INTEGRATION OF MULTI-LOOK BEAMFORMERS FOR MULTI-CHANNEL KEYWORD SPOTTING

Xuan Ji[1*], *Meng Yu*[2*], *Jie Chen*[1], *Jimeng Zheng*[1], *Dan Su*[1], *Dong Yu*[2]

[1]Tencent AI Lab, Shenzhen, China
[2]Tencent AI Lab, Bellevue, WA, USA

ABSTRACT

Keyword spotting (KWS) is in great demand in smart devices in the era of Internet of Things. Albeit recent progresses, the performance of KWS, measured in false alarms and false rejects, may still degrade significantly under the far field and noisy conditions. In this paper, we propose integrating multiple beamformed signals and a microphone signal as input to an end-to-end KWS model and leveraging the attention mechanism to dynamically tune the model's attention to the reliable input sources. We demonstrate, on our large simulated and recorded noisy and far-field evaluation sets, that our proposed approach significantly improves the KWS performance and reduces the computation cost against the baseline KWS systems.

Index Terms: KWS, multi-look beamforming, attention

1. INTRODUCTION

With the proliferation of smart homes and mobile and automotive devices, speech-based human-machine interaction becomes prevailing. To achieve hands-free speech recognition experience, the system continuously listens for specific wake-up words, a process often called keyword spotting (KWS)[1], to initiate speech recognition. For the privacy concern, the wake-up KWS typically happens completely on the device with low footprint and power consumption requirement.

The KWS systems usually perform very well under clean-speech conditions. However, their performance degrades significantly under noisy conditions. In order to improve the robustness to the background noises, two major techniques, namely multi-condition training and frontend enhancement, have been proposed in recent years. Multi-condition training [2, 3, 4] pools data under different environments to train neural networks and often leads to more robust systems. Nevertheless, the feature representation learned in this way, and thus the performance of KWS, is still worse than desired because the size of KWS networks is limited by the platform memory and processing power. The frontend enhancement technique, on the other hand, filters out the interference signals from the noisy stream before feeding it to the KWS system. The recent deep learning based techniques have been developed to address the speech enhancement problem [5, 6, 7]. The far field speech processing suffers from the reverberation which blurs speech spectral cues and degrades the single-channel speech enhancement. When sound sources are spatially separated, with microphone array inputs one may localize sound sources and then extract the source from the target direction. Multi-channel systems have been widely deployed for speech recognition. Such systems usually separate speech enhancement (including localization, beamforming and noise suppression) from acoustic modeling. Recently, joint optimization on multi-channel enhancement and acoustic model has been developed in deep neural network framework [8, 9, 10, 11]. Those acoustic models take raw multi-channel signals as input, i.e. without any preceding speech enhancement, and learns the multi-channel filtering and feature representation through the supervised training. However, small footprint models often lack the power of spatial filtering and feature learning ability from raw microphone data. A beamformed channel is incorporated in the multi-channel acoustic model in [11]. Nevertheless, the knowledge of the true target speaker direction is not available in real applications.

In this work, we propose a novel, effective yet simple multi-channel KWS model (MC-KWS). The multi-channel processing is handled by fixed beamformers with multiple fixed beams, equally sampled in space. After combining the beamformed signals at multiple look directions and a microphone signal as the input, the proposed end-to-end KWS model incorporates an attention mechanism to softly tune its attention to more reliable input sources. Unlike traditional multi-channel frontend pre-processing strategy, our system jointly optimizes the multi-channel feature mapping and the detection model to improve the keyword recognition accuracy. By using beamforming, the spatial information is utilized to pre-enhance the signal. The additional channel from one microphone is leveraged to preserve target speech quality in high signal-to-noise-ratio (SNR) conditions as beamformed signals are often degraded due to the misalignment between the target direction and the beam's look direction. As a result, the proposed "multi-beam + mic" model leads to better performance under both noisy and clean

[*] Equal contribution

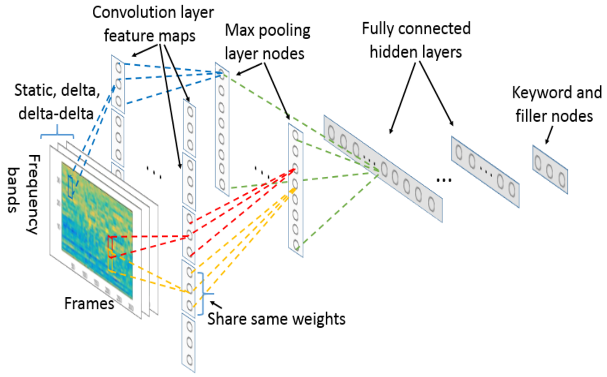

Fig. 1. Baseline KWS architecture.

conditions. To the best of our knowledge, this is the first ever end-to-end multi-channel keyword spotting model.

The rest of the paper is organized as follows. In Section 2, the baseline KWS system is reviewed. In Section 3, we present details of the proposed multi-channel KWS model and discuss its advantages. We describe our experimental setup in Section 4, and evaluate the effectiveness of the proposed system in Section 5. We conclude this work in Section 6.

2. BASELINE KEYWORD DETECTION

A block diagram of the baseline KWS system employed in this work is shown in Figure 1. The network weights are trained to optimize a cross-entropy criterion using stochastic gradient descent with momentum. Finally, in the posterior handling module, individual frame-level posterior scores from the neural network are combined into a single score corresponding to the keyword(s). We refer the readers to our previous work [12] for the implementation in details.

3. MULTI-CHANNEL KWS MODEL

Beamforming shows its advantage for speech preservation through its linear spatial filter design and processing [13, 14, 15, 16]. The conventional beamforming approaches steer the "look-direction" according to the target direction, and thus enhances the signals from a certain direction while suppressing interference from other directions. In practice, the target direction estimation is infeasible under noisy conditions, particularly when the interfering sources are background speech or competing talkers. The idea of "multi-look direction" has been applied to multi-channel acoustic model [9] and speech separation [17, 18], respectively, where a small number of spatial look directions cover all possible target speaker locations. A beamformer is applied to the multi-channel microphone signals to provide the subsequent MC-KWS network with a set of beamformed signals. The beamformer is designed such that each beam has a different look direction. More specifically, we uniformly sample the space of direction of arrival with a fixed set of beamformers. Since this is feasible for any microphone arrays, we would expect to obtain a system that is not very sensitive to the geometry of the microphone array to be used.

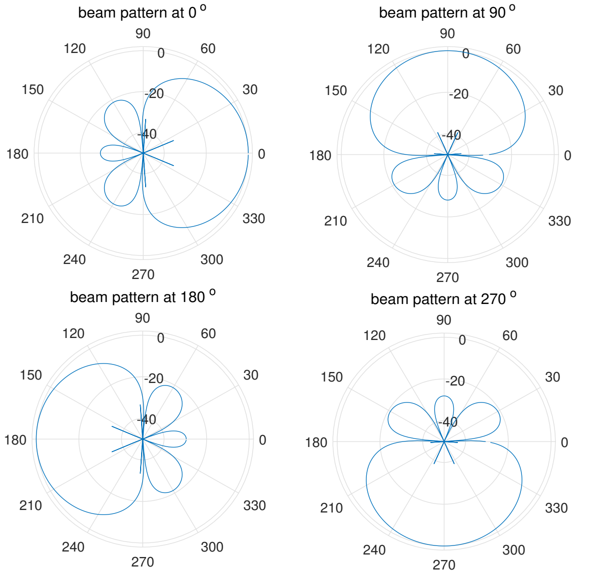

Fig. 2. Spatial gains in dB for 4 look-directions at 500Hz

The beamformer directivity patterns are optimally designed in advance for a target microphone array geometry. We propose to utilize differential beamforming to define the beamformer set. Because differential beaformers can explicitly form acoustic nulls, they can better suppress interfering sources than the additive beamformers when these interfering sources are spatially isolated from the target speaker direction [19, 20]. In our experiments, we simply used a set of four second-order differential beamformers and empirically designed the directivity patterns for the four beams to cover 360°. The four acoustic beams are targeted at 0°, 90°, 180° and 270°, respectively. An illustration of the beam pattern at the 4 desired look-directions is shown in Figure 2. In principle, there is a trade-off regarding the number of beams to use. The more beams we have, the more likely one of them is targeted at the target speaker direction. Unfortunately, this may complicate the proposed MC-KWS model. The target speaker direction is unknown in noisy environments. However, performing keyword detection on each of the beamformed signals is computationally very expensive and not tolerable for many applications. Therefore, we propose a multi-channel KWS model that efficiently integrates fixed beamformers and neural network KWS model. Since the space resolution of the sampled beams are not necessarily sufficient enough to cover the target direction, the mismatch between the target speaker direction and the look-direction causes speech distortion in the beamformed signals. This distortion degrades the KWS performance particularly in high SNR conditions when the KWS system is more sensitive to the speech distortion than residual noises. An additional

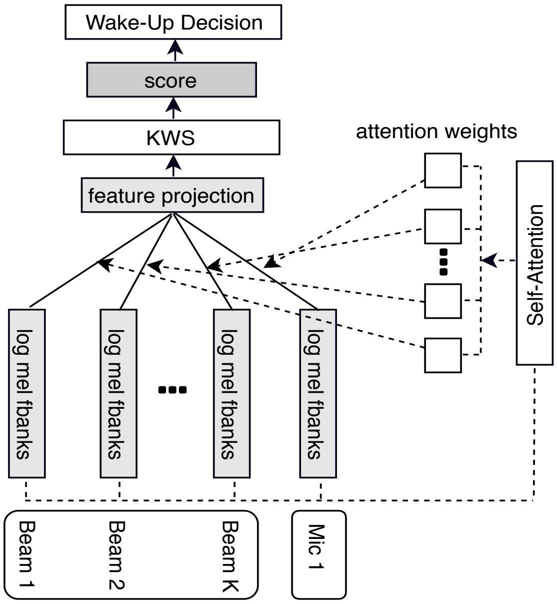

Fig. 3. Multi-channel KWS model

channel from one microphone is thereby leveraged to preserve target speech quality in the multi-channel training. A diagram of the proposed system is shown in Figure 3.

Inspired by the application of attention mechanism in speech recognition [10], speaker verification [21] and single channel keyword spotting [22], we propose an attention-based end-to-end model for multi-channel KWS. The attention mechanism we use is the soft self-attention. For each time-step, we compute a $K+1$ dimensional attention weight vector α for input feature $\mathbf{x} = [x_1, x_2, \ldots, x_{K+1}]$ as:

$$e_i = v^T tanh(Wx_i + b) \tag{1}$$

$$\alpha_i = \frac{exp(e_i)}{\sum_{k=1}^{K+1} exp(e_k)} \tag{2}$$

where a shared-parameter non-linear attention with the same W, b and v is used for each channel i of all $K+1$ channels. $\mathbf{x}$ is a 5-channel input feature tensor in our implementation, corresponding to 4 beamformed signals and 1 microphone signal. W is a $128 \times D$ weight matrix where D is the input feature size defined in Section 2, b is a 128-dimension bias vector, and v is a 128-dimension vector.

A weighted sum of multi-channel inputs is computed as

$$\hat{x} = \sum_{i=1}^{K+1} \alpha_i x_i \tag{3}$$

As a result, the multi-channel input features are projected to a feature vector $\hat{x}$ of the same size as the single-channel one, which serves as the input to the baseline single-channel KWS neural network. The KWS neural network and the multi-channel feature projection are then jointly optimized towards improving the keyword recognition accuracy.

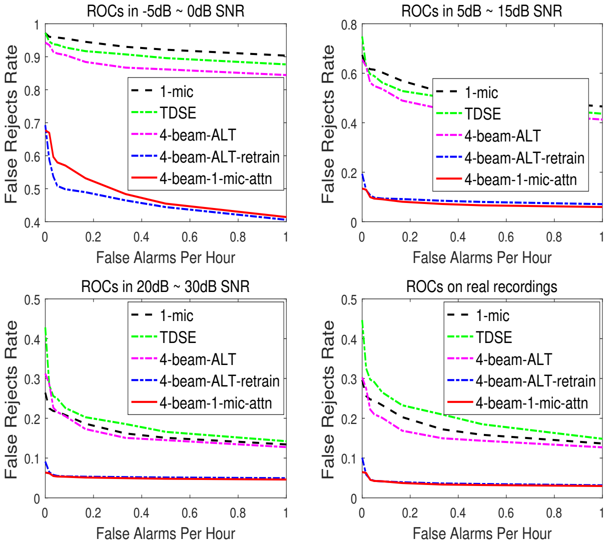

Fig. 4. ROC curves comparing single-channel KWS approaches with the presented MC-KWS model.

4. EXPERIMENTAL SETUP

A keyword of four Chinese characters is employed in this work, with their representation in pinyin as “ni3”, “hao3”, “wei1” and “ling2”. The keyword data were collected in quiet and noisy regular living rooms by a headset microphone and a circular microphone array of 7cm diameter placed at the distance $0.5m$, $1m$, $3m$, $5m$ and $7m$ to the speaker, respectively. Such recordings are used for preparing the following training and evaluation data sets.

Training Corpora of Baseline KWS: 45K utterances from headset recordings (near-field clean data) and 179K utterances from the distant microphone array (far-field noisy data) served as 200-hour positive samples. A 139-hour dataset of 100K non-keyword utterances from a Mandarin speech database served as negative samples.

Training Corpora of MC-KWS: The MC-KWS model was initialized from the baseline KWS model except the attention based multi-channel projection layer. 45K headset utterances were used to create the reverberant and noisy training data in desired circular array. The room simulator based on the image method [23] generated 15K room impulse responses (RIRs) with reverberation time RT_{60}s ranging from 0 to 600 ms. Noise signals, including environmental noises and non-keyword background speech signals, were mixed with the clean keyword utterances at uniformly distributed SNRs ranging from 5 to 30 dB. In total, we have 90K positive training samples. A smaller set of 10K negative samples was collected for fine-tuning MC-KWS model.

Evaluation Corpora: We evaluated our models using simulated and real far-field noisy data, respectively. The SNR of simulated data ranges from -5 to 30 dB in 5 dB increment. 4789 utterances are simulated for each SNR category. The

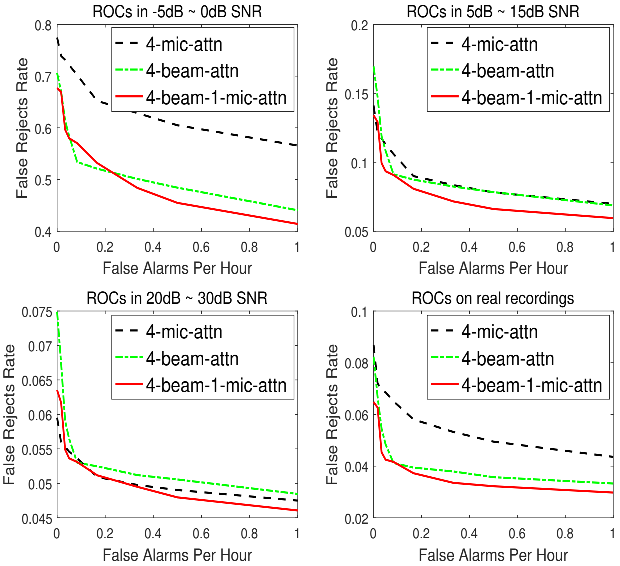

Fig. 5. ROC curves comparing multi-channel approaches

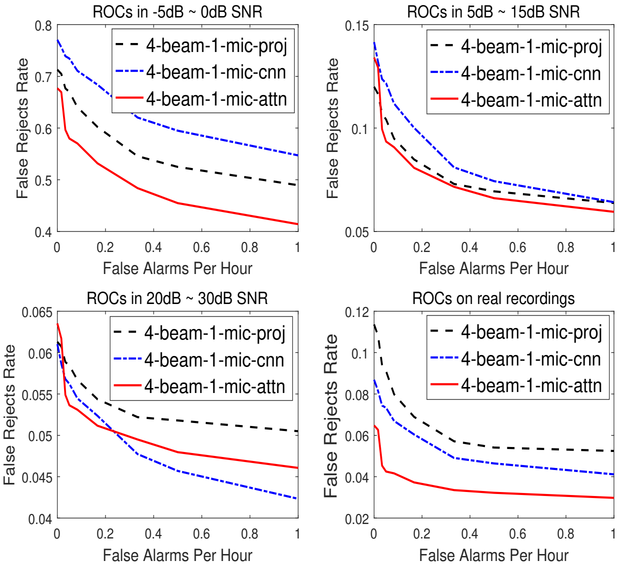

Fig. 6. ROC curves comparing feature projections

real recording set consists of 4570 utterances higher SNRs ranging from 5dB to 20dB. 60 hours of negative samples are incorporated for evaluation. No overlap of speakers and noise sources exists between training and testing.

5. RESULTS AND DISCUSSION

While the baseline single-channel KWS model contains 700K parameters, the presented MC-KWS model has 940K parameters, only 240K over the baseline system. It is thus favored for the embedded application due to the small memory footprint and low computational cost. KWS performance is measured by plotting the receiver operating curve (ROC), which calculates the false reject (FR) rate per false alarm (FA) rate. The lower the FR per FA rate, the better the system.

First, we compare the presented "multi-beam + mic" MC-KWS model with the single-channel KWS approaches. Figure 4 shows that the MC-KWS method (4-beam-1-mic-attn) outperforms the baseline KWS model with one microphone input (1-mic), neural network based monaural speech enhancement [12] (TDSE) and 4 beamformed inputs with one channel input at a time (4-beam-ALT) by a significant margin, respectively. We further retrained the baseline KWS model by using the beamformed channel that is closest to the target direction as the input. Same as scheme 4-beam-ALT, the new model (4-beam-ALT-retrain) detects keywords by evaluating 4 beamformed channels alternately and indicating a successful detection if any of the 4 trials triggers the threshold. The retrained model reduced the acoustic channel mismatch between training and testing and improved the single channel KWS performance in all conditions. Nevertheless, the MC-KWS approach saves over 70% run-time computation cost compared to 4-beam-ALT and 4-beam-ALT-retrain approaches since the model is only called once in MC-KWS approach.

Next, we illustrate the efficacy of the one additional microphone channel in the MC-KWS training. Based on the proposed multi-channel training scheme in Figure 3, Figure 5 shows the results obtained with 3 different input channel assembling approaches, namely 4-mic-attn for training with 4 microphone channels, 4-beam-attn for training with 4 beam channels and 4-beam-1mic-attn for training with 4 beam plus 1 microphone channels, respectively. The best system 4-beam-1mic-attn performs well in both low SNR conditions and relatively clean conditions, while a performance drop is observed in high SNR conditions if the microphone channel is not employed in training (4-beam-attn).

Finally, we investigate various channel mapping approaches. Besides the attention based channel projection (4-beam-1-mic-attn), a linear projection by a fully connected layer (4-beam-1-mic-proj), and channel-wise attention in the KWS model's CNN feature mapping [24] (4-beam-1-mic-cnn) are implemented for comparison as shown in Figure 6. The results indicate that the proposed attention-based channel mapping is more effective than other two methods.

6. CONCLUSIONS

We proposed to integrate multi-look beamformers for robust keyword spotting which efficiently exploits multi-channel microphone signals. Experimental results show that the proposed MC-KWS model significantly outperforms the baseline KWS system and the KWS systems with front-end enhancement under far-field noisy conditions. We believe that the integration of MC-KWS and neural network based monaural speech enhancement is desirable for future exploration.

7. REFERENCES

[1] J. Rohlicek, W. Russell, S. Roukos, and H. Gish, "Continuous hidden markov modeling for speaker-independent word spotting," in *the Proceedings of ICASSP*, 1989, pp. 627–630.

[2] Y. Wang, P. Getreuer, T. Hughes, R. Lyon, and R. Saurous, "Trainable frontend for robust and far-field keyword spotting," in *the Proceedings of ICASSP*, 2017, pp. 5670–5674.

[3] R. Prabhavalkar, R. Alvarez, C. Parada, P. Nakkiran, and T. N. Sainath, "Automatic gain control and multi-style training for robust small-footprint keyword spotting with deep neural networks," in *the Proceedings of ICASSP*, 2015, pp. 4704–4708.

[4] T. Sainath and C. Parada, "Convolutional neural networks for small-footprint keyword spotting," in *Proc. Interspeech*, 2015.

[5] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[6] Y. Xu, J. Du, L. Dai, and C. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 1849–1858, 2015.

[7] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. L. Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr," *in Latent Variable Analysis and Signal Separation, Springer*, pp. 91–99, 2015.

[8] Y. Hoshen, R. J. Weiss, and K. W. Wilson, "Speech acoustic modeling from raw multichannel waveforms," in *the Proceedings of ICASSP*, 2015, pp. 4624—4628.

[9] T. N. Sainath, R. J. Weiss, K. W. Wilson, B. Li, A. Narayanan, E. Variani, M. Bacchiani, I. Shafran, A. Senior, K. Chin, A. Misra, and C. Kim, "Multichannel signal processing with deep neural networks for automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2017.

[10] S. Kim and I. Lane, "End-to-end speech recognition with auditory attention for multi-microphone distance speech recognition," in *Interspeech*, 2017, pp. 3867–3871.

[11] S. Chen, A. Subramanian, H. Xu, and S. Watanabe, "Building state-of-the-art distant speech recognition using the chime-4 challenge with a setup of speech enhancement baseline," in *Interspeech*, 2018, pp. 1571–1575.

[12] M. Yu, X. Ji, Y. Gao, L. Chen, J. Chen, J. Zheng, D. Su, and D. Yu, "Text-dependent speech enhancement for small-footprint robust keyword detection," in *Interspeech*, 2018, pp. 2613–2617.

[13] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, "Robust mvdr beamforming using time-frequency masks for online/offline asr in noise," in *ICASSP*. IEEE, 2016, pp. 5210–5214.

[14] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *ICASSP*. IEEE, 2016, pp. 196–200.

[15] H. Erdogan, J. R. Hershey, S. Watanabe, M. I. Mandel, and J. Le Roux, "Improved mvdr beamforming using single-channel mask prediction networks." in *Interspeech*, 2016, pp. 1981–1985.

[16] X. Xiao, C. Xu, Z. Zhang, S. Zhao, S. Sun, S. Watanabe, L. Wang, L. Xie, D. L. Jones, E. S. Chng *et al.*, "A study of learning based beamforming methods for speech recognition," in *CHiME 2016 workshop*, 2016, pp. 26–31.

[17] Z. Chen, J. Li, X. Xiao, T. Yoshioka, H. Wang, Z. Wang, and Y. Gong, "Cracking the cocktail party problem by multi-beam deep attractor network," in *IEEE Workshop on ASRU*, 2017.

[18] Z. Chen, T. Yoshioka, X. Xiao, J. Li, M. L. Seltzer, and Y. Gong, "Efficient integration of fixed beamformers and speech separation networks for multi-channel far-field speech separation," in *the Proceedings of ICASSP*, 2018.

[19] G. Elko, *Differential Microphone Arrays*. US: Springer, 2004.

[20] G. Elko and J. Meyer, "Microphone arrays," *Springer Berlin Heidelberg*, pp. 1021–1041, 2008.

[21] F. Chowdhury, Q. Wang, I. L. Moreno, and L. Wan, "Attention based models for text-dependent speaker verification," *arXiv preprint arXiv:1710.10470*, 2017.

[22] C. Shan, J. Zhang, Y. Wang, and L. Xie, "Attention-based end-to-end models for small-footprint keyword spotting," in *Interspeech*, 2018, pp. 1571–1575.

[23] J. B. Allen and D. A. Berkley, "Image method for efficiently simulation room-small acoustic," *Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[24] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T. Chua, "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning," in *CVPR*, 2017.