS dataset show that the accuracy of AVKT exceeded 99% in clean scenes and 85% in extremely noisy conditions. The code is available at https://github.com/jialeren/AVKT.

KEYWORDS

artificial intelligence, multimodal approaches, natural language processing, neural network, speech processing

1 | INTRODUCTION

Keyword Spotting (KWS) is the task of detecting one or more specific words in a corpus, usually in the context of human–machine dialogue. KWS plays a crucial role in human–robot interaction [1] and is mainly used in voice assistants [2] and service robots [3].

In recent years, with the rapid development of deep learning techniques, various neural networks have been adopted to improve the performance of KWS [4–12]. Deep learning-based methods are shown to achieve excellent performance under experimental conditions. However, the performance of uni-modal KWS models based on speech signals degrades rapidly in noisy scenarios, which are common in real-world scenarios. Therefore, reducing the impact of noise on performance is the focus of many speech-related tasks, including KWS. For the most typical speech task, Audio Speech Recognition (ASR), audio–visual fusion has been proven to be a promising technique to tackle the noise problem, since the visual information is not affected by acoustic distortions [13–21]. As the most common forms of perception in human communication, hearing and watching have received increasing attention from researchers in the multi-modal field. On one hand, sound and images form a set of naturally interrelated modalities. On the other hand, the human processing of sound signals is often influenced by visual information. When listening to a person speaking, one also notices its movements, especially the movements of the mouth. In praxiology, McGurk discusses the interactions and effects of lip movements and sound in human understanding of multi-modal information [22]. Combining the complementary information between audio and vision corresponds to an increased amount of information, which is an essential factor for the success of Audio–Visual Speech Recognition (AVSR)

[23]. Following this intuition, the successful experience of AVSR can also be applied to KWS. However, relatively little research has been done on Audio–Visual Keyword Spotting (AVKWS). A lip descriptor is proposed as the visual feature extractor for a Hidden Markov Model (HMM)-based AVKWS model [24]. In Ref. [25], an audio–visual neural network based on Multidimensional Convolutional Neural Network (MCNN) is proposed to perform AVKWS. The above methods are limited to handling cases where the model input is an isolated word. However, in practical scenarios, keywords may appear anywhere in an unconstrained sentence. An audio–visual Query-by-Example (QbE) model is proposed to query whether a word is in a sentence or not [26]. The model maps audio, video, and keyword text to a common space and queries whether the keyword text exists in the sentence through the similarity map. This is a zero-shot method, which means that the querying keywords can be arbitrary. However, the zero-shot method cannot achieve high accuracy in specific scenes. Moreover, the keyword is used as a query in text, which essentially introduces text as an additional modality. More recently, Y. Su el al. proposed a hybrid fusion-based keyword recognition method to achieve sentence-level recognition [27]. However, in the existence of noise, the hybrid fusion does not handle the visual modality well, so the model still lacks robustness.

In contrast, this work focuses on the detection of specific keywords in unconstrained speech in real-world conversation. In this paper, we propose an Audio–Visual Keyword Transformer (AVKT) network. Based on the basic transformer encoder followed by a classifier head, AVKT can precisely determine the presence and position of pre-specified keywords in unconstrained variable-length speech or video. The process is characterised as a classification model. Using CLS tokens, the transformer encoder embeds variable-length audio and visual inputs into a fixed-length vector containing global information. During the calculation of the self-attention scores, the model automatically attends to the parts of the keywords. In addition, for the keywords present in the sentence, the position of them appearing in the whole utterance is indicated by the attention map. Note that the frame-by-frame position labels are not utilised during the training. The proposed label-free approach is suitable for simpler and more accessible datasets. The contribution of this paper are summarised as follows:

- A novel audio–visual transformer network is proposed for keyword spotting, which focuses on processing unconstrained sentence-level inputs rather than common isolated words.
- A transformer classifier with the learnable token is introduced to classify the keyword and explore the keyword position without additional position labels.
- A new Mandarin dataset PKU-KWS is collected for the AVKWS task. The excellent performance on the PKU-KWS and the LRS2-KWS datasets demonstrates the accuracy and robustness of the model, especially in extremely noisy environments.

The rest of this paper is organised as follows: Section 2 introduces three related works. Section 3 details the proposed AVKT model. The experiments and discussion in Section 4 present the datasets and implementation details. The accuracy and robustness of the proposed model are demonstrated by reporting detailed experimental results. Conclusions are finally given in Section 5.

2 | RELATED WORKS

In this section, we briefly review the related works in three areas: keyword spotting, audio–visual fusion, and transformer-based model.

2.1 | Keyword spotting

Keyword spotting is a classical task in the field of speech control systems for detecting specific words from the audio stream. In the early days, with the development of speech recognition technology, the most commonly used technology for KWS is based on Large-Vocabulary Continuous Speech Recognition (LVCSR) [28, 29]. Speech signals are decoded into phonemes, syllables and other speech units, or directly into text, and then keywords are retrieved from the decoded information. Decoding the entire sentence complicates the model and wastes a lot of resources. Later on, this inefficient method is gradually replaced by methods based on Hidden Markov Mode (HMM) and Gaussian Mixture Model (GMM) [30, 31]. These acoustic models divide the speech signal into several windows and encode and decode the signals by probability modelling to determine whether the current segment belongs to a keyword.

With the development of deep learning, the first deep learning-based keyword spotting model is proposed [32], which replaces probabilistic modelling with neural networks and outputs recognition results directly through the Convolutional Neural Network (CNN). Since then, feature extractors and acoustic models have been gradually replaced by neural networks, and deep learning-based methods become mainstream for KWS task. The deep learning models are broadly classified into two categories. One is the end-to-end method, which aims to directly predict the audio segment through the model and output the prediction result (with or without a keyword). The other is the QbE-based method [33], which calculates the similarity between the speech to be recognised and the keyword to query whether the word appears in the current speech segment. In recent years, more and more studies have proposed models with various network architectures [34–38], pushing the performance of KWS models to a new level. In this work, we solve the KWS task as a classification problem in an end-to-end manner. In addition to the use of audio, there are also a few studies using video signals for keyword spotting. Ref, [39] only uses video as input and spots given words unseen during training

through LSTMs and Query-by-Text method. Neural visual KWS is a very meaningful innovation, and this work obtains very promising results.

2.2 | Audio–visual fusion

The audio–visual fusion model refers to the synergistic usage of information from both visual and auditory modalities to process multi-modal tasks. As a complement to audio information, the presence of visual information can significantly enhance the robustness of the model in noisy scenes. The most common application of the audio–visual fusion model is the audio–visual speech recognition task. With the first audio–visual speech recognition system proposed [40], the task has gradually become a research hotspot in recent years.

The core of the audio–visual fusion method is how to effectively fuse the heterogeneous information from the two modalities of audio and visual. At present, common modal fusion methods are divided into two categories: decision fusion and feature fusion. Decision fusion refers to two independent models outputting their respective predictions and then fusing the results of both with a specific strategy to obtain the final result. The classic audio–visual fusion speech recognition model TMseq2seq [13] adopts this architecture. Different from decision fusion, feature fusion combines feature vectors extracted from audio and video streams by concatenation, averaging, or encoding. In the end-to-end audio–visual fusion speech recognition model proposed in Ref. [15], a feature extraction network is used to extract features from audio and visual streams, and then the features are directly concatenated and input into a speech recognition model, and the model directly outputs the prediction results. In addition to these two basic approaches, some studies have proposed more sophisticated fusion strategies to fully combine information from visual and audio modalities. A hybrid fusion method [16] is proposed to combine the advantages of both feature fusion and decision fusion. A multi-modal perception attention network is designed to measure the reliability and effectiveness of intermittent audio and visual streams disturbed by noise [41].

Nowadays, audio–visual fusion has become one of the most important solutions to solve the problem of noise interference [42, 43]. The performance of some audio–visual speech recognition models has surpassed human capabilities. Nevertheless, only very few works have focussed on using audio–visual fusion methods for KWS [24–26], which are limited to solving isolated word recognition tasks or query-based approaches to spot keywords. Ref. [27] proposed a sentence-level task, but its audio–visual fusion method is still not robust in loud-noise scenes. In this paper, our goal is to improve the accuracy and robustness of KWS in complex scenes through audio–visual fusion.

2.3 | Transformer-based model

As a new paradigm in the field of Natural Language Processing (NLP), the powerful modelling capabilities of the transformer model are gradually being applied to other research areas [44–46]. The Vision Transformer (ViT) [47] directly transfers the standard transformer model to the computer vision field. Due to the self-attention mechanism, ViT has a completely different learning paradigm and can even show better performance than CNN on many tasks. More recently, ViT and its variants [48–52] have achieved great success in the field of computer vision.

In the KWS task, the model is required to pay attention to keywords that appear locally in the audio and video sequences. Transformer models the global correlation of the input through the attention mechanism, which is suitable for the KWS task.

The attention mechanism can be naturally introduced into audio and visual tasks, as well as audio–visual fusion tasks [14, 53–56]. However, the transformer architecture applied to AVKWS has yet to be studied. Inspired by ViT, the transformer-based encoder is employed in our model to perform keyword spotting for audio and visual branches, respectively. We use a special learnable CLS token to map variable-length audio and video inputs to feature embeddings with a uniform dimension, and a subsequent Multi-Layer Perceptron (MLP) to accomplish the classification task. In addition, we implement keyword localisation through the attention matrix to better demonstrate the effect of the self-attention mechanism in this specific task.

3 | METHOD

The architecture of our AVKT model is depicted in Figure 1. In this framework, audio and visual streams are fed into the model simultaneously and generate respective keyword intermediate predictions. The resulting joint prediction is then produced by the decision fusion layer. For audio and visual streams, the KWS model is designed as the same structure. First, features from audio signals and video frames are extracted by two pre-trained models. Following the feature extractor is the transformer classifier. The attention map of the last attention layer is used to infer the location of keywords.

3.1 | Feature extractor

Representing the input signal as a useful feature is an essential role of the audio and visual model. Traditionally, widely used feature extractors are divided into two categories. In the early days, handcrafted features designed based on prior knowledge were usually used. For example, MFCC for audio and HOG for video. With the development of deep learning, handcrafted features were gradually replaced by deep learning-based features, which aim to use deep neural networks to learn the ideal feature representation for the corresponding task.

In recent years, Self-Supervised Learning (SSL) has achieved great success [57, 58]. Self-supervised learning on large datasets enables the model to obtain a common representation

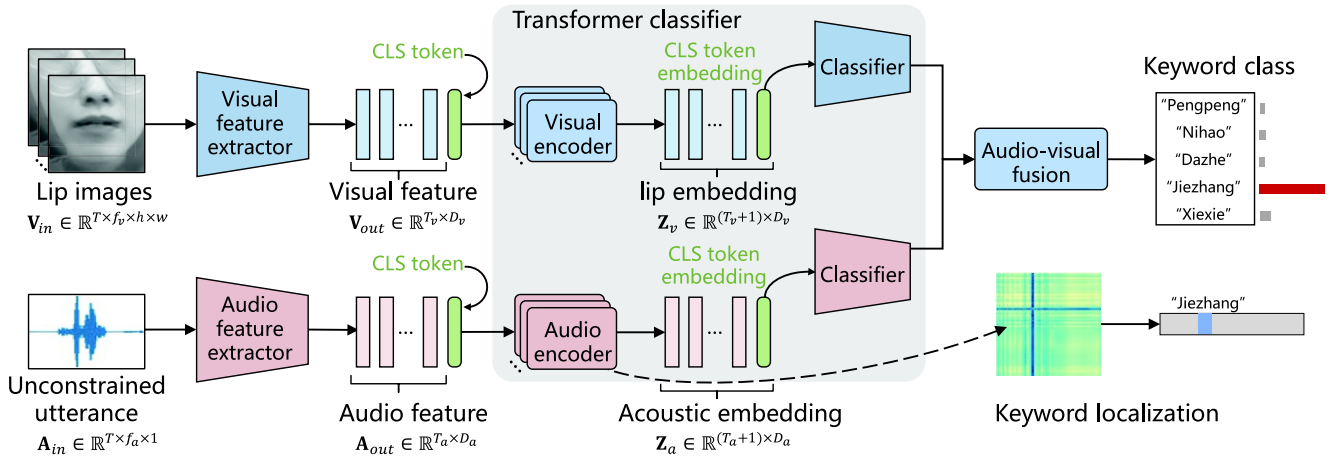


FIGURE 1 The framework of the proposed AVKT network. First, lip images and unconstrained utterances are input to the corresponding feature extractor in the visual and audio branches. CLS tokens are connected to features as the additional learnable parameters for the feature encoder. The CLS token embeddings are fed into the following classifier head. The predictions of the two branches are fused in the audio–visual fusion module. The attention map generated by the self-attention layer is used to locate keywords in the sentence.

of the features. Therefore, the pre-trained model based on large-scale data is employed as a feature extractor. The idea has recently been successfully used for speech and video analysis [59, 60]. The experimental results demonstrate that the general features extracted by the pre-trained model have strong generalisation capabilities and achieve remarkable results on a variety of downstream tasks, outperforming numerous hand-craft and deep features.

In this paper, WavLM [61] is employed as the audio feature extractor, and the visual part of AV-HuBERT [60] is used as the visual feature extractor. WavLM is a pre-trained model that jointly learns masked speech prediction and denoising speech to solve full-stack downstream speech tasks. It consists of a seven-layer convolutional feature encoder and a transformer encoder with gated relative position bias. AV-HuBERT is a self-supervised representation learning framework for audio–visual speech recognition based on HuBERT. The model benefits from an offline clustering step to provide target labels for a BERT-like prediction loss [62]. The audio features extracted by WavLM are more universal, while HuBERT focuses more on the ASR task. However, there is no general pre-trained model for lip feature extraction. Therefore, the visual model of AV-HuBERT is used as the visual feature extractor, which encodes masked image sequences into visual features via a hybrid ResNet-transformer architecture to predict the pre-determined sequence of discrete cluster assignments.

The inputs of the model are raw audio signal and grayscale images, which are denoted as follows:

$$\mathbf{A}_{in} \in \mathbb{R}^{T\times f_a\times 1}, \mathbf{V}_{in} \in \mathbb{R}^{T\times f_v\times h\times w}, \tag{1}$$

where T denotes the duration of the input signal, f_a and f_v denote the frame rate of audio and video signal, respectively. The image sequence is the lip Region of Interest (ROI) obtained by cropping the original video frames through face detection and face alignment. $h \times w$ is the size of each lip image. The features output by the extractor are represented as follows:

$$\mathbf{A}_{out} \in \mathbb{R}^{T_a\times D_a}, \mathbf{V}_{out} \in \mathbb{R}^{T_v\times D_v}, \tag{2}$$

where T_a and T_v represent the lengths of the time dimension, which varies with data. D_a and D_v denote the output feature lengths of the pre-trained model.

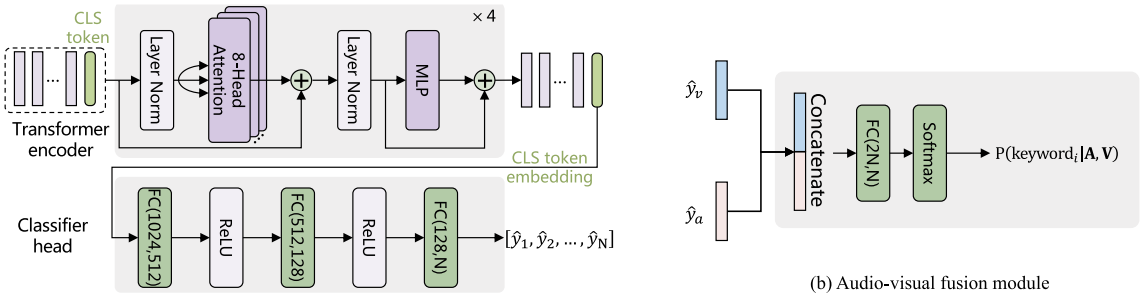
3.2 | Transformer classifier

After feature extractors, the downstream task is to classify sentences containing different keywords. To enable the model to focus on whether a keyword appears in a sentence or not, we used a transformer-based model as a keyword classifier. The model is able to handle variable-length sequences and obtain representations that contain extensive semantic information. We adopt the standard transformer architecture [63] as the fundamental component in designing our classifier network. The audio and visual features are passed through a multi-head attention encoder with eight heads and four layers, as shown in Figure 2a. Moreover, a learnable CLS token is employed as a global representation of the entire sentence. Before feeding into the transformer encoder, the CLS token is concatenated to the feature along the time dimension. The processing of the transformer encoder is formulated as follows:

$$\mathbf{Z}_a = \text{Encoder}_a([\mathbf{CLS}_a; \mathbf{A}_{out}]) \in \mathbb{R}^{(T_a+1)\times D_a}, \tag{3}$$

$$\mathbf{Z}_v = \text{Encoder}_v([\mathbf{CLS}_v; \mathbf{V}_{out}]) \in \mathbb{R}^{(T_v+1)\times D_v}. \tag{4}$$

After the transformer encoder, the CLS token embeddings are fed to the subsequent MLP-based classifier head. The audio



(a) Transformer classifier

(b) Audio-visual fusion module

FIGURE 2 Network architectures of transformer classifier and audio–visual fusion module. Transformer encoder is based on multi-head self-attention in a residual form. The output of the transformer classifier is the probability of each keyword, and N denotes the number of keyword categories. (a) Transformer classifier. (b) Audio–visual fusion module.

and visual CLS token embeddings that represent global features are denoted as follows:

$$\mathbf{Z}_a[0,:] \in \mathbb{R}^{1\times D_a}, \mathbf{Z}_v[0,:] \in \mathbb{R}^{1\times D_v}. \tag{5}$$

The classifier head is designed as a three-layer fully connected feed-forward network that consists of three linear transformations with the ReLU as an activation function in between. The final classification prediction is derived as follows:

$$\widehat{y}_a = \mathrm{MLP}(\mathbf{Z}_a[0,:]), \widehat{y}_v = \mathrm{MLP}(\mathbf{Z}_v[0,:]), \tag{6}$$

where $\widehat{y}_a$ and $\widehat{y}_v$ denote the prediction results of audio and visual branches, respectively.

3.3 | Audio-visual fusion

The decision fusion method is utilised in AVKT framework. The classifier for each of the modalities derives the corresponding predictions, and the decision layer generates the final fusion prediction probability. In general, the simplest method of decision fusion is a weighted summation of the prediction for every single model, which is formulated as follows:

$$\widehat{y}_{a,v} = \beta\widehat{y}_a + (1-\beta)\widehat{y}_v, \tag{7}$$

where β is a hyperparameter or learnt parameter and $\widehat{y}_{a,v}$ represents the fusion result.

To allow the model to focus on the internal interaction between two modalities, a linear layer is used as the fusion layer. The raw prediction results from the two modalities of audio and visual are one-dimensional vectors of length N. N denotes the number of keyword categories. $\widehat{y}_a$ and $\widehat{y}_v$ are concatenated to a vector of length 2N, which is fed into the fusion layer:

$$\widehat{y}_{av} = \mathbf{W}[\widehat{y}_a; \widehat{y}_v] + \mathrm{b}, \tag{8}$$

where the symbols $\mathbf{W}$ and b stand for the weight matrix and the basis vector, respectively. The predicted probability of each category is obtained through a softmax function:

$$\mathrm{P}(\mathrm{keywords}_i|\mathbf{A},\mathbf{V}) = \mathrm{Softmax}(\widehat{y}_{av}). \tag{9}$$

3.4 | Keyword localisation

In computer vision tasks, visualisation of the results of many transformer-based studies is used to validate the model. Self-attentive mechanisms are good at capturing the internal correlations of data or features, focussing the weights on more important information. In our classifier, the mechanism paid more attention to the keyword information in the input sequence. The feature extractor does not change the relative positional relationship of the original signal. The size of the weights in each self-attention layer of the transformer can be used to determine the focus of the model. Thus, it is possible to determine where the keywords actually appear in the phrase. In the transformer framework, the input signal is first multiplied by three projection matrices to get query ($\mathbf{Q}$), key ($\mathbf{K}$), and value ($\mathbf{V}$). Let $\mathbf{X} \in \mathbb{R}^{T\times d}$ denote the input vector, the processing is formulated as follows:

$$[\mathbf{Q};\mathbf{K};\mathbf{V}] = \mathbf{X} \times \left[\mathbf{W}_\mathbf{Q};\mathbf{W}_\mathbf{K};\mathbf{W}_\mathbf{V}\right]^\mathsf{T}, \tag{10}$$

where $\mathbf{W}$ denotes the different learnt linear projections, and $\mathbf{Q}, \mathbf{K} \in \mathbb{R}^{T\times D}$. The attention map is derived as follows:

$$\mathbf{M} = \mathbf{Q} \times \mathbf{K}^\mathbf{T} \in \mathbb{R}^{T\times T}, \tag{11}$$

where $\mathbf{M}$ represents the attention distribution for the current input. The position with the higher attention score indicates the relative position of that keyword in the input features. More specifically, $\mathbf{M}$ is first aggregated into a one-dimensional vector by averaging pooling in the time dimension. This vector is then compared with a threshold τ_0. The place of $tau \geq 0$ is estimated as the time interval where the keyword appears.

FIGURE 3 Example frames of the PKU-KWS dataset. From left to right are single, double, and triple speaker scenes. All videos are recorded in an indoor scene.

4 | EXPERIMENTS AND DISCUSSIONS

4.1 | Datasets

In this section, the proposed AVKT model is evaluated on two datasets. Our task aims to detect keywords in videos containing unconstrained utterances, which requires a large number of videos including different classes of keywords as datasets, but there is no acceptable public dataset that can be used directly. Therefore, we adopt two approaches to obtain two different datasets: selecting data from an existing dataset and collecting a new dataset.

4.1.1 | LRS2-KWS

LRS2 [15] is a dataset consisting of thousands of spoken sentences from BBC television and is commonly used in audio–visual speech recognition as well as lip reading. By counting the frequency of occurrence of each word in the entire dataset through the text labels corresponding to each video, five words ('About', 'When', 'My', 'Have', and 'One') that are common in daily life and appear with similar frequency are selected as the keywords. We collect 6,187 samples in LRS2 to form the new dataset, named LRS2-KWS, for keyword spotting task.

4.1.2 | PKU-KWS

We collected a dataset named PKU-KWS[1] for audio–visual keyword spotting, which is the first audio–visual corpus containing unconstrained Mandarin utterances of multiple speakers. It consists of 1,347 non-overlapped Mandarin dialogue videos with five different keywords: 'Pengpeng', 'Nihao', 'Dazhe', 'Jiezhang', and 'Xiexie'. The videos are recorded by 25 people in an indoor laboratory environment, and the recordings are constructed to real-life dialogue scenarios with varying speech rates and accents. 700 of these videos are scenes of multiple people talking, where two or three people are in the image simultaneously. Figure 3 shows some examples of typical frames in the PKU-KWS dataset. Table 1 shows specifications of the LRS2-KWS and the PKU-KWS datasets, including the data format and the number of video samples for each keyword.

TABLE 1 Specifications of the LRS2-KWS and the PKU-KWS datasets.

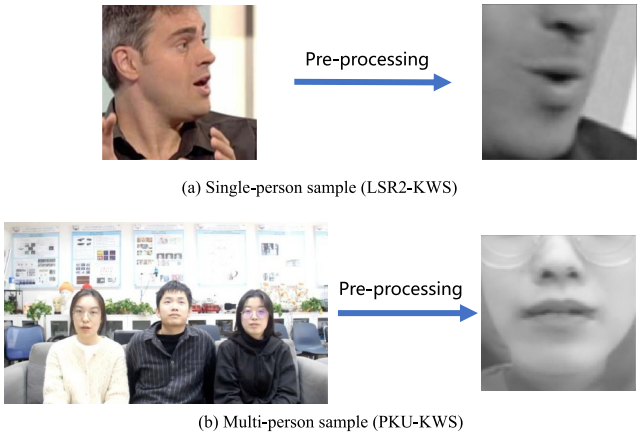
Dataset	Format	Number of samples	Keywords	
LRS2-KWS	mov, 480P	6187	'About'	1416
			'When'	1083
			'Have'	1147
			'My'	1329
			'One'	1212
PKU-KWS	mp4, 1080P	1347	'Pengpeng'	328
			'Nihao'	283
			'Jiezhang'	246
			'Dazhe'	260
			'Xiexie'	230

4.2 | Implementation details

4.2.1 | Pre-processing

The audio–visual data in the LRS2-KWS and the PKU-KWS datasets are processed as follows before being fed into the model. For audio signal, the audio is collected individually at a 16 kHz sample rate, leaving only a single channel. For visual signal, the Mouth Region of Interest (ROI) is first extracted for each video frame. During the extraction, each face is aligned with the alignment method in Ref. [64] to reduce the effect of different angles and deformations on the appearance. When multiple people are present in the image, the mouth ROI corresponding to each person is averaged pixel by pixel. As seen from the extracted lip area, multiple faces are overlapping and the lips are in the same position. The video sequences carry pixel motion information between frames, so the overlay of the multiplayer ROI retains a portion of the lip motion information in the presence of participants speaking. The extracted mouth ROI is uniformly converted to a grayscale image with the size of 96 × 96. Figure 4 shows two examples of visual processing.

[1]This dataset is available at https://zenodo.org/record/6792058.



(a) Single-person sample (LSR2-KWS)

(b) Multi-person sample (PKU-KWS)

FIGURE 4 Two examples of mouth ROI extraction. The upper part and the lower part show the processing of the single (from LRS2-KWS) and multi-person (from PKU-KWS) video scenes respectively. In the processing of the single-person scene, the side-facing face is corrected to a frontal view. In the multi-person scene, the ROIs of multiple faces are averaged to a single image. (a) Single-person sample (LSR2-KWS). (b) Multi-person sample (PKU-KWS).

In the process of data loading, the method of padding the data with zeros is used so that each batch of data has the same length. In the same batch, the data are padded in the time dimension to the maximum length of all data in the current batch. Due to the computational nature of the self-attention mechanism, the additional zeros do not affect the properties of the original data.

4.2.2 | Training procedure

The staged training strategy is adopted in our method. First, the classifiers for audio and visual modalities are pre-trained. The uni-modal models can also be used independently as the audio-only and visual-only KWS models. After that, the two models are combined to train the audio–visual fusion model. During training, the original dataset is first used, and then random noise is added to samples after the convergence of the first stage. The method of adding noise is similar to that in Ref. [5]. For each audio signal, a segment with the same length is randomly intercepted in a one-minute-long white noise audio clip. The noise fragment is then added to the original audio signal according to a selected Signal-to-Noise Ratio (SNR). Besides, the cross-entropy loss is used as the loss function and the parameters of the pre-trained extractor are fixed during training. The loss function is defined as follows:

$$\mathcal{L} = -\sum_{i=1}^{N} y_i \log \widehat{y}_i. \tag{12}$$

For the training of audio and video models, 35 epochs are required to be carried out at a learning rate of 0.01 (the learning rate is set to 0.001 when random noise is added). Only five epochs are needed to achieve good results for the training of the audio–visual fusion model. All training is conducted on one NVIDIA 3090 GPU with the batch size set to 32.

During the process of training and testing, all the samples in the dataset are randomly shuffled and used as training, validation, and testing set with a ratio of 8:1:1. According to the staged training strategy, the model is trained on the two datasets separately and then test on the test set. In the experiment, comparison experiments are conducted to compare the performance of audio-only models, visual-only models, and audio–visual fusion models under different noise conditions (SNRs vary from 10 dB to −10 dB within intervals of 5 dB). A series of supplementary experiments are conducted to test the performance of the model in real complex scenarios, of which details are explained in the next part.

4.3 | Evaluation metrics

For multi-classification task, accuracy is used as an evaluation metric. During testing, the average accuracy of five tests is reported as the experimental result. For the binary classification task, recall rate, precision, and F1-score are reported for comprehensive evaluation. Considering four parameters namely True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN), the above mentioned criteria are defined as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{13}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{14}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{15}$$

$$\begin{aligned} \text{F1 - Score} &= 2 \times \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \\ &= \frac{2\text{TP} + \text{TN}}{2\text{TP} + \text{FP} + \text{FN}} \end{aligned} \tag{16}$$

4.4 | Results

4.4.1 | Comparison with the state-of-the-art

We compare the AVKT model with two state-of-the-art methods proposed in Refs. [25, 27]. For the MCNN-based model [25], its relatively simple structure is difficult to converge on the sentence-level dataset. Considering that the model is designed for isolated word spotting, we construct a subset of PKU-KWS for a fair comparison. This subset contains only short sentences with lengths of less than 5 seconds for audio and video. The accuracy under different noise conditions is shown in Figure 5. The result shows that the original

MCNN-based model in Ref. [25] achieves 55.88% accuracy in clean scenes on the PKU-KWS sub-dataset. The performance decreases rapidly with the increase in noise. We further improve the model structure by adding convolutional and residual layers, which is denoted as MCNN*. The accuracy of MCNN* can reach 85.29%, but its poor robustness cannot cope with noise scenarios. With the noise increasing, the accuracy of MCNN* is maintained at about 58%. Since the noise is more disturbing to the audio model, its accuracy mainly comes from the design of the visual model. On the PKU-KWS subset, our proposed model is more accurate and robust than the other two comparison models under different noises. Even under noisy conditions with a SNR of −10 dB, the accuracy of AVKT reaches 89.41%, which exceeds the performance of the improved MCNN model in clean scenes (85.29%). The above results show that the MCNN-based model [25] is not applicable to complex sentence-level tasks. Nevertheless, our AVKT model designed for sentence-level tasks can still be effectively used for isolated keyword spotting.

The proposed method is also compared with the hybrid-fusion audio–visual keyword spotting model proposed in Ref. [27]. The results of Ref. [27] are reported in their papers. Note that the accuracy at −10 dB is not reported in the experimental results of Ref. [27]. The experiments are performed on the PKU-KWS dataset. Table 2 reports the accuracy of uni-modal and multi-modal methods under different noise conditions. Note that the visual-only method is not disturbed by acoustic noise variations. As shown in Table 2, with the advantage of the proposed audio–visual transformer architecture, our AVKT model achieves much higher accuracy than Ref. [27]. The accuracy of the audio-only model of Ref. [27] is highly disturbed by noise. Our audio-only model is both accurate and robust, with higher accuracy than the visual-only model (except under adverse noise conditions with SNR of −10 dB). Audio-visual fusion further enhances the accuracy of each model. The comparison results in Figure 5 and Table 2 demonstrate that the AVKT method achieves state-of-the-art results, making our method a powerful baseline for transformer-based audio–visual keyword spotting.

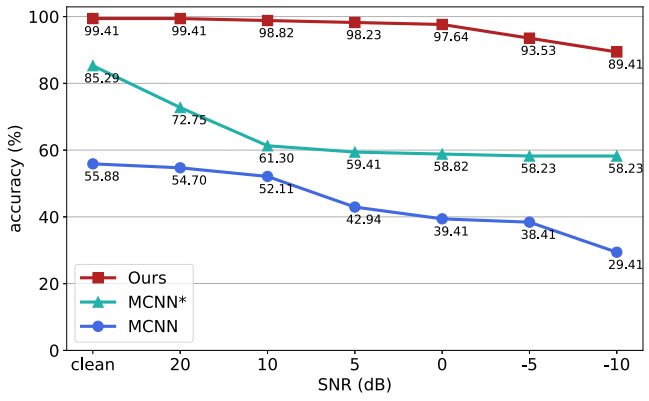


FIGURE 5 Accuracy (%) of the proposed AVKT and MCNN-based models [25] on the PKU-KWS sub-dataset. MCNN* denotes the improved MCNN model.

4.4.2 | Result on the LRS2-KWS dataset

We evaluate the proposed model on the LRS2-KWS dataset. As shown in Table 3, the classification accuracy of the audio-only model decreases significantly with the increase of noise, while the visual-only model is not affected by noise. The audio–visual fusion model effectively combines the information of two modalities and achieves excellent classification accuracy under various noise conditions, especially in the case of small SNRs. Compared with LRS2-KWS, audio and videos in the PKU-KWS dataset are clearer. Therefore, the performance on PKU-KWS (in Table 2) is slightly higher than LRS2-KWS. The results on these two datasets demonstrate that our model can effectively distinguish sentences containing different keywords in the presence of strong noise.

In addition, we add negative samples that do not contain any keywords to the five categories of keywords. In terms of noise, besides white noise, four types of background noises used in Ref. [5] (pink noise, exercise bike, dishwashing, and running tap) are also applied. The results in Table 4

TABLE 2 Accuracy (%) in different noise cases with the method proposed by [27] on the PKU-KWS dataset.

SNR (dB)		Clean	15	10	5	0	−5	−10
AO	Su et al.	98.50	96.00	92.50	85.50	65.50	36.00	-
	Ours	**98.72**	**98.72**	**98.42**	**97.64**	**96.58**	**85.82**	**62.20**
VO	Su et al.	82.00	82.00	82.00	82.00	82.00	82.00	82.00
	Ours	**85.19**	**85.19**	**85.19**	**85.19**	**85.19**	**85.19**	**85.19**
AV	Su et al.	98.00	98.00	97.50	96.50	89.50	71.00	-
	Ours	**99.54**	**99.54**	**99.05**	**98.89**	**96.58**	**88.50**	**85.98**

Note: Bold text indicates the best results.

Abbreviations: AO, audio-only; VO, visual-only; AV, audio–visual; Su et al: [27].

TABLE 3 Accuracy (%) of the proposed AVKT on the LRS2-KWS dataset under SNRs vary from 10 dB to −10 dB.

SNR (dB)	Clean	10	5	0	−5	−10
AO	98.39	97.90	97.25	91.83	71.36	41.75
VO	85.43	85.43	85.43	85.43	85.43	85.43
AV	**99.07**	**98.15**	**97.54**	**95.34**	**90.19**	**86.44**

Note: Bold text indicates the best results.

Abbreviations: AO, audio-only; VO, visual-only; AV, audio-visual.

TABLE 4 Keyword spotting accuracy (%) of the proposed AVKT model on the LRS2-KWS dataset under four types of noise after adding negative samples without keywords.

SNR (dB)	Clean	10	5	0	−5	−10
White noise	96.62	96.24	94.29	91.66	84.01	79.42
Pink noise	96.62	95.82	94.30	90.68	83.72	78.44
Bike	96.62	96.24	93.05	89.01	83.03	78.16
Dishes	96.62	94.85	91.66	85.25	81.36	77.75
Running tap	96.62	96.24	94.71	91.37	87.90	82.06

demonstrate that our model can distinguish sentences without keywords. As the task becomes more difficult, the performance is slightly lower than the previous keyword classification (in Table 3). Overall, our model can accurately distinguish sentences containing keywords in noise scenarios of different types and intensities, which demonstrates the generalisation, accuracy, and robustness of our AVKT model.

4.4.3 | Result of binary classification

In practical scenarios, more attention is paid to the presence of keywords rather than their categories. To verify the performance of the model in this case, an additional binary classification experiment is conducted. The video in the LRS2-KWS dataset containing five classes of keywords are considered as positive classes containing keywords, and the utterances that do not contain keywords are defined as negative samples. The performance of the model is evaluated from the perspective of binary classification. Experimental results in terms of the recall rate and precision are listed in Tables 5 and 6. As can be seen from the two tables, the audio–visual fusion model performs extremely well in different environmental noise situations when the task is simplified to two classifications. In the absence of noise, recall, and accuracy can reach more than 99% and 97%, respectively. Especially under the noise condition of −10 dB, the recall rate can still reach more than 95%. In most practical scenarios, only the presence or absence of keywords need to be judged, not the type of keywords. So in terms of recall, the proposed model has practical application value and can maintain a fairly high recall rate in extremely noisy scenarios. In the same noisy environment, the precision is slightly lower than the recall, which indicates that the model suffers from some misclassification. Although it can accurately identify sentences containing keywords, it sacrifices a certain misclassification rate.

TABLE 5 Binary recall rate (%) of our AVKT model on the LRS2-KWS dataset under different noise types.

SNR (dB)	Clean	10	5	0	−5	−10
White noise	99.50	99.01	98.67	98.32	96.28	96.21
Pink noise	99.50	99.50	98.83	98.29	96.48	95.36
Bike	99.50	99.17	98.65	98.24	96.98	96.10
Dishes	99.50	98.81	97.85	97.81	96.57	95.29
Running tap	99.50	99.49	99.33	98.47	96.95	95.44

TABLE 6 Binary precision (%) of our AVKT model on the LRS2-KWS dataset under different noise types.

SNR (dB)	Clean	10	5	0	−5	−10
White noise	97.25	97.08	96.60	94.66	92.23	90.45
Pink noise	97.25	96.60	95.63	93.53	93.20	93.04
Bike	97.25	96.28	94.82	90.61	88.19	87.70
Dishes	97.25	93.85	90.12	86.89	86.56	83.98
Running tap	97.25	97.24	96.27	94.01	92.71	91.58

4.4.4 | Visualisation of keyword localisation

In our method, the attention map of the last attention layer in the transformer classifier of the audio stream is taken for keyword localisation. Figure 6 shows some visualisation examples for keyword localisation in the form of heatmap. The text content is only used to show the effect of localisation, and the text modality is not considered in this work. The red parts in the audio waveform and text represent the position of the manually marked keywords. Note that the position labels are not used as supervised signals in our model. As shown in the figure, the horizontal axis of the attention graph corresponds to the time T dimension, where the dark areas indicate the locations of keyword occurrences, which are automatically learnt in the transformer classifier model.

5 | CONCLUSIONS

In this paper, an audio–visual transformer network is presented for keyword spotting, which not only implements common isolated word spotting but also focuses on more complex unconstrained sentence-level input. We introduce the transformed classifier with learnable CLS tokens to handle variable-length audio and visual features and design a decision fusion module to combine the prediction results of audio and visual branches. In particular, the position of the keyword is localised by the maximum value in the attention map, which demonstrates the interpretability of the model. Additionally, we collect and publish a new sentence-level multi-speaker AVKWS dataset that will be made available to the research community.

The significance of our AVKT lies in the ability to handle unlimited recordings, which makes it applicable in the real world and in human–computer dialogue scenarios. In addition,

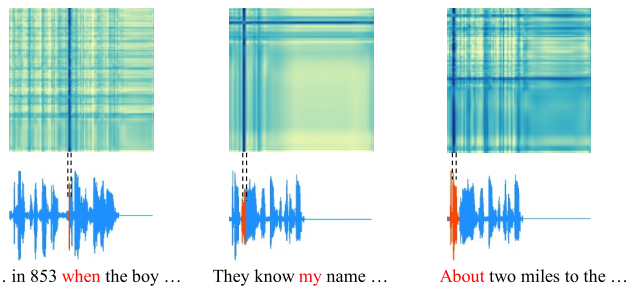


FIGURE 6 Visualisation of keyword localisation. The text at the bottom is the content of the input utterance. The heatmap is the attention map obtained from the transformer classifier model. The red part of the audio waveform is the manually labelled position of the real keywords. The length of the audio waveform is the same as the dimension of the attention heatmap.

the use of visual information from mouth movements provides additional cues that can help identify keywords, especially in cases where the audio quality is poor or noisy. Future work focuses on the challenging problem of discovering multiple keywords in highly complex multi-speaker environments.

ACKNOWLEDGEMENTS

This work is supported by the National Natural Science Foundation of China (No. 62073004), and the Science and Technology Plan of Shenzhen (No. JCYJ20200109140410340).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in https://zenodo.org/record/6792058 and https://www.robots.ox.ac.uk/~vgg/data/lip_reading/lrs2.html.

ORCID

Yidi Li https://orcid.org/0000-0002-5236-7010

REFERENCES

1. Ask, T.F., et al.: Human-human communication in cyber threat situations: a systematic review. In: International Conference on Human-Computer Interaction, pp. 21–43. Springer (2021).
2. Hoy, M.B.: Alexa, Siri, Cortana, and more: an introduction to voice assistants. Med. Ref. Serv. Q. 37(1), 81–88 (2018). https://doi.org/10.1080/02763869.2018.1404391
3. Lee, D.-W., Lee, D.W.: A study on current states and required technologies of smart speaker in service robot. J. Adv. Inf. Technol. Converg. 8(2), 129–137 (2018). https://doi.org/10.14801/jaitc.2018.8.2.129
4. Arik, S.O., et al.: Convolutional recurrent neural networks for small-footprint keyword spotting (2017). *arXiv preprint arXiv:1703.05390*
5. Warden, P.: Speech commands: a dataset for limited-vocabulary speech recognition (2018). *arXiv preprint arXiv:1804.03209*
6. Tang, R., Lin, J.: Deep residual learning for small-footprint keyword spotting. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5484–5488. (2018).
7. Higuchi, T., Gupta, A., Dhir, C.: Multi-task learning with cross attention for keyword spotting (2021). *arXiv preprint arXiv:2107.07634*
8. Berg, A., O'Connor, M., Cruz, M.T.: Keyword transformer: a self-attention model for keyword spotting (2021). *arXiv preprint arXiv:2104.00769*
9. Neti, C., et al.: Audio visual speech recognition. IDIAP, Tech. Rep. (2000).
10. Nefian, A.V., et al.: Dynamic Bayesian networks for audio-visual speech recognition. EURASIP J. Appl. Signal Process. 2002(11), 1–15 (2002). https://doi.org/10.1155/s1110865702206083
11. Dupont, S., Luettin, J.: Audio-visual speech modeling for continuous speech recognition. IEEE Trans. Multimed. 2(3), 141–151 (2000). https://doi.org/10.1109/6046.865479
12. Mroueh, Y., Marcheret, E., Goel, V.: Deep multimodal learning for audio-visual speech recognition. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2130–2134. IEEE (2015).
13. Afouras, T., et al.: Deep audio-visual speech recognition. IEEE Trans. Pattern Anal. Mach. Intell. 44(12), 1–8727 (2018). https://doi.org/10.1109/tpami.2018.2889052
14. Ma, P., Petridis, S., Pantic, M.: End-to-end audio-visual speech recognition with conformers. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7613–7617 (2021).
15. Petridis, S., et al.: End-to-end audiovisual speech recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6548–6552 (2018).
16. Liu, H., Li, W., Yang, B.: Robust audio-visual speech recognition based on hybrid fusion. In: International Conference on Pattern Recognition (ICPR), pp. 7580–7586 (2021).
17. Liu, H., Wang, Y., Yang, B.: Mutual alignment between audiovisual features for end-to-end audiovisual speech recognition. In: International Conference on Pattern Recognition (ICPR), pp. 5348–5353 (2021).
18. Liu, H., Xu, W., Yang, B.: Audio-visual speech recognition using a two-step feature fusion strategy. In: International Conference on Pattern Recognition (ICPR), pp. 1896–1903 (2021).
19. Parnami, A., Lee, M.: Few-shot keyword spotting with prototypical networks. In: 2022 7th International Conference on Machine Learning Technologies (ICMLT), pp. 277–283 (2022).
20. Ahmed, S., et al.: Towards more robust keyword spotting for voice assistants. In: 31st USENIX Security Symposium (USENIX Security 22) (2022).
21. Ghosh, A., et al.: Low-resource low-footprint wake-word detection using knowledge distillation (2022). *arXiv preprint arXiv:2207.03331*
22. McGurk, H., MacDonald, J.: Hearing lips and seeing voices. Nature 264(5588), 746–748 (1976). https://doi.org/10.1038/264746a0
23. Yang, S., Guan, Y.-p.: Audio–visual perception-based multimodal hci. J. Eng. 2018(4), 190–198 (2018). https://doi.org/10.1049/joe.2017.0333
24. Wu, P., et al.: A novel lip descriptor for audio-visual keyword spotting based on adaptive decision fusion. IEEE Trans. Multimed. 18(3), 326–338 (2016). https://doi.org/10.1109/tmm.2016.2520091
25. Ding, R., Pang, C., Liu, H.: Audio-visual keyword spotting based on multidimensional convolutional neural network. In: IEEE International Conference on Image Processing (ICIP), pp. 4138–4142 (2018).
26. Momeni, L., et al.: Seeing wake words: audio-visual keyword spotting (2020). *arXiv preprint arXiv:2009.01225*
27. Su, Y., Miao, Z., Liu, H.: Audio-visual multi-person keyword spotting via hybrid fusion. In: CAAI International Conference on Artificial Intelligence (CICAI) (2022).
28. Chen, B.: Word topic models for spoken document retrieval and transcription. ACM Trans. Asian Lang. Inf. Process 8(1), 1–27 (2009). https://doi.org/10.1145/1482343.1482345
29. Karakos, D., et al.: Score normalization and system combination for improved keyword spotting. In: 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 210–215. IEEE (2013).
30. Rose, R.C., Paul, D.B.: A hidden Markov model based keyword recognition system. In: International Conference on Acoustics, Speech, and Signal Processing, pp. 129–132. IEEE (1990).
31. Rohlicek, J.R., et al.: Continuous hidden Markov modeling for speaker-independent word spotting. In: International Conference on Acoustics, Speech, and Signal Processing, pp. 627–630. IEEE (1989).
32. Wilpon, J.G., et al.: Automatic recognition of keywords in unconstrained speech using hidden Markov models. IEEE Trans. Acoust. Speech Signal Process. 38(11), 1870–1878 (1990). https://doi.org/10.1109/29.103088
33. Mandal, A., Prasanna Kumar, K., Mitra, P.: Recent developments in spoken term detection: a survey. Int. J. Speech Technol. 17(2), 183–198 (2014). https://doi.org/10.1007/s10772-013-9217-1
34. Chen, G., Parada, C., Heigold, G.: Small-footprint keyword spotting using deep neural networks. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4087–4091. IEEE (2014).
35. Sainath, T.N., Parada, C.: Convolutional neural networks for small-footprint keyword spotting. Proc. Interspeech 2015, 1478–1482 (2015)
36. Stewart, D., et al.: Robust audio-visual speech recognition under noisy audio-video conditions. IEEE Trans. Cybern. 44(2), 175–184 (2013). https://doi.org/10.1109/tcyb.2013.2250954
37. Abdelaziz, A.H., Zeiler, S., Kolossa, D.: Learning dynamic stream weights for coupled-hmm-based audio-visual speech recognition. In: IEEE/ACM Transactions on Audio, Speech, and Language Processing, (Vol, 23(5), pp. 863–876) (2015).

38. Zheng, H., Wang, M., Li, Z.: Audio-visual speaker identification with multi-view distance metric learning. In: 2010 IEEE International Conference on Image Processing, pp. 4561–4564. IEEE (2010).
39. Stafylakis, T., Tzimiropoulos, G.: Zero-shot keyword spotting for visual speech recognition in-the-wild. In: European Conference on Computer Vision (2018).
40. Petajan, E.D.: Automatic Lipreading to Enhance Speech Recognition (Speech Reading). University of Illinois at Urbana-Champaign (1984).
41. Li, Y., Liu, H., Tang, H.: Multi-modal perception attention network with self-supervised learning for audio-visual speaker tracking. In: AAAI Conference on Artificial Intelligence, pp. 1456–1463 (2022).
42. Han, H., et al.: Noise-tolerant learning for audio-visual action recognition (2022). *arXiv preprint arXiv:2205.07611*
43. Shi, B., Hsu, W.-N., Mohamed, A.: Robust self-supervised audio-visual speech recognition (2022). *arXiv preprint arXiv:2201.01763*
44. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021).
45. Zhao, H., et al.: Point transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 16259–16268 (2021).
46. Strudel, R., et al.: Segmenter: transformer for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7262–7272 (2021).
47. Dosovitskiy, A. et al.: An image is worth 16x16 words: transformers for image recognition at scale (2020). *arXiv preprint arXiv:2010.11929*
48. Chen, C.-F.R., Fan, Q., Panda, R.: Crossvit: cross-attention multi-scale vision transformer for image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 357–366 (2021).
49. Li, Y., et al.: Contextual transformer networks for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. 45(2), 1489–1500 (2022). https://doi.org/10.1109/tpami.2022.3164083
50. Mao, X., et al.: Towards robust vision transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12042–12051 (2022).
51. Chen, C.-F., Panda, R., Fan, Q.: Regionvit: regional-to-local attention for vision transformers. In: International Conference on Learning Representations (2022).
52. Yao, T., et al.: Wave-vit: unifying wavelet and transformers for visual representation learning. In: European Conference on Computer Vision, pp. 328–345 (2022).
53. Serdyuk, D., Braga, O., Siohan, O.: Audio-visual speech recognition is worth 32 × 32 × 8 voxels. In: 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 796–802. IEEE (2021).
54. Chang, F.-J., et al.: End-to-end multi-channel transformer for speech recognition. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5884–5888. IEEE (2021).
55. Li, J., et al.: Recent advances in end-to-end automatic speech recognition. APSIPA Transactions on Signal and Information Processing 11(1) (2022). https://doi.org/10.1561/116.00000050
56. Prajwal, K., Afouras, T., Zisserman, A.: Sub-word level lip reading with visual attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5162–5172 (2022).
57. Devlin, J., et al.: Bert: pre-training of deep bidirectional transformers for language understanding (2018). *arXiv preprint arXiv:1810.04805*
58. Radford, A., et al.: Language models are unsupervised multitask learners. OpenAI blog 1(8), 9 (2019)
59. Yang, S.-w., et al.: Superb: speech processing universal performance benchmark (2021). *arXiv preprint arXiv:2105.01051*
60. Shi, B., et al.: Learning audio-visual speech representation by masked multimodal cluster prediction (2022). *arXiv preprint arXiv:2201.02184*
61. Chen, S., et al.: Wavlm: large-scale self-supervised pre-training for full stack speech processing (2021). *arXiv preprint arXiv:2110.13900*
62. Tseng, L.-H., et al.: Mandarin-English code-switching speech recognition with self-supervised speech representation models (2021). *arXiv preprint arXiv:2110.03504*
63. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 5998–6008. (2017).
64. Bulat, A., Tzimiropoulos, G.: How far are we from solving the 2d and 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 1021–1030. (2017).

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: Li, Y., et al.: Audio–visual keyword transformer for unconstrained sentence-level keyword spotting. CAAI Trans. Intell. Technol. 9(1), 142–152 (2024). https://doi.org/10.1049/cit2.12212