Audio-Visual Multi-person Keyword Spotting via Hybrid Fusion

Yuxin Su[1,2], Ziling Miao[1], and Hong Liu[1](✉)

[1] Key Laboratory of Machine Perception, Peking University, Shenzhen Graduate School, Shenzhen, China
yuxinsu@stu.pku.edu.cn, {zilingmiao,hongliu}@pku.edu.cn
[2] Xidian University, Xian, China

Abstract. As an important research method for speech recognition tasks, audio-visual fusion has achieved good performances in improving the robustness of keyword spotting (KWS) models, especially in a noisy environment. However, most related studies are implemented under the single-person scenarios, while ignoring the application in multi-person scenarios. In this work, an audio-visual model using the hybrid fusion is proposed for multi-person KWS. In detail, a speaker detection model based on the attention mechanism is firstly used in the visual frontend to select the key visual signals corresponding to the speaker. Then, semantic features of audio signals and visual signals are extracted by using two pre-trained feature extraction networks. Finally, in order to exploit the complementarity and independence of the signals from two modalities from the feature and decision level, the features are fed into the proposed hybrid fusion module. In addition, the first Chinese keyword spotting dataset named PKU-KWS is recorded. Experiments on this dataset demonstrate the reliability of the proposed method for practical applications. Meanwhile, the model also shows stable performance under different noise intensities.

Keywords: Audio-visual fusion · Multi-person · Keyword spotting · Hybrid fusion

1 Introduction

Keyword spotting refers to spotting keywords in a continuous audio stream [1–3]. Traditional speech-based keyword spotting has achieved good performance in pure speech conditions [4]. However, when in complex scenes such as noisy and reverberant environments or multi-speaker crossover, the performance of audio model drops significantly. Therefore, visual information, which can't be affected by audio noise, has been used to compensate for the degradation of performance due to noise in audio models [5–8].

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

L. Fang et al. (Eds.): CICAI 2022, LNAI 13605, pp. 327–338, 2022.
https://doi.org/10.1007/978-3-031-20500-2_27

Audio-visual fusion has been an active research area in speech recognition in recent years. However, there is little relevant work proposed on the implementation of audio-visual keyword spotting (AV-KWS) [9]. Wu et al. are the earliest to propose an AV-KWS model based on the Hidden Markov filler model (HMM-filler) [10]. In the study by Ding et al. [11], an AV-KWS model based on multidimensional convolutional neural network (MCNN) as the main architecture is built and obtains a better result. More recent research [12] has used CNN combined with long short-term memory (LSTM) for feature extraction, with a better focus on longer-term sequence correlation. What is more, the deep learning model is increasingly being used for extracting the visual features from the original speaker's face images, which seems to be the preferred approach [9].

According to recent studies, deep networks gradually replace traditional methods for keyword spotting. Most studies have been conducted on single-person scenes, while few on multi-person scenes. Moreover, most existing models based on audio-visual fusion use decision fusion, which is easy to handle asynchrony and ensures that signals from different modalities are not interfered with by each other. However, it does not make use of the correlation of the signal of the two modalities at the feature level [13].

In this paper, an audio-visual model using the hybrid fusion is proposed for multi-person keyword spotting [14]. First, a speaker detection model based on attention mechanism [15] is constructed. It enables the capture of speaker facial images in multi-person scenes and minimizes the interference of non-speakers, as well as visual background noise. Second, two pre-trained models are separately used to extract semantic features of the two modalities. Finally, the features are fed into the designed hybrid fusion network for classification. The main contributions of our work are the following:1) the first AV-KWS model which is also adapted to multi-person scenes is proposed, 2)hybrid fusion is used to exploit the complementarity and independence of visual and audio signals from the feature and decision level, 3) the first Chinese dataset PKU-KWS is recorded for multi-person keyword spotting. Eventually, experimental results on the PKU-KWS dataset show that the designed model has high reliability and stronger robustness to noise in practical applications.

2 AV-KWS Model

2.1 Architectures

As shown in Fig. 1, the designed audio-visual model using hybrid fusion for keyword spotting uses a joint learning strategy. The independence of each modality is learnt separately by the audio stream and visual stream, and the relevance between the audio modality and visual modality is focused on in the parallel feature fusion stream. Finally, hybrid fusion is used to combine the advantages of independence and relevance of the two modalities.

The model consists of three main components: speaker detection, semantic feature extraction and classification. In the speaker detection model, a bilinear attention network [16] is used to calculate match scores after extracting features

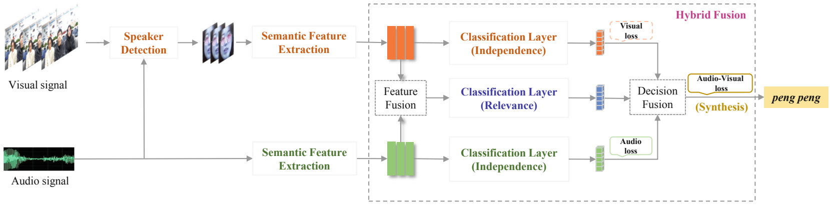

Visual signal
Audio signal
Speaker Detection
Semantic Feature Extraction
Semantic Feature Extraction
Hybrid Fusion
Classification Layer (Independence)
Feature Fusion
Classification Layer (Relevance)
Classification Layer (Independence)
Visual loss
Audio loss
Decision Fusion
Audio-Visual loss
(Synthesis)
peng peng

Fig. 1. The audio-visual model using hybrid fusion for muilt-person keyword spotting.

by Resnet-18. The face images with the highest match score will be selected as the output of the speaker detection model. After that, the speaker's facial images and audio signals are fed separately into the pre-trained feature extraction model to obtain high-level semantic features, and then features of the two modalities are aligned by using fully connected networks. Next, the features of the two modalities are then concatenated and fed into the posterior classification network. Meanwhile, audio features and visual features are also fed separately to two same classification networks. Finally, decision fusion is used before judgement to combine the result from the feature fusion stream and the two single modality models.

2.2 Speaker Detection

For multi-person scenes, visual background noise is very detrimental to the model, especially the facial images of non-critical speakers. In order to minimize the impact of irrelevant visual signals, a speaker detection model based on an attention mechanism is implemented in this paper [17]. In this model, the inputs are audio signals and visual signals. The speaker's original facial images will be selected as the output through the match scores between audio features and visual features. The addition of the audio signal makes the model detect the key speaker accurately through match scores even in the multi-speaker environment. The details of the model are shown in Fig. 2.

First, visual features $V \in \mathbb{R}^{B \times N \times T \times L_v}$ and auditory features $A \in \mathbb{R}^{B \times T \times L_a}$ are extracted from the inputs, where B is the batch size, N is the number of the characters, L equals 512 for the length of the feature tensor of each frame and T represents the number of time steps. For the video stream, a 3D convolutional layer with a kernel size of $5 \times 5 \times 7$ followed by a ResNet-18 [18] is used to obtain a high-level feature representation of each image stream. The audio feature is extracted by using a 1D convolutional layer with a kernel size of 80 followed by Resnet-18 [19].

A bilinear attention network is used after feature extraction to calculate the match scores S of each visual feature of each person to the audio feature. In this process, we make the attention query $Q = A$ and the attention key $K = V$. The original images corresponding to the N faces form the space to be selected.

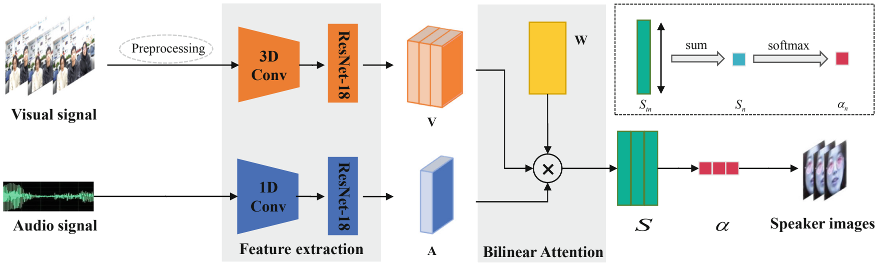


Fig. 2. The specific structure of the speaker detection model.

The match scores of each frame corresponding to the n_{th} person S_{tn} will be calculated by an additional learnable parameter matrix W:

$$S_{tn} = Q_t W V_{tn},\ with\ S \in \mathbb{R}^{1\times T\times N}, \tag{1}$$

where Q_t and V_{tn}, respectively, denote the audio features and the visual features of the nth person at t.

Next, all the match scores S_{tn} of each person are summed for only one person is speaking in each video.:

$$S_n = \sum\nolimits_t S_{tn}. \tag{2}$$

Then, the match scores S_{tn} of each person are used to calculate the normalized attention score α_n of the n_{th} person through the softmax function:

$$\alpha_n = \frac{e^{s_n}}{\sum_m e^{s_n}},\ with\ \alpha_n \in \mathbb{R}^{1\times N}. \tag{3}$$

Finally, the original images of the highest match score are selected as the output of the speaker detection model.

2.3 Semantic Feature Extraction Based on Pre-trained Models

In this part, two pre-trained models are used to extract semantic features in our model. The pre-trained WavLM [20] which learns universal speech representations from massive speech data, is applied in the frontend of the audio stream to acquire semantic features. Meanwhile, the visual modality also makes use of the Audio-Visual HuBERT (AV-HuBERT) [21] model to obtain high-level visual feature representations. In addition, both of the two pre-trained models enable to extract features from indeterminate length signals.

The inputs to the above two pre-trained models are the original audio and the grey-scale face image stream which is selected in the speaker detection model. The final frame-level feature representation is obtained with a length of 768. However, due to the differences in the processing of the two modalities, the final features are not consistent in the temporal dimension. But we found that the number of the time steps of audio modality T_a with that of visual modality T_v

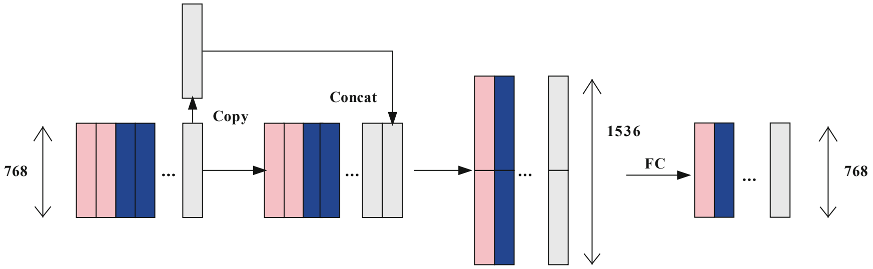


Fig. 3. Feature alignment. Align the dimension of the audio features and visual features by processing audio features. Each rectangle in this figure represents the audio feature of each frame.

always satisfies the relationship $T_a = 2T_v - 1$. It is very important for feature fusion to ensure the alignment in the time dimension between the features from two modalities. Therefore, the process shown in Fig. 3 is used to obtain audio features with the same dimension as video features. We first take the audio feature of the last time step and splice it behind the original features. Then, the audio feature of every two frames will be spliced together and fed into the fully connected network to convert the length of each splice feature to 768.

2.4 Classification Layer

First, two Bidirectional Gate Recurrent Units (BiGRUs) with a hidden layer of dimension 1,024 are used to further learn the feature representation adapted to the KWS. Then, convolutional layers and a fully connected layer are used to generate a one-dimensional vector. Finally, a softmax function is used to calculate the posterior probability of each keyword.

2.5 Hybrid Fusion

Both the feature fusion and decision fusion are used in the proposed hybrid fusion model. First, features from different modalities are concatenated together besides being fed into the classification layer directly. Before judgment, a fully connection layer is used to learn the validity of the results from each channel.

3 Experiments and Discussions

3.1 Dataset

The dataset used in our experiments, the PKU-KWS dataset, is collected in a quiet environment with controlled normal light. It contains 2,700 video segments (unequal in length) recorded in Chinese at 25 frames per second. The resolution of the video segments is 1920 × 1080 and the audio sample rate of 48 kHz. In each segment, 1, 2, or 3 characters with clear facial images can be seen, while only one person is speaking at the same moment. Some example frames have shown in Fig. 4. In addition, we set five keywords and select 1,040 segments to be applied to experiments where only one keyword is used. These video segments are randomly divided into three parts, 600 for the training set, 240 for the validation set, and the remaining 200 for testing. The distribution of keywords is shown in Fig. 5.

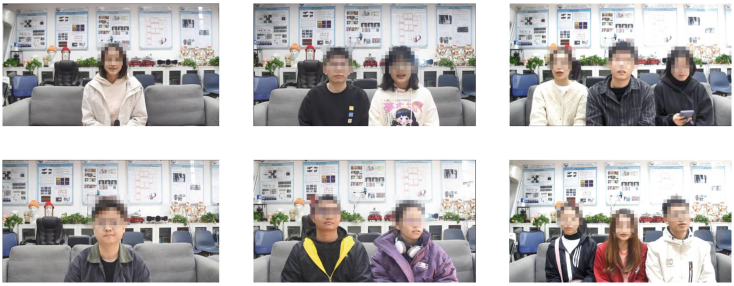


Fig. 4. Example frames from the PKU-KWS dataset.

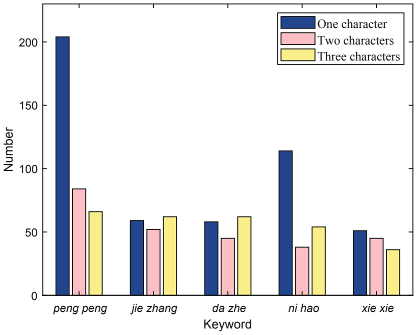
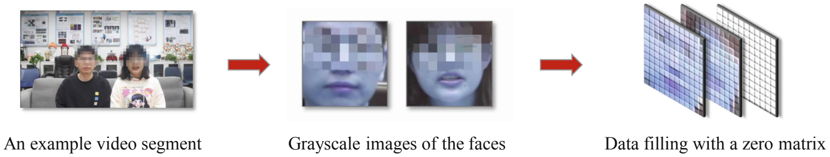


Fig. 5. The distribution of the five keywords in the PKU-KWS including *peng peng*, *jie zhang*, *da zhe*, *ni hao* and *xie xie*.

3.2 Preprocessing

In the visual modality, OpenCV followed by Dlib is used to detect the faces. Then the face regions in the frames are cropped down and sorted by spatial location. Finally, all the cropped images are converted to a grey-scale image with a resolution of 112 × 112. In addition, each group of visual images will be expanded to a three-person sample by zero-filling. A data filling example for a frame with two characters is shown in Fig. 6. After extracting the face images from each frame, a zero matrix will be added behind the grayscale images. Similar to the sample with two characters, the single-person sample will be supplemented with two additional zero matrices.


An example video segment

Grayscale images of the faces

Data filling with a zero matrix

Fig. 6. An example of the data filling for visual signals. A zero matrix is added behind the grayscale images arranged according to the spatial position.

For audio modality, the audio signals are extracted from the video segments at a sampling rate of 16 kHz. During training, it is worth mentioning that the audio signal will be randomly added with Gaussian white noise with different signal to noise ratio (SNR) which follows a uniform distribution between −5 dB and 15 dB.

3.3 Training

The training process can be divided into three stages: First, the speaker detection model is trained to select the speaker by learning the input. Then, two single modality models and a feature fusion based model are trained to learn the parameters of the classification network. Finally, a hybrid fusion-based model is trained to combine the judgements from different streams.

All experiments are implemented through PyTorch, using an NVIDIA Graphics RTX 3090 GPU. We use the stochastic gradient descent in all experiments to train the parameters and calculate the cross entropy loss based on the output and the label. Meanwhile, the learning rate in the hybrid fusion model is set to 0.01, while in all other training the learning rate is 0.2. As training continues, the learning rate will gradually decrease. The specific update strategy is:

$$Lr_{new} = Lr_{init} \times \frac{1}{2^{(\frac{epoch-4}{5})}} \tag{4}$$

where Lr_{new} is the updated learning rate, Lr_{init} is the initial value of the learning rate, and *epoch* represents the current training count.

3.4 Results and Analysis

Speaker Detection Model. From the examples of the distribution of attention scores shown in Fig. 7, the speaker's face images always obtain the highest match scores and more attention in each frame. The speaker is detected accurately whether in single-person or multi-speaker scenes. Experimental results of the speaker detection model are shown in Table 1. It shows that the speaker detection model has achieved recognition accuracy of 99.5% under a clean condition. Meanwhile, we find that the addition of noise has little effect on the model's recognition, which means the model is not very sensitive to audio signals. In our opinion, the reason is that there is only one character speaking in each segment, and the model prefers to detect the speaker through a simpler visual feature such as lip movement.

Table 1. Results of the speaker detection model.

SNR (dB)	Clean	15	10	5	0	−5
ACC	99.5%	99.5%	99.5%	99.5%	99.5%	99.0%

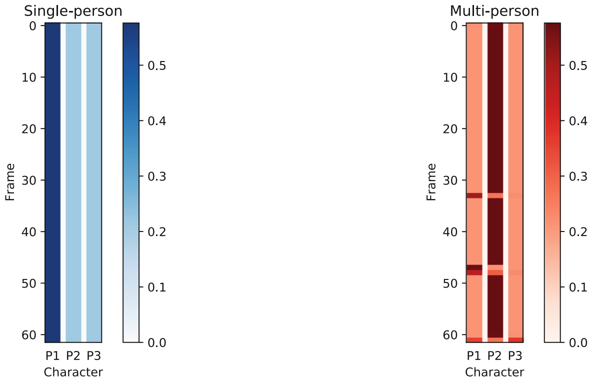


Fig. 7. Example results about distribution of match scores.

Validation of the Effectiveness of Data Enhancement Strategies. As mentioned in Sect. 3.2, data enhancement is used on the audio signals in the training set. To demonstrate the effectiveness of the data enhancement method we used, we trained the audio model separately using enhanced and unenhanced audio data. Then the two models were tested under different noise conditions and the accuracy curves with noise for both conditions are shown in Fig. 8. At low noise levels (SNR $\geq$ 5 dB), the difference in performance between the two models is not significant. However, as the noise continues to increase, the model using data enhancement (Red solid line) performs significantly better than the model without data enhancement (Blue solid line), showing a higher degree of robustness to noise.

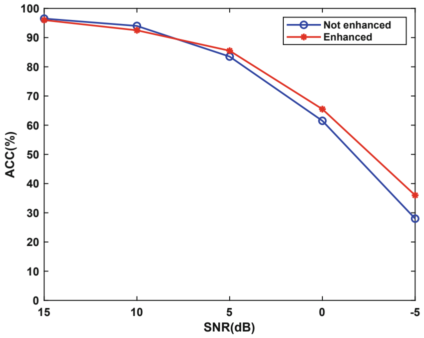


Fig. 8. Curves of recognition accuracy with noise for audio models with and without data enhanced. (Color figure online)

Validation of the Effectiveness of The Audio-Visual Model Using a Hybrid Fusion. In order to verify the contribution of the audio-visual model based hybrid fusion, the audio-visual fusion model and the single modality model are trained separately, and tested in six different noise environments. Table 2 shows the recognition accuracy of the single modality models and the fusion model in different environments. The visual model has a recognition accuracy of 82.0% and the recognition accuracy of the audio model is significantly higher than that of the visual model when the noise is not very strong (SNR $\geq$ 5 dB). As the noise is further enhanced, the original audio data is severely disturbed and the recognition accuracy decreases substantially. In an environment with clean audio, the fusion model performs slightly lower than the audio model. It is confirmed that the involvement of too much visual information leads to a decrease in recognition accuracy in the case of pure speech signals. However, with the continuous addition of noise, it can be seen that the recognition performance of the designed audio-visual fusion model obtains a better performance compared to the audio model. Especially at the SNR of -5 dB, the recognition accuracy of the audio-visual model is 71.0% which is significantly higher than that of the audio model. It means that the introduction of visual information which is not affected by audio noise compensates for the decline in recognition accuracy due to noise effectively.

Table 2. Test results of the Audio model(A), Visual model (V) and the designed Hybrid fusion based Audio-visual fusion model (AV-H) in different noise environments.

SNR (dB)	Clean	15	10	5	0	−5
A	98.5%	96.0%	92.5%	85.5%	65.5%	36.0%
V	82.0%	82.0%	82.0%	82.0%	82.0%	82.0%
AV-H	98.0%	98.0%	97.5%	96.5%	89.5%	71.0%

A Comparative Experiment for Different Modality Fusion Learning Strategies. In order to further verify the effectiveness of the designed modality synthesis learning, a comparative experiment for modality fusion learning strategies is set on the PKU-KWS dataset. The AV-KWS models with the different learning strategies methods are trained with the same database and experimental setup. As is shown in Table 3, the proposed hybrid fusion model using the modality synthesis learning obtains the best performance under both clean conditions and low noise levels (SNR $\geq$ 5 dB). It suggests that the use of the hybrid fusion makes the best use of the advantages and independence of audio-visual fusion in a low noise environment. However, as the noise increases, the recognition accuracy of the model based on feature fusion which focuses on relevance declines rapidly. Especially in the case where the SNR is -5 dB, the recognition accuracy is just 56.5% and significantly lower than that of the other two methods, while the model based on decision fusion which focuses on the independence of the two modalities performs best. It means that for KWS under high noise conditions, it may be a better choice to ensure independence between modalities than to learn relevance between features.

Table 3. Test results for the AV-KWS models focused on the independence, relevance and synthesis.

SNR (dB)	Clean	15	10	5	0	−5
Relevance	97.0%	96.0%	94.0%	89.0%	83.0%	56.5%
Independence	97.5%	97.0%	96.5%	95.0%	90.5%	73.0%
Synthesis	98.0%	98.0%	97.5%	96.5%	89.5%	71.0%

4 Conclusion

In this paper, we propose an audio-visual model based on the hybrid fusion for multi-person keyword spotting. Experiments on the PKU-KWS dataset show that the designed hybrid fusion model obtains recognition accuracy comparable to that of the audio modality model under clean conditions. Meanwhile, compared with the single modality models and the audio-visual models with other fusion methods, it achieves the best performance in low noise conditions. Even in a high noise environment, It still shows stable performance. The next step is to extend the function of the model to enable to detect the sentences with more than one keyword.

References

1. Pang, C., Liu, H., Zhang, J., Li, X.: Binaural sound localization based on reverberation weighting and generalized parametric mapping. IEEE/ACM Trans. Audio Speech Lang. Process. **25**(8), 1618–1632 (2017). https://doi.org/10.1109/TASLP.2017.2703650
2. Wöllmer, M., Eyben, F., Keshet, J., Graves, A., Schuller, B., Rigoll, G.: Robust discriminative keyword spotting for emotionally colored spontaneous speech using bidirectional LSTM networks, pp. 3949–3952 (2009). https://doi.org/10.1109/ICASSP.2009.4960492
3. 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 (2013). https://doi.org/10.1109/ASRU.2013.6707731
4. Kim, B., Chang, S., Lee, J., Sung, D.: Broadcasted residual learning for efficient keyword spotting. arXiv preprint arXiv:2106.04140 (2021)
5. Li, Y., Liu, H., Tang, H.: Multi-modal perception attention network with self-supervised learning for audio-visual speaker tracking. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 1456–1463 (2022)
6. Zheng, H., Wang, M., Li, Z.: Audio-visual speaker identification with multi-view distance metric learning, pp. 4561–4564 (2010). https://doi.org/10.1109/ICIP.2010.5653016
7. Stewart, D., Seymour, R., Pass, A., Ming, J.: Robust audio-visual speech recognition under noisy audio-video conditions. IEEE Trans. Cybern. **44**(2), 175–184 (2014). https://doi.org/10.1109/TCYB.2013.2250954
8. Miao, Y., Gowayyed, M., Metze, F.: EESEN: end-to-end speech recognition using deep RNN models and WFST-based decoding. In: 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp. 167–174. IEEE (2015)
9. López-Espejo, I., Tan, Z.H., Hansen, J., Jensen, J.: Deep spoken keyword spotting: an overview. IEEE Access (2021)
10. Wu, P., Liu, H., Li, X., Fan, T., Zhang, X.: A novel lip descriptor for audio-visual keyword spotting based on adaptive decision fusion. IEEE Trans. Multimedia **18**(3), 326–338 (2016)
11. Ding, R., Pang, C., Liu, H.: Audio-visual keyword spotting based on multidimensional convolutional neural network. In: 2018 25th IEEE International Conference on Image Processing (ICIP), pp. 4138–4142. IEEE (2018)
12. Momeni, L., Afouras, T., Stafylakis, T., Albanie, S., Zisserman, A.: Seeing wake words: audio-visual keyword spotting. arXiv preprint arXiv:2009.01225 (2020)
13. Katsaggelos, A.K., Bahaadini, S., Molina, R.: Audiovisual fusion: challenges and new approaches. Proc. IEEE **103**(9), 1635–1653 (2015). https://doi.org/10.1109/JPROC.2015.2459017
14. Liu, H., Li, W., Yang, B.: Robust audio-visual speech recognition based on hybrid fusion. In: 2020 25th International Conference on Pattern Recognition (ICPR), pp. 7580–7586 (2021). https://doi.org/10.1109/ICPR48806.2021.9412817
15. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
16. Kim, J., Jun, J., Zhang, B.: Bilinear attention networks. CoRR (2018). arXiv:1805.07932
17. Braga, O., Siohan, O.: A closer look at audio-visual multi-person speech recognition and active speaker selection. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6863–6867. IEEE (2021)

18. Afouras, T., Chung, J.S., Senior, A., Vinyals, O., Zisserman, A.: Deep audio-visual speech recognition. IEEE Trans. Pattern Anal. Mach. Intell. (2018)
19. Petridis, S., Stafylakis, T., Ma, P., Cai, F., Tzimiropoulos, G., Pantic, M.: End-to-end audiovisual speech recognition. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6548–6552. IEEE (2018)
20. Chen, S., et al.: WavLM: large-scale self-supervised pre-training for full stack speech processing. arXiv preprint arXiv:2110.13900 (2021)
21. Shi, B., Hsu, W.N., Mohamed, A.: Robust self-supervised audio-visual speech recognition. arXiv preprint arXiv:2201.01763 (2022)