Multi-Resolution Stacked 1D-CNN for Small-Footprint keyword Spotting with Two-Stage Detection

Jian Tang and Shaofei Xue

AIspeech Ltd., Suzhou, China

{jian.tang, shaofei.xue}@aispeech.com

Abstract

Keyword spotting (KWS) is an important technique to free users' hands in man-machine communication. It is quite challenging to build a system with both low False Reject Ratio (FRR) and low False Alarm Ratio (FAR) for real scenarios, especially when computational resources are limited. In this paper, we propose a two-stage KWS system to obtain the trade-off between low computation and high performance. To meet the low-computation requirement, we propose an acoustic model based on multi-resolution GLU stacked 1D convolutional neural network (MRG-S1D). The second requirement is achieved by a second stage classification strategy, in which the neural network features are selected as classifier input for final wakeup word detection. Without increasing the relative FRR, it can reduce the FAR by introducing a few network parameters only. Experiments on a 10K hours Mandarin dataset show that the proposed model can achieve a 39.8% relative FRR reduction compared to the traditional Stacked 1D-CNN. With the second stage classifier, we are further able to reduce the FAR relatively by about 70%. In total, our proposed system significantly leads to a 62.1% relative FRR reduction at 0.1 false alarm per hour.

Index Terms: Keyword spotting, convolution neural network, multi-resolution, two-stage system

1. Introduction

Keyword Spotting (KWS) has become an important technology since it provides hands-free, natural, and ubiquitous access to the interacting device. A streaming production-quality KWS module must minimize the false rejection ratio at a low false alarm ratio. Meanwhile, it must be highly accurate, low-latency, small-footprint, and run in computationally constrained environments such as modern mobile devices. Despite the significant advancement made in KWS after the introduction of deep neural network (DNN) based models [1, 2, 3, 4, 5] , such as convolution neural network (CNN), the KWS remains a challenging task due to various reasons [6]. In the real application scenarios, the speech signal is often captured by microphone arrays located far away from the speaker. It is susceptible to distortions such as environmental noises, reverberations, and unexpected energy variations.

A series of extended technologies have been developed to support the KWS task. Most existing methods could be grouped into three categories. Firstly, many previous works adapt the large vocabulary continuous speech recognition (LVSCR) techniques for detecting keywords [7, 8, 9]. However, the LVCSR based systems need to generate rich lattices and high computational resources are required for the keyword search. Secondly, the combination of acoustic model and post-processing module is widely applied in more recent approaches. The acoustic model computes the frame-wise posterior probabilities and the post-processing module uses posteriors to compute keyword detection scores. Several kinds of models have been successfully applied for acoustic modeling [10, 11, 12, 13]. More recently, models like singular value decomposition filters (SVDFs) [14] and stacked 1D convolutional neural network (S1DCNN) [15] further speed up the computation. A commonly used posterior handling method is proposed in [16]. No sequence search algorithm requirement leads to a simpler implementation, but the temporal information within wakeup words has not been considered. In [17], a two-stage system is proposed, where multiple artificially defined features are fed into a SVM classifier in the second stage. However, these features are too artificially designed, and the variety of application scenarios is limited. At last, end-to-end approaches like attention-based models (ABM) [18] and recurrent neural network transducer (RNN-T) [19] have also been used, where all components of the detection system are jointly optimized to produce the detection likelihood score.

Among the above-mentioned methods, the combination of KWS model and post-processing module is the most mature one. Nevertheless, there is still space for improvement. The S1DCNN has achieved a balance between computation and performance, but the model parameters are potential to be further reduced. In post-processing, the existing two-stage method requires the artificial setting of features [17, 20]. We attempt to select the input features from the acoustic model directly.

Based on the above-mentioned considerations, we propose a two-stage KWS system. In the first stage, a novel acoustic model named multi-resolution stacked 1D convolutional neural network (MRG-S1D) is implemented. It enables us to build a streaming production-quality KWS system with a small memory footprint, low computational cost, and high precision. Next, in the second stage, a posterior correction (PC) classifier is utilized to alleviate the false alarms (FA) problem, which introduces only a few network parameters. Besides, we raise a dimensionality reduction method to reduce the model parameters' increment. Experimental results on a 10K hours Mandarin dataset show that our proposed model significantly leads to up 62.1% relative FRR reduction at the same 0.1 FA per hour.

The remainder of this paper is organized as follows. Section 2 discusses related works. Section 3 explores the proposed acoustic model. In Section 4, the proposed two-stage KWS system is detailed, including the MRG-S1D based acoustic model and the PC classifier. Experimental results are presented in Section 5. Finally, Section 6 concludes this work.

2. Related Work

2.1. Stacked 1D Convolutional Layer

S1DCNN, as shown in Fig. 1, could efficiently aggregate information in time and frequency dimensions through two 1D-convolution layers. Let us assume that the kernel size of the

979-8-3503-9796-3/22/$31.00

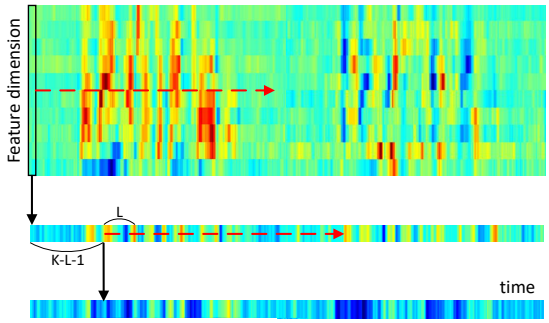

Figure 1: *The illustration of stacked 1D CNN layer [15].*

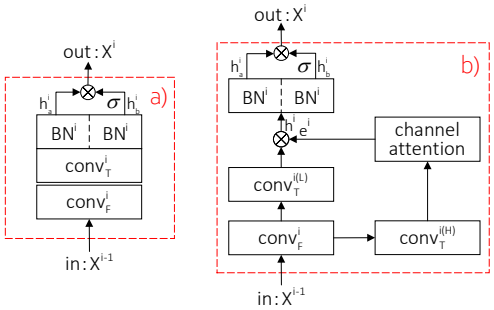

Figure 2: *Illustration of two kinds of hidden layer: a) S1DCNN with GLU activation (short as S1DGLU), b) S1DCNN with GLU activation and multi-resolution (MRG-S1D).*

convolution layer in the frequency and time domain are F and K respectively, and the first convolution layer is composed of N filters of filter size K. The n-th filter output of the first convolution ($\mathrm{conv_F}$) can be written as

$$a_t^{(n)} = \sum_{f=1}^{F} w_f^n x_{ft}, \tag{1}$$

where w_f^n denotes the n-th filter weight in frequency dimension of an input vector X, and the bias variable in this section is omitted for clarity. Thus, this first convolution accumulates information over the feature dimension.

Set the second 1D convolution ($\mathrm{conv_T}$) has N filter of size K. The filter is applied to the outputs of the first 1D convolution layer in a so-called "depth-wise" manner [21], where the n-th filter of the second 1D CNN is applied only to the n-th filter output of the first 1D CNN. By performing depth-wise 1D convolution, an output of the n-th filter can be expressed as

$$h_t^{(n)} = \sigma\Big(\sum_{f=1}^{K} u_k^{(n)} a_{t-K+k+L}^{(n)}\Big), \tag{2}$$

where σ denote the sigmoid function, L denote time offset, and $u_k^{(n)}$ denote k-th component of n-th CNN filter weight. The second layer looks up $K-L-1$ past outputs and L future outputs from the first layer.

2.2. DNN-based AWUWSR system

In [17], an automatic wakeup-word speech recognition (AWUWSR) system was adopted to integrate the phonetic knowledge and model-based classification into detecting wakeup words. Six effective confidence measures from different perspectives are presented. All the confidence measures for each phoneme and the whole segment are concatenated into a confidence vector, and utilized as the input of one SVM classifier.

Our proposed two-stage LWS system also uses a classifier in the second stage to control the emergence of false alarms. Unlike AWUWSR, we utilize a neural network as the classifier, and directly use the hidden layer's outputs of the KWS model to select specific moments. Hence, there is no need to define the confidence measures artificially.

3. The Proposed Acoustic Models

3.1. GLU Activation

The S1DCNN can be regarded as a factorization of a 2D CNN. An $F\times K$ filter of the 2D CNN layer is factorized into a 1D CNN layer with $F\times 1$ filters and another layer with $1\times K$ filters. F is consistent with the channel numbers of previous layer. Since F is much larger than K, halving F value could further decrease the number of model parameters. We add GLU activation [22] after i-th $\mathrm{conv_T^i}$ layer, the output of which is marked as h^i. The GLU is expressed as follows:

$$X^i = h_a^i \otimes \sigma(h_b^i), \tag{3}$$

where h^i is split in half along the channel dim to form h_a^i and h_b^i, σ is the sigmoid function and $\otimes$ is the element-wise product.

This activation could not only bring nonlinear operation but also decrease the model parameters by halving the channel numbers. Thus we can obtain the Stacked 1D GLU Convolution Layer (S1DGLU), as shown in Fig.2 a).

3.2. Multi-resolution GLU S1DCNN

In S1DGLU, the time domain convolution in the second layer has a low proportion of parameters. The increasement of the time domain receptive kernel K, influences the number of parameters lightly. Through channel-wise attention mechanism [23], the contextual information of high receptive fields is utilized to weigh the information of low receptive fields. The structure of the temporal multi-resolution version is shown in Fig.2 b). The channel-wise attention mechanism could integrate both the contextual information in the long and short scales. In the i-th layer,two group convolutions are added after the frequency domain convolution layer ($\mathrm{conv_F^i}$) . Thus, high and low receptive fields,$\mathrm{conv_T^{iH}}$ and $\mathrm{conv_T^{iL}}$, in time domain are obtained respectively [24] through two different dilation rates. The output of those two time domain convolution layers are denoted as $h^{i\mathrm{H}}$ and $h^{i\mathrm{L}}$.

Each $h^{i\mathrm{L}} \in R^{D\times 1}$ represents the output of $\mathrm{conv_T^{iL}}$. All the channels are split into G groups, those group has D/G channels, denoted as D'. Thus, the layer architecture can be expressed in the following way:

$$e^i = v^{\mathrm{T}} \tanh(Wh^{i\mathrm{H}} + Uh^{i\mathrm{L}} + b^i), \tag{4}$$

$$h^i = h^{i\mathrm{L}} \mathrm{softmax}(e^i), \tag{5}$$

where $W \in \mathbb{R}^{D\times D'}$ and $U \in \mathbb{R}^{D'\times D'}$ are transformation matrices that map the feature and high-resolution state into a same dimension, $v \in \mathbb{R}^{D'\times 1}$ and b^i are vectors.

4. Two-stage KWS System

The proposed two-stage system is shown as Fig. 3. In the first stage, our models, S1DGLU and MRG-S1D, are utilized to de-

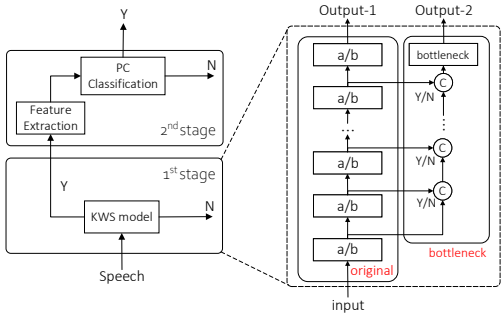

Figure 3: *The illustration of the two-stage KWS system.*

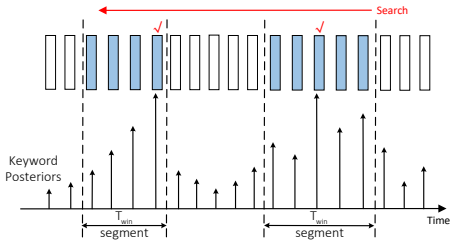

Figure 4: *The feature extraction method of posterior correction (PC) classifier.*

tect pre-defined keywords. Once the word is detected, our system enters the second stage, where a binary classifier is used to further determine whether the audio contains wakeup words. The double-check could reduce FA for the system.

Specifically, The 1^{st} stage refer to the decoding process of the conventional KWS model. The decoding process utilizes its posterior probability to decide whether the current statement contains the wakeup word. The classifier in the 2^{nd} stage is trained based on feature vectors, which are extracted from the wakeup word candidates, and then makes the final decision. Here we use a small feed-forward neural network (NN) with a single hidden layer.

The pre-defined keyword is assumed with S characters. The local maximum position of the posterior probability within $\mathrm{T_{win}}$ frame for each character is searched sequentially in reverse order. Thus, S positions can be found. The outputs of the first and the last layer, which correspond to S positions, are concatenated and used as input features for a classifier, which we called the posterior correction (PC) classifier. This feature extraction process is shown in Fig. 4. The output of the PC classifier directly indicates whether the audio signal contains the wakeup word.

The second stage classification brings additional parameters, and the incresement is positively correlated with the dimension of the hidden layer in the KWS model. To reduce the model parameter, we propose a dimensionality reduction method based on multi-task learning. While reducing parameter increment, this method could maintain the high recall.

The left and right parts in Fig. 3 are marked as the original and the bottleneck, respectively. During the training stage, two sets of parameters are updated : the original weights, the bottleneck weights. For the model optimization, we use a multiple phases training strategy (abbr. multi-phase), which consists of three phases. In the first phase, the original weights are trained without the inserted bottleneck weights, which is the same as the training process of a traditional KWS model. Then, the bottleneck layers are added and the additional weights are all randomly initialized in the second phase. The original weights are fixed when bottleneck weights are trained. Finally, all the weights are further fine-tuned to optimize the whole network. In contrast with multi-phase, the joint learning procedure (abbr. joint-learn) is applied, in which all original weights and bottleneck weights are jointly learned from scratch for comparison.

5. Experimental evaluation

About 15 million utterances (10k hours in total) in Mandarin are used for training. The utterances are composed of common and custom training samples, which are recorded by smartphones, tablet devices, TV sets, etc. Custom training samples indicate the utterances with the target phrase, in this case "ni hao xiao chi". Besides, S is equal to 4 and $\mathrm{T_{win}}$ is set to be 400ms. Common training samples do not contain the specific phrase, irrelevant contents instead (negative samples). The custom training samples were augmented mainly in two ways. The first one is audio cross-channel data simulation, which is realized by convolving room impulse responses; the other is adding noise in different scenes (living room, car music, street, etc.) on the recordings, and then reducing noises by beamforming algorithms. The ratio of common and custom training samples is about 75:1. For the training data in the 2^{nd} stage, we configured the ratio of positive and negative training samples between 3:1 and 4:1. The positive ones are extracted from the positive training samples for the KWS model in the first stage. Correspondingly, the negative samples are from the audios, which lead to false alarms.

For evaluation, we re-recorded 7500 relative clean utterances in the office scene with three signal noise ratios, 0dB, 5dB, and 10dB, and used them as positive samples. In addition, 110 hours audio without the trigger phrase, are obtained from different noise scenes and utilized for negative samples.

We use 40-dimensional log-mel filterbank coefficients, with sentence mean normalization. M and D' are set to 32 and 30 throughout this paper. For all experiments, features at 1 past frame and 1 future frame were concatenated with the current frame. The experiments were repeated four times, and we report the mean result for each. We compare several variant KWS models, including:

- S1DCNN: the baseline is trained with the size of 7×300 (7 layers × 300 nodes for each layer) S1DCNN. This system shares the same training criteria and strategies as the following systems.
- S1DCNN+: the baseline extended by increasing the number of hidden nodes (7×320) to eliminate the effects of model size.
- S1DGLU: using S1DGLU for the acoustic model, which has the size of 10×360.
- MRG-S1D: This KWS model consisted of 10 MRG-S1D layers, each of which consists of 330 nodes.

5.1. Evaluation of the model structure

This experiment was made for comparing the performance of S1DCNN, S1DGLU, and MRG-S1D. The result in Table 1, where we also measure the model size in Kbytes (KB), indicates that the FRR of S1DGLU is lower than the traditional S1DCNN. Due to the compression capability of the GLU function, S1DGLU is superior to traditional S1DCNN in terms of model parameters and computation. Besides, by comparing

Table 1: *Comparison False reject ratios (FRRs in %) at 0.1 false alarm per hour, model parameters (#Para) and multiply-accumulate operation (MACC per frame) of various models.*

Model	#Para(KB)	MACC	FRR			
			0dB	5dB	10dB	Avg
S1DCNN	716	720	18.0	8.1	5.0	10.4
S1DCNN+	800	807	17.8	8.2	4.9	10.3
S1DGLU	725	729	11.7	6.4	3.4	7.2
MRG-S1D	755	744	10.4	5.5	2.8	6.2

Table 2: *The FRRs, RFRRs (in %) and #Para (additional Parameters/cache) with/without the PC classifier. RFRR means the relative FRR.*

Model	FAR	#Para(KB)	RFRR	FRR
S1DGLU	0.10	-	-	7.2
+ PC classifier	0.03	46.1+18	0.1	7.3
MRG-S1D	0.10	-	-	6.2
+ PC classifier	0.03	42.2+16.5	0.1	6.3

this result with a larger baseline (S1DCNN+), we conduct that the performance gain does not come from the increasement of model parameters. On the other hand, MRG-S1D obtains the best performance among all model performance, which brings about 39.8% relative improvement than S1DCNN+. These results indicate that the proposed multi-resolution model is helpful to improve performance.

5.2. Evaluation of the PC Classifier

For the PC classifier in the 2^{nd} stage, we first apply the outputs of the first and last layers in the KWS acoustic model as input features. Table 2 compares performance with and without the PC classifier on both S1DGLU and the MRG-S1D. It turns out that, when the FRR increases lightly, about 70% relative reduction of the FA could be achieved at 0.1 FA per hour.

The PC classifier increases the model parameters and the cache. The cache is utilized to save the input features of the PC classifier. For the KWS model, we need to consider how to further reduce its memory usage. Here, we measured the performances of multi-phase and joint-learn optimization methods. Four system settings for the PC classifier are compared. The first one is the joint-learn optimization with 5 hidden layers. The other three methods are about the multi-phase strategy with 5, 3, and 2 hidden layers. Among those, five indicates 5 hidden layers at equal intervals, three means the first, middle and last three hidden layers, and two refers to the first and last 2 layers.

The results in the first three rows of Table 3 indicate that the performance of two optimization strategies, where five layers are concatenated directly and compressed, is slightly decreased than MRG-S1D, and the multi-phase method outperforms the joint-learn one. Besides, the results in row 3-5 show that, the information obtained from the two layers is enough for the PC classifier. Simultaneously, additional model parameters and cache are both greatly reduced. The similar phenomenons can be observed on the S1DGLU. In conclusion, we finally choose

Table 3: *Comparison of variant optimization methods at 0.03 false alarm per hour.*

Model	#Para(KB)	RFRR	FRR
MRG-S1D	42.2+16.5	0.1	6.3
+ joint-learn(5)	30.5+1.6	0.9	7.0
+ multi-phase(5)	30.5+1.6	0.3	6.5
+ multi-phase(3)	20.0+1.6	0.4	6.6
+ multi-phase(2)	14.7+1.6	0.3	6.5
S1DGLU	46.1+18	0.1	7.3
+ joint-learn(5)	32.9+1.6	1.0	8.1
+ multi-phase(5)	32.9+1.6	0.8	7.9
+ multi-phase(3)	21.3+1.6	0.2	7.4
+ multi-phase(2)	15.6+1.6	0.3	7.5

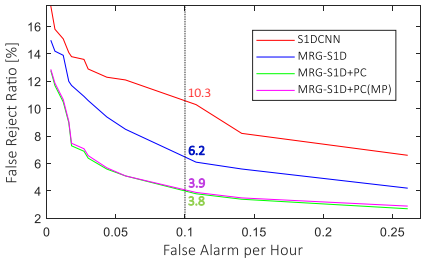

Figure 5: *DET curves obtained by various KWS systems. The individual abbreviations mean: "S1DCNN+": the larger baseline; "MRG-S1D": using the proposed MRG-S1D as the acoustic model; "MRG-S1D + PC": the MRG-S1D plus the PC classifier; "MRG-S1D + PC(MP)":using the dimensionality reduction method and the multi-phase optimization in 2^{st} stage.*

the 2-layer multi-phase method, as it could obtain the trade-off between parameter increment and performance.

5.3. System Performance

Fig. 5 shows detection error trade-off (DET) curves obtained by various KWS systems. Table 3 shows FRRs at an operating point, i.e., 0.1 FA per hour. Compared with the S1DCNN+, the MRG-S1D achieved about 39.8% relative FRR reduction at 0.1 FA per hour. By simply adding 2% percent parameters, the two stage KWS system ("MRG-S1D + PC(MP)") achieved further performance gain by 62.1% relative to the S1DCNN+.

6. Conclusions

In this work, we propose a two-stage KWS system that consists of an acoustic model and a PC classifier. To reduce the model parameters and obtain high precision, this paper proposes a novel two-stage KWS system. Through the novel acoustic model and the PC classifier, our system can obtain high performance while only introducing a few network parameters. We found that both GLU activation and the multi-solution architecture could bring high benefits for our KWS system, and the PC classifier with multi-phase strategy can reduce the FAR by introducing a few network parameters only. In the future, we will consider combining the acoustic model with the PC classifier to form an end-to-end KWS system.

7. References

[1] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 1, pp. 30–42, 2011.

[2] O. A. H., A. R. M., H. J., L. D., G. P., and D. Y., "Convolutional neural networks for speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 10, pp. 1533–1545, 2014.

[3] T. N. S. and C. P., "Convolutional neural networks for small-footprint keyword spotting," pp. 1478–1482, 2015.

[4] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural networks*, vol. 18, no. 5-6, pp. 602–610, 2005.

[5] S. Xue and Z. Yan, "Improving latency-controlled blstm acoustic models for online speech recognition," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2017, pp. 5340–5344.

[6] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech & Language*, vol. 46, pp. 535–557, 2017.

[7] S. Parlak and M. Saraclar, "Spoken term detection for turkish broadcast news," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2008, pp. 5244–5247.

[8] G. Chen, O. Yilmaz, J. Trmal, D. Povey, and S. Khudanpur, "Using proxies for oov keywords in the keyword search task," in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop, (ASRU)*, 2013, pp. 416–421.

[9] C. Ni, C.-C. Leung, L. Wang, H. Liu, F. Rao, L. Lu, N. F. Chen, B. Ma, and H. Li, "Cross-lingual deep neural network based submodular unbiased data selection for low-resource keyword search," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2016, pp. 6015–6019.

[10] N. P., A. R., P. R., and P. C., "Compressing deep neural networks using a rank-constrained topology," in *ISCA Interspeech*, 2015, pp. 1473–1477.

[11] M. Sun, D. Snyder, Y. Gao, V. K. Nagaraja, M. Rodehorst, S. Panchapagesan, N. Strom, S. Matsoukas, and S. Vitaladevuni, "Compressed time delay neural network for small-footprint keyword spotting." in *Interspeech*, 2017, pp. 3607–3611.

[12] Y. Zhuang, X. Chang, Y. Qian, and K. Yu, "Unrestricted vocabulary keyword spotting using lstm-ctc." in *ISCA Interspeech*, 2016, pp. 938–942.

[13] M. Chen, S. Zhang, M. Lei, Y. Liu, H. Yao, and J. Gao, "Compact feedforward sequential memory networks for small-footprint keyword spotting." in *ISCA Interspeech*, 2018, pp. 2663–2667.

[14] H.-J. Park, P. Violette, and N. Subrahmanya, "Learning to detect keyword parts and whole by smoothed max pooling," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2020, pp. 7899–7903.

[15] T. Higuchi, M. Ghasemzadeh, K. You, and C. Dhir, "Stacked 1d convolutional networks for end-to-end small footprint voice trigger detection," *arXiv preprint arXiv:2008.03405*, 2020.

[16] G. Chen, C. Parada, and G. Heigold, "Small-footprint keyword spotting using deep neural networks," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2014, pp. 4087–4091.

[17] F. Ge and Y. Yan, "Deep neural network based wake-up-word speech recognition with two-stage detection," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2017, pp. 2761–2765.

[18] C. Shan, J. Zhang, Y. Wang, and L. Xie, "Attention-based end-to-end models for small-footprint keyword spotting," in *ISCA Interspeech*, 2018, pp. 2037–2041.

[19] Y. He, R. Prabhavalkar, K. Rao, W. Li, A. Bakhtin, and I. McGraw, "Streaming small-footprint keyword spotting using sequence-to-sequence models," in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop, (ASRU)*, 2017, pp. 474–481.

[20] M. Wu, S. Panchapagesan, M. Sun, J. Gu, R. Thomas, S. N. P. Vitaladevuni, B. Hoffmeister, and A. Mandal, "Monophone-based background modeling for two-stage on-device wake word detection," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2018, pp. 5494–5498.

[21] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[22] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *Int. Conf. Machine Learning (ICML)*, 2017, pp. 933–941.

[23] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning," in *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5659–5667.

[24] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.