INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Robust Keyword Spotting for Noisy Environments by Leveraging Speech Enhancement and Speech Presence Probability

Chouchang Yang, Yashas Malur Saidutta, Rakshith Sharma Srinivasa, Ching-Hua Lee, Yilin Shen, Hongxia Jin

Samsung Research America, Mountain View, CA, USA

{c.yang1, ym.saidutta, r.srinivasa, chinghua.l, yilin.shen, hongxia.jin}@samsung.com

Abstract

Although various deep keyword spotting (KWS) systems have demonstrated promising performance under relatively noiseless environments, accurate keyword detection in the presence of strong noise remains challenging. Room acoustics and noise conditions can be highly diverse, leading to drastic performance degradation if not handled carefully. In this paper, we propose a noise management front-end called SE-SPP Net performing speech enhancement (SE) and speech presence probability (SPP) estimation jointly for robust KWS in noise. The SE-SPP Net estimates both the denoised Mel spectrogram and the position of the speech utterance in the noisy signal, where the latter is estimated as the probability of a particular time-frequency bin containing speech. Further, it comes at relatively no cost in model size when compared to a model estimating the denoised speech. Our SE-SPP Net can improve noisy KWS performance by up to 7% compared to a similar sized state-of-the-art model at SNR -10dB.

Index Terms: keyword spotting, speech commands, speech presence probability, noise robust, speech enhancement

1. Introduction

Driven by various mobile devices and smart home applications, keyword spotting (KWS) systems have gained considerable attention and form a cornerstone of human device interaction [1]. KWS systems continuously process audio streams to detect keywords. In most scenarios, the devices are constrained in their memory and power budget. Hence, it is paramount to design KWS systems with an emphasis on small memory footprint and low power consumption. On the other hand, background noise is ubiquitous in our daily lives. Therefore, designing KWS systems that are robust to noise is equally important.

It is evident that existing KWS systems based on deep neural networks (DNNs) usually perform well in relatively clean conditions, but often degrade significantly under noisy environments [2]. Without carefully accounting for the noise, a less sensitive KWS system might fail to detect the keyword in low signal-to-noise ratio (SNR) environments, leading to low detection rates. On the other-hand, a more sensitive system might mistake background noise for keywords and accidentally trigger the device, resulting in high false acceptance rates. Further, it is even more challenging for small models to achieve robustness to noise [3, 4].

In this paper, we focus on improving the robustness of KWS in both noisy near and far-field environments while maintaining a small memory footprint. We find that the key is to jointly perform speech enhancement (SE) to denoise the signal and at the same time incorporate positional information of the keyword utterance in the 2-D time-frequency (T-F) domain. To this end, we propose a novel noise management front-end called SE-SPP Net that simultaneously performs SE and speech presence probability (SPP) [1, 5, 6] estimation. To our knowledge, current KWS systems that cope with noise [7–12] have not considered utilizing such combination. The SE-SPP Net is trained to predict a denoising mask and an SPP map. The denoising mask is used to generate an estimate of the clean (noise-free) input signal. The learned SPP map has values ranging from 0 to 1 to represent the likelihood of the presence of speech in the noisy mixture at the input, effectively encoding the positional information of the utterance in the T-F domain. The learned SPP map is then passed to the keyword detection (KWD) module along with the denoised signal estimate. We present experimental results to show that by incorporating SE+SPP into the KWS system, the detection accuracy can be improved in noisy settings (both far-field and near-field) even when the model size is small.

The contributions of our work are:

- We propose a modular design of KWS systems for both noisy near and far-field environments comprising of two modules: a novel combined SE+SPP noise management module (SE-SPP Net), and a KWD module.
- We show dramatic performance improvement in accuracy over state-of-the-art (SOTA) models with similar model size.
- Our proposed SPP module can boost performance with a very negligible increment in model size.

2. Related work

KWS systems based on DNNs have shown promising performance compared to the traditional approaches [13]. In 2017, Google released the first large-scale speech command dataset [14] consisting of 65000 one-second long utterances of 30 short words collected through thousands of speakers. Since then, various deep learning frameworks have been proposed to push the performance limit further [4,15–23]. However, as we shall show in the experimental results, models that achieve SOTA results in relatively clean keyword utterance detection fail in noisy conditions, even when exposed to noise during training. Thus, these solutions are far from satisfactory for various speech command applications in the real world where environmental noise and interference are ubiquitous. This indicates the need to train specialized KWS systems capable of handling such circumstances.

Several works have started to tackle the noisy KWS problem. In [24], a novel loss function based on the N-pair loss function is proposed to improve KWS performance in noisy conditions. In [12], a small-footprint KWS model is introduced by constructing a novel convolutional neural network encoder with a mixer module, along with the use of curriculum training to learn better from data with SNR variety. Other works like

10.21437/Interspeech.2023-2222

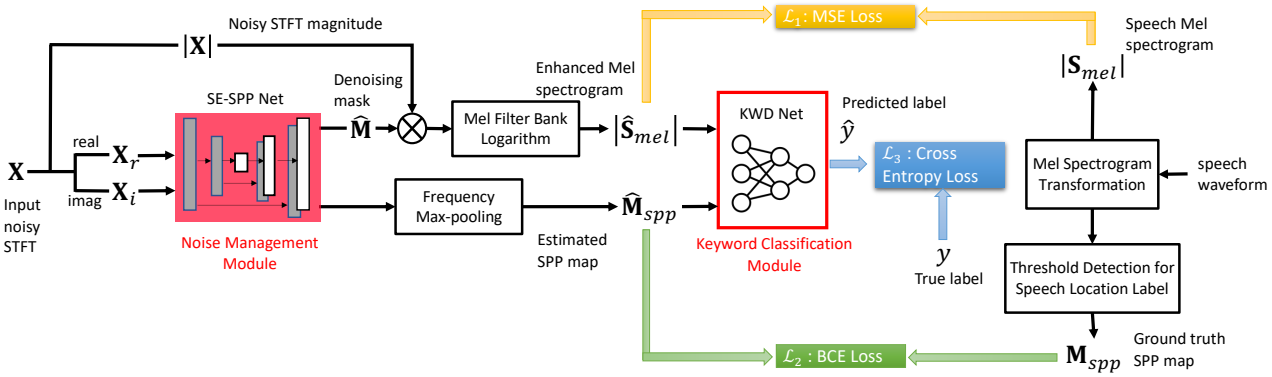


Figure 1: *The proposed noise robust KWS system consisting of an SE-SPP Net based noise management module and a keyword classification module incorporating SPP-based positional information in addition to the denoised signal for improved keyword detection.*

[7, 10] use a speech denoising front-end followed by a KWD module which are jointly optimized during training. However, the denoising module is designed to estimate the entire clean speech while the subsequent KWD module only takes the Mel spectral magnitude as input features. This in turn could result in reduced efficiency due to the redundancy for phase estimation, leading to the model easily getting over 1M parameters. Although the above KWS works have demonstrated certain robustness to noise to some extent, we still see that existing KWS systems often trade off model conciseness for increased detection rate, especially in the presence of strong noise.

3. Proposed method

3.1. Overview

Figure 1 depicts the proposed KWS system where a noise management module (SE-SPP Net) is cascaded with a KWD module. The model takes the short-time Fourier transform (STFT) [25] of a noisy audio signal $\mathbf{X} \in \mathbb{C}^{F \times T}$ as input, where F is the number of frequency bins and T is the number of time frames. The SE-SPP Net takes the real and imaginary parts of the noisy STFT, i.e., $\mathbf{X}_r \in \mathbb{R}^{F \times T}$ and $\mathbf{X}_i \in \mathbb{R}^{F \times T}$ as inputs to jointly perform denoising and SPP estimation. It provides two outputs: a mask to enhance the input noisy spectrogram and the other represents the SPP. The denoising mask $\hat{\mathbf{M}}$ is a matrix of size $F \times T$ with elements ranging between 0 and 1. The mask is then multiplied with $|\mathbf{X}|$ to generate the enhanced magnitude spectrogram. It is then transformed to the Mel spectral domain. Additionally, the estimated SPP map (denoted by $\hat{\mathbf{M}}_{spp}$) is computed directly in the Mel spectral domain with the dimension matching that of the enhanced Mel spectrogram. Subsequently, both the enhanced Mel spectrum and the estimated SPP map are fed into the second stage of KWS system, the KWD module. The KWD module performs the classification task to predict the class label y. Note that the proposed methodology can be applied to any existing deep KWS models.

3.2. U-Net based noise management module (SE-SPP Net)

The SE-SPP module is based on a U-Net [26] architecture which leverages a series of encoders and decoders based on 2-D convolutions with skip connections for enhanced feature extraction. The U-Net has been widely adopted in many T-F domain audio processing tasks including SE [27–29]. In this paper, we modify the U-Net to enhance the input signal as well as compute the SPP map to aid the downstream KWD module.

The proposed noise management module takes the real and imaginary spectrograms of the noisy STFT as inputs and predicts i) a denoising mask to be multiplied with the noisy STFT magnitude to enhance the input noisy speech magnitude and ii) an SPP map that represents the likelihood of the speech being present in each T-F bin in the Mel spectrogram domain. The two outputs, i.e., the masked speech magnitude spectrogram and the predicted SPP map are both passed to the KWD stage for performing classification.

To learn to estimate the SPP, the network is trained to predict a pre-computed binary map that represents speech presence (=1) and absence (=0) obtained from the ground truth clean speech. The SPP map thus inherently carries information regarding the position of the speech utterance on the 2-D spectrogram which provides useful information for the later detector to better focus on capturing the keyword characteristics.

3.3. BC-ResNet based KWD module utilizing SE and SPP

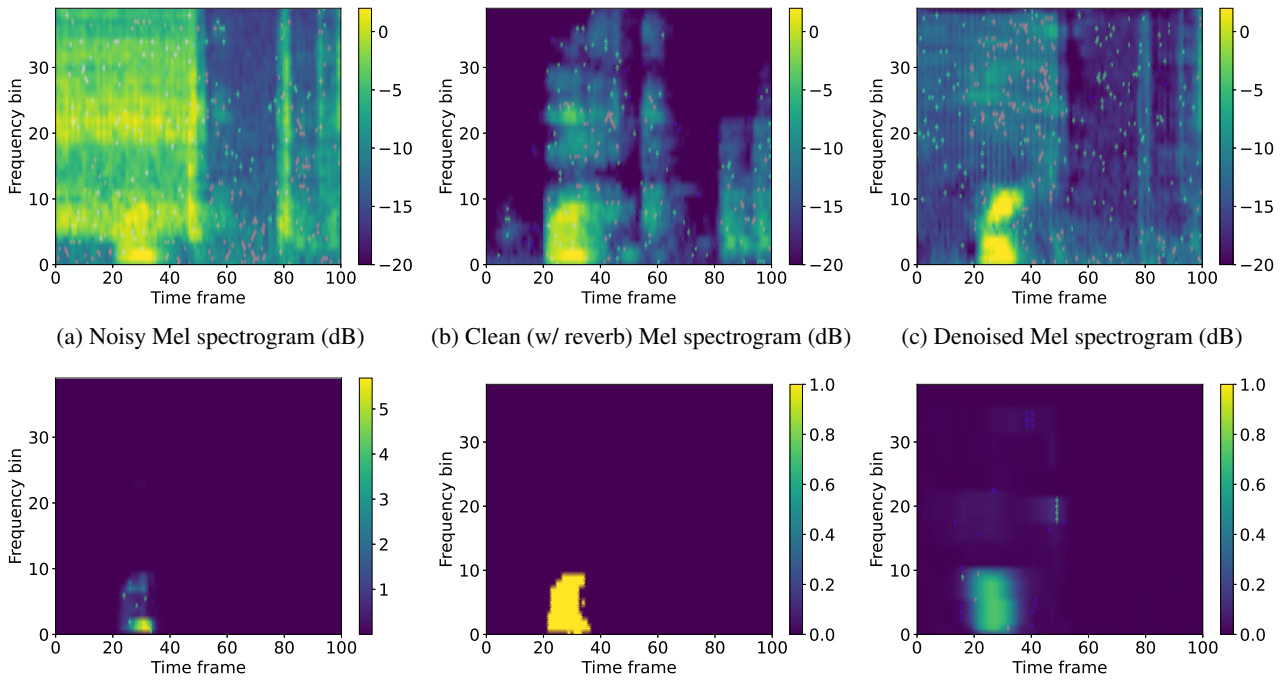
For the KWD module we adopt the BC-ResNet [18] which is developed based on broadcasted residual learning that utilizes the advantage of 1-D temporal and 2-D convolution while minimizing the increase of computation for KWS. In our system, the BC-ResNet takes the denoised Mel spectrogram and the estimated SPP map from the previous SE-SPP Net as input features and predict the probability of the keyword being uttered at the output. The advantages of the proposed method are i) the KWD module sees a relatively noiseless signal rather than seeing the original noisy mixture and ii) the KWD module sees the positional information of the speech utterance distributed in the T-F domain and can hence learn to better detect keywords.

3.4. Training loss:

The entire noise management + keyword detection system is trained in a two-step manner. In the first step, only the U-Net (SE-SPP Net) is trained to perform proper denoising. Then in the second step the U-Net and the BC-ResNet are jointly trained to perform the final classification. To be more specific, in the first step we optimize the following loss function:

$$\mathcal{L}_{pre} = \lambda \mathcal{L}_1(|\hat{\mathbf{S}}_{mel}|, |\mathbf{S}_{mel}|) + \mathcal{L}_2(\hat{\mathbf{M}}_{spp}, \mathbf{M}_{spp}), \quad (1)$$

where $\mathcal{L}_1(\cdot,\cdot)$ and $\mathcal{L}_2(\cdot,\cdot)$ are some criteria for measuring the distance between the two arguments and $\lambda > 0$ is a hyperparameter for weighting the Mel spectrogram regression loss.



(a) Noisy Mel spectrogram (dB) (b) Clean (w/ reverb) Mel spectrogram (dB) (c) Denoised Mel spectrogram (dB)

(d) Clean (w/o reverb) Mel spectrogram (e) Ground truth SPP map (f) Estimated SPP map

Figure 2: *Mel spectrograms and SPP maps of a particular noisy keyword "left" at -10 dB SNR to illustrate the proposed framework in Figure 1. (a) The noisy signal in Mel spectrogram domain. (b) The corresponding clean signal with room reverberation* $\mathbf{S}_{mel}$*. (c) The denoised signal* $\hat{\mathbf{S}}_{mel}$ *from SE-SPP Net. (d) the corresponding clean far-field signal without room reverberation (non dB). (e) The ground truth SPP map that can be obtained by thresholding either (b) or (d). (f) The estimated SPP map* $\hat{\mathbf{M}}_{spp}$ *from SE-SPP Net.*

In this paper, we use the MSE loss for $\mathcal{L}_1$ and the binary cross-entropy (BCE) loss for $\mathcal{L}_2$. Note that in this phase the SE-SPP Net is trained on a set of noisy-clean pairs of speech signals. Specifically, the target $\mathbf{M}_{spp}$ is obtained from the clean signal without reverberation whereas the $\mathbf{S}_{mel}$ is derived from the clean speech signal with reverberation.

After the SE-SPP Net is well-trained, we then jointly train the whole system along with the KWD module with a combined loss function that includes the loss pertaining to the keyword classification task and also the denoising and SPP loss used during pretraining. The loss function in this step is written as:

$$\mathcal{L}_{full} = \mathcal{L}_3(\hat{y}, y) + \gamma \mathcal{L}_{pre}, \tag{2}$$

where $\mathcal{L}_3(\cdot, \cdot)$ is the cross entropy loss and $\gamma > 0$ is a hyper-parameter for weighting the combined denoising and SPP loss with respect to the cross entropy loss.

4. Experiments

Datasets: We use the 12 class Google Speech Command V2 (GSCV2) dataset [14] to evaluate our proposed method. Our goal is to design a KWS system for both near-field and far-field data. To achieve this, we pre-process the dataset in a two step manner. In the first step, following a procedure similar to [12], we convert 50% of the clean speech data to far field. The conversion to far field is performed by using the BUT Speech@FIT Reverb Database's room impulse responses [30]. In the second step we add noise profiles from the "noise" subset of MUSAN [31] to all the files. Noise is added (after the previous step) by randomly selecting one of the 930 noise profiles and adding them at an SNR randomly sampled from the set [20, 15, 10, 5, 0, -3, -5, -7, -9, -10, -12]. We use 80:10:10 split of the dataset to form the training, validation and testing datasets. In addition to the GSCV2 data, we also use the VoiceBank corpus [32]. In particular, we use 11000 speech samples from this corpus to train the first stage of our model to perform denoising. This allows us to leverage more generic speech data that do not contain keywords to train the denoising module. We process the VoiceBank data in the same way as GSCV2, by adding reverb to 50% of the data and then adding noise sampled from MUSAN at various SNRs.

To generate the ground truth SPP map for all speech data, we use their corresponding clean speech data without reverb and convert it into Mel-scale magnitude map as shown in Figure 2(d). Each T-F bin with magnitude greater than 0.15 is labeled as '1' (contains speech) and otherwise '0' (speech absent). Note that it is also possible to use reverb Mel spectrum in Figure 2(b) with threshold to generate the ground truth SPP label. To generate the target signal for denoising, we use the clean speech with reverb for the far field data and clean speech without reverb for the near field data. This way, we train the denoising module only to enhance speech by removing noise and not overcome reverb.

Implementation details: Our proposed framework can use any denoising architecture. Here, we adopted the U-Net[1] for the purpose of demonstration, restricted to only having 40K parameters. The number of input and output channels are set to 2 as shown in Figure 1. One of the output channels is the denoising

[1]We used the implementation from `https://github.com/milesial/Pytorch-UNet`.

Table 1: *Results over training on the 12 class GSC-V2 dataset with ten keywords, an unknown, and a silence class. Results for the ConvMixer are obtained from the paper [12]. The rest are trained by us.*

Model	Acc (%)	Num. Params	20 dB	0 dB	-5 dB	-10 dB	-15 dB	-20 dB
SE + SPP + BCResNet-1 (Ours)	80.21%	50.5K	88.61%	83.60%	77.08%	71.16%	60.30%	50.14%
SE + SPP + BCResNet-3 (Ours)	82.63%	100K	90.94%	85.67%	81.20%	73.78%	64.46%	53.24%
SE + SPP + BCResNet-5 (Ours)	84.29%	190K	91.37%	86.03%	81.81%	75.30%	65.10%	54.49%
BCResNet-1 [18]	70.69%	9.8K	72.16%	70.66%	64.87%	56.35%	43.24%	41.15%
BCResNet-3 [18]	77.28%	58.9K	76.69%	70.67%	68.62%	59.64%	44.51%	43.70%
BCResNet-5 [18]	78.84%	147K	78.74%	70.54%	58.80%	68.56%	56.47%	37.69%
BCResNet-7 [18]	80.23%	275K	80.79%	80.78%	77.47%	67.00%	61.88%	42.97%
ConvMixer [12]	-	119K	87.85%	78.10%	72.78%	66.50%	-	-
ConvMixer (with Curriculum Training) [12]	-	119K	90.83%	83.04%	78.39%	71.88%	-	-

Table 2: *Ablation study showcasing the improvement from Speech Presence Probability (SPP) module.*

Model	Acc (%)	Num. Params	20 dB	0 dB	-5 dB	-10 dB	-15dB	-20dB
SE + SPP + BCResNet-1	**80.21%**	50.5K	**88.61%**	**83.60%**	**77.08%**	**71.16%**	**60.30%**	**50.14%**
SE + BCResNet-1	77.61%	50.1K	86.03%	80.82%	75.70%	68.87%	58.75%	48.63%
SE + SPP + BCResNet-3	**82.63%**	100K	**90.94%**	**85.67%**	**81.20%**	**73.78%**	**64.46%**	**53.24%**
SE + BCResNet-3	81.59%	99.2K	89.34%	83.24%	79.31%	71.98%	62.27%	50.84%
SE + SPP + BCResNet-5	**84.29%**	190K	**91.37%**	**86.03%**	**81.81%**	**75.30%**	**65.10%**	**54.49%**
SE + BCResNet-5	81.91%	188K	89.82%	84.10%	80.33%	74.11%	63.88%	52.70%

mask. The other output channel is passed through a 1-D max-pooling of kernel size 45 and stride 12 to convert a frequency dimension of 513 down to 40. This 40-dimensional entity becomes the estimated SPP map. For the KWD module, there are multiple possible architectures, e.g., [18, 21, 22]. Here, we use BC-ResNet [18] as it achieved SOTA results on clean keyword detection and provides an easy mechanism to vary the model size by changing the scaling factor.

For computing STFT, we use FFT of size 1024 to process frame of size 30 ms with a hop length of 10 ms. The Mel Filter Bank used to convert the denoised STFT spectrum (obtained by multiplying the noisy STFT magnitude with the denoising mask, referred to Figure 1) to Mel spectrogram is also set to 40.

Training details: We divide our training procedure into two parts as mentioned earlier. The first part uses only the noisy VoiceBank data to train the denoised module as decribed in Sec. 3.2. The batch size is set to 4, λ to 0.01, and an initial learning rate of 1e-3 is decayed by 0.1 every 50 epochs. The total training is carried out for 250 epochs and the best model (based on validation set) is chosen for the next training stage. For the second part, we use noisy GSC to jointly train both the noise management module and the KWD module as illustrated in Figure 1. The batch size is set to 100 and we apply the the one-cycle learning rate scheduler [33] where the learning rate goes from 0.004 to 0.1 over seven epochs and decays to 4×10^{-6} over the next 18 epochs. γ is set to 1.

Results: Table 1 showcases the classification accuracy of of different models. The first set of models we consider are the proposed models denoted as "UNet + SPP + BCResNet." Here, the UNet is fixed and only the size of the KWD is varied by varying the scaling factor of the BC-ResNet [18]. We also consider two external baselines. In the first, we perform multi-condition training [34] on various BC-ResNet models. The second is SOTA for noisy keyword detection, the ConvMixer [12]. The value in the "Acc (%)" represents the case where the test samples are subject to SNRs from the same range that was used for the training data. We also study cases where all test samples are subject to a single SNR indicated by the column heading.

We first compare our proposed models with the BC-ResNet baseline. **We find that the smallest proposed model "UNet + SPP + BCResNet-1" achieves almost 3% improvement in accuracy compared to BCResNet-3 while being slightly smaller. In fact, it achieves the same performance as BCResNet-7 which is five times larger.** Second, we compare our proposed model ("UNet + SPP + BCResNet-3") with the current SOTA and comparably sized ConvMixer. **We find that our proposed models with slightly lesser parameters (19% lesser) outperform ConvMixer significantly, obtaining improvements in the range of 3% to 8% over the range of SNRs tested.** Note, for the ConvMixer, we do not have the results of overall accuracy over the test set with randomly sampled SNRs, since the set of SNR choices was constructed by us and not [12]. **Further, even though our models are not trained with curriculum learning, our models match the performance of the curriculum trained ConvMixer at high SNR and surpass it by 2% to 3% in the lower SNRs.**

In Table 2 we perform ablation studies by studying the importance of the SPP component where both models were trained with the same hyper-parameters and same procedures. **With very negligible increase in number of parameters ($<$1%), the SPP module yields upto 3% improvement in accuracy.** Table 2 showcases the qualitative results corresponding to the denoising and the SPP estimation components individually.

5. Conclusion

In this paper, we presented a deep KWS system robust to environmental noise based on a novel noise management front-end, i.e., SE-SPP Net, that combines SE and SPP estimation. This front-end module performs denoising and at the same time encodes the positional information of the speech utterance in the T-F domain to facilitate the subsequent KWD model learning. We demonstrated that the SE-SPP Net front-end module is able to improve keyword detection robustness in very noisy environments, especially in the small memory-footprint KWS regime.

6. References

[1] R. Haeb-Umbach, S. Watanabe, T. Nakatani, M. Bacchiani, B. Hoffmeister, M. L. Seltzer, H. Zen, and M. Souden, "Speech processing for digital home assistants: Combining signal processing with deep-learning techniques," *IEEE Signal Processing Magazine*, vol. 36, no. 6, pp. 111–124, 2019.

[2] A. Mohanty, A. Frischknecht, C. Gerum, and O. Bringmann, "Behavior of keyword spotting networks under noisy conditions," in *Proceedings of International Conference on Artificial Neural Networks (ICANN)*, 2021, pp. 369–378.

[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 *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4704–4708.

[4] S. Arık, M. Kliegl, R. Child, J. Hestness, A. Gibiansky, C. Fougner, R. Prenger, and A. Coates, "Convolutional recurrent neural networks for small-footprint keyword spotting," in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, 2017, pp. 1606–1610.

[5] T. Gerkmann and R. C. Hendriks, "Unbiased mmse-based noise power estimation with low complexity and low tracking delay," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393, 2011.

[6] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, 2017.

[7] 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 *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, 2018, pp. 2613–2617.

[8] Y. Huang, T. Hughes, T. Z. Shabestary, and T. Applebaum, "Supervised noise reduction for multichannel keyword spotting," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5474–5478.

[9] Y. Gu, Z. Du, H. Zhang, and X. Zhang, "A monaural speech enhancement method for robust small-footprint keyword spotting," *arXiv preprint arXiv:1906.08415*, 2019.

[10] D. Bonet, G. Cámbara, F. López, P. Gómez, C. Segura, J. Luque, and M. Farrús, "Speech enhancement for wake-up-word detection in voice assistants," in *Proceedings of IberSPEECH*, 2021, pp. 41–45.

[11] I. López-Espejo, Z.-H. Tan, and J. 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.

[12] D. Ng, Y. Chen, B. Tian, Q. Fu, and E. S. Chng, "Convmixer: Feature interactive convolution with curriculum learning for small footprint and noisy far-field keyword spotting," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 3603–3607.

[13] I. López-Espejo, Z.-H. Tan, J. Hansen, and J. Jensen, "Deep spoken keyword spotting: An overview," *IEEE Access*, 2021.

[14] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," *CoRR*, vol. abs/1804.03209, 2018. [Online]. Available: http://arxiv.org/abs/1804.03209

[15] S. Choi, S. Seo, B. Shin, H. Byun, M. Kersner, B. Kim, D. Kim, and S. Ha, "Temporal convolution for real-time keyword spotting on mobile devices," in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, 2019, pp. 3372–3376.

[16] X. Li, X. Wei, and X. Qin, "Small-footprint keyword spotting with multi-scale temporal convolution," in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, 2020, pp. 1987–1991.

[17] M. Xu and X.-L. Zhang, "Depthwise separable convolutional ResNet with squeeze-and-excitation blocks for small-footprint keyword spotting," in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, 2020, pp. 2547–2551.

[18] B. Kim, S. Chang, J. Lee, and D. Sung, "Broadcasted residual learning for efficient keyword spotting," in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, 2021, pp. 4538–4542.

[19] M. Zeng and N. Xiao, "Effective combination of DenseNet and BiLSTM for keyword spotting," *IEEE Access*, vol. 7, pp. 10 767–10 775, 2019.

[20] O. Rybakov, N. Kononenko, N. Subrahmanya, M. Visontai, and S. Laurenzo, "Streaming keyword spotting on mbile devices," in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, 2020, pp. 2277–2281.

[21] A. Berg, M. O'Connor, and M. T. Cruz, "Keyword transformer: A self-attention model for keyword spotting," in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, 2021, pp. 4249–4253.

[22] Y. Bai, J. Yi, J. Tao, Z. Wen, Z. Tian, C. Zhao, and C. Fan, "A time delay neural network with shared weight self-attention for small-footprint keyword spotting," in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, 2019, pp. 2190–2194.

[23] X. Chen, S. Yin, D. Song, P. Ouyang, L. Liu, and S. Wei, "Small-footprint keyword spotting with graph convolutional network," in *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 539–546.

[24] K. Sohn, "Improved deep metric learning with multi-class N-pair loss objective," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 29, 2016.

[25] M. Parchami, W.-P. Zhu, B. Champagne, and E. Plourde, "Recent developments in speech enhancement in the short-time Fourier transform domain," *IEEE Circuits and Systems Magazine*, vol. 16, no. 3, pp. 45–77, 2016.

[26] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015, pp. 234–241.

[27] R. Giri, U. Isik, and A. Krishnaswamy, "Attention wave-u-net for speech enhancement," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019, pp. 249–253.

[28] H.-S. Choi, J.-H. Kim, J. Huh, A. Kim, J.-W. Ha, and K. Lee, "Phase-aware speech enhancement with deep complex U-Net," in *Proceedings of International Conference on Learning Representations (ICLR)*, 2019.

[29] R. Gao and K. Grauman, "Co-separating sounds of visual objects," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, 2019, pp. 3879–3888.

[30] I. Szöke, M. Skácel, L. Mošner, J. Paliesek, and J. Černockỳ, "Building and evaluation of a real room impulse response dataset," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 863–876, 2019.

[31] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[32] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *Proceedings of International Conference Oriental COCOSDA held jointly with Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, 2013.

[33] L. N. Smith and N. Topin, "Super-convergence: Very fast training of neural networks using large learning rates," in *Artificial intelligence and machine learning for multi-domain operations applications*, vol. 11006. SPIE, 2019, pp. 369–386.

[34] D. Yu, M. L. Seltzer, J. Li, J.-T. Huang, and F. Seide, "Feature learning in deep neural networks-studies on speech recognition tasks," *arXiv preprint arXiv:1301.3605*, 2013.