Interspeech 2022
18-22 September 2022, Incheon, Korea

# Audio-Visual Wake Word Spotting in MISP2021 Challenge: Dataset Release and Deep Analysis

*Hengshun Zhou[1], Jun Du[1,*], Gongzhen Zou[1], Zhaoxu Nian[1], Chin-Hui Lee[2], Sabato Marco Siniscalchi[2,3], Shinji Watanabe[4], Odette Scharenborg[5], Jingdong Chen[6], Shifu Xiong[7], Jian-Qing Gao[7]*

[1]University of Science and Technology of China, China [2]Georgia Institute of Technology, USA
[3]Kore University of Enna, Italy [4]Carnegie Mellon University, USA [7] iFlytek, China
[5]Delft University of Technology, The Netherlands [6]Northwestern Polytechnical University, China

✉jundu@ustc.edu.cn

## Abstract

In this paper, we describe and release publicly the audio-visual wake word spotting (WWS) database in the MISP2021 Challenge, which covers a range of scenarios of audio and video data collected by near-, mid-, and far-field microphone arrays, and cameras, to create a shared and publicly available database for WWS. The database [1] and the code [2] are released, which will be a valuable addition to the community for promoting WWS research using multi-modality information in realistic and complex conditions. Moreover, we investigated the different data augmentation methods for single modalities on an end-to-end WWS network. A set of audio-visual fusion experiments and analysis were conducted to observe the assistance from visual information to acoustic information based on different audio and video field configurations. The results showed that the fusion system generally improves over the single-modality (audio- or video-only) system, especially under complex noisy conditions.

**Index Terms**: Wake word spotting, audio-visual database, data augmentation, analysis

## 1. Introduction

Wake word spotting (WWS) can be considered as a specific case of keyword spotting (KWS), which plays a very important role in smart device terminals like mobile phones and digital assistants, concerning the identification of predefined wake word(s) in the input utterances [1, 2]. The usual WWS systems are based on audio modality and require a large amount of keywords data to train [3, 4, 5, 6, 7]. These WWS systems usually perform well in clean speech conditions. However, their performance will degrade significantly under noisy conditions [8, 9, 10, 11]. In addition, WWS is still challenging in the far-field due to the interference in signal transmission and the complexity of acoustic environment [12]. In order to activate the interactions between devices and users, a robust wake word detection module is particularly important.

Significant progress has been made in the past few years in mitigating the challenges under a variety of acoustic conditions and improving the far-field for audio wake word detection. A number of works have been done to improve the robustness to noise by introducing a speech enhancement module [10, 12, 13], the secondary network [14], data augmentation techniques [15], the new training method [16] and investigating novel network structure [17, 18, 19]. Researchers also use various training techniques, such as data augmentation, semi-supervised learning and audio front-end (AFE) algorithms to mitigate the challenges under far-field conditions [12, 20, 21, 22]. In [23], a multi-look enhancement network (MLENet), which enhances the acoustic sources with multiple look directions simultaneously, is utilized to improve KWS performance in large noisy and far-field conditions. Despite the above research progress, WWS is yet a challenging task, especially in realistic environments with a low signal-to-noise ratio (SNR) and has attracted the attention of speech researchers. In [24], the authors demonstrate that if audio is available, visual keyword spotting improves the performance for both a clean and noisy audio signal. They also pointed out some applications for visual KWS, such as assisting people with speech impairment or aphonia.

One of the main challenges in real-world audio-visual KWS applications is the increased computational cost brought by the visual modality. In our previous work [25], a neural network pruning strategy via the lottery ticket hypothesis in an iterative fine-tuning manner (LTH-IF) provides a solution for the lightweight and low computational. Another factor that hinders the development of audio-visual KWS is the lack of publicly-available large-scale datasets. Zheng-hua Tan et al. reported a review on existing keyword spotting databases and approaches [26]. A limitation of the existing public WWS database is that most of them contain only audio. Therefore, those databases cannot be used to study the information from other modalities. The corpus containing rich modalities and scenarios will also greatly promote the research of this task.

In this paper, we describe and release publicly the wake word spotting database in the MISP2021 Challenge [27], which involves a distant multi-microphone conversational audio-visual data recorded in the home TV scenario. Specifically, we have double-checked all data and deleted a few asynchronous samples in the training set and development set. In order to improve the challenge in the evaluation set, we have also added some confusing-word data. Different from the challenge, we also supplemented the data of near- and mid-field to all researchers. Moreover, extensive experiments were provided for deep analyses of the database, involving different data augmentation methods of audio and video modalities. We also designed a set of audio-visual fusion experiments under different audio and video field configurations. Finally, a deep analysis was conducted to present the assistance and complementarity of visual information to acoustic information in complex noisy environments.

*corresponding author

[1]https://challenge.xfyun.cn/misp_dataset

[2]https://github.com/mispchallenge/MISP2021-AVWWS

Copyright © 2022 ISCA

10.21437/Interspeech.2022-10650

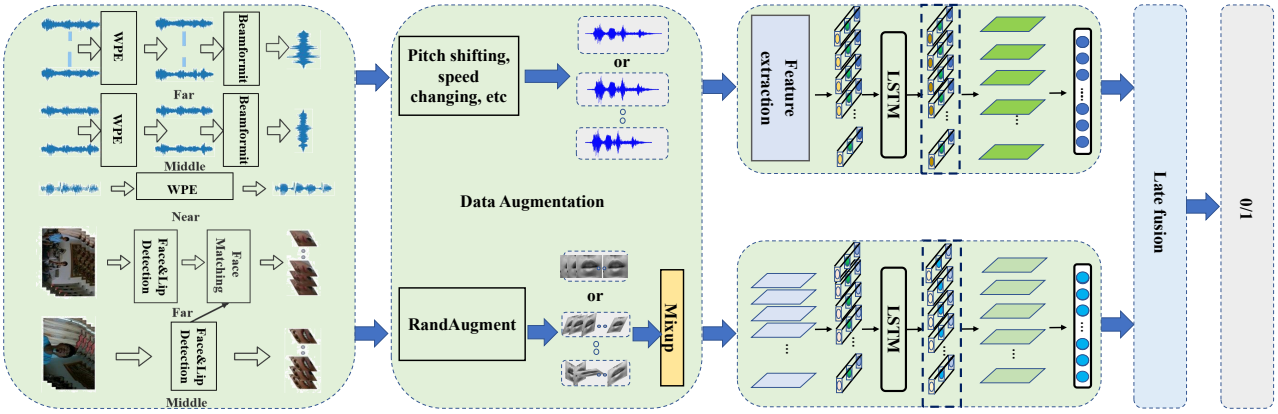



Figure 1: *The architecture of our audio-visual wake word spotting research.*

The database is available online, which will be an extremely valuable resource for researchers in wake word spotting and will push forward the endeavour in multi-modal information processing under realistic and complex conditions.

## 2. Database

### 2.1. Existing databases

Data are an essential ingredient for any deep learning system, which is not only used to train the parameters of the algorithm but also used to validate it. Zheng-hua Tan et al. discussed the existing keyword spotting databases [26], focusing on the audio corpus. The authors point out that one of the existing problems is that the majority of the speech corpora of interest for KWS research and development are not publicly available, and they are for (company) internal use only. One obvious manifestation is the number of keywords per dataset is set mostly 1 or 2 for this is that datasets mainly fit the application of KWS, such as voice assistants. According to [26], only seven are public, from the 26 datasets of statistics. They also highlight the importance that the dataset acoustic conditions should be as close as possible to real-world KWS deployment scenarios.

Most audio-visual databases are not publicly available. As it is discussed in [26], one main problem with the databases in [24, 28] is that they were not developed for KWS, and, therefore, they do not standardize a way of utilization facilitating KWS technology comparison. Recording scenarios consist of home environments with background TV noise or little visual occlusion since this is the target scenario of many KWS applications. Other requirements for a good database include: a decent number of speakers, synchronized multi-modal data, and good coverage of complex realistic scenarios. To the best of our knowledge, there are currently few WWS databases that satisfy these core requirements.

### 2.2. The MISP2021-AVWWS database

The MISP2021-AVWWS database described in this paper was designed to overcome some of these basic limitations. In contrast to the challenge corpus, the updated corpus has the following features:

- During the challenge, only far-field audio/video data in the evaluation set were released. In the updated corpus, all evaluation data were released to support various research, such as mid-field video could be used for lipreading tasks, etc.
- We have double-checked all training/development data and deleted asynchronous samples;
- More than 800 confusing words utterances have been added to the evaluation set for increasing the challenge of the task.

The database contains about 125 hours of audio-visual data. A main problem with existing WWS corpora is that they do not standardize a way to facilitate WWS technology reproducibility and comparison. We divide the data set such that there are non-overlapping speakers across the training, development and evaluation sets. Table 1 shows the division of the audio-visual data and indicates details regarding the number of sessions, the type of room, and the number of male/female speakers. The wake word is "Xiao T Xiao T". The data set includes 115 sessions. The number of speakers within one conversation session ranges from 1 to 6. The total number of speakers in the data set is 327. All speakers are native Chinese speaking Mandarin without strong accents. Various conversation topics were recommended during the recording. Some real noise data is also provided. In addition, for video data, the lip region of interest (ROI) is also provided, which is extracted by using our internal detection tools.

Table 1: *Overview of the MISP2021-AVWWS database. [P: for presence of wake word, N: for absence of wake word]*

| Dataset | Training | | Dev | | Eval | |
|---|---|---|---|---|---|---|
| | P | N | P | N | P | N |
| Duration (h) | 5.49 | 107.89 | 0.48 | 2.28 | 1.68 | 7.83 |
| Session | 88 | 88 | 8 | 8 | 19 | 19 |
| Room | 25 | 25 | 5 | 5 | 8 | 8 |
| Participant | 252 | 252 | 28 | 28 | 47 | 47 |
| Male | 78 | 78 | 9 | 9 | 27 | 27 |
| Female | 174 | 174 | 19 | 19 | 20 | 20 |

## 3. Framework

The overall flowchart of the audio-visual WWS architecture is shown in Figure. 1, which mainly consists of three parts: audio stream, video stream and fusion stream. The details are elaborated in the following subsections.

### 3.1. Audio stream and data augmentation

Data augmentation methods are proven to be very effective in the WWS task of the MISP2021 Challenge [29, 30]. Inspired by [31], we adopt several data augmentation methods for audio data to improve the generalization capabilities of the models in the training stage, which are listed below:

- Noise/Reverberation adding: It is performed on audio waveforms to add random interference;
- Pitch shifting: It is performed on audio waveforms to randomly shift the pitch based on the uniform distribution;
- Speed changing: The speed of audio is randomly changed on utterance level to perform data augmentation.

For the audio stream, the input is the 40-dimensional filter bank (FBank) features computed with a window size of 25ms and a window shift of 10ms. They are normalized by the global mean and variance are selected as input features. A frontend consisting of weighted prediction error (WPE) dereverberation [32] and weighted delay-and-sum beamforming (BeamformIt) [33] is applied to the far-field 6-channel speech and the middle-field 2-channel speech before the FBank feature extraction. Given the raw input audio data $I_{\mathrm{A}}$, we can calculate normalized FBank features $f_{\mathrm{A}}$ through the FBank extractor $F_{\mathrm{A}}$:

$$f_{\mathrm{A}} = F_{\mathrm{A}}(I_{\mathrm{A}}) \tag{1}$$

### 3.2. Video stream and data augmentation

Several data augmentation methods are also adopted for video data, which are listed below:

- Mixup: It was proposed in [34]. Two batches of data are randomly mixed in each step along with the corresponding labels when training audio models. And 90% of the training data are employed with mixup when building the visual WWS models;
- RandAugment: It was proposed in [35] for visual object detection. We adopt three sub-policies in the search space, namely Sharpness, Color and IdentityMapping. For each image, two operations are randomly selected to be applied in sequence;

For the video stream, we select ResNet-18 [36] as our video embedding extractor. The gray scale lip $f_{\mathrm{V}}$ is used as the ResNet-18 input, which is pre-trained on a word-level lip reading task. For details, please refer to [27]. The lip feature $f_{\mathrm{V}}$ is calculated with the lip feature extractor $F_{\mathrm{V}}$:

$$f_{\mathrm{V}} = F_{\mathrm{V}}(I_{\mathrm{V}}) \tag{2}$$

### 3.3. Fusion stream

For the audio-visual fusion stream, the final decision is formed from the combination of the decisions from separate audio and visual WWS network.

$$P_{\mathrm{AV}} = \alpha \times P_{\mathrm{A}}(y_a|\boldsymbol{f}_a) + \beta \times P_{\mathrm{V}}(y_v|\boldsymbol{f}_v) \tag{3}$$

where, $P_{\mathrm{A}}(y_a|\boldsymbol{f}_a)$ and $P_{\mathrm{V}}(y_v|\boldsymbol{f}_v)$ are the posterior of $y_a$/$y_v$ given $\boldsymbol{f}_a$/$\boldsymbol{f}_v$ generated by the audio-only model and the video-only model, respectively. $\alpha$ and $\beta$ are the weights of audio system and video system respectively.

The output value of these models is compared with the preset threshold ($th_A, th_V, th_{AV}$) after the sigmoid operation. "1" indicates that the current sample contains wake word, and "0" indicates the opposite.

### 3.4. Evaluation

Following the MISP2021 challenge [27], the combination of false reject rate (FRR) and false alarm rate (FAR) is adopted as the criterion, which is defined as follows.

$$Score = FRR + FAR = \frac{N_{\mathrm{FR}}}{N_{\text{wake}}} + \frac{N_{\mathrm{FA}}}{N_{\text{non-wake}}} \tag{4}$$

where $N_{\text{wake}}$ and $N_{\text{non-wake}}$ denote the number of samples with the wake word and without the wake word in the evaluation set, respectively. $N_{\mathrm{FR}}$ denotes the number of samples that include the wake word but where the WWS system erroneously did not detect it and $N_{\mathrm{FA}}$ is the number of samples that do not contain the wake word but where the WWS system erroneously detected it. The lower $Score$, the better the system performance.

Table 2: *Performance comparison of audio-only WWS systems.*

| Field | DA | Dev | | | Eval | | |
|---|---|---|---|---|---|---|---|
| | | $FAR$ | $FRR$ | $Score$ | $FAR$ | $FRR$ | $Score$ |
| Far | No | 0.078 | 0.104 | 0.182 | 0.092 | 0.236 | 0.328 |
| Mid | No | 0.039 | 0.043 | 0.082 | 0.060 | 0.159 | 0.219 |
| Near | No | 0.006 | 0.006 | 0.012 | 0.010 | 0.123 | 0.133 |
| Far | Yes | 0.104 | 0.060 | 0.164 | 0.147 | 0.115 | 0.262 |
| Mid | Yes | 0.048 | 0.027 | 0.075 | 0.065 | 0.098 | 0.163 |
| Near | Yes | 0.006 | 0.006 | 0.012 | 0.016 | 0.071 | 0.087 |

## 4. Experiments

### 4.1. Data simulation and implementation details

For audio data simulation, the Room Impulse Response (RIR) is generated according to the actual room size and microphone position by using an open-source toolkit, i.e. pyroomacoustic [37]. In addition, we also add noise with 7 different signal-to-noise ratios (from -15dB to 15dB with a step of 5dB) by using the tools officially provided by MISP2021.

We employ PyTorch to train all models and minimize the loss function using the Adam optimization method. The batch size is 64 for the audio-only WWS system and 8 for the video-only system. The learning rates are set to 0.0002, 0.00005 for audio-only, video-only and systems respectively.

### 4.2. Results for audio-only WWS systems

First, we evaluate the performance of the audio-only WWS systems. Table 2 reports detailed results of different fields. "DA" indicates data augmentation. According to the upper block of Table 2, it is noted that the audio-based system achieves the best performance on the near-field and the worst performance in the far-field with the $Score$ is 0.328 on the evaluation set. We further evaluate audio-only WWS systems by applying data augmentation, and the results are shown in the bottom block of Table 2. From Table 2, we can observe that the performance of the audio-only system has been improved in all three fields when data augmentation is applied, especially in the far-field, the score of 0.262 has been achieved with the absolute $Score$ gain of 0.066 compared with the $Score$ without data augmentation.

### 4.3. Results for video-only WWS systems

We then evaluate video-only WWS systems with the results presented in Table 3. In the upper block of Table 3, we show the performance of video-only WWS systems based on different fields. Clearly, model performance is better for mid-field. According to the results in the bottom block of Table 3, when the data augmentation methods introduced in Section 3.2 are applied to training the video-only model, the systems performance is improved for the mid-field and far-field compared to the original model, which demonstrates the effectiveness of the data augmentation strategy.

Table 3: *Performance comparison of video-only WWS systems.*

| Field | DA | Dev | | | Eval | | |
|---|---|---|---|---|---|---|---|
| | | $FAR$ | $FRR$ | $Score$ | $FAR$ | $FRR$ | $Score$ |
| Far | No | 0.147 | 0.494 | 0.641 | 0.113 | 0.601 | 0.714 |
| Mid | No | 0.167 | 0.104 | 0.271 | 0.246 | 0.364 | 0.610 |
| Far | Yes | 0.384 | 0.087 | 0.471 | 0.317 | 0.267 | 0.584 |
| Mid | Yes | 0.151 | 0.091 | 0.242 | 0.229 | 0.274 | 0.513 |

### 4.4. Results for audio-visual fusion systems

Based on the above positive results, we conduct audio-visual fusion, and the results under multiple hybrid configurations are shown in Table 4. We can see that introducing the visual information consistently improves the system performance, especially for the combination of far-field audio and mid-field video with a $Score$ gain of 0.043 compared to audio-only system. Here [27] is the baseline scheme of MISP2021. It is noted that the performance of the audio-visual fusion system has been significantly improved by applying data augmentation.

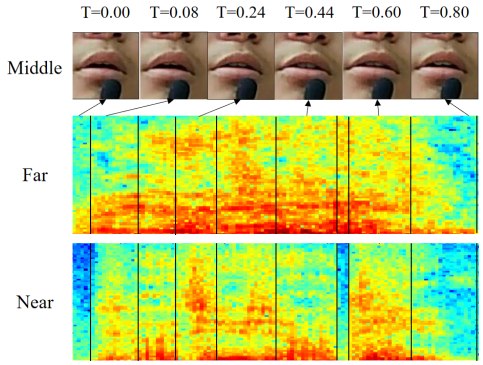
Table 4: *Performance comparison of audio-visual WWS systems under multiple hybrid configurations.*

| Audio-Visual | Dev | | | Eval | | |
|---|---|---|---|---|---|---|
| | $FAR$ | $FRR$ | $Score$ | $FAR$ | $FRR$ | $Score$ |
| Near + Middle | 0.003 | 0.002 | 0.005 | 0.031 | 0.038 | 0.069 |
| Far + Middle | 0.049 | 0.076 | 0.125 | 0.100 | 0.119 | 0.219 |
| Far + Far | 0.073 | 0.068 | 0.141 | 0.101 | 0.150 | 0.251 |
| Far + Far [27] | 0.140 | 0.120 | 0.260 | 0.129 | 0.314 | 0.443 |

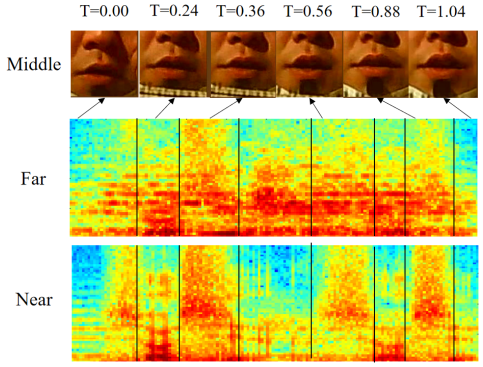
To better demonstrate the advantages of introducing the visual modality, we plot the video frame and spectrogram for two picked examples from the evaluation set in Figure. 2. The first part of the figure is the lip of the target speaker and the corresponding time in the mid-field video. The middle and bottom of the figure represent the far-field and near-field spectrograms of the target speaker respectively.

The example in Figure. 2 (a) is a positive sample, but the audio is seriously damaged by the noise under far-field conditions, which can also be observed by comparing the far-field and near-field spectrograms. This example was also misclassified as "Negative" by the audio-only system. However, the visual information is not influenced by acoustic noise. Accordingly, the example was correctly classified as "Positive" after audio-visual fusion, which demonstrates that visual information can improve the performance of the system, especially in noisy environments.

For the example in Figure. 2 (b), which is a negative sample, the far-field audio, the audio-only system classified it as




(a) R10_S182183184185_C0009_I1_Far_185_176044_176864




(b) R81_S505_C0008_I0_Far_505_145215_146255

Figure 2: *Analysis on two randomly selected examples.*

"Positive". We analyzed this sample. On the one hand, it is seriously damaged by noise. On the other hand, it's a confusing word ("Xiao Xie Xiao Xie"), and the video system also misclassified it due to the similar lip shape changes. Although the visual lip does not output accurate semantic information at this time, it still helps maintain the anti-noise ability. After fusion with acoustic information, the correct prediction is obtained, which demonstrated the good coupling between audio and video modalities.

## 5. Conclusion

In this paper, we presented and fully released the audio-visual wake word spotting database in the MISP2021 Challenge. This database can play an important role in developing WWS for realistic and complex environments, which is expected to bring better human-machine interaction. Moreover, We investigated the effective data augmentation techniques for audio and video modalities respectively. We further implemented audio-visual fusion for this database based on different audio and video field configurations. The experimental result and misclassification analysis of the systems show the advantages of introducing the visual modality.

## 6. Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant No. 62171427 and the Strategic Priority Research Program of Chinese Academy of Sciences under Grant No. XDC08050200.

# 7. References


[1] R. Rose and D. Paul, "A hidden markov model based keyword recognition system," in *Proc. ICASSP 1990*, 1990, pp. 129–132.

[2] D. Leroy, A. Coucke, T. Lavril, T. Gisselbrecht, and J. Dureau, "Federated learning for keyword spotting," in *Proc. ICASSP 2019*, 2019, pp. 6341–6345.

[3] G. Chen, C. Parada, and G. Heigold, "Small-footprint keyword spotting using deep neural networks," in *Proc. ICASSP 2014*, 2014, pp. 4087–4091.

[4] I. López-Espejo, Z.-H. Tan, and J. Jensen, "Improved external speaker-robust keyword spotting for hearing assistive devices," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 28, pp. 1233–1247, 2020.

[5] C. Jose, Y. Mishchenko, T. Sénéchal, A. Shah, A. Escott, and S. N. P. Vitaladevuni, "Accurate detection of wake word start and end using a CNN," in *Interspeech*, 2020, pp. 3346–3350.

[6] S. Tabibian, "A survey on structured discriminative spoken keyword spotting," *Artificial Intelligence Review*, vol. 53, 04 2020.

[7] Y. Wang, H. Lv, D. Povey, L. Xie, and S. Khudanpur, "Wake word detection with streaming transformers," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 5864–5868.

[8] X. Ji, M. Yu, J. Chen, J. Zheng, D. Su, and D. Yu, "Integration of multi-look beamformers for multi-channel keyword spotting," in *Proc. ICASSP 2020*, 2020, pp. 7464–7468.

[9] I. López-Espejo, Z.-H. Tan, and J. Jensen, "A novel loss function and training strategy for noise-robust keyword spotting," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2254–2266, 2021.

[10] Y. A. Huang, T. Z. Shabestary, and A. Gruenstein, "Hotword cleaner: Dual-microphone adaptive noise cancellation with deferred filter coefficients for robust keyword spotting," in *ICASSP*, 2019, pp. 6346–6350.

[11] E. Yılmaz, Özgür Bora Gevrek, J. Wu, Y. Chen, X. Meng, and H. Li, "Deep Convolutional Spiking Neural Networks for Keyword Spotting," in *Proc. Interspeech 2020*, 2020, pp. 2557–2561.

[12] Y. Gao, Y. Mishchenko, A. Shah, S. Matsoukas, and S. Vitaladevuni, "Towards data-efficient modeling for wake word spotting," in *Proc. ICASSP 2020*, 2020, pp. 7479–7483.

[13] M. Jung, Y. Jung, J. Goo, and H. Kim, "Multi-Task Network for Noise-Robust Keyword Spotting and Speaker Verification Using CTC-Based Soft VAD and Global Query Attention," in *Proc. Interspeech 2020*, 2020, pp. 931–935.

[14] R. Kumar, M. Rodehorst, J. Wang, J. Gu, and B. Kulis, "Building a Robust Word-Level Wakeword Verification Network," in *Proc. Interspeech 2020*, 2020, pp. 1972–1976.

[15] Y. Wang, H. Lv, D. Povey, L. Xie, and S. Khudanpur, "Wake Word Detection with Alignment-Free Lattice-Free MMI," in *Proc. Interspeech 2020*, 2020, pp. 4258–4262.

[16] K. Łopatka and T. Bocklet, "State Sequence Pooling Training of Acoustic Models for Keyword Spotting," in *Proc. Interspeech 2020*, 2020, pp. 4338–4342.

[17] C. Yang, X. Wen, and L. Song, "Multi-Scale Convolution for Robust Keyword Spotting," in *Proc. Interspeech 2020*, 2020, pp. 2577–2581.

[18] T. Higuchi, S. Saxena, M. Souden, T. D. Tran, M. Delfarah, and C. Dhir, "Dynamic curriculum learning via data parameters for noise robust keyword spotting," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6848–6852.

[19] B. Liu, S. Nie, Y. Zhang, S. Liang, Z. Yang, and W. Liu, "Loss and double-edge-triggered detector for robust small-footprint keyword spotting," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6361–6365.

[20] Y. Gao, N. D. Stein, C.-C. Kao, Y. Cai, M. Sun, T. Zhang, and S. N. P. Vitaladevuni, "On Front-End Gain Invariant Modeling for Wake Word Spotting," in *Proc. Interspeech 2020*, 2020, pp. 991–995.

[21] H.-J. Park, P. Zhu, I. L. Moreno, and N. Subrahmanya, "Noisy Student-Teacher Training for Robust Keyword Spotting," in *Proc. Interspeech 2021*, 2021, pp. 331–335.

[22] A. Hard, K. Partridge, C. Nguyen, N. Subrahmanya, A. Shah, P. Zhu, I. L. Moreno, and R. Mathews, "Training Keyword Spotting Models on Non-IID Data with Federated Learning," in *Proc. Interspeech 2020*, 2020, pp. 4343–4347.

[23] M. Yu, X. Ji, B. Wu, D. Su, and D. Yu, "End-to-End Multi-Look Keyword Spotting," in *Proc. Interspeech 2020*, 2020, pp. 66–70.

[24] L. Momeni, T. Afouras, T. Stafylakis, S. Albanie, and A. Zisserman, "Seeing wake words: Audio-visual keyword spotting," in *British Machine Vision Association*, 2020.

[25] H. Zhou, J. Du, C.-H. H. Yang, S. Xiong, and C.-H. Lee, "A study of designing compact audio-visual wake word spotting system based on iterative fine-tuning in neural network pruning," in *ICASSP*, 2022.

[26] I. López-Espejo, Z.-H. Tan, J. H. L. Hansen, and J. Jensen, "Deep spoken keyword spotting: An overview," *IEEE Access*, vol. 10, pp. 4169–4199, 2022.

[27] H. Chen, H. Zhou, J. Du, C.-H. Lee, J. Chen, S. Watanabe, S. M. Siniscalchi, O. Scharenborg, D.-Y. Liu, B.-C. Yin, J. Pan, J.-Q. Gao, and C. Liu, "The first multimodal information based speech processing (misp) challenge: Data, tasks, baselines and results," in *Proc. ICASSP 2022*, 2022.

[28] P. Wu, H. Liu, X. Li, T. Fan, and X. Zhang, "A novel lip descriptor for audio-visual keyword spotting based on adaptive decision fusion," *IEEE Transactions on Multimedia*, vol. 18, no. 3, pp. 326–338, 2016.

[29] M. Cheng, H. Wang, Y. Wang, and M. Li, "The duk audio-visual wake word spotting system for the 2021 misp challenge," in *Proc. ICASSP 2022*, 2022.

[30] Y. Xu, J. Sun, Y. Han, S. Zhao, C. Mei, T. Guo, S. Zhou, C. Xie, W. Zou, and X. Li, "Audio-visual wake word spotting system for misp challenge 2021," in *Proc. ICASSP 2022*, 2022.

[31] Q. Wang, S. Zheng, Y. Li, Y. Wang, Y. Wu, H. Hu, C.-H. H. Yang, S. M. Siniscalchi, Y. Wang, J. Du, and C.-H. Lee, "A model ensemble approach for audio-visual scene classification," DCASE2021 Challenge, Tech. Rep., June 2021.

[32] L. Drude, J. Heymann, C. Boeddeker, and R. Haeb-Umbach, "Nara-wpe: A python package for weighted prediction error dereverberation in numpy and tensorflow for online and offline processing," in *Speech Communication; 13th ITG-Symposium*, 2018, pp. 1–5.

[33] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2011–2022, 2007.

[34] H. Zhang, M. Cissé, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *6th International Conference on Learning Representations, ICLR 2018*. OpenReview.net, 2018. [Online]. Available: https://openreview.net/forum?id=r1Ddp1-Rb

[35] E. D. Cubuk, B. Zoph, J. Shlens, and Q. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 18 613–18 624.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[37] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *Proc. ICASSP 2018*, 2018, pp. 351–355.