# A STUDY OF DESIGNING COMPACT AUDIO-VISUAL WAKE WORD SPOTTING SYSTEM BASED ON ITERATIVE FINE-TUNING IN NEURAL NETWORK PRUNING

*Hengshun Zhou*[1], *Jun Du*[1,*], *Chao-Han Huck Yang*[2], *Shifu Xiong*[3], *Chin-Hui Lee*[2]

[1]University of Science and Technology of China, Hefei, Anhui, P. R. China
[2]Georgia Institute of Technology, Atlanta, GA. USA
[3]iFlytek Research, Hefei, Anhui, P. R. China

zhhs@mail.ustc.edu.cn, ✉jundu@ustc.edu.cn, huckiyang@gatech.edu, sfxiong@iflytek.com, chl@ece.gatech.edu

## ABSTRACT

Audio-only based wake word spotting (WWS) is challenging under noisy conditions due to the environmental interference in signal transmission. In this paper, we investigate on designing a compact audio-visual WWS system by utilizing the visual information to alleviate the degradation. Specifically, in order to use visual information, we first encode the detected lips to fixed-size vectors with MobileNet and concatenate them with acoustic features followed by the fusion network for WWS. However, the audio-visual model based on neural network requires a large footprint and a high computational complexity. To meet the application requirements, we introduce a neural network pruning strategy via the lottery ticket hypothesis in an iterative fine-tuning manner (LTH-IF), to the single-modal and multi-modal models, respectively. Tested on our in-house corpus for audio-visual WWS in a home TV scene, the proposed audio-visual system achieves significant performance improvements over the single-modality (audio-only or video-only) system under different noisy conditions. Moreover, LTH-IF pruning can largely reduce the network parameters and computations with no degradation of WWS performance, leading to a potential product solution for the TV wake-up scenario.

***Index Terms***— Wake word spotting, noisy environments, audio-visual, LTH pruning, iterative fine-tuning

## 1. INTRODUCTION

Wake word spotting (WWS) can be considered as a specific case of keyword spotting (KWS), concerning the identification of predefined wake word(s) in utterances, often used for the wake-up of speech-enabled devices, such as "Hey Siri" in iPhone, "Alexa" in Amazon Echo, and "Ok Google" in Google Home [1, 2, 3, 4], etc. In order to activate the interactions between devices and users, a standby wake word detection module is particularly important [5].

Traditional approaches in keyword spotting tasks involve the keyword/filler hidden Markov model (HMM) [6, 7], namely training an HMM for the keyword and a filler HMM for the non-keyword segments, respectively. Recently, deep learning based keyword spotting have attracted much attention. Chen et al. proposed a simple discriminative keyword spotting approach based on deep neural networks which have improved the performance of system [8]. The first attempt to use convolutional neural networks (CNNs) for keyword spotting, by Sainath and Parada [9], was recently improved by jointly integrating deep residual learning and dilated convolutions [10]. Arik et al. [11] also applied the convolutional recurrent neural network (CRNN) architecture to single English keyword detection. With the achievements of Transformer [12] in the field of deep learning, several variants of Transformers for wake word detection are explored in [13]. Besides, more efficient networks have been also investigated by leveraging recent advances in differentiable neural architecture search [14].

Despite the above research progress, KWS is still a challenging task and has attracted the attention of speech researchers. On the one hand, the KWS systems usually perform very well under clean speech conditions. However, the systems suffer from sharp performance degradation under noisy environments. The authors in [15] propose integrating multiple beamformed signals and leveraging the attention mechanism to dynamically tune the model's attention to the reliable input sources to improve the KWS performance under noisy and far-field conditions. The data-efficient solutions are presented in [16] to improve the model robustness in WWS modeling under noisy conditions. A multi-task network that performs KWS and speaker verification (SV) simultaneously is also proposed in [17] to fully utilize the interrelated domain information aiming at performance improvement in challenging conditions. In addition, the authors in [18] have developed a novel tuple-based loss function along with a training strategy for noise-robust keyword spotting.

On the other hand, the WWS system usually runs on smart devices, it's critical to design the model with a small footprint and low computational power. The application of multi-scale temporal modeling to the small-footprint keyword spotting task has been explored in [19]. The authors also explore the latency and accuracy of KWS models in streaming and non-streaming modes for simplifying model deployment on mobile devices [20]. In [21], the researchers designed different models and neural architectures for small footprint keyword spotting. Different loss functions for the training of a small-footprint KWS system have also been explored in [22]. Recently, in order to optimize towards memory footprint and execution time, power-consuming audio preprocessing and data transfer steps are eliminated in [23] by directly classifying from raw audio.

In order to alleviate the performance degradation of WWS under noisy conditions, in this paper, we investigate an audio-visual WWS system by utilizing the visual information. First, the detected lips are encoded to fixed-size vectors with MobileNet, and concatenated with acoustic features. Next, a neural network pruning strategy, i.e the lottery ticket hypothesis based iterative fine-tuning (LTH-IF) is introduced to the WWS systems. Finally, tested on our in-house corpus for audio-visual WWS in a home TV scene, the proposed audio-visual system achieves significant performance improvements over the single-modality system under different noisy conditions. Moreover, LTH-IF pruning can largely reduce the network parameters and

*corresponding author

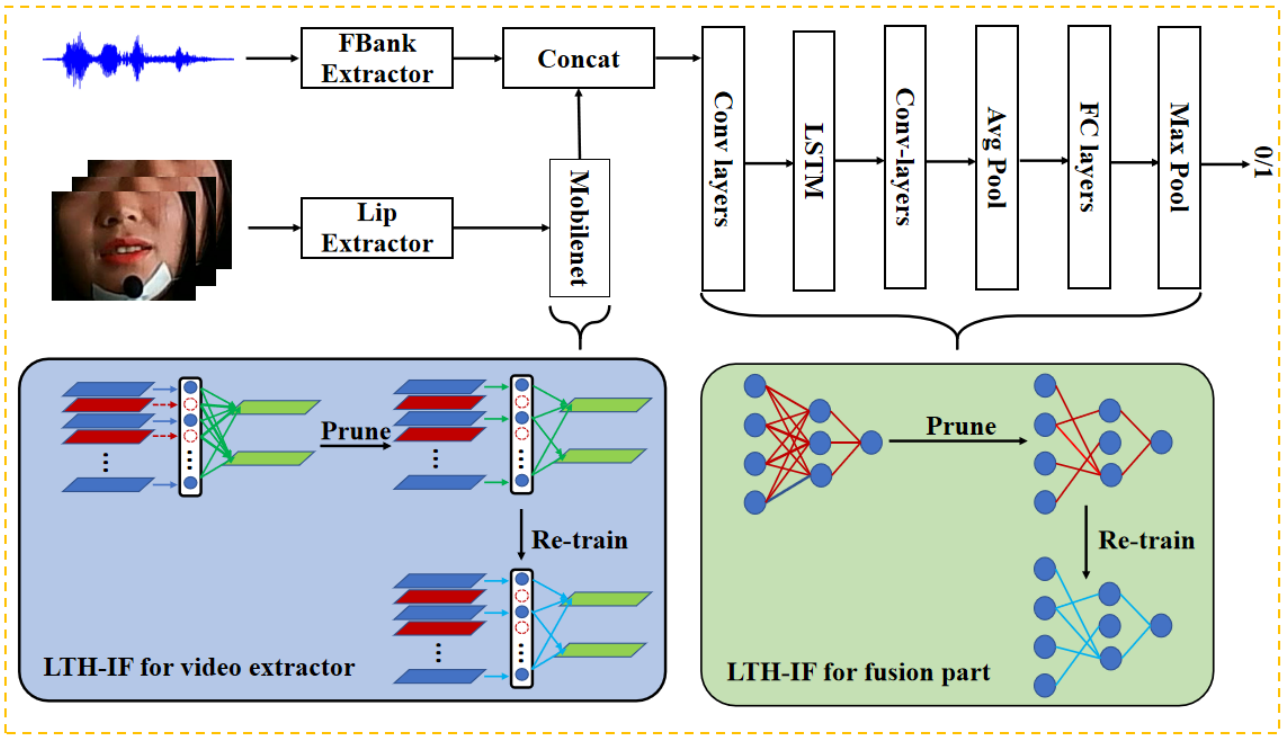


**Fig. 1**. The architecture of proposed audio-visual wake word spotting with neural network pruning using LTH-IF.

computations with no degradation of WWS performance, leading to a potential product solution for the TV wake-up scenario.

## 2. PROPOSED APPROACH

### 2.1. Audio-Visual Model for Wake Word Spotting

Inspired by the work in [24], we design the proposed audio-visual wake word spotting (WWS) architecture in an end-to-end manner. The main difference from [24] lies on that we have the only one wake word, so we do not need to use the phonetic sequence or calculate the similarity matrix. Accordingly we directly adopt the classification task through the fully connected (FC) layers. The overall flowchart of proposed audio-visual WWS architecture is shown in Fig. 1, which mainly consists of three parts: audio stream, video stream and fusion stream. The details will be elaborated in the following subsections.

#### *2.1.1. Audio Stream*

For the audio stream, the acoustic features extracted frame by frame are selected as input features. Here, we employ 40-dimensional filter bank (FBank) features normalized by global mean and variance. Given the raw input audio data $I_{\mathrm{A}}$, we can calculate normalized FBank features $F_{\mathrm{A}}$ through the FBank extractor $f_{\mathrm{A}}$:

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

#### *2.1.2. Video Stream*

For the video stream, considering practicality and lightweight, we select a combination of MobileNetV2 [25] as our video embedding extractor. In addition, we replace bidirectional long short-term memory (BLSTM) with LSTM to further reduce the model size and latency. In this study, 13 linear bottlenecks are first adopted as a lip feature extractor $f_{\mathrm{V}}$. Given the input image $I_{\mathrm{V}}$, the lip feature $F_{\mathrm{V}}$ is calculated through $f_{\mathrm{V}}$:

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

The gray scale lip $F_{\mathrm{V}}$ reshaped to $88 \times 88$ is used as the MobileNetV2 input, and the output is an encoded vector by using average pooling. The high-level video embeddings $E_{\mathrm{V}}$ can be obtained by MobileNetV2 $f_{\mathrm{Mobile}}$:

$$E_{\mathrm{V}} = f_{\mathrm{Mobile}}(F_{\mathrm{V}}) \tag{3}$$

For more details, please refer to [25].

#### *2.1.3. Fusion Stream*

For the audio-visual fusion stream, a direct concatenation for audio-visual fusion at the encoder is first considered to integrate information from two sources $F_{\mathrm{A}}$ and $E_{\mathrm{V}}$:

$$F_{\mathrm{AV}} = [F_{\mathrm{A}}, E_{\mathrm{V}}] \tag{4}$$

Then a mixture of convolutional layers, LSTM layers and FC layers are designed to generate the final output as shown in Fig. 1. For the single-modality network, the above audio-visual features $F_{\mathrm{AV}}$ are replaced by audio acoustic features $F_{\mathrm{A}}$ and visual embeddings $E_{\mathrm{V}}$, respectively. In this study, 5 epochs ($E = 5$) are selected to train these three WWS models, namely audio-only model, video-only model, and audio-visual model. The final output of these models is compared with the preset threshold after sigmoid operation,

'1' indicates that the current sample contains wake word, and '0' indicates the opposite. Given a sample, the model ($G$) outputs a probability $p(y = 1|\Theta)$ representing the possibility that the wake word is included, where $\Theta$ represents the model parameter set. The optimisation objective is a binary cross-entropy loss between this prediction and the ground truth label $y^*$:

$$L_{\text{WWS}} = -y^* \log\ p(y=1|\Theta) - (1-y^*)\ \log(1-\ p(y=1|\Theta)) \quad (5)$$

**Algorithm 1** LTH with Iterative Fine-tuning

**Input:** A model, $G_0$

1: Randomly initialize weights $(\Theta_0)$
2: Initialize model: $G_0(\Theta_0) \rightarrow G_1$
3: For $t = 1, \cdots, T : \#\ Pruning\ searching\ iterations$
4: If $t = 1$:
5: For $e = 1, \cdots, E : \#\ Training\ epochs$
6: $\Theta_e \rightarrow \Theta$ : Train $G_t$ for its final weights $(\Theta_t)$
7: Else:
8: For $e = 1 : \#\ Training\ epoch$
9: $\Theta_e \rightarrow \Theta$ : Train $G_t$ for its final weights $(\Theta_t)$
10: If $t < T : \#\ LTH\ pruning\ strategy$
11: Mask$(\Theta_t)$ to get a pruned graph $G_p$ from $G_{t-1}$
12: Load weights $\Theta_p \in \Theta_t$ from $G_p$
13: Update target model $G_p(\Theta_p) \rightarrow G_{t+1}$

**Output:** A well-trained pruned model $G_T(\Theta_T)$

### 2.2. Audio-Visual Model Pruning Using LTH-IF

The recent "Lottery Ticket Hypothesis" [26] showed an encouraging phenomenon that some subnetworks (winning tickets) could be obtained by pruning the original network through specific methods, and then they can be trained to achieve the performance equal to or better than the untrimmed original model. Although convolutional neural networks have shown to be effective to the small-footprint WWS problem, they still need hundreds of thousands of parameters to achieve good performance [27]. Although LTH-based low-complexity neural models have proven competitive prediction performance on several image classification tasks, machine translation [28, 29] and acoustic scene classification [30], and recently have been supported with some theoretical findings [31] related to overparameterization, the effect of LTH on our multimodal task of audio-visual WWS is not unknown. In this study, with a high demand of designing a compact audio-visual WWS model with low-latency for real applications, we investigate on neural network pruning based on LTH with an iterative fine-tuning strategy.

**LTH-IF Algorithm Design:** In Algorithm 1, we detail our approach: First, our WWS model with its original neural architecture $G_0$ is initialized with the weights parameters $(\Theta_0)$. Different from [26], the complete number of iterations ($E = 5$) is selected to train the model only before pruning (i.e. $t = 1$). At the end of each training phase, a pruning iteration is started if the current iteration $t$ is less than $T$. And the final weights $\Theta_T$ are used for the new initialization to fine-tune the model. The LTH pruning searches for a low-complexity model from steps (10) to (13).

For the audio-only WWS model, we prune the whole network directly based on LTH-IF. However for video-only WWS model, unlike audio-only WWS, it includes an additional lip encoder, and we also prune it jointly with the back-end module based on LTH-IF.

Interestingly, for the audio-visual WWS model, we found that separate pruning for the lip encoder first will yield better performance than pruning the whole model directly at the same degree of pruning. Therefore, for the audio-visual WWS model, we first prune the lip encoder using LTH-IF. Then the lip encoder is fixed and initialized using the weights and mask obtained above. Finally, the model is trained for pruning back-end network of fusion stream based on LTH-IF.

## 3. EXPERIMENTS

### 3.1. Databases and Implementation Details

We conduct experiments on an audio-visual dataset collected in smart home TV scenes. There are 250 speakers in total, with a male-to-female ratio of 1:1. The wake word is "Xiao T Xiao T". The speakers in the training set, development set and test set do not overlap, which are 210, 20 and 20 respectively. For the training set, in addition to the original positive and negative samples, we also added several types of noises for data augmentation. Our final training set includes 50 hours of positive samples and 50 hours of negative samples, respectively. For the development/test set, in order to facilitate comparison, we only add noise with three signal-to-noise ratios (i.e. -5dB, 0dB, 5dB), including 2 hours of positive samples and 2 hours of negative samples, respectively. Each positive sample contains only one wake word. The duration of each sample is 1.3 seconds.

We use false reject rate (FRR) and false alarm rate (FAR) on test set as the criterion of the WWS performance. Suppose the test set consists of $N_{\text{wake}}$ examples with wake word and $N_{\text{non-wake}}$ examples without wake word, FRR and FAR are defined as follows:

$$\text{FRR} = \frac{N_{\text{FR}}}{N_{\text{wake}}} \quad (6)$$

$$\text{FAR} = \frac{N_{\text{FA}}}{N_{\text{non-wake}}} \quad (7)$$

where $N_{\text{FR}}$ denotes the number of examples including wake-up word but the WWS system gives a negative decision. $N_{\text{FA}}$ is the number of examples without wake-up word but the WWS system gives a positive decision.

We employ pytorch to train all models and minimize the loss function using the Adam optimization method. The batch size is 64 for audio-only WWS system and 16 for video-only and audio-visual systems. The learning rates are set to 0.0001, 0.0002, 0.0002 for audio-only, video-only and audio-visual systems respectively.

### 3.2. Results on Audio-Visual Wake Word Spotting

First we evaluate the performance of the single-modal systems. We train the audio-only and the video-only system respectively, corresponding to the audio stream and video stream of the audio-visual model shown in Fig. 1. Our results are presented in Table 1. We can observe that the better results were achieved by audio-only system compared with video-only especially in less-noisy environments. In low-SNR adverse acoustic environments, video-only system achieves better FAR performance, which indicates potential of audio-visual system that integrates the advantages of audio-only and video-only systems. Moreover, a direct concatenation for audio-visual fusion at the encoder part is further implemented, which yielded remarkable improvements compared with the single-modal system under various signal-to-noise ratios. For example, the performance gap for -5dB is 4.78% of FAR between audio-only model and audio-visual model.

**Table 1**. Test set performance comparison of different systems.

| Modality | 1-FRR (%) | FAR (%) | | |
|---|---|---|---|---|
| | | -5dB | 0dB | 5dB |
| Audio | 98.78 | 8.03 | 2.95 | 1.60 |
| Video | 98.78 | 6.92 | 6.92 | 6.92 |
| Audio-visual | 98.78 | 3.25 | 1.06 | 0.56 |

**Table 2**. Parameter statistics of different systems.

| Network | Param. (M) | FLOPs (M) |
|---|---|---|
| Audio | 0.35 | 11.29 |
| Lip Encoder | 0.39 | 642.56 |
| Video | 1.19 | 651.68 |
| Audio-Visual | 1.55 | 656.48 |

The statistics of the parameters and FLOPs of these three models are shown in Table 2. We can observe that after adding video modality, the number of parameters and FLOPs of the audio-visual WWS model is greatly increased. In particular, the parameters and FLOPs of lip encoder exceed those of audio-only network, which also promotes us to explore more effective pruning methods.

**Table 3**. Test set performance of single-modalities and multi-modality systems under different pruning degrees. [1-FRR : 98.78%]

| Modality | Method | FAR (%) | | | Pruned (%) |
|---|---|---|---|---|---|
| | | -5dB | 0dB | 5dB | |
| Audio | No Pruning | 8.03 | 2.95 | 1.60 | 0.00 |
| | LTH [26] | 34.00 | 23.67 | 18.3 | 71.90 |
| | LTH-IF | 7.71 | 2.26 | 1.15 | 80.21 |
| Video | No Pruning | 6.92 | 6.92 | 6.92 | 0.00 |
| | LTH [26] | 16.11 | 16.11 | 16.11 | 29.86 |
| | LTH-IF | 6.89 | 6.89 | 6.89 | 55.65 |
| Audio-Visual | No Pruning | 3.25 | 1.06 | 0.56 | 0.00 |
| | LTH-IF | 2.29 | 0.84 | 0.53 | 42.52 |

### 3.3. Results on LTH-IF Pruning

We first implemented LTH-based pruning on single-modal systems described in [26] with the results shown in Table 3. When LTH-based pruning in [26] is firstly applied to the single-modal models, the performance degrades rapidly in both audio and video modalities. Compared to original unpruned model, using LTH with iterative fine-tuning (LTH-IF) achieves better performance especially in the audio modality even though over 80% model parameters are pruned, which demonstrates the effectiveness of the iterative fine-tuning strategy.

Based on these positive results, we next apply LTH-IF pruning to the audio-visual WWS model, and the results are shown in last row of Table 3. According to Table 2, most of the parameters and FLOPs in audio-visual system come from the lip encoder. Thus we design an experiment by only applying LTH-IF pruning for the lip encoder. The pruned model achieves better performance than the audio-visual unpruned model after pruning about 80% of the parameters in lip encoder. Fig. 2 shows the results comparison of applying LTH-IF pruning (blue) to the audio-visual WWS model and original unpruned model (red). According to the experimental results above,

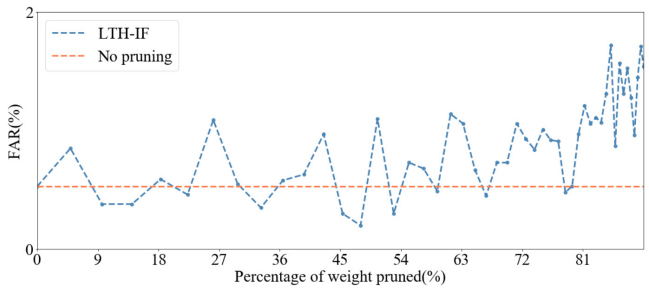


**Fig. 2**. Test set performance on audio-visual WWS model with the iterative pruning during the training process.

the system performance gradually deteriorates after about 80% of the model parameters are pruned. Finally, we initialize the mask in LTH-IF according to the lip encoder result, and then apply LTH-IF to the whole audio-visual model for pruning. According to Table 3, after pruning 42.52% of the parameters, we achieve better performance compared to the model without pruning under all three signal-to-noise ratios.

We randomly select a specific iteration (e.g. $t = 20$) and list the pruned parameters of different layers of the model which is shown in Table 4. It can be observed that for both single-modality and multi-modality systems, the pruning proportions of different types of layers are similar.

However, during the pruning process, we observe that the performance of a video-only system is often unstable (degradation with potentially important nodes pruned) and more sensitive to the pruning proportion setting compared with the audio-visual system. So the audio-visual system seems more robust to pruning, which might be explained by that the audio-visual fusion leads to the selection of more suitable nodes without being pruned.

**Table 4**. The comparison of pruned parameters after the same iterations for single-modalities and multi-modality systems. [$t$=20]

| Modality | Conv layers(%) | LSTM(%) | FC layers(%) |
|---|---|---|---|
| Audio | 65.54 | 65.24 | 65.72 |
| Video | 65.54 | 65.86 | 65.73 |
| Audio-Visual | 65.53 | 65.89 | 65.70 |

## 4. CONCLUSION

In this paper, we investigate on designing a compact audio-visual WWS system under noisy conditions by utilizing video information to alleviate the performance degradation. Tested on our in-house corpus for audio-visual WWS in a home TV scene, the proposed audio-visual system achieves significant performance improvements over the single-modality system under different noisy conditions. Furthermore, a neural network pruning strategy via LTH in an iterative fine-tuning manner is introduced, which can largely reduce the network parameters and computations with no degradation of WWS performance.

## 5. ACKNOWLEDGEMENT

This work was supported in part by the National Natural Science Foundation of China under Grant No. 62171427, and Alibaba Group.

## 6. REFERENCES


[1] David Leroy, Alice Coucke, Thibaut Lavril, Thibault Gisselbrecht, and Joseph Dureau, "Federated learning for keyword spotting," in *ICASSP*, 2019, pp. 6341–6345.

[2] Iván López-Espejo, Zheng-Hua Tan, and Jesper Jensen, "Improved external speaker-robust keyword spotting for hearing assistive devices," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 28, pp. 1233–1247, 2020.

[3] Yiteng Arden Huang, Turaj Z. Shabestary, and Alexander Gruenstein, "Hotword cleaner: Dual-microphone adaptive noise cancellation with deferred filter coefficients for robust keyword spotting," in *ICASSP*, 2019, pp. 6346–6350.

[4] Christin Jose, Yuriy Mishchenko, Thibaud Sénéchal, Anish Shah, Alex Escott, and Shiv Naga Prasad Vitaladevuni, "Accurate detection of wake word start and end using a CNN," in *INTERSPEECH*, 2020, pp. 3346–3350.

[5] Xiong Wang, Sining Sun, Changhao Shan, Jingyong Hou, Lei Xie, Shen Li, and Xin Lei, "Adversarial examples for improving end-to-end attention-based small-footprint keyword spotting," in *ICASSP*, 2019, pp. 6366–6370.

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

[7] Ashish Shrivastava, Arnav Kundu, Chandra Dhir, Devang Naik, and Oncel Tuzel, "Optimize what matters: Training dnn-hmm keyword spotting model using end metric," in *ICASSP*, 2021, pp. 4000–4004.

[8] Guoguo Chen, Carolina Parada, and Georg Heigold, "Small-footprint keyword spotting using deep neural networks," in *ICASSP*, 2014, pp. 4087–4091.

[9] Tara Sainath and Carolina Parada, "Convolutional neural networks for small-footprint keyword spotting," in *INTERSPEECH*, 2015, p. 1478–1482.

[10] Raphael Tang and Jimmy Lin, "Deep residual learning for small-footprint keyword spotting," in *ICASSP*, 2018, pp. 5484–5488.

[11] Sercan Ömer Arik, Markus Kliegl, Rewon Child, Joel Hestness, Andrew Gibiansky, Christopher Fougner, Ryan Prenger, and Adam Coates, "Convolutional recurrent neural networks for small-footprint keyword spotting," in *INTERSPEECH*, 2017, pp. 1606–1610.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

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

[14] Bo Zhang, Wenfeng Li, Qingyuan Li, Weiji Zhuang, Xiangxiang Chu, and Yujun Wang, "Autokws: Keyword spotting with differentiable architecture search," in *ICASSP*, 2021, pp. 2830–2834.

[15] Xuan Ji, Meng Yu, Jie Chen, Jimeng Zheng, Dan Su, and Dong Yu, "Integration of multi-look beamformers for multi-channel keyword spotting," in *ICASSP*, 2020, pp. 7464–7468.

[16] Yixin Gao, Yuriy Mishchenko, Anish Shah, Spyros Matsoukas, and Shiv Vitaladevuni, "Towards data-efficient modeling for wake word spotting," in *ICASSP*, 2020, pp. 7479–7483.

[17] Myunghun Jung, Youngmoon Jung, Jahyun Goo, and Hoirin Kim, "Multi-task network for noise-robust keyword spotting and speaker verification using ctc-based soft VAD and global query attention," in *INTERSPEECH*, 2020, pp. 931–935.

[18] Iván López-Espejo, Zheng-Hua Tan, and Jesper Jensen, "A novel loss function and training strategy for noise-robust keyword spotting," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 29, pp. 2254–2266, 2021.

[19] Ximin Li, Xiaodong Wei, and Xiaowei Qin, "Small-footprint keyword spotting with multi-scale temporal convolution," in *INTERSPEECH*, 2020, pp. 1987–1991.

[20] Oleg Rybakov, Natasha Kononenko, Niranjan Subrahmanya, Mirkó Visontai, and Stella Laurenzo, "Streaming keyword spotting on mobile devices," in *INTERSPEECH*, 2020, pp. 2277–2281.

[21] Théodore Bluche and Thibault Gisselbrecht, "Predicting detection filters for small footprint open-vocabulary keyword spotting," in *INTERSPEECH*, 2020, pp. 2552–2556.

[22] Bin Liu, Shuai Nie, Yaping Zhang, Shan Liang, Zhanlei Yang, and Wenju Liu, "Loss and double-edge-triggered detector for robust small-footprint keyword spotting," in *ICASSP*, 2019, pp. 6361–6365.

[23] Simon Mittermaier, Ludwig Kürzinger, Bernd Waschneck, and Gerhard Rigoll, "Small-footprint keyword spotting on raw audio data with sinc-convolutions," in *ICASSP*, 2020, pp. 7454–7458.

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

[25] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *CVPR*, 2018, pp. 4510–4520.

[26] Jonathan Frankle and Michael Carbin, "The lottery ticket hypothesis: Finding sparse, trainable neural networks," in *ICLR*, 2018.

[27] Menglong Xu and Xiao-Lei Zhang, "Depthwise separable convolutional resnet with squeeze-and-excitation blocks for small-footprint keyword spotting," in *INTERSPEECH*, 2020, pp. 2547–2551.

[28] Alex Renda, Jonathan Frankle, and Michael Carbin, "Comparing rewinding and fine-tuning in neural network pruning," in *ICLR*, 2019.

[29] Xiaohan Chen, Yu Cheng, Shuohang Wang, Zhe Gan, Jingjing Liu, and Zhangyang Wang, "The elastic lottery ticket hypothesis," *arXiv preprint arXiv:2103.16547*, 2021.

[30] Chao-Han Huck Yang, Hu Hu, Sabato Marco Siniscalchi, Qing Wang, Yuyang Wang, Xianjun Xia, Yuanjun Zhao, Yuzhong Wu, Yannan Wang, Jun Du, et al., "A lottery ticket hypothesis framework for low-complexity device-robust neural acoustic scene classification," *arXiv preprint arXiv:2107.01461*, 2021.

[31] Eran Malach, Gilad Yehudai, Shai Shalev-Schwartz, and Ohad Shamir, "Proving the lottery ticket hypothesis: Pruning is all you need," in *ICML*. PMLR, 2020, pp. 6682–6691.