THE XIAOMI-TALKFREELY SYSTEM FOR AUDIO-VISUAL WAKE WORD SPOTTING OF MISP CHALLENGE 2021

Quandong Wang[1],*Weiji Zhuang*[1*], *Yuxiang Kong*[1*], *Yongqing Wang*[1*], *Junnan Wu*[1*], *Dongbo Li*[1], *Zhiyong Yan*[1], *Mingshuang Luo*[1], *Xinyu Tang*[1], *Xinyu Cai*[2], *Liyong Guo*[1], *Zhigao Chen*[1], *Yuquan Liang*[1], *Shijie Deng*[1], *Lichun Fan*[1], *Junbo Zhang*[1], *Peng Gao*[1], *Yujun Wang*[1], *Ying Huang*[1], *Zhiyong Wu*[2]

[1]Xiaomi Inc., Beijing, China
[2] Shenzhen International Graduate School, Tsinghua University, Shenzhen, China

ABSTRACT

This paper presents our submitted system to task1 of the multimodal information based speech processing (MISP) challenge 2021. For task1 of audio-visual wake word spotting, our main technical points include several kinds of data augmentation methods and a two-stage model strategy. For the first stage, a frequency-domain heterogeneous-input multi-branch model is proposed to deal with the challenging acoustic conditions. For the second stage, an alignment-free lattice-free MMI model is used suppress the confusing words. Tested on the development set and the evaluation set, our best system showed absolute score reduction of 0.219 and 0.264 respectively, compared to the official baseline system and got the first place in the challenge.

Index Terms— wake word spotting, adversarial samples, heterogeneous-input, multi-branch model

1. INTRODUCTION

With severe multi-speaker speech overlap, background noise, and reverberation in far-field home and meeting speech interactive scenarios, the performance of keyword spotting decreases significantly.

In this context, the multimodal information based speech processing (MISP) challenge aims to tackle these problems by utilizing both audio and visual modal information. As we empirically found in our experiments with audio-visual wake word spotting, the visual modal could indeed improve the performance. However, it brought marginal gains upon our audio-only systems. So we only report our proposed audio-only systems for the challenge.

2. PROPOSED SYSTEM

2.1. System description

In this session, we briefly introduce our main ideas and the system design. The task1 results are evaluated by a combination score of false reject rate (FRR) and false alarm rate (FAR). Due to the challenging recording environments, the score of development data for the baseline system was 0.259. During analyzing the results of the baseline system, we found three main reasons for the badcases:

- The amount of training data is limited and the utterance context is restricted as only short utterances are provided.

- The signal-to-noise ratio (SNR) is very low due to reverberation, strong noise and overlapping speakers.

- There are many confusing words that sound similar to the wake word.

Accordingly, we explored different data augmentation methods, leveraged multi-channel information and borrowed the idea of two-stage strategy [1, 2, 3, 4] to alleviate these salient issues. Fig. 1 shows our system workflow. We will introduce the details in the next sections.

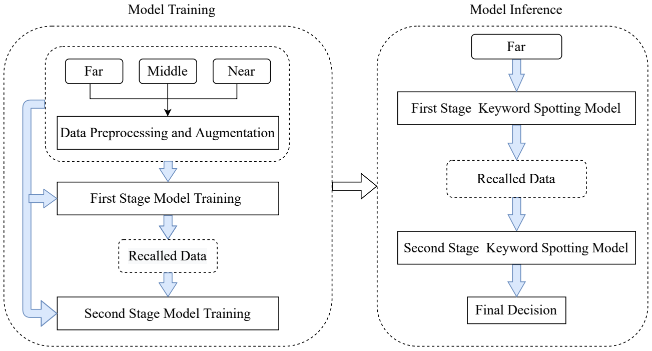
Model Training
Far
Middle
Near
Data Preprocessing and Augmentation
First Stage Model Training
Recalled Data
Second Stage Model Training
Model Inference
Far
First Stage Keyword Spotting Model
Recalled Data
Second Stage Keyword Spotting Model
Final Decision

Fig. 1. An illustration of our proposed system on task1.

2.2. Data Preprocessing and Augmentation

We explored data preprocessing and augmentation approaches for improving robustness against noisy conditions and confusing words.

2.2.1. Speech Enhancement and Separation

WPE (weighted prediction error) [5] and IVA (independent vector analysis) [6] have been tried on the baseline system.

* stands for equal contribution

Though the methods achieved few gains and we did not use them during the inference, output of speech enhancement and separation methods can be viewed as perturbation of the training data and facilitated the robustness of our models when used as part of our training data.

2.2.2. *Negative Sub-segmentation*

The duration distribution of positive data and negative data are not exactly the same. There is severe overfitting when training acoustic model without sub-segmentation [7]. The duration distribution of different positive and negative samples will result in different ratios of positive and negative samples in different length mini-batches. Models are easier to overfit to the audio length rather than to learn specific keyword. For example, the sample with a duration of more than 3 seconds would be all determined as negatives and ones with a duration of less than 0.6 seconds would be decided as positives. In order to avoid this problem, we sub-segment the negative data in the same way as [7] in data preprocessing.

2.2.3. *Speed Perturbation*

Our model architecture can handle audio of different lengths. Alignment-free lattice-free MMI (AF-LF-MMI) [7] model is trained on whole utterances rather than chunks. But the training utterances are of different lengths and we have no reliable time aligned boundaries. All utterances in a mini-batch should have the same length. In practice, we choose a set of numbers whose range covers the lengths of positive training samples, which can be called as a duration set, and we approximate all durations of training samples to the duration set by perturbing speed.

2.2.4. *Adversarial samples generation*

For suppressing confusing words, some novel augmentation strategies [8, 9] for generating adversarial samples have been proposed recently, including concatenating samples, synthesize samples and mask samples. We don't adopt the waveform concatenation method because it leads to worse performance on real data. It is very effective to suppress confusing words by generating synthesized samples. However the training data for task1 doesn't have text-level annotations, we couldn't train a speech synthesis system, so we can't use the text-to-speech augmentation approach. To obtain training samples of confused words, we applied random masking on keyword samples like [9] and used them as the adversarial negative data in training to improve the robustness of our second stage model which is discussed in next section. In addition, negative samples that are false alarmed by the first-stage are also used as adversarial samples.

2.2.5. *Multi-channel data simulation*

In order to increase the robustness of the acoustic models, we adopt several data simulation techniques as stated in [10], including simulating far-field data by convolving the near-field speech with simulated room impulse response (RIR) [11], and augmenting the far-field data by adding far-field noise.

2.2.6. *Other augmentation*

Augmenting the training data is an effective way to improve the performance of the model on small data sets. Common augmentation also includes volume perturbation, SpecAugment [12] and trimming the beginning or end of the recording slightly. We also applied these augmentation techniques to the training data.

2.3. Two-stage strategy

2.3.1. *First stage*

The first stage model focuses on handling the noisy far-field multi-channel speech. Multi-channel recordings of speech contain spatial information, providing a supplement upon time-frequency domain information for keyword spotting especially in noisy environments. We proposed a frequency domain heterogeneous-input multi-branch model, which is a combination of frequency domain multi-channel [13] and multi-branch acoustic model [14].

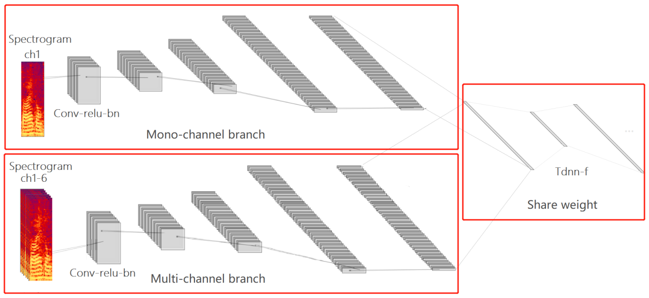

Fig. 2. An illustration of frequency domain heterogeneous-input multi-branch model.

As shown in Fig.2, for the mono-channel branch, there is only one channel per time-frequency bin. For the multi-channel branch, there are six channels per time-frequency bin. The first stage model structure is CNN-TDNNF [15, 16]. It was trained with LF-MMI [17]. We chose a working point with a relatively low graph cost to ensure that the first stage model has a high recall rate.

The first stage model of task1 was trained with pooled data of both far-field, middle-field, close-talking and augmented data. Our experiments were conducted based on Kaldi [18, 19]. Sequence level training tends to overfit, so the cross-entropy loss function is used together with the LF-MMI loss function as a regularization. Although negative samples are without text-level transcription, we can still train a simple GMM-HMM model using the positive and negative label for forced alignment and lattice generation.

Empirical analysis [20] shows that it would bring significant improvements (8% relative reduction in word error rate) in far-field speech recognition systems by training far-field recognition model with higher quality alignments generated by model trained on parallel close-talk microphone recordings. So we used the alignments from close-talk models.

2.3.2. *Second stage*

In the second stage, it is a TDNNF-based [21] single channel model, which is optimized with AF-LF-MMI [7]. The second stage model focuses on judging whether speech recalled by the first stage model contains a confusing word. The second stage model was trained with pooled data of both far-field, middle-field, close-talking, augmented data and adversarial samples. In the second stage, multi-channel data is split into multiple single channel data. All six channels of far-field speech are evaluated by the second stage single-channel model. Finally, channel fusion is adopted to combine results from different channels. If more than half of the channels are triggered, it determines that the utterance contains a keyword.

3. EXPERIMENTS AND RESULTS

Since only the far-field data is provided for evaluation set, we mainly consider the far-field results of development and evaluation set. Table 1 shows the performance of our task1 system. To compare with the official baseline system, we trained TDNNF models based on LF-MMI, denoted as LF-MMI-baseline. With the official training set (orig), the LF-MMI baseline obtained a much better score than the official baselines. Using data augmentation (orig+aug), the LF-MMI baseline system achieved about 10% relative reduction in score with respect to the one trained without augmentations. Our first-stage-only model and second-stage-only model both performed better than the LF-MMI baseline system. By deploying the two-stage strategy, the two-stage-fusion model further reduced the score by about 22% relatively upon the second-stage-only model.

Model	Data	Dev-Middle	Dev-Far	Eval-Far
Official A	-	0.15	0.27	-
Official A-V	-	0.13	0.26	0.322
LF-MMI-baseline	orig	-	0.12	0.132
LF-MMI-baseline	orig+aug	-	0.092	0.118
First-stage-only	orig+aug	-	0.081	0.107
Second-stage-only	orig+aug	-	0.057	0.075
Two-stage-fusion	-	-	0.041	**0.058**

Table 1. Performance of our task1 system in terms of score. The last line is our submitted result.

4. CONCLUSION

In this technical report, we present our system for task1 of the MISP challenge. To handle multiple interference in the recording environments and suppress the confusing words, we used many kinds of augmentation methods and proposed a two-stage model fusion strategy containing a heterogeneous-input multi-branch model and an alignment-free lattice-free MMI model. Submitting such an effective system, we obtained the first place in task1 of the challenge.

5. REFERENCES

[1] Ming Sun, Varun Nagaraja, Björn Hoffmeister, and Shiv Vitaladevuni, "Model shrinking for embedded keyword spotting," in *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2015, pp. 369–374.

[2] Assaf Hurwitz Michaely, Xuedong Zhang, Gabor Simko, Carolina Parada, and Petar Aleksic, "Keyword spotting for google assistant using contextual speech recognition," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 272–278.

[3] Minhua Wu, Sankaran Panchapagesan, Ming Sun, Jiacheng Gu, Ryan Thomas, Shiv Naga Prasad Vitaladevuni, Bjorn Hoffmeister, and Arindam Mandal, "Monophone-based background modeling for two-stage on-device wake word detection," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5494–5498.

[4] Yougen Yuan, Zhiqiang Lv, Shen Huang, and Lei Xie, "Verifying deep keyword spotting detection with acoustic word embeddings," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 613–620.

[5] Takuya Yoshioka and Tomohiro Nakatani, "Generalization of multi-channel linear prediction methods for blind mimo impulse response shortening," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 10, pp. 2707–2720, 2012.

[6] Toru Taniguchi, Nobutaka Ono, Akinori Kawamura, and Shigeki Sagayama, "An auxiliary-function approach to online independent vector analysis for real-time blind source separation," in *2014 4th Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, 2014, pp. 107–111.

[7] Yiming Wang, Hang Lv, Daniel Povey, Lei Xie, and Sanjeev Khudanpur, "Wake word detection with alignment-free lattice-free mmi," *arXiv preprint arXiv:2005.08347*, 2020.

[8] Yan Jia, Zexin Cai, Murong Ma, Zeqing Zhao, Xuyang Wang, Junjie Wang, and Ming Li, "Training wake word detection with synthesized speech data on confusion words," *arXiv preprint arXiv:2011.01460*, 2020.

[9] Haoxu Wang, Yan Jia, Zeqing Zhao, Xuyang Wang, Junjie Wang, and Ming Li, "Generating adversarial samples for training wake-up word detection systems against confusing words," *arXiv preprint arXiv:2201.00167*, 2022.

[10] Feng Ma, Li Chai, Jun Du, Diyuan Liu, Zhongfu Ye, and Chin-Hui Lee, "Acoustic model ensembling using effective data augmentation for chime-5 challenge.," in *INTERSPEECH*, 2019, pp. 1258–1262.

[11] Robin Scheibler, Eric Bezzam, and Ivan Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 351–355.

[12] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[13] Wu Minhua, Kenichi Kumatani, Shiva Sundaram, Nikko Ström, and Björn Hoffmeister, "Frequency domain multi-channel acoustic modeling for distant speech recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6640–6644.

[14] Naoyuki Kanda, Yusuke Fujita, Shota Horiguchi, Rintaro Ikeshita, Kenji Nagamatsu, and Shinji Watanabe, "Acoustic modeling for distant multi-talker speech recognition with single-and multi-channel branches," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6630–6634.

[15] Cătălin Zorilă, Christoph Boeddeker, Rama Doddipatla, and Reinhold Haeb-Umbach, "An investigation into the effectiveness of enhancement in asr training and test for chime-5 dinner party transcription," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 47–53.

[16] Li Chai, Jun Du, Di-Yuan Liu, Yan-Hui Tu, and Chin-Hui Lee, "Acoustic modeling for multi-array conversational speech recognition in the chime-6 challenge," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 912–918.

[17] Zhehuai Chen, Yanmin Qian, and Kai Yu, "Sequence discriminative training for deep learning based acoustic keyword spotting," *Speech Communication*, vol. 102, pp. 100–111, 2018.

[18] Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur, "Purely sequence-trained neural networks for asr based on lattice-free mmi.," in *Interspeech*, 2016, pp. 2751–2755.

[19] Hossein Hadian, Hossein Sameti, Daniel Povey, and Sanjeev Khudanpur, "End-to-end speech recognition using lattice-free mmi.," in *Interspeech*, 2018, pp. 12–16.

[20] Pranay Dighe, Afsaneh Asaei, and Hervé Bourlard, "Far-field asr using low-rank and sparse soft targets from parallel data," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 581–587.

[21] Daniel Povey, Gaofeng Cheng, Yiming Wang, Ke Li, Hainan Xu, Mahsa Yarmohammadi, and Sanjeev Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks.," in *Interspeech*, 2018, pp. 3743–3747.