On-device audio-visual multi-person wake word spotting

Journal Article

Author(s):
Li, Yidi; Wang, Guoquan; Chen, Zhan; Tang, Hao; Liu, Hong

Publication date:
2023-12

Permanent link:
https://doi.org/10.3929/ethz-b-000603879

Originally published in:
CAAI Transactions on Intelligence Technology 8(4), https://doi.org/10.1049/cit2.12189

Received: 18 October 2022 | Revised: 14 December 2022 | Accepted: 9 January 2023

CAAI Transactions on Intelligence Technology

DOI: 10.1049/cit2.12189

ORIGINAL RESEARCH

On-device audio-visual multi-person wake word spotting

Yidi Li[1] | Guoquan Wang[1,2] | Zhan Chen[1] | Hao Tang[3] | Hong Liu[1]

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

[2]College of Computer and Information, Hefei University of Technology, Hefei, China

[3]Computer Vision Lab, ETH Zurich, Zurich, Switzerland

Correspondence

Guoquan Wang, Key Laboratory of Machine Perception, Peking University, Shenzhen Graduate School, Shenzhen, China.
Email: guoquanwang@stu.pku.edu.cn

Funding information

Science and Technology Plan of Shenzhen, Grant/Award Number: No. JCYJ20200109140410340; National Key R&D Program of China, Grant/Award Number: No. 2020AAA0108904

Abstract

Audio-visual wake word spotting is a challenging multi-modal task that exploits visual information of lip motion patterns to supplement acoustic speech to improve overall detection performance. However, most audio-visual wake word spotting models are only suitable for simple single-speaker scenarios and require high computational complexity. Further development is hindered by complex multi-person scenarios and computational limitations in mobile environments. In this paper, a novel audio-visual model is proposed for on-device multi-person wake word spotting. Firstly, an attention-based audio-visual voice activity detection module is presented, which generates an attention score matrix of audio and visual representations to derive active speaker representation. Secondly, the knowledge distillation method is introduced to transfer knowledge from the large model to the on-device model to control the size of our model. Moreover, a new audio-visual dataset, PKU-KWS, is collected for sentence-level multi-person wake word spotting. Experimental results on the PKU-KWS dataset show that this approach outperforms the previous state-of-the-art methods.

KEYWORDS

audio-visual fusion, human-computer interfacing, speech processing

1 | INTRODUCTION

Wake Word Spotting (WWS) aims to detect pre-registered wake words by classifying utterances into a pre-defined set of words. In recent years, due to the rapid development of artificial intelligence technology, WWS [1–4] is widely used in various fields, such as mobile phone voice assistants [5–9], intelligent robots [10, 11], and smart home devices [12–15]. For example, virtual assistants such as Microsoft's Cortana and Amazon's Alexa [16, 17] rely on specific wake words for activation and further human-computer interaction. Meanwhile, WWS is usually performed on portable devices for a faster response time. Compared to Automatic Speech Recognition (ASR) [12, 14], WWS does not require recognition of the entire input utterance, which significantly reduces the computational cost of unnecessary operations.

Early research studies in the field of WWS typically use audio-only information to spot wake words [1–4, 6–9, 12–14]. In a noise-free audio-only environment, the performance of WWS has far exceeded the level that can be achieved by the human auditory system. Despite audio-only methods [12, 14, 17, 18] commonly used in daily life, the approach that relies exclusively on audio modality faces unfixable drawbacks. For instance, in a variety of sophisticated acoustic scenarios, especially in noisy environments, WWS performance degrades significantly. In addition, factors such as the number of speakers, gender, age, and speaking style also impact the spotting of wake words.

As multi-modal technology has advanced, audio-visual fusion methods [19–22] are considered the most promising solution for robust WWS. The human speech perception system is bimodal and relies on audio and visual information. Therefore, in daily face-to-face communication, observable pronunciation organs such as lips are also important sources of information. Similarly, the speech processing system can also use visual and audio information to improve the performance of WWS in various complex acoustic scenarios. Visual information is especially effective in real-world scenes with severe acoustic distortions, such as strong background noise and sound mixing, since it is not affected by acoustic distortions. For the Audio-Visual Wake Word Spotting (AV-WWS) task, we use lip motion information and audio information to identify the presence or absence of wake words in the visual stream.

In recent decades, researchers have made much exploration in the field of AV-WWS. Traditional methods often take the manual lip motion features and audio spectrum features as the input of the Hidden Markov Model (HMM) [23–25] to achieve audio-visual fusion. However, because these systems rely on hand-crafted elements or rules to evaluate spatial patterns, the accuracy of traditional methods used to develop these systems is far from adequate for real-world applications. It is challenging to modify the hand-crafted models described above, which are created for specific objectives. To address the above issues, AV-WWS based on deep learning methods has attracted extensive attention. An AV-WWS method based on multi-dimensional CNN is proposed in Ref. [26]. In particular, Refs. [27, 28] apply the zero-shot method for in-the-wild videos rather than the methods evaluated with keywords used during training. Deep learning methods can implicitly capture audio and visual features, surpassing the limitations of traditional models, which is the strength of deep neural networks. The proposed audio-visual model achieves superior performance compared to previous methods using hand-crafted parts, with considerably less effort in manual design.

Up until recently, the existing works of AV-WWS have been carried out for single-person scenarios. However, in daily life, multiple people are often present in a human-–computer interaction situation. When multiple people appear in the scene, it is difficult for existing AV-WWS models to accurately identify wake words utilising the Region-Of-Interest (ROI) sequence of the mouth area. As a result, the performance of the AV-WWS models designed for the single-speaker case degrades drastically due to the intractable additional visual redundancy. A multi-task training model [29] is proposed to close the gap between speech activity detection and AV-WWS by a jointly trained model with a multi-task loss. However, the performance of this work is not sufficient for complex multi-person scenarios. To address the WWS problem in multi-person scenarios, we propose an attention-based Audio-Visual Voice Activity Detection (AV-VAD) module. An attention mechanism is introduced to compute the temporal score matrix and detect potential speakers by comparing the audio-visual score results of each frame.

AV-WWS models are commonly equipped on mobile and portable devices. Therefore, the size and inference time of the model is critical. In general, the models trained by current AV-WWS methods are large models whose prediction speed is slow on mobile devices. To this end, the Knowledge Distillation (KD) method is designed to incorporate the temporal knowledge embedded in attention weights of large models to on-device models [30]. In this work, we leverage knowledge distillation [31] to transfer the knowledge from the large model (teacher model) to the on-device model (student model) to achieve the purpose of compressing the model.

The contributions of our work are summarised as follows:

- We propose a novel on-device audio-visual network for the challenging multi-person wake word spotting task.
- We design an audio-visual voice activity detection model for multi-person active feature extraction, which combines the audio and visual representations via an attention module.
- We introduce the knowledge distillation approach to compress the AV-WWS model to meet the on-device demands with low computational complexity.
- A new annotated multi-modal dataset is collected for audio-visual multi-person wake word spotting. The proposed model achieves superior performance compared to the previous state-of-the-art method.

The rest of this paper is organised as follows: Section 2 provides a literature review of the on-device multi-person AV-WWS method and discusses the voice activity detection method and model compressing method in AV-WWS. Section 3 discusses the proposed method in detail, including the visual front-back module, audio front-back module, voice activity detection module, audio-visual fusion module, and knowledge distillation module. Section 4 describes the content and configuration of our newly collected PKU-KWS dataset. Section V shows the implementation details and reports the results of the experimental validation and robustness test of our approach. Finally, conclusions are given in Section 6.

2 | RELATED WORKS

2.1 | Audio-visual model for wake word spotting

Figure 1 shows the general schematic of the representative WWS system based on audio-visual fusion. The audio and visual signals are fed into a two-branch framework for pre-processing, feature extraction, and temporal modelling. In previous AV-WWS works, the manually designed feature extractor is commonly used for visual feature extraction. Recently, deep learning models that directly process the raw image of the speaker's lips have become the preferred method for visual feature extraction [27]. Meanwhile, with respect to speech feature extraction, deep learning models are gradually replacing the traditional Hidden Markov Models (HMM) to obtain more representative features. Temporal modelling is used for wake word location detection, and the obtained local features are enhanced for wake word spotting. Finally, the extracted audio-visual information is fused in the fusion module. The fusion features are passed through a classifier to determine the presence or absence of wake words in the dialog segments.

Ref. [28] is the first to study query-by-text visual KWS for words that are not seen during training, which designs an end-to-end architecture that employs RNN to learn correlations between visual features and keyword representations. Extending on work [28], a CNN-based AV-WWS architecture is proposed in Ref. [27]. Their best audio-visual models can correctly spot wake words under different background noises on the LRS2 [32] dataset, which indicates that the combination of audio and visual modalities enables the model to adapt to the various complex acoustic scenes. It is also proved that the network can be

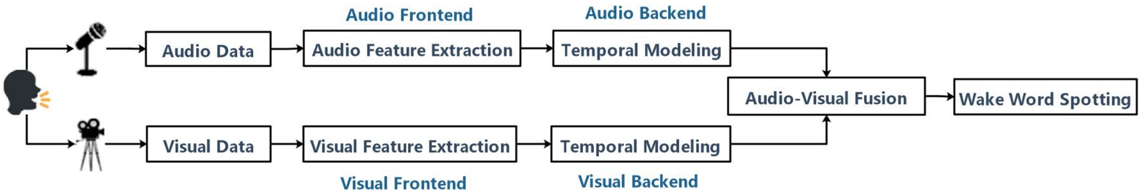


FIGURE 1 Illustration of the representative audio-visual wake word spotting framework. The audio and visual data are pre-processed, feature extracted, and temporal modelled separately, and then integrated with the fusion module. Finally, a classifier is applied to determine the presence or absence of wake words.

extended to languages other than English. An AV-WWS method based on a multi-dimensional Convolutional Neural Network (CNN) is proposed in Ref. [26]. To make full use of the dimensional information in the audio-visual features, a two-dimensional CNN is designed to simultaneously learn the time-frequency features of the logarithmic spectrogram. Meanwhile, the temporal and spatial features of the lip region sequences are fully learned by the 3D CNN operation. The previously mentioned AV-WWS methods do not consider the global dependencies in the features, and it is difficult to solve the WWS of unconstrained long utterances. To this end, we proposed a model based on an enhanced attention mechanism to discover wake words in long sentences without location-specific annotations.

2.2 | Audio-visual model for voice activity detection

To close the gap between the active speaker detection task and the audio-visual speech recognition task, a multi-task training model is proposed in Ref. [29]. It applies a single model to be jointly trained with a multi-task loss. The architecture provides a signal for on-screen speakers without requiring an explicit model. By combining these two tasks in training, the proposed model increases the accuracy of voice activity classification while improving automatic speech recognition performance compared to a multi-person baseline trained specifically for automatic speech recognition. An extension of the work of Refs. [33, 34] proposes a method for active speaker detection using an attention mechanism, which is a soft selection method. After adding the active speaker recognition module, the performance of audio-visual speech recognition improves greatly. Experiments with more than 50,000 h of YouTube public videos as training data show that the system can perform well under various noise conditions. The previous voice activity detection model is mainly learned as an independent task or used as an upstream speech recognition task. In this paper, we design a multi-person AV-VAD module based on the attention mechanism, effectively improving the performance of AV-WWS.

2.3 | Lightweight audio-visual model

Lightweight model design [35] is important in real-world application scenarios where computational resources are limited. In the field of knowledge distillation, some researchers focus on the compression method of the audio-only speech recognition model. An attention distillation method is proposed in Ref. [30], which incorporates the temporal knowledge embedded in attention weights of a large transformer-based model into an on-device model. The MHAtt-RNN [36] and the TC-ResNet14 [37] are considered as their student models. As a result of the experiments, when using the multi-label Experimental results show that the accuracy of MHAtt-RNN and TC-ResNet14 improves when tested with multi-label datasets, which displays the temporal knowledge from large teacher model is transferred to the on-device student model for audio classification. With the widespread use of multi-modal models, researchers have focused more on designing lightweight audio-visual fusion models. Ref. [21] proposes a compact audio-visual WWS system by introducing a neural network pruning strategy via the lottery ticket hypothesis in an iterative fine-tuning manner. The proposed audio-visual system achieves significant performance in an indoor corpus that is collected in a home TV scenario. The method introduced above is either to distill the knowledge of the audio-only model or to compress the AV-WWS model by pruning. In this paper, we propose a method to distill and compress the knowledge of the AV-WWS model. The student model after distillation not only has a faster response time but also better noise resistance.

3 | THE PROPOSED METHOD

In this paper, we propose an audio-visual fusion network to determine the active speaker in multi-person scenes and detect wake words in unconstrained utterances. The overview of the proposed AV-WWS architecture is illustrated in Figure 2. The architecture consists of the following parts: a visual stream to process the visual signals, an audio stream to process the speech signals, the AV-VAD module to choose the active speaker, a fusion module to fuse the audio and visual representations, a classifier module to predict the wake word, and a KD structure to compress the AV-WWS model. The model takes mouth ROI sequences and speech sequences as the input signals of the front-back processing module, which includes a frontend block for feature extraction and a backend block for temporal modelling. The modified ResNet-18 [38] network pre-trained on LRW [39] is used for visual frontend feature extraction. The wav2vec [40] model pre-trained on Librispeech [41] is used for audio feature extraction. The audio time-frequency feature is fed to an attention module along with the visual feature for active

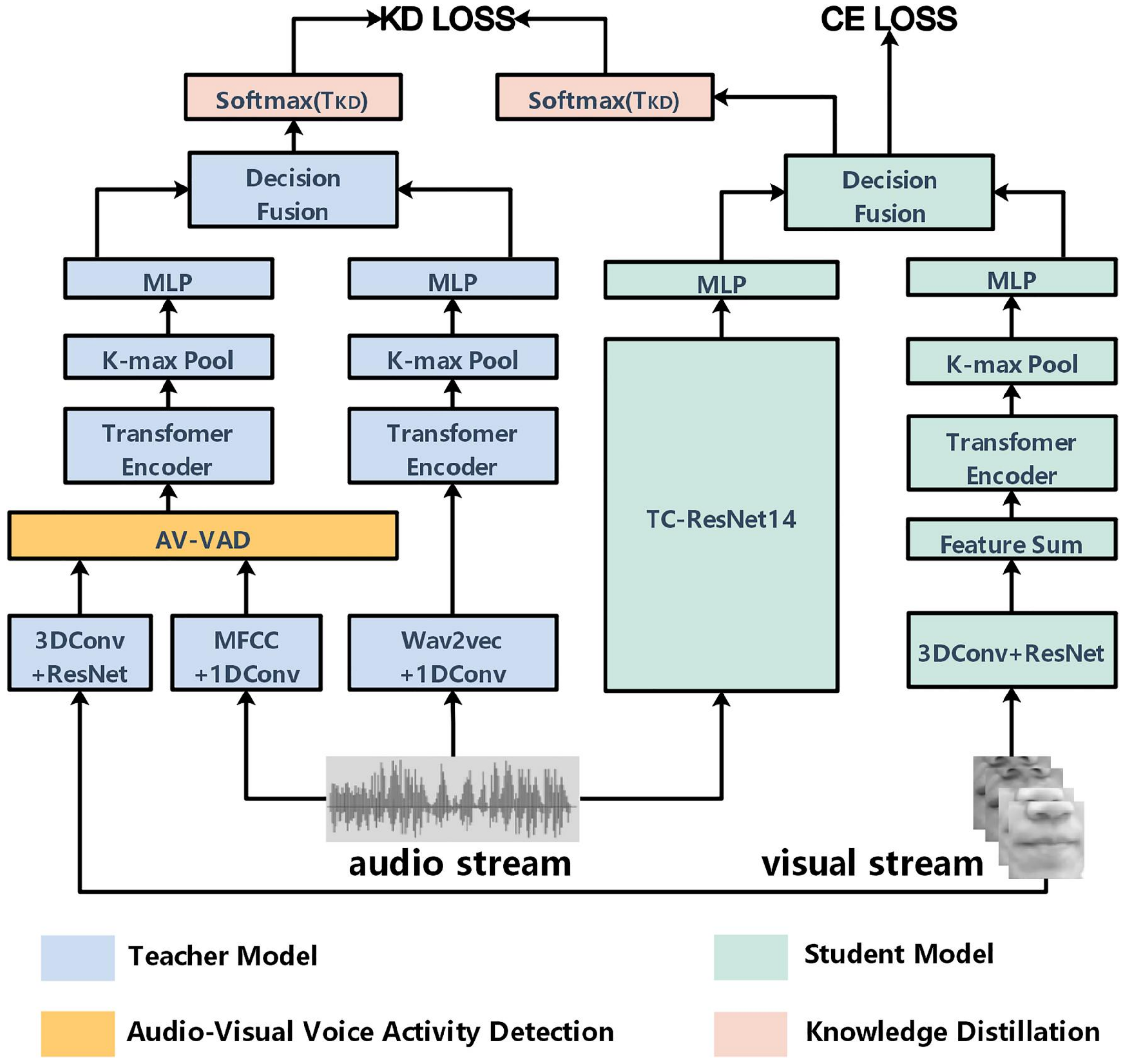


FIGURE 2 Illustration of the proposed on-device audio-visual multi-person wake word spotting network, including a visual stream, an audio stream, an AV-VAD module, a fusion module, and a classifier module. In addition, the KD structure is utilised to compress the AV-WWS model.

speaker detection. Then, a transformer encoder is applied to model and enhance the feature of each stream. Finally, the information of different streams is fused to estimate the posterior probability of each wake word. Moreover, in order to transfer knowledge from the teacher model to the student model, we employ the KD method to compress the model.

3.1 | Visual stream

The frontend architecture of the visual modality is used to capture lip motion, and its output representation reflects lip position differences. The architecture is similar to the common encoder in Ref. [32]. In order to capture the informative lip position differences between frames, a 3D convolutional layer is used to extract the temporal and spatial information of lip motion. Then, a modified ResNet-18 is applied to extract the internal information of each frame. In the visual backend, a stack of four layers of transformer encoder is employed to model the mouth ROI sequences in the time dimension. Through the visual feature extraction module, the input greyscale image of each frame is converted into a feature vector with a size of 512. Table 1 shows detailed information about the network architecture of the visual stream.

3.2 | Audio stream

The unsupervised model wav2vec [40] pre-trained on the large speech dataset Librispeech [41] is employed in our audio frontend module, as it is normally used for feature extraction of WWS tasks. In the audio backend, a stack of four layers of transformer encoder is employed to model the audio feature sequence in the time dimension. Through the feature extraction of audio modality, the original audio signal is converted into a feature vector with a size of 512. Table 2 shows detailed information about the audio stream, where T_S is the length of the original audio sequence.

TABLE 1 Detail information about the network architecture of the visual stream including kernel size, stride, and padding size

Stage	Layer name		Output size	Detail
Visual frontend	Input layer		$T_V \times 112 \times 112$	
	Conv3d		$T_V \times 28 \times 28 \times 64$	Kernel = 5 × 7 × 7, Stride = 1 × 2 × 2, Padding = 2 × 3 × 3, max pool
	Conv2d-1	ResNet18	$T_V \times 28 \times 28 \times 64$	Kernel = 3 × 3, Stride = 1 × 1
	Conv2d-2		$T_V \times 14 \times 14 \times 128$	Kernel = 3 × 3, Stride = 2 × 2
	Conv2d-3		$T_V \times 7 \times 7 \times 256$	Kernel = 3 × 3, Stride = 2 × 2
	Conv2d-4		$T_V \times 4 \times 4 \times 512$	Kernel = 3 × 3, Stride = 2 × 2
	Average pool		$T_V \times 1 \times 1 \times 512$	Kernel = 4 × 4, Stride = 1 × 1
	Out-layer		$T_V \times 512$	
Visual backend	Conv1d		$T_V \times 512$	Kernel = 1 × 1, Stride = 1 × 1
	Transformer encoder		$T_V \times 512$	Hidden Dim = 1024, layer num = 4

TABLE 2 Detail information about the network architecture of the audio stream including kernel size, stride, and padding size

Stage	Layer name	Output size	Detail
Audio frontend	Input layer	$T_S \times 1$	
	Wav2vec	$T_A \times 512$	Pretrained on librispeech
Audio backend	Padding	$2T_V \times 512$	Padding Number = 0
	Conv1d	$T_V \times 512$	Kernel = 2, Stride = 2
	Transformer encoder	$T_V \times 512$	Hidden Dim = 1024, layer num = 4

Note: The bolded values show the best experimental results.

3.3 | Audio-visual voice activity detection

The AV-VAD module is designed to select the active speaker when multiple speakers appear on the screen at the same time. The inputs of the AV-VAD module include visual and audio features, and the attention scores they generate are used to determine who is speaking. Here, we take two people who appear on the screen simultaneously as an example. The network architecture of the AV-VAD module is illustrated in Figure 3. Firstly, the network extracts the audio representation $A \in \mathbb{R}^{T_A \times D_A}$ and the visual representation, respectively, and then stacks the extracted visual representations on a new dimension to derive $V \in \mathbb{R}^{N \times T_V \times D_V}$. N denotes the number of people in the scene, and $N = 2$ in this example. The visual representation V and audio representation A are multiplied with a weighted feature matrix to convert them into the new feature matrix $K \in \mathbb{R}^{N \times T_V \times D_K}$ and $Q \in \mathbb{R}^{T_A \times D_Q}$, respectively. In addition, the audio representation is expanded in dimension through the broadcast mechanism to ensure the same size as the visual dimension. Then, the matrix multiplication operation is performed on the two representations to derive the time-score matrix $S \in \mathbb{R}^{N \times T_A \times T_V}$. It is formulated as follow:

$$S_{nt_at_v} = Q_{nt_aq} W_{qk} K_{nkt_v}, \quad S \in \mathbb{R}^{N \times T_A \times T_V}, \tag{1}$$

where $W \in \mathbb{R}^{D_Q \times D_K}$ is a matrix used to unify the feature dimensions of K and Q. Adding the time score matrix S at the audio time dimension, the score matrix is derived as $S' \in \mathbb{R}^{N \times T_V}$. It estimates the correlation of each frame of the visual representation with the overall audio sequences. We perform softmax over the N dimension of S' to derive the normalised score matrix, which is formulated as follows:

$$S'_{nt_v} = \sum_{t_a} S_{nt_at_v}, \quad S' \in \mathbb{R}^{N \times T_V}, \tag{2}$$

$$\alpha_{nt_v} = \frac{e^{S'_{nt_v}}}{\sum_n e^{S'_{nt_v}}}, \quad \alpha \in \mathbb{R}^{N \times T_V}. \tag{3}$$

We calculate the average score on the visual dimension, where the index of the maximum value $N_{\max}$ is derived. The index is used to select the input visual representation $V' \in \mathbb{R}^{T_V \times D_V}$. The max index is calculated as follows:

$$N_{\max} = \underset{N}{argmax} \frac{1}{T_V} \sum_{t_v} \alpha_{nt_v}, \quad N_{\max} \in 1, 2, \ldots, N, \tag{4}$$

where $N_{\max}$ is used to derive the active visual representation V' from the total visual inputs V. The AV-VAD module is regarded as a classification network, which is trained with the Cross-Entropy loss.

3.4 | Audio-visual fusion and classifier

LayerNorm [42] is applied to the feature dimension of each modality before fusion to balance the multi-modal

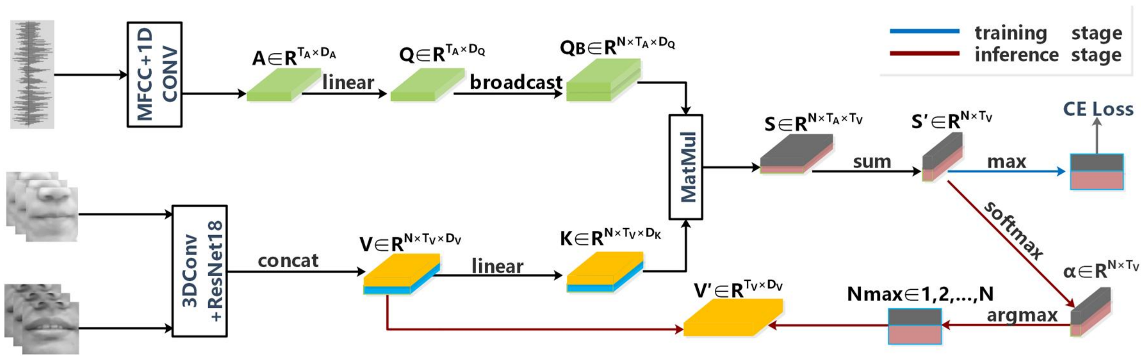


FIGURE 3 Network architecture of the proposed audio-visual voice activity detection module. The colours of the arrows indicate different stages: blue for training, red for inference, and black for both training and inference stages.

representations and avoid one modality covering the whole representation with a large variance. Decision-level audio-visual fusion is used in this work. The fusion weights of audio and visual modalities are the same and set to 0.5.

The classifier module includes a K-max pooling layer and a fully connected layer. In the K-max pooling layer, the feature vectors are summed along the channel dimension, and then the top K-frames with large values in the time dimension are selected as the input to the fully connected layer. For long-term unconstrained input utterances, the wake words appear only in local segments of the sequence. Thus, K-max pooling in the time dimension avoids introducing unnecessary redundant information by sampling in the time dimension. In the fully connected layer, the AV-WWS network is compressed into five output units, representing the number of wake words. The output represents the posterior probability of each wake word. The Cross-Entropy loss is applied for training the network.

3.5 | Knowledge distillation

Figure 2 shows the teacher model, the student model, and the knowledge distillation process. The front-back processing module of the student visual stream is similar to the teacher visual stream. Differently, the number of feature output channels of the 3D convolutional layer changes from 64 to 16, and the number of output channels of the 2D convolutional layer changes from [64,128,256,512] to [16,32,64,128]. In the audio feature extraction module, we employ a lightweight CNN-based model TC-ResNet [37], where the backend input size of the student model changes from 512 to 128. In the training stage, we take the one-hot vector corresponding to the real label as the hard label and use the distillation temperature T_{KD} in a softmax function. Softening the hard label reduces the gradient difference between the training examples of the model to provide more information for the student model.

$$q_i = \frac{\exp\left(\frac{z_i}{T_{KD}}\right)}{\sum_j \exp\left(\frac{z_j}{T_{KD}}\right)}, \tag{5}$$

where q_i is the ith soft target and z_i represents the ith hard target. In the training process of the student model, the Cross-Entropy loss and the Kullback-Leibler loss are applied. In the evaluation process of the student model, we only use Cross-Entropy loss. The training loss function of the student network model is defined as follows:

$$\mathcal{L}_{student} = \lambda \mathcal{L}_{CE} + (1 - \lambda)\mathcal{L}_{KL}, \tag{6}$$

where λ is a hyperparameter to control the importance of each loss item.

4 | THE PKU-KWS DATASET

Most of the audio-visual speech datasets [32, 43] are focused on speech recognition tasks. The datasets for speech recognition lack wake word labels. However, WWS datasets with wake word labels are usually word level [39, 44] rather than sentence and dialog. In addition, these datasets are recorded in simple single-speaker scenarios. To address the limitations of these datasets, we collected a sentence-level audio-visual wake word spotting dataset, including multi-person dialogs in a supermarket scenario, called PKU-KWS https://zenodo.org/record/6792058, which is publicly available for academic research. The dataset is collected in a relatively quiet acoustic environment with a colour camera recorded at 25 frames per second (fps). The video resolution is 1080 × 1920. The audio is synchronously recorded at the sampling rate of 16,000 Hz with 16 bits for each sampling. We define five wake words commonly used in supermarket shopping ('Pengpeng', 'Nihao', 'Xiexie', 'Dazhe', 'Jiezhang'). These words are used to wake up service robots, ask for checkout, inquire about discounts at supermarkets, etc. Unlike other datasets, the PKU-KWS dataset is the first audio-visual dataset of wake word spotting based on Mandarin Chinese in a multi-person environment. The dataset contains 500 single-speaker conversations, 300 double-speaker conversations, and 200 three-speaker conversations. The content of each conversation is a common phrase in everyday life, and there is no limit to its duration or sentence length. Conversations may or may not contain wake words.

The above scenarios are set closer to practical situations, bring more difficulty to AV-WWS.

Moreover, in the process of data augmentation, the supermarket background sound in the TAU-urban-acoustic-scenes-2020-mobile-development dataset https://doi.org/10.5281/zenodo.3685828 is used as added noise. The sampling frequency of noise audio is 22,500 Hz, and it is downsampled to the audio settings in the PKU-KWS. The noise signal is added to the original speech signals with different Signal-to-Noise Ratios (SNRs) to evaluate the robustness of the algorithm.

5 | EXPERIMENTS AND DISCUSSIONS

5.1 | Implementation details

In the pre-processing step, we use dlib [45] to detect and track 68 facial landmarks for each video. The features are normalised according to the overall mean and variance. Then, a bounding box of 112 × 112 is set to crop the mouth ROI of each person and stack them according to the face position in the video. The size of the mouth ROI sequence is $N \times 112 \times 112$, where N represents the number of faces on the screen. To avoid the influence of colour difference information on the model, we grey the obtained colour image and then splice the output of each frame to obtain the preprocessed mouth ROI sequence. For each raw audio, the features are normalised according to the overall mean and variance. During the training process, instead of setting all sentences to a uniform duration, we standardise the sequence duration in each mini-batch by padding.

In the feature extraction step, we use the pre-trained model to extract the front-back features of the teacher network. In the audio stream, the unsupervised learning model wav2vec [40] pre-trained on the Librispeech [41] dataset is set for the audio frontend feature extraction. In the visual stream, the pre-trained model on the LRW dataset [38] is employed to extract the frontend features for AV-VAD and AV-WWS. In the classifier module, the parameter of the K-max Pooling layer is set to $K = 25$. The Dropout [46] with a probability of 0.1 is used to reduce the overfitting caused by the convergence of different modalities at different speeds.

In the process of knowledge distillation, we only learn the audio and visual backend, feature fusion, and classifier parameters of the teacher model. In the training process, the Cross-Entropy loss is applied for the teacher model, and the loss function defined in Eq. (6) is applied for the student model. Moreover, a series of experiments are carried out on the hyperparameters selection of the knowledge distillation. The experimental results are shown in Figure 4. In the training process of the student model, the hyperparameter λ is set to 0.5, and the distillation temperature T_{KD} is set to 3 to make the student model perform better.

Our implementation is based on the Pytorch https://pytorch.org library and trained on NVIDIA GeForce GTX1660 Ti GPU with 6 GB memory. The batch size of all experiments is set to 8. The network is trained using the Adam optimiser [47] with decay factors $\beta_1 = 0.9$, $\beta_2 = 0.999$, and the initial learning rate is 0.0001. The dataset is split with a ratio of 8:2 for training and testing, respectively.

5.2 | Evaluation metrics

We evaluate the performance in terms of the Accuracy (ACC), Receiver Operating Characteristic (ROC) Curve, Area Under Curve (AUC), and Equal Error Rate (EER), which are commonly used in WWS tasks. ACC is the most intuitive metric to evaluate the WWS model, which represents the proportion of positive cases to the total samples. The AUC is the area of the curve enclosed by the horizontal coordinate of False Positives (FP) and the vertical coordinate of True Positives (TP). EER is the value when False Reject (FR) and False Alarm (FA) are equal. The lower the EER value, the higher the accuracy of the model. The metrics are also applied in papers [48–50]. To extend the AUC to the multi-class task, we compute the micro-averaging and macro-averaging for multiple classes. The macro-

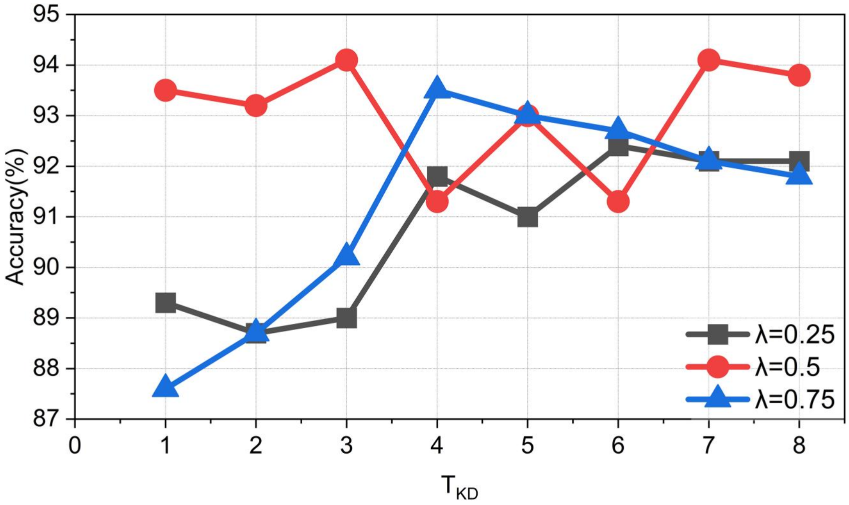


FIGURE 4 Hyperparameters selection of KD, where T_{KD} is the distillation temperature to soft labels and λ is to balance the loss of the student model.

averaging means that the metrics are calculated within each class, and the resulting metrics are averaged across classes. The micro-averaging method aggregates outcomes across all classes and then compute metric with aggregate outcomes. In the experiment results, AUC stands for micro-averaging AUC if not otherwise specified.

5.3 | Baseline model

The isolated word recognition based on the audio-visual fusion model described in Ref. [51] is employed as the baseline model. Differently, our baseline model takes the Mel Frequency Spectrum Coefficient (MFCC) representations of the audio as an input of the audio stream. In the audio frontend feature extraction module, a 1D convolutional layer is applied to replace the modified ResNet-18 network.

The strategies of decision fusion and feature fusion are commonly used for multi-modal fusion models. We compare the two fusion strategies on the baseline model. The architecture of the decision fusion method and feature fusion method is shown in Figure 5. The experimental results in Table 3 show that the absolute accuracy of the AV-WWS model using decision fusion is 4.2% higher than that using feature fusion, and its EER and AUC are also significantly higher. Through the analysis of the experimental results, the reason why the recognition performance of AV-WWS based on decision fusion is better is that in the process of feature fusion the model needs to conduct secondary temporal modelling of the fused information through the attention mechanism. However, decision fusion only needs to consider the results of the final classifier. In the case of limited data, the performance is more robust using the decision fusion-based WWS model. Therefore, in the subsequent experiments, decision fusion is chosen as the fusion method for our AV-WWS model.

5.4 | Experimental results

In this section, we evaluate the effectiveness of each component in an ablation experiment. The proposed teacher and student models are compared with the uni-modal method and state-of-the-art audio-visual methods. Finally, we evaluate the robustness of our AV-WWS model and each module under different noise conditions.

Ablation Study. Results of the ablation study on the PKU-KWS dataset are shown in Table 4. The audio frontend is initialised with a model pre-trained on Librispeech [41] instead of training a modified ResNet-18 [39] audio frontend. In the backend module, a further absolute gain of 0.56% is obtained by replacing the Bi-GRUs module with a stacked 4-layer transformer encoder, which implies the advantage of the attention mechanism. Additionally, K-max Pooling further enhances the features to extract effective information related to the wake word from the high-dimensional redundant information, which results in an absolute increase of 1.21%. In particular, an improvement of 1.89% is obtained by using the VAD module, showing the necessary role of the proposed VAD module in the multi-person task.

TABLE 3 Comparison of Baseline Models based on the decision fusion method and feature fusion method

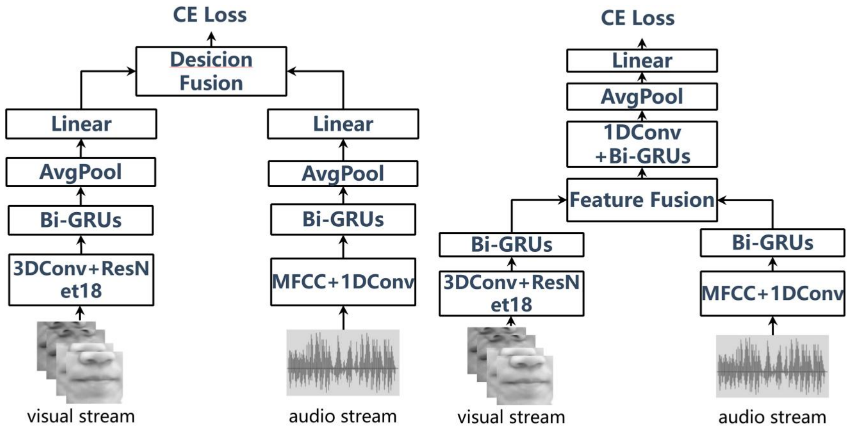
Fusion method	Param.(M) ↓	ACC (%) ↑	EER (%) ↓	AUC (%) ↑
Decision fusion	**2.39**	**95.52**	**2.53**	**99.68**
Feature fusion	2.92	91.32	3.94	99.29

Note: The bolded values show the best experimental results.

TABLE 4 Ablation Study of the proposed AV-WWS model on the PKU-KWS dataset (The last column lists the absolute improvement relative to the baseline model.)

Method	ACC (%) ↑	
Baseline [51]	95.52	
+ Wav2vec frontend	96.06	+0.54
+ transformer encoder backend	96.62	+1.1
+ K-max pooling	97.83	+2.31
+ voice activity detection	**99.72**	**+4.2**

Note: The bolded values show the best experimental results.



（a） decision fusion （b） feature fusion

FIGURE 5 Architecture of the baseline model with the (a) decision fusion method and (b) feature fusion method.

Comparison with Audio-Only Method. We compare the proposed Audio-Visual (AV) method with the Audio-Only (AO) method under the noiseless condition in Table 5. The AV method achieves better results than the AO method, especially with the VAD module. Further, we compare the AO and AV methods under different noise conditions in terms of more metrics in Figure 6. For a fair comparison, the AV-WWS model is not equipped with VAD and KD modules. As shown in Figure 6, the difference in performance between the AO and AV models is 9.28% ACC at the noise condition of -5 dB. This indicates that lip motion from the visual modality plays a significant role in the low SNR acoustic environment. Similarly, the AV model is more robust in WWS by comparison to other evaluation metrics, demonstrating the effectiveness of audio-visual fusion.

Comparison with State-of-the-Art Methods. Our method is compared with the previous state-of-the-art method MCNN [26]. Unlike word-level works, our work is for more complex unconstrained sentence-level wake words. For the fairness of comparison, we fine-tune the feature extraction network of MCNN proposed in Ref. [26] and add the K-max pooling layer to the backend of MCNN. As shown in Table 6, compared with the modified MCNN, our teacher model has higher ACC (99.72% vs. 87.57%), lower EER (0.28% vs. 6.49%), and higher AUC (99.99% vs. 98.25%). In addition, our student model also performs better than MCNN. In the pre-processing process, we use the same ROI extraction method as MCNN. Therefore, we only consider the number of parameters of the WWS model. MCNN has fewer parameters because it is designed for word-level WWS and thus does not need to deal with sequence data in long sentences, which makes it perform poorly in real dialog scenarios.

TABLE 5 Comparison of Audio-Visual Method with Audio-Only Method under noiseless condition (with/without VAD or KD Module)

Modality	VAD	KD	ACC (%) ↑	EER (%) ↓	AUC (%) ↑
Audio-only	×	×	97.46	1.12	99.79
Audio-visual	×	×	97.83	1.19	99.94
	✓	×	**99.72**	**0.28**	**99.99**
	✓	✓	94.08	2.81	99.38

Note: The bolded values show the best experimental results.

Knowledge Distillation. We evaluated the effectiveness of the KD module in Table 6. After the KD process, the parameters of the model decreased significantly, and the parameters of the student model are 7.53 M less than the teacher model. The reduction in accuracy is very limited compared to the compression of parameters. Figure 6 shows the performance of the teacher model (VAD + AV-WWS) and the student model (KD + VAD + AV-WWS) under different noise conditions. Although the performance of the student model is lower than the teacher model when the audio signal is clean, the ACC of the student and teacher models reaches 91.27% and 85.07% when the SNR level is 0 dB. This indicates that the performance of the student model is better than the teacher model in some high-noise situations. With limited resources, the compressed and effective knowledge has been successfully transferred from the large teacher model to the smaller student model on the device.

Robustness Test. We test the proposed method under different noise conditions in terms of ROC curves and report the individual categories and the total results in Figure 7. Unlike other experiments using babble noise to evaluate the robustness of the model, we test the model in a more challenging supermarket background noise environment. As shown in Figure 7a, the detection performance of the first wake word in the five categories is lower than the other four categories under all noise

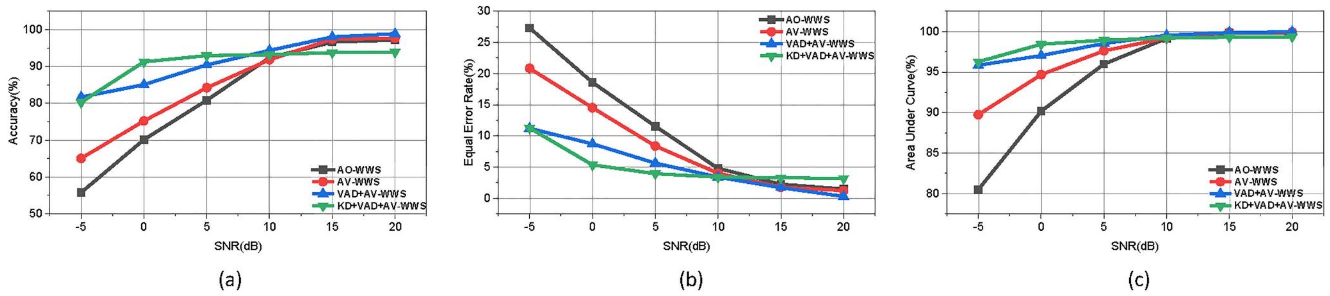
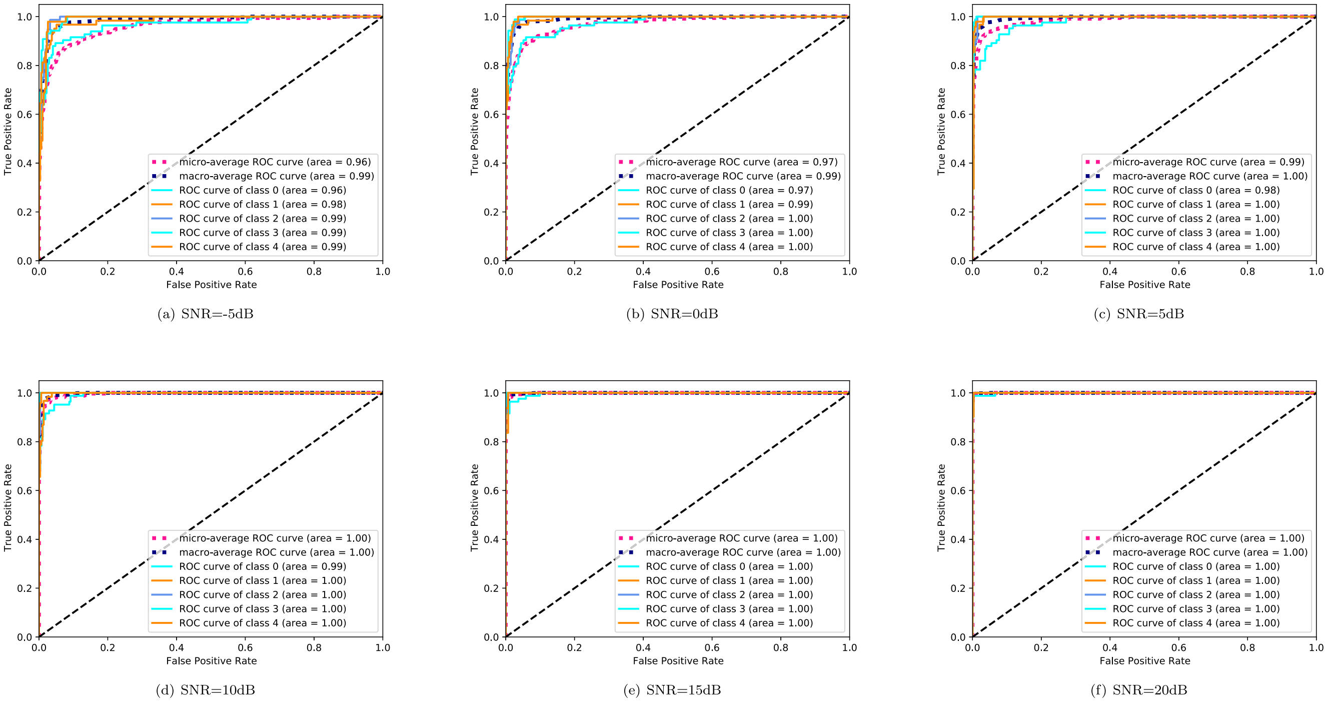


FIGURE 6 Performance under different noise conditions in terms of (a) Accuracy, (b) Equal Error Rate, and (c) Area Under Curve. AO-WWS: audio-only model, AV-WWS: audio-visual model, VAD + AV-WWS: teacher model, KD + VAD + AV-WWS: student model.

TABLE 6 Performance and parameters comparison with state-of-the-art methods

Method	Param. (M) ↓	ACC (%) ↑	EER (%) ↓	AUC (%) ↑
Teacher model	9.45	**99.72**	**0.28**	**99.99**
Student model with KD	1.92	94.08	2.81	99.38
Student model without KD	1.92	90.42	4.22	99.17
MCNN [26]	**0.83**	87.57	6.49	98.25



(a) SNR=-5dB

(b) SNR=0dB

(c) SNR=5dB

(d) SNR=10dB

(e) SNR=15dB

(f) SNR=20dB

FIGURE 7 ROC curves for each category and all categories under different noise conditions. Micro-averaging and macro-averaging calculate the overall performance of multiple categories using two statistical methods. The larger the area under the ROC curve, the better the reliability of the classification.

conditions, exhibiting a lower ROC curve. The first wake word is a reduplicated word "Pengpeng", which is pronounced continuously or quickly by speakers in daily use, and different speakers have different pronunciation habits for reduplicated words, resulting in a lower recognition accuracy than other words. We further test the robustness of the VAD module under different noise conditions, and the results are shown in Figure 6. The VAD + AV-WWS model (teacher model) results in an absolute improvement of 16.64%, 6.1%, and 9.64% over the AV-WWS model in terms of ACC, EER, and AUC under high noise levels (-5 dB), respectively. Because the VAD module greatly reduces the redundancy of visual information and thus performs much better in multi-person scenarios, especially in low SNR acoustic environments.

5.5 | Visualisation results

Figure 8 shows the visualisation of the proposed AV-WWS model. Here, take the case of three people appearing on the screen at the same time as an example. The top left shows the input visual stream. The upper right shows the detection result of the whole model. The bottom left shows the extraction result of the face lips. Due to the limited display space, we only show a few frames of the lip image. The lower right shows the result of the selection of the speaker's face obtained by the attention-based AV-VAD module. In the AV-VAD results, the figure shows the active speaker score for each frame, where different speakers are represented by different colours. As shown in the detection result (green bounding box) of the model in Figure 8a, the active speaker *p*2 in the video is speaking. Meanwhile, the corresponding AV-VAD results show that the green marker representing *p*2 fills most of the time frames. The visualisation demonstrates the accuracy of the proposed active speaker detection module. The performance of AV-WWS can be significantly improved by reducing unnecessary redundancy by removing visual features of non-speakers. As seen from the AV-WWS results, the posterior probability of the correct category is considerably higher than other categories, showing the effectiveness of the proposed model.

6 | CONCLUSION

In this paper, we propose a novel audio-visual model for the challenging multi-person wake word spotting task. We design an attention-based audio-visual voice activity detection module to reduce redundant visual information from irrelevant persons in multi-person scenes. We introduce the knowledge distillation method to compress the model parameters to meet the low computational complexity requirements on the on-device level. The compressed student model maintains a competitive detection accuracy rate while significantly reducing parameters. In noisy environments, the audio-visual fusion method can significantly improve the performance of WWS with the benefit of the lip information provided by the visual stream. Unlike word-level WWS, our approach achieves accurate detection in unconstrained utterances or conversations,

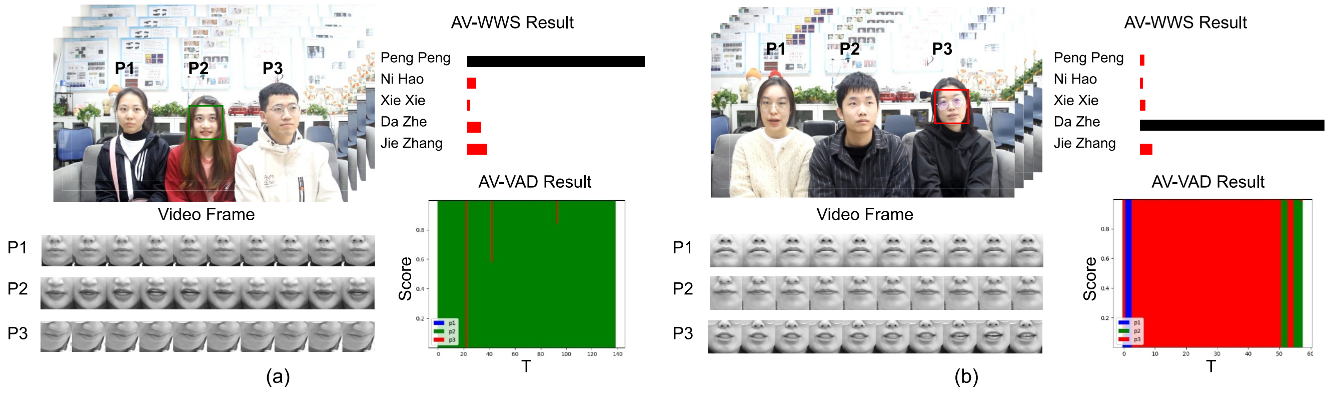


FIGURE 8 Visualisation of the proposed AV-WWS model. In each sample, the top left shows the input visual stream. The upper right shows the detection results. The bottom left shows the extracted lip sequence. The lower right shows the AV-VAD result, where different speakers are represented by different colours.

which is closer to practical application scenarios. Future work will focus on large-scale wake word spotting, rather than recognising a limited number of words.

ACKNOWLEDGEMENT

This work is supported by the National Key R&D Program of China (No. 2020AAA0108904), and the Science and Technology Plan of Shenzhen (No. JCYJ20200109140410340).

CONFLICT OF INTEREST

The author declares that there is no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available at https://zenodo.org/record/6792058. All data included in this study are available upon request by contact with the corresponding author.

ORCID

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

REFERENCES

1. Tsai, T.-H., Hao, P.-C.: Customized wake-up word with key word spotting using convolutional neural network. In: International SoC Design Conference, pp. 136–137 (2019)
2. Zhang, S., Liu, W., Qin, Y.: Wake-up-word spotting using end-to-end deep neural network system. In: International Conference on Pattern Recognition, pp. 2878–2883 (2016)
3. Gao, Y., et al.: Towards data-efficient modeling for wake word spotting. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 7479–7483 (2020)
4. Gao, Y., et al.: On front-end gain invariant modeling for wake word spotting. arXiv preprint arXiv:2010.06676 (2020)
5. López-Espejo, I., et al.: Deep spoken keyword spotting: an overview. IEEE Access, 4169–4199 (2021)
6. Peter, D., Roth, W., Pernkopf, F.: End-to-end keyword spotting using neural architecture search and quantization. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 3423–3427 (2022)
7. Kim, K., et al.: A 23μw solar-powered keyword-spotting asic with ring-oscillator-based time-domain feature extraction. In: IEEE International Solid-State Circuits Conference, vol. 65, pp. 1–3 (2022)
8. Liu, L., et al.: Keyword spotting techniques to improve the recognition accuracy of user-defined keywords. Neural Netw. 139, 237–245 (2021). https://doi.org/10.1016/j.neunet.2021.03.012
9. Kim, B., et al.: Broadcasted residual learning for efficient keyword spotting. arXiv preprint arXiv:2106.04140 (2021)
10. Yan, Y., et al.: Low-power low-latency keyword spotting and adaptive control with a spinnaker 2 prototype and comparison with Loihi. arXiv preprint arXiv:2009.08921 (2020)
11. Shree, V., et al.: Learning to assess danger from movies for cooperative escape planning in hazardous environments. arXiv preprint arXiv:2207.13791 (2022)
12. Nouza, J., Červa, P., Žd'ánskỳ, J.: Very fast keyword spotting system with real time factor below 0.01. In: International Conference on Text, Speech, and Dialogue, pp. 426–436 (2020)
13. Retsinas, G., et al.: An alternative deep feature approach to line level keyword spotting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12658–12666 (2019)
14. Wang, X., Sun, S., Xie, L.: Virtual adversarial training for DS-CNN based small-footprint keyword spotting. In: IEEE Automatic Speech Recognition and Understanding Workshop, pp. 607–612 (2019)
15. Zeng, M., Xiao, N.: Effective combination of densenet and bilstm for keyword spotting. IEEE Access 7, 10767–10775 (2019). https://doi.org/10.1109/access.2019.2891838
16. Leroy, D., et al.: Federated learning for keyword spotting. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 6341–6345 (2019)
17. López-Espejo, I., Tan, Z.-H., Jensen, J.: Keyword spotting for hearing assistive devices robust to external speakers. arXiv preprint arXiv:1906.09417 (2019)
18. Liu, H., et al.: Binaural sound source localization based on weighted template matching. CAAI Trans. Intellig. Technol. 6(2), 214–223 (2021). https://doi.org/10.1049/cit2.12009
19. Cheng, M., et al.: The dku audio-visual wake word spotting system for the 2021 misp challenge. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 9256–9260 (2022)
20. Xu, Y., et al.: Audio-visual wake word spotting system for MISP challenge 2021. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 9246–9250 (2022)
21. Zhou, H., et al.: A study of designing compact audio-visual wake word spotting system based on iterative fine-tuning in neural network pruning. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 7572–7576 (2022)
22. Li, Y., Liu, H., Tang, H.: Multi-modal perception attention network with self-supervised learning for audio-visual speaker tracking. Proc. AAAI Conf. Artif. Intell. 36(2), 1456–1463 (2022). https://doi.org/10.1609/aaai.v36i2.20035

23. 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 (1990)
24. Shrivastava, A., et al.: Optimize what matters: training DNN-HMM keyword spotting model using end metric. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4000–4004 (2021)
25. 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
26. Ding, R., Pang, C., Liu, H.: Audio-visual keyword spotting based on multidimensional convolutional neural network. In: IEEE International Conference on Image Processing, pp. 4138–4142 (2018)
27. Momeni, L., et al.: Seeing wake words: audio-visual keyword spotting. arXiv preprint arXiv:2009.01225 (2020)
28. Stafylakis, T., Tzimiropoulos, G.: Zero-shot keyword spotting for visual speech recognition in-the-wild. In: Proceedings of the European Conference on Computer Vision (2018)
29. Braga, O., Siohan, O.: Best of both worlds: multi-task audio-visual automatic speech recognition and active speaker detection. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 6047–6051 (2022)
30. Choi, K., et al.: Temporal knowledge distillation for on-device audio classification. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 486–490 (2022)
31. Hinton, G. et al.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531, vol. 2, no. (7), (2015)
32. Afouras, T., et al.: Deep audio-visual speech recognition. IEEE Trans. Pattern Anal. Mach. Intellig. (2018)
33. Braga, O., et al.: End-to-end multi-person audio/visual automatic speech recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 6994–6998 (2020)
34. Braga, O., Siohan, O.: A closer look at audio-visual multi-person speech recognition and active speaker selection. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 6863–6867 (2021)
35. Li, X., Li, S.: Transformer help CNN see better: a lightweight hybrid apple disease identification model based on transformers. Agriculture 12(6), 884 (2022). https://doi.org/10.3390/agriculture12060884
36. Rybakov, O., et al.: Streaming keyword spotting on mobile devices. arXiv preprint arXiv:2005.06720 (2020)
37. Choi, S., et al.: Temporal convolution for real-time keyword spotting on mobile devices. arXiv preprint arXiv:1904.03814 (2019)
38. Stafylakis, T., Tzimiropoulos, G.: Combining residual networks with LSTMs for lipreading. Interspeech 2017 (2017)
39. Chung, J.S., Zisserman, A.: Lip Reading in the Wild. Springer, Cham (2016)
40. Schneider, S., et al.: wav2vec: unsupervised pre-training for speech recognition. arXiv preprint arXiv:1904.05862 (2019)
41. Panayotov, V., et al.: Librispeech: an ASR corpus based on public domain audio books. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 5206–5210 (2015)
42. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
43. Afouras, T., Chung, J.S., Zisserman, A.: Lrs3-ted: a large-scale dataset for visual speech recognition. arXiv preprint arXiv:1809.00496 (2018)
44. Warden, P.: Speech commands: a dataset for limited-vocabulary speech recognition. arXiv preprint arXiv:1804.03209 (2018)
45. King, D.E.: Dlib-ml: a machine learning toolkit. J. Mach. Learn. Res. 10, 1755–1758 (2009)
46. Srivastava, N., et al.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15(1), 1929–1958 (2014)
47. Kingma, D.P. and Ba, J.: Adam: a method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014)
48. Kumar, R., Yeruva, V., Ganapathy, S.: On convolutional lstm modeling for joint wake-word detection and text dependent speaker verification. In: Interspeech, pp. 1121–1125 (2018)
49. Zhao, Z., Zhang, W.-Q.: End-to-end keyword search system based on attention mechanism and energy scorer for low resource languages. Neural Netw. 139, 326–334 (2021). https://doi.org/10.1016/j.neunet.2021.04.002
50. Snyder, D., et al.: Speaker recognition for multi-speaker conversations using x-vectors. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 5796–5800 (2019)
51. Petridis, S., et al.: End-to-end audiovisual speech recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 6548–6552 (2018)

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.: On-device audio-visual multi-person wake word spotting. CAAI Trans. Intell. Technol. 8(4), 1578–1589 (2023). https://doi.org/10.1049/cit2.12189