THE DKU AUDIO-VISUAL WAKE WORD SPOTTING SYSTEM FOR THE 2021 MISP CHALLENGE

Ming Cheng[1,2] *Haoxu Wang*[1,2] *Yechen Wang*[2] *Ming Li*[1,2†]

[1]School of Computer Science, Wuhan University, Wuhan, China
[2]Data Science Research Center, Duke Kunshan University, Kunshan, China

ABSTRACT

This paper describes the system developed by the DKU team for the MISP Challenge 2021. We present a two-stage approach consisting of end-to-end neural networks for the audio-visual wake word spotting task. We first process audio and video data to give them a similar structure and then train two unimodal models with unified network architecture separately. Second, we propose a Hierarchical Modality Aggregation (HMA) module that fuses multi-scale audio-visual information from pre-trained unimodal models. Our system has a clear and concise framework consisting of end-to-end neural networks. With this framework and extensive data augmentation methods, our presented system achieves a false reject rate of 3.85% and a false alarm rate of 3.42% on far-field audio in the development set of the competition database, which ranks 2nd in the wake word spotting track of the MISP challenge.

***Index Terms*—** MISP Challenge, Audio-visual Wake Word Spotting, Deep Neural Network, Multimodal Fusion

1. INTRODUCTION

Wake Word Spotting (WWS), also known as keyword detection, is a task that aims at detecting the occurrences of the pre-defined wake word in a continuous audio stream. With the rapid development of various speech-enabled applications, WWS systems are increasingly interested.

Traditional WWS approaches mainly rely on statistical models (e.g., HMMs [1, 2, 3]) and Vertibi search algorithms [4] to calculate the likelihood ratio of the wake word occurrence. Recently, many researchers turn to focus on deep neural network based WWS systems, including convolutional neural networks [5], temporal convolutional neural networks [6, 7], recurrent neural networks [8, 9], and Transformers [10]. These methods demonstrate a large potential and achieve better performance than traditional methods.

One difficulty in WWS applications is that the false alarm rate often increases in complex acoustic environments (far-field audio, background noises, and reverberations) and conversational multi-speaker interactions with a large portion of speech overlap. Therefore, many methods have been proposed to tackle these challenges. Wu et al. [11] incorporate domain knowledge into network training and improve the performance of the wake word classifier on far-field conditions. Park et al. [12] utilize a smoothed max-pooling loss to mitigate the inaccurate alignments by employing Large Vocabulary Continuous Speech Recognition (LVSCR) in complex acoustic environments. Furthermore, it is reasonable to tackle these problems by introducing information from additional modalities, such as video and text, yielding boosted robustness against various environments and speaker identities in practical applications. For instance, Ding et al. [13] propose an audio-visual neural network based on a multi-dimensional convolutional neural network (MCNN) to perform audio-visual WWS.

This paper focuses on building a straightforward but robust audio-visual wake word spotting system in complex environments. We first train unimodal neural networks for audio and video data in an end-to-end manner, respectively. Next, a multimodal fusion layer is proposed to process the embedding vectors extracted from each unimodal branch and output the final prediction. In Task-1 of the MISP Challenge 2021 [14], our audio-visual WWS system shows considerable robustness in the home-living environments.

2. METHODS

This paper adopts a two-stage framework to cope with audio-visual wake word spotting. Generally, we first investigate methods of training strong unimodal backbone models to process audio and video signals individually. Then, we study how to fuse multimodal information to optimize the final system.

2.1. Unimodal Models

Residual connections [15] and three-dimensional convolutional layers [16] are proven to be effective in deep learning and have been widely used in different tasks (e.g., Video Action Recognition [17, 18, 19], Speaker Identification [20]). Hence, we choose to construct the neural networks based on residual 3-D CNNs (ResNet-3D [21]). To simplify the problem, we design audio and video backbone models to have the same network architecture, except that the dimensions of model inputs are different. Table 1 describes the structure of the backbone model.

Unlike the classical ResNet, ResNet-3D replaces 2-D convolutional kernels with 3-D ones with an extra dimension T. Therefore, some preprocessing methods for raw audio and video signals are needed to adapt to the required shape of the model inputs. As video data comprises consecutive images, it is easy to use the dimensions $(H, W, 3)$ to represent each RGB image and T to indicate the number of frames. By this method, the video inputs are organized to have the shape of $(T, H, W, 3)$. For audio signals, the 1-D waveforms are firstly transformed to 2-D features in the frequency domain (e.g., DCT, Filter Bank, MFCC [22]). Then, we use a sliding window to split the feature map along the time axis, and each sliced part is regarded as a frame-level acoustic feature with a shape of $(H, W, 1)$. By stacking all sliced T frames, we organize the audio inputs in the shape of $(T, H, W, 1)$. Finally, both audio and video inputs have the same structure (T, H, W, C), so we can implement similar neural network architectures for audio and video backbone models.

† Corresponding Author, E-mail: ming.li369@dukekunshan.edu.cn

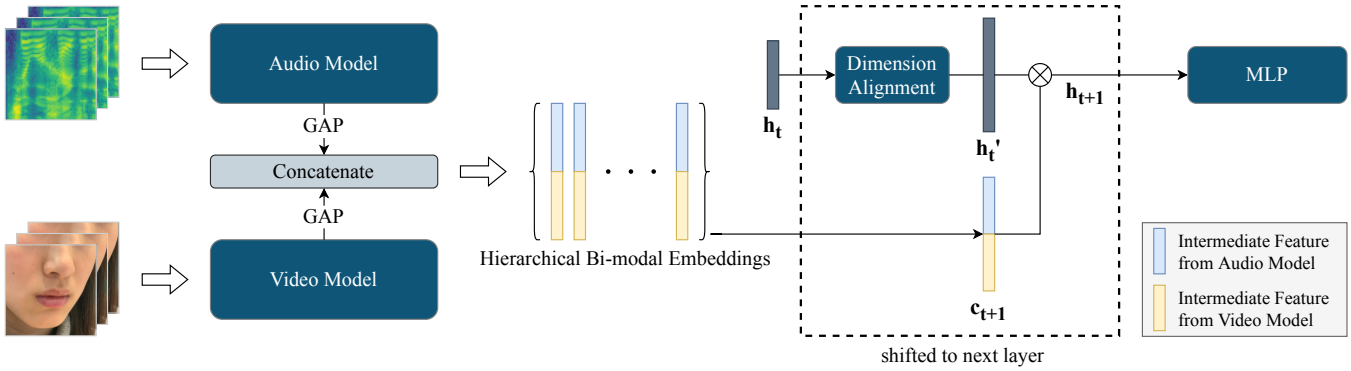

Fig. 1. Proposed Framework of Hierarchical Modality Aggregation (HMA)

2.2. Hierarchical Modality Aggregation

As the signals from one modality are sensitive to disturbances in complex environments, it is essential to take advantage of complementary information in multimodal data to build a robust system. Therefore, we propose a hierarchical fusion mechanism to extract audio-visual information from the pre-trained unimodal models.

Figure 1 depicts the whole framework, namely Hierarchical Modality Aggregation (HMA). Each unimodal neural network consists of several residual blocks which can output intermediate feature maps at different levels. We apply the global average pooling to these feature maps and concatenate the obtained embedding vectors at the same level from two modalities. Let $\mathbf{c_t}$ denote the concatenated embedding vector at level t, and $\mathbf{h_t}$ represent the hidden feature vector that contains the bi-modal information extracted from previous layers no later than level t.

At the first stage, we set the initial feature vector $\mathbf{h_1}$ to be the first concatenated embedding vector $\mathbf{c_1}$. In the following, each $\mathbf{h_t}$ is processed by a dimension alignment module (a linear transformation with a sigmoid activation) to generate $\mathbf{h_t'}$, which has the same shape aligned to the next bi-modal embedding vector $\mathbf{c_{t+1}}$. Then, we apply element-wise multiplication to $\mathbf{h_t'}$ and $\mathbf{c_{t+1}}$ to update the hidden feature vector from $\mathbf{h_t}$ to $\mathbf{h_{t+1}}$. We repeat this structure to aggregate the intermediate audio-visual features from unimodal models step by step. At the last stage, the feature vector containing multi-scale audio-visual information is fed into an MLP to make the final prediction. This way, our proposed method can extract multimodal information hierarchically only through a series of fully-connected layers, and it outperforms two widely-used fusion baselines.

3. EXPERIMENTS

3.1. Database and Evaluation Metrics

The training data is from Task-1 of the 1st Multimodal Information Based Speech Processing Challenge (MISP 2021 [14]). This competition targets the home scenarios in which people speak Chinese or interact with smart devices while TV is playing in the living room. In this case, a WWS system is needed to detect the wake word "Xiao T, Xiao T" spoken by participants.

The released database has two subsets: training set (47k+ negative samples and 5K+ positive samples) and development set (2k+ negative samples and 600+ positive samples). Moreover, an evaluation set (8K +) without annotations is provided to competition participants. The testing results (the score of WWS) can be obtained by uploading model predictions to the official competition webpage. Each sample includes five categories of synchronous data captured by different devices:

- 6-channel far-field microphone array
- 2-channel mid-field microphone array
- single-channel near-field microphone
- mid-field high-definition camera
- far-field wide-angle camera

In such a classification task, the positive class represents the existence of the wake word in a given sample, and the negative class indicates the opposite. Following the requirements of the competition committee, we use False Reject Rate (FRR), False Alarm Rate (FAR), and the Score of WWS as the evaluation criteria. Let N_{wake} denote the number of samples that contain the wake word, and N_{non_wake} represent the number of samples without the wake word. The FRR and FAR are defined as follows:

$$FRR = \frac{N_{FR}}{N_{wake}}, \quad FAR = \frac{N_{FA}}{N_{non\_wake}} \tag{1}$$

where N_{FR} denotes the number of samples containing the wake word while not recognized by the system. N_{FA} denotes the num-

Table 1. Structure of Backbone Model

Layer Name	Parameters
Inputs	–
Conv 1	$\begin{bmatrix}\text{conv } 3\times3\times3, & 32\\ \text{pool } 3\times3\times1, & -\end{bmatrix}\times 1$
ResBlock 1	$\begin{bmatrix}\text{conv } 3\times3\times3, & 32\\ \text{conv } 3\times3\times3, & 32\end{bmatrix}\times 3,\ /2$
ResBlock 2	$\begin{bmatrix}\text{conv } 3\times3\times3, & 64\\ \text{conv } 3\times3\times3, & 64\end{bmatrix}\times 3,\ /2$
ResBlock 3	$\begin{bmatrix}\text{conv } 3\times3\times3, & 64\\ \text{conv } 3\times3\times3, & 64\end{bmatrix}\times 3,\ /2$
ResBlock 4	$\begin{bmatrix}\text{conv } 3\times3\times3, & 128\\ \text{conv } 3\times3\times3, & 128\end{bmatrix}\times 3,\ /2$
ResBlock 5	$\begin{bmatrix}\text{conv } 3\times3\times3, & 256\\ \text{conv } 3\times3\times3, & 256\end{bmatrix}\times 3,\ /2$
Global Avg Pool	–
Dropout	$\text{p} = 0.2$
Linear 1	256×32
Linear 2	32×2

ber of samples containing no wake words while predicted to be positive by the system. Hence, the final score of Wake Word Spotting (WWS) is defined as:

$$Score^{WWS} = FRR + FAR \tag{2}$$

3.2. Audio Backbone Model

3.2.1. Preprocessing

To fit the required inputs of the audio backbone model, we reshape each audio clip to (T, H, W, C).

- The raw signal is converted into a 2-D spectral feature map by Filter Bank (implemented in the Torchaudio toolkit [23]) with a filter number of 80, a frame length of 25 ms, and a frameshift of 10 ms.
- The full feature map is split into a series of frame-level blocks by a sliding window along the time axis. The window size is set to 80 with a stride of 4. Then, we stack all sliced feature blocks consecutively to obtain the data with a shape of $(T, 80, 80, 1)$
- In our experiments, the T is set to the constant (64). We utilize the random clipping or zero padding techniques for each sample to give it a fixed block number. Thus, the shape of each audio sample becomes $(64, 80, 80, 1)$.
- Before being fed into a neural network, each sample should be standardized to have a mean of 0 and a standard deviation of 1.

3.2.2. Data Augmentation

From the perspective of data augmentation, we assemble a set of transformations based on sox effects. The audio sample could randomly undergo the following procedures with a probability of 0.5.

- Each audio is randomly selected to change its tempo or volume. The tempo variation is in the range $[0.9, 1.1]$, and the volume changes in the range $[0, 20]$.
- Each audio is randomly selected to change its speed to become faster or slower, with the ratio in the range $[0.8, 1.2]$.
- For samples with more than T feature blocks (described in Section 3.2.1), we utilize a clipping method that chooses a random portion from the original sequence.

3.2.3. Ablation Experiments

We train the audio backbone model with the Adam optimizer, utilizing an initial learning rate of 0.001. To tackle the imbalance between positive and negative samples, we adopt the weighted CrossEntropy Loss (negative:positive=1:5) with the label smoothing method [24].

Table 2 shows the experimental results of the audio backbone model. The baseline represents the model trained without any extra tricks. In addition, we introduce considerable data augmentation that brings a significant performance improvement. Furthermore, we implement beamforming (MVDR [25]) to multi-channel audio signals to exhaust the potential of microphone arrays. Instead of single-channel audio, incorporating beamforming-enhanced audio samples into the training data can improve the performance further.

Table 2. Experimental results of the audio backbone model. DA and BF represent the data augmentation and beamforming preprocess, respectively. The symbol + denotes the cumulative addition of the current method based on the above ones, and $*$ indicates the best model tested on the evaluation set.

Method	Field	Dev (%)			Eval (%)
		FRR	FAR	Score	Score
Baseline	Near	0.96	1.78	2.74	-
	Mid	3.37	5.24	8.61	-
	Far	9.77	6.83	16.61	18.7
+ DA	Near	1.92	1.64	3.56	-
	Mid	3.37	4.67	8.04	-
	Far	8.17	5.34	13.51	15.8
+ BF*	Near	1.12	1.87	2.99	-
	Mid	2.56	5.39	7.95	-
	Far	6.41	6.01	12.42	12.2

3.3. Video Backbone Model

3.3.1. Preprocessing

This work considers only lip regions of the videos as model inputs instead of full faces. Similar to the audio backbone model, there are multiple steps for transforming raw videos to lip-region videos in the shape of (T, H, W, C).

- The face detector (RetinaFace [26]) extracts all face images and the corresponding 5 facial landmarks in each video.
- We assume that a talking face will not move dramatically in a short time window. Based on the sequential coordinates of the detected faces, the K-means algorithm in Scikit-learn toolkit [27] is used to cluster faces of the same person in one given video.
- The far-field wide-angle video may contain multiple people while only one is the target speaker. We deploy a face recognizer (ArcFace [28]) to select the target speakers from pairwise mid-field videos that only have one identity in each video.

After obtaining the facial images of each target speaker, we crop the lip regions based on the detected facial landmarks. Let $\mathbf{p_1}$, $\mathbf{p_2}$, and $\mathbf{p_3}$ represent coordinates of the detected nose, left corner of mouth and right corner of mouth, respectively. The Region of Interest (RoI) bounding box is defined as:

$$x_{center}, y_{center} = \frac{\mathbf{p_2} + \mathbf{p_3}}{2} \tag{3}$$

$$width = min\left\{3.2 \times d_{MN}, 2 \times max\left\{d_{MN}, d_{p1p2}\right\}\right\} \tag{4}$$

where d_{p1p2} denotes the distance between $\mathbf{p_1}$ and $\mathbf{p_2}$, and d_{MN} denotes the euclidean distance between $\mathbf{p_1}$ and the box center. This formula is an empirical setting introduced in a lip-reading database (CAS-VSR-W1k [29]), which is a widely used protocol.

Each extracted lip-region video is resized to have a resolution of 112×112 with 3 RGB channels. The dimension T is set to 64, which means each video is sampled to contain 64 frames. Therefore, the shape of the video sample becomes $(64, 112, 112, 3)$. Lastly, each video is normalized to be within the range of $[0, 1]$.

3.3.2. Data Augmentation

We also use some video-based data augmentation techniques. The video sample could randomly undergo the following procedures with a probability of 0.5.

- Each video is randomly selected to change its playing speed (FPS) by adding or removing some redundant frames. The scale of speed variation is in the range $[0.75, 1.25]$.
- Each video is randomly selected to take a frame-wise rotation with an angle in the range $[1, 15]$.
- Each video is randomly selected to be flipped horizontally (frame by frame). Moreover, the frame-level cropping is implemented with the random scale within $[0.8, 1]$.
- Each video is randomly selected to have a color transformation in terms of contrast, brightness, and saturation. Besides, each video has a probability of 0.2 to be converted to a grayscale one.

3.3.3. Ablation Experiments

For the video backbone model, we employ the same basic settings for model training as the audio backbone model. Table 3 shows the experimental results of the video backbone model. The baseline represents the model trained without any extra tricks. When adding the described video data augmentations, the generalization capability of the trained model is enhanced. Last but not least, we also pre-train the backbone model on a lip-reading database named CAS-VSR-W1k database [29] and have it fine-tuned on the MISP database with the above training setup. It is found that the pre-training further improves the model performance on the evaluation set.

3.4. Multimodal Fusion

3.4.1. Preprocessing and Data Augmentation

Since the multimodal model takes both audio and video data as inputs, the described data processing and augmentation methods for audio and video are still workable. This part introduces two additional data augmentation methods specified for the fusion stage.

- Paired audio and video do not necessarily come from the same field. For instance, the far-field audio (from different channels) and mid-field video can be joined to make a new training sample.
- Paired audio and video do not necessarily come from the same identity. The audio and video samples can be paired to be fed into the model as long as they belong to the same category.

3.4.2. Ablation Experiments

We again adopt the same training setup as the previous models. Table 4 shows the performances of fusion-based models. First, we implement the late fusion method by simply averaging the score-level predictions from two unimodal models. Furthermore, we test the early fusion method, which means two unimodal models are concatenated at the intermediate feature vectors and followed by a series of fully-connected layers. Finally, it is found that our proposed Hierarchical Modality Aggregation (HMA) method obtains the best performance in mid-field and far-field cases and the highest score on the evaluation set.

Table 3. Experimental results of the video backbone model. DA represents the data augmentation.The symbol + denotes the cumulative addition of the current method based on the above ones, and $*$ indicates the best model tested on the evaluation set.

Method	Field	Dev (%)			Eval (%)
		FRR	FAR	Score	Score
Baseline	Mid	9.13	10.44	19.57	-
	Far	15.06	13.66	28.72	29.0
+ DA	Mid	6.41	4.81	11.22	-
	Far	8.81	8.71	17.52	26.4
+ Pre-train*	Mid	4.81	8.32	13.13	-
	Far	8.65	8.41	17.06	21.7

Table 4. Experimental results of the multimodal fusion. LF and EF represent the Late Fusion method and Early Fusion method, respectively. Due to the lack of near-field videos, each near-field audio is tested by pairing with its corresponding mid-field video.

Method	Field	Dev (%)			Eval (%)
		FRR	FAR	Score	Score
LF	Near	1.92	9.48	11.40	
	Mid	0.16	9.67	9.83	-
	Far	1.92	9.48	11.4	11.8
EF	Near	1.12	1.06	2.18	
	Mid	0.48	3.03	3.51	-
	Far	4.01	4.09	8.1	8.3
Ours	Near	0.8	1.25	2.05	
	Mid	0.64	2.74	3.38	-
	Far	3.85	3.42	7.27	7.1

4. CONCLUSION

This paper presents the audio-visual wake word spotting system developed by the DKU team for the MISP Challenge 2021. We design an end-to-end neural network based system free from tedious front-end preprocessing and feature extractions, whose unimodal backbone networks consist of 3-D convolutional layers with residual connections. Moreover, we propose a concise framework to hierarchically extract and fuse multi-scale multimodal information to boost our system's robustness against the noise in home-living scenarios. In the end, our proposed system obtains a false reject rate of 3.85% and a false alarm rate of 3.42% on the far-field development set of the MISP database as well as the WWS score of 7.1% on the evaluation set.

5. ACKNOWLEDGMENTS

This research is funded in part by the National Natural Science Foundation of China (62171207), the Fundamental Research Funds for the Central Universities (2042021kf0039), Science and Technology Program of Guangzhou City (202007030011). Many thanks for the computational resource provided by the Advanced Computing East China Sub-Center.

6. REFERENCES

[1] Richard C Rose and Douglas B Paul, "A hidden markov model based keyword recognition system," in *Proc. ICASSP 1990.*

IEEE, 1990, pp. 129–132.

[2] Jay G Wilpon, Lawrence R Rabiner, C-H Lee, and ER Goldman, "Automatic recognition of keywords in unconstrained speech using hidden markov models," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 11, pp. 1870–1878, 1990.

[3] Jay G Wilpon, Laura G Miller, and P Modi, "Improvements and applications for key word recognition using hidden markov modeling techniques," in *Proc. ICASSP 1991*. IEEE, 1991, pp. 309–312.

[4] G David Forney, "The viterbi algorithm," *Proceedings of the IEEE*, vol. 61, no. 3, pp. 268–278, 1973.

[5] Tara N Sainath and Carolina Parada, "Convolutional neural networks for small-footprint keyword spotting," in *Proc. Interspeech 2015*, 2015.

[6] Seungwoo Choi, Seokjun Seo, Beomjun Shin, Hyeongmin Byun, Martin Kersner, Beomsu Kim, Dongyoung Kim, and Sungjoo Ha, "Temporal convolution for real-time keyword spotting on mobile devices," in *Proc. Interspeech 2019*, 2019, pp. 3372–3376.

[7] Somshubra Majumdar and Boris Ginsburg, "Matchboxnet: 1d time-channel separable convolutional neural network architecture for speech commands recognition," in *Proc. Interspeech 2020*, 2020, pp. 3356–3360.

[8] Santiago Fernández, Alex Graves, and Jürgen Schmidhuber, "An application of recurrent neural networks to discriminative keyword spotting," in *Proc. ICANN 2007*, 2007, pp. 220–229.

[9] Martin Woellmer, Bjoern Schuller, and Gerhard Rigoll, "Keyword spotting exploiting long short-term memory," *Speech communication*, vol. 55, no. 2, pp. 252–265, 2013.

[10] Yiming Wang, Hang Lv, Daniel Povey, Lei Xie, and Sanjeev Khudanpur, "Wake word detection with streaming transformers," in *Proc. ICASSP 2021*. IEEE, 2021, pp. 5864–5868.

[11] Haiwei Wu, Yan Jia, Yuanfei Nie, and Ming Li, "Domain aware training for far-field small-footprint keyword spotting," in *Proc. Interspeech 2020*, 2020, pp. 2562–2566.

[12] Hyun-Jin Park, Patrick Violette, and Niranjan Subrahmanya, "Learning to detect keyword parts and whole by smoothed max pooling," in *Proc. ICASSP 2020*, 2020, pp. 7899–7903.

[13] Runwei Ding, Cheng Pang, and Hong Liu, "Audio-visual keyword spotting based on multidimensional convolutional neural network," in *Proc. ICIP 2018*, 2018, pp. 4138–4142.

[14] Hang Chen, Hengshun Zhou, Jun Du, Chin-Hui Lee, Jingdong Chen, Shinji Watanabe, Sabato Marco Siniscalchi, Odette Scharenborg, Di-Yuan Liu, Bao-Cai Yin, Jia Pan, Jian-Qing Gao, and Cong Liu, "The first multimodal information based speech processing (misp) challenge: data, tasks, baselines and results," in *Proc. ICASSP 2022*, 2022.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proc. CVPR 2016*, 2016, pp. 770–778.

[16] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proc. ICCV 2015*, 2015, pp. 4489–4497.

[17] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proc. CVPR 2018*, 2018, pp. 6450–6459.

[18] Huijuan Xu, Abir Das, and Kate Saenko, "R-c3d: Region convolutional 3d network for temporal activity detection," in *Proc. ICCV 2017*, 2017, pp. 5783–5792.

[19] Zhaofan Qiu, Ting Yao, and Tao Mei, "Learning spatio-temporal representation with pseudo-3d residual networks," in *Proc. ICCV 2017*, 2017, pp. 5533–5541.

[20] Danwei Cai, Xiaoyi Qin, and Ming Li, "Multi-channel training for end-to-end speaker recognition under reverberant and noisy environment," in *Proc. Interspeech 2019*, 2019, pp. 4365–4369.

[21] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh, "Learning spatio-temporal features with 3d residual networks for action recognition," in *Proc. ICCVW*, 2017, pp. 3154–3160.

[22] Alan V Oppenheim, Alan S Willsky, Syed Hamid Nawab, Gloria Mata Hernández, et al., *Signals & systems*, Pearson Educación, 1997.

[23] Yao-Yuan Yang, Moto Hira, Zhaoheng Ni, Anjali Chourdia, Artyom Astafurov, Caroline Chen, Ching-Feng Yeh, Christian Puhrsch, David Pollack, Dmitriy Genzel, et al., "Torchaudio: Building blocks for audio and speech processing," *arXiv preprint arXiv:2110.15018*, 2021.

[24] Rafael Müller, Simon Kornblith, and Geoffrey Hinton, "When does label smoothing help?," *arXiv preprint arXiv:1906.02629*, 2019.

[25] Mehrez Souden, Jacob Benesty, and Sofine Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 260–276, 2010.

[26] Jiankang Deng, J. Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou, "Retinaface: Single-stage dense face localisation in the wild," *ArXiv*, vol. abs/1905.00641, 2019.

[27] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011.

[28] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proc. CVPR 2019*, 2019, pp. 4690–4699.

[29] Shuang Yang, Yuanhang Zhang, Dalu Feng, Mingmin Yang, Chenhao Wang, Jingyun Xiao, Keyu Long, Shiguang Shan, and Xilin Chen, "Lrw-1000: A naturally-distributed large-scale benchmark for lip reading in the wild," in *Proc. FG 2019*, 2019, pp. 1–8.