INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Small Footprint Multi-channel Network for Keyword Spotting with Centroid Based Awareness

Dianwen Ng[1,2], *Yang Xiao*[2], *Jia Qi Yip*[1,2], *Zhao Yang*[2], *Biao Tian*[1]
Qiang Fu[1], *Eng Siong Chng*[2], *Bin Ma*[1]

[1]Speech Lab of DAMO Academy, Alibaba Group
[2]School of Computer Science and Engineering, Nanyang Technological University, Singapore
{dianwen.ng, b.ma}@alibaba-inc.com

Abstract

Spoken Keyword Spotting (KWS) in noisy far-field environments is challenging for small-footprint models, given the restrictions on computational resources (e.g., model size, running memory). This is even more intricate when handling noises from multiple microphones. To address this, we present a new multi-channel model that uses a CNN-based network with a linear mixing unit to achieve local-global dependency representations. Our method enhances noise-robustness while ensuring more efficient computation. Besides, we propose an end-to-end centroid-based awareness module that provides class similarity awareness at the bottleneck level to correct ambiguous cases during prediction. We conducted experiments using real noisy far-field data from the MISP challenge 2021 and achieved SOTA results compared to existing small-footprint KWS models. Our best score of 0.126 is highly competitive against larger models like 3D-ResNet, which is 0.122, but ours is much smaller at 473K compared to 13M.

Index Terms: Small Footprint, Keyword Spotting, Multi-channel, Noisy Far-field, Centroid Awareness

1. Introduction

Voice assistant applications in smart devices experiencing increasing adoption due to the recent success of automatic speech recognition. Keywords such as "Alexa" or "Hey Siri" are frequently utilized to activate hands-free applications. The task of identifying these predefined words within a continuous utterance is referred to as keyword spotting (KWS). It is crucial to develop KWS systems with small footprint models and low inference latency, as they are commonly deployed on-device.

Recent works on small footprint KWS [1, 2, 3, 4, 3, 5, 6, 7, 8] have demonstrated promising outcomes when dealing with clean and close-talking audio datasets. However, the performance of these models significantly deteriorates when applied to far-field utterances. This decline is particularly evident in multi-talker environments with low signal-to-noise ratios (SNRs). While conventional techniques like multi-conditioning [9] and front-end speech enhancement [10, 11] are employed to address this issue, models utilizing multi-conditioning often struggle with a wide range of noise levels [12], resulting in poor performance. Furthermore, in far-field speech processing, factors such as reverberation and multiple sources of interference blur spectral cues, compromising the quality of single-channel speech enhancement.

Recent work on small footprint Keyword Spotting (KWS), ConvMixer [13], addresses the challenge of applying KWS in noisy far-field environments. In order to enhance the performance of small KWS systems, the authors propose a novel encoder architecture based on convolutional neural networks (CNN). This architecture incorporates a mixer module as an alternative to attention mechanisms. The mixer unit computes weighted feature interactions across different channels, allowing for the efficient flow of information with varying degrees of importance. While the previous work has shown promising performance in small footprint Keyword Spotting (KWS) for noisy far-field conditions, it is specifically tailored to single-channel speech data. In contrast, multi-channel speech data comprises recordings captured from multiple microphones placed in different locations, introducing various noise sources and acoustic environments. The variability among channels can lead to inter-channel discrepancies in the audio signals, which can have a negative impact on the performance of single-channel keyword spotting models

Multi-channel systems have been extensively studied to improve the noise robustness of speech recognition [14, 15]. Recent advances [16, 17] utilizing architectures resembling beamforming or masking networks have achieved success in jointly optimizing multi-channel enhancement and acoustic modeling. However, small footprint models often encounter challenges in effectively learning spatial filtering and noise-robust feature extraction from audio data. To tackle this problem, [18] introduced a low latency model architecture that incorporates a three-dimensional single value decomposition filter layer. This innovation enables the model to process raw microphone arrays for on-device multi-channel Keyword Spotting (KWS). Nonetheless, this approach comes with increased computational costs. Despite the potential benefits, research on small footprint multi-channel KWS systems remains relatively limited, with only a limited amount of literature available on the subject.

In this paper, we build on [13] to address the limitations of small footprint models. Specifically, we modify the single-channel framework to design a multi-channel KWS that is more robust in noisy and far-field environments, while still maintaining a small footprint for on-device applications. Our model architecture builds on the success of the convolutional-based networks in modeling local dependencies on clean utterances. However, the performance degenerates when the sample is noisy and reverberated. The proposed networks enhance the existing architecture by using a convolution-mixer module that computes the global dependency with a linear unit implemented after the CNN block across temporal, frequency and audio channels. The linear unit captures the global information to improve the understanding of the noisy features. Most importantly, our setup employs a direct and simple linear block that is highly efficient, requiring less memory and computational power compared to other methods. In addition, we suggest improving the model's performance by using a trainable codebook with the gradient descent algorithm to learn the estimation of the key-

10.21437/Interspeech.2023-1210

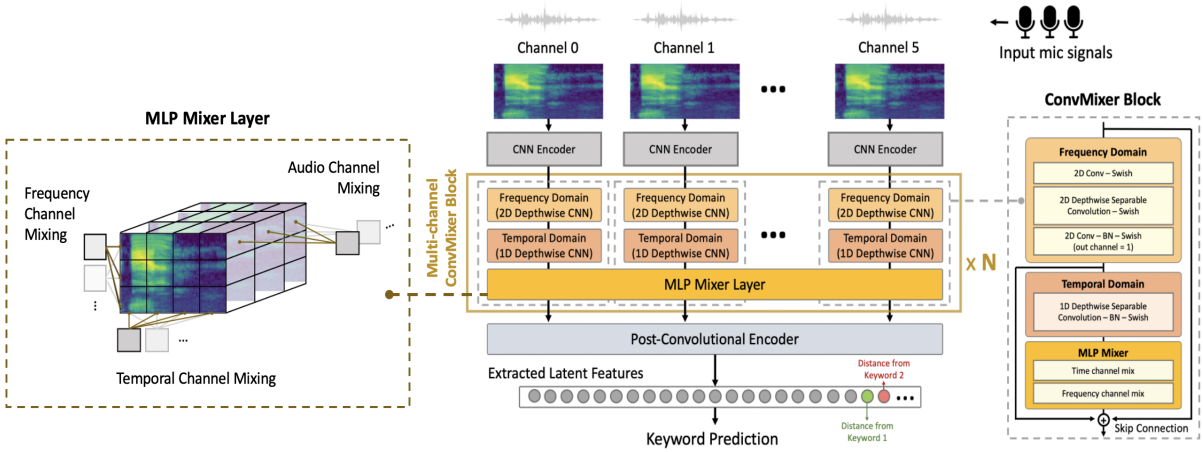

Figure 1: *Model Architecture for Multi-channel ConvMixer with Centroid Awareness (An example of a 6-channel model)*

word centroids. Finally, our results show that applying multi-look beamforming and weighted Prediction Error (WPE) to our proposed structure achieves the new SOTA for small multi-channel KWS. In general, our model performs competitively with popular models like Conformers and 3D-ResNet, which are at least 27 times larger than ours.

2. Methodology

2.1. KWS Model: Multi-channel ConvMixer

While many small-footprint KWS have succeeded with the CNN architecture, they often fail when audio is badly distorted by noise and reverberation. To address this issue, we propose using a linear unit (mixer), that is attached after the convolutional block to compute global information. As shown in [19], using simply two linear operators over the channel, and the token level can effectively learn global information in image classification. Thus, we developed our mixer module to consist of 2 linear layers (i.e., a downward projection, GELU [20] and an upward projection), and applied it to the time, frequency and microphone channels. This allows the model to look into the global frames in these three dimensions and prioritize relevant attributes to embed noise-robust representations for the KWS task. It is important to note that the objective of the mixer unit is not to perform audio enhancement, but rather to exploit linear computation to extract global contextual dependency information to increase the capacity of our networks. Furthermore, our proposed mixer unit has a lower computational complexity compared to a typical attention module, making the overall architecture small and maintaining low latency.

$$\begin{aligned} u_{*,t,*} &= x_{*,t,*} + W_2 \cdot \delta[W_1 \cdot \text{LayerNorm}(x)_{*,t,*}] \\ y_{f,*,*} &= u_{f,*,*} + W_4 \cdot \delta[W_3 \cdot \text{LayerNorm}(u)_{f,*,*}] \\ z_{*,*,c} &= y_{*,*,c} + W_6 \cdot \delta[W_5 \cdot \text{LayerNorm}(y)_{*,*,c}] \end{aligned} \quad (1)$$

where x, u, y, and z refer to the intermediate features of the layer output respectively. δ represents the GELU unit. W_1 and W_2 are the learnable weights of the linear layer for temporal channel shared across all frequency frame f, for $f \in \{1, 2.., F\}$. W_3 and W_4 are the learnable weights of the linear layer for frequency channel shared across all time frame t for $t \in \{1, 2.., T\}$. And W_5 and W_6 are the learnable weights of the linear layer for audio channel shared across c microphones, for $c \in \{1, .., 6\}$ in a six-channel KWS model.

Our proposed multi-channel ConvMixer model comprises a c-parallel (non-sharing) convolutional encoder where c is determined by the number of given microphones. A graphical representation is presented in Figure 1. The model first encodes simple convolutional features, and we pass it to our multi-channel ConvMixer Block ($N = 4$) to execute the proposed computation. At first, time-frequency information is extracted independently using the 2D and 1D depthwise separable convolution. Then, the mixer unit combines the c microphones to learn and construct our global dependency information. Finally, a post-convolutional encoder is used to aggregate the microphones with a pointwise convolutional layer, and we obtain a D-dimensional vector after a global max pooling. Before sending for prediction, our final latent representations are appended with the centroid awareness, as discussed in the following subsection.

2.2. Centroid Based Awareness

Under a noisy environmental setting, audio can be heavily distorted by background noises. The latent embedding of such an instance tends to lie near the hyper-planes of the decision boundary and is often projected into a wrong keyword cluster. To improve the latent embeddings of the utterance, we proposed to consider an additional term that computes the L2-norm Euclidean distance of the utterance to the mean of each keyword. The L2-norm distance is measured at a higher dimensional space of 320, which is an increase from the original dimension of 80, where we pass the latent vector of the utterance to a linear projection function. The purpose of this term is to measure the "similarity" of the noisy sample at a more fine-grained higher dimensional space to the clustering keywords and provide additional informative prior that helps in making better decision.

Formally, we understand this by the optimization process of a typical cross-entropy loss. The learning function is commonly written as:

$$H(q, \hat{q}) = \sum_{k=1}^{C} q^{(k)} \log(\hat{q}^{(k)})$$

This can be decomposed based on Bregman divergence [21] as:

$$\mathbb{E}(H(q, \hat{q})) = \underbrace{D_{\text{KL}}(q||\bar{q})}_{\text{Bias}^2} + \underbrace{\mathbb{E}(D_{\text{KL}}(\bar{q}||\hat{q}))}_{\text{Variance}} \quad (2)$$

where q is the one-hot target label, $\hat{q}^{(k)}$ denotes the prediction probability of the k-th class, $\bar{q}$ represents the mean of log-probability after normalization, resembling the mean probability distribution of the classes, and D_{KL} is the KL divergence between the two conditioning distribution.

$$\bar{q}^{(k)} \propto \exp\{\mathbb{E}(log(\hat{q}^{(k)}))\}, \text{for } k = 1, .., K \text{ (of K-classes)}$$

Here, the model error consists of the two KL divergence terms in (2) and the predictive outcome of $\hat{q}$ is given by the softmax of the dense layer

$$\hat{q} \propto \exp(W_{feat}X_{feat} + W_{L_2\text{-norm}}X_{L_2\text{-norm}}) \tag{3}$$

where X_{feat} is the pooled D-dimensional latent representation of intermediate output, X. $X_{L_2\text{-}norm}$ is the L2-norm distance of the pooled X_{feat} to the centroid mean of keyword classes. W_{feat} and $W_{L_2\text{-norm}}$ are the learnable weights of the X_{feat} and $X_{L_2\text{-norm}}$. We claim that the performance improves over the centroid-based awareness as it lowers the model error rate by exploiting the distancing information from the extra $L_2\text{-}norm$ term as it provides, in simple explanation, how close the inferring utterance is to each of the predicting classes with the similarity distancing.

Specifically, we expect $\hat{q}$ to be more confident with the addition of the "similarity" information between the class centroids and the divergence term of $\bar{q}$ and $\hat{q}$ to decrease. Similarly, the average class probability map of $\bar{q}$ tends to get closer to the label from the vanilla decision boundary, and the divergence of q and $\bar{q}$ decreases. This reduces estimation bias and variance of our model, as presented in (2), in which the model converges to a better minimum with a lower model error rate.

Since estimating the keyword centroids with the training data can be computationally expensive, we propose to learn the estimation of the keyword centroids using a trainable codebook with the gradient descent algorithm. This approach jointly trains the networks and avoids the need to collect the updated latent features for mean computation in every training iteration. We initialize the codebook with the respective keywords at the dimension of 320. Then, we extract the latent features of every minibatch and update the embedding vectors by minimizing the mean square error (MSE) within the class labels.

3. Experiments

3.1. Experimental Setup

Experimental Dataset. We perform our experiments using the task 1 dataset from the MISP challenge 2021 [22]. In our work, we only consider the audio data of 120hrs of training, where we will build a KWS model that is robust to the home TV scenario, i.e., noisy and far-field. In particular, a family will be seated 3-5m away from the TV, and there may be conversations while someone is interacting with the television. A linear microphone array (6 channels) is placed near the TV at a distance of a far-field (3-5m) condition, and our task aims to detect the keyword "Xiao-T Xiao-T" from the recorded utterances. In addition, parallel recordings for mid-field (1-1.5m, 2 channels) and near-field close-talking (0m, 1 channel) are provided.

Input Feature. We convert the wav utterance to a 40-dimensional log Mel filterbank (FBank) with a 30ms window size and a 10ms shift. We fixed the length of our FBank at 2s, with shorter utterances being right-padded with zeros. During training, data augmentation is performed with random time shift of range between -100 to 100ms. Additionally, we apply SpecAugmentation, which involves applying two–frequency- and time-maskings of 25 and 7 to the audio.

Training details. We use the original noisy audio for all multi-channel modelling. All models are trained on a batch size of 64 and an initial learning rate (LR) of 6e-4. The LR decays with cosine annealing, reaching a lower bound of 1e-12. We use Adam and binary cross-entropy loss during optimization. To account for data imbalance, we utilized oversampling with bootstrapping to reduce the bias from class imbalance. We trained our model with curriculum learning in three phases, where the first phase uses near-field followed by mid-field and then far-field dataset. No additional noise perturbation is done during preprocessing. Since there is not much literature on small footprint multi-channel KWS, we compare our small footprint model with a popular approach by enhancing the noisy far-field audio with a beamformer [1] before fitting them to SOTA small models. This reduces the noise distortion and improves the detection, creating a more challenging ground to surpass, showing the significance of our work.

Metrics. We evaluate all models based on the ***score***, i.e. sum between the false alarm rate (FAR) and false rejection rate (FRR). This metric offset the likelihood of an over-optimistic assessment derived from the highly imbalanced class distribution, where FAR and FRR are defined as follows

$$FAR = \frac{FP}{FP + TN} \qquad FRR = \frac{FN}{FN + TP}$$

3.2. Results

From Table 1, we can observe that the eval set is more challenging with a consistently higher error score achieved by several SOTA models, and the official baseline (CNN-LSTM) from the challenge has achieved a decent performance of 0.34. However, our single-channel ConvMixer with 124K parameter size has evidently outperformed the former with a 48% boost in performance, indicating the robustness of the network. Our proposed multi-channel model (without centroid awareness) has also consistently beaten the single-channel ConvMixer with an additional relative gain of 13.9% for the dev set and 0.5% for eval set. Most importantly, after adding the centroid awareness, the score has a further gain of 5.6% on the eval set. From this, we can see that the centroid awareness is more beneficial for more noisy and challenging audio like the eval set, and even surpassing the large 3D-ResNet model [24].

Overall, our proposed final system has achieved a total of 54.5% improvement from the dev set and 55.8% for the eval set in comparison to the official baseline. Besides, the model is merely 23% in size of the baseline. In addition, we have compared our proposed networks against recent SOTA small footprint KWS models (i.e., MatchboxNet and Keyword Transformer-KWT). To improve the quality of noisy far-field audio, we employ the MVDR beamforming technique with a 90° beamformer, before fitting them to SOTA small footprint KWS models. Although both models have outperformed the official baseline, their best performance on the eval set is above the score of 0.2. In comparison, we have outdo theirs with the gain of at least 26.6% for the eval set and 26% for the dev set.

[1] https://github.com/AkojimaSLP/Beamforming-for-speech-enhancement

Table 1: *Performance of our experimental models with Task 1, MISP challenge 2021 development and evaluation set.* ***DA*** *refers to the set of data augmentations implemented in their paper.* [1] *estimated size from their proposed architecture*

Model Architectures	Params (K)	Development Set			Evaluation Set		
		FAR (↓)	FRR (↓)	Score (↓)	FAR (↓)	FRR (↓)	Score (↓)
Official Baseline [22]	2,682	0.181	0.094	0.275	0.261	0.083	0.344
90° Beamform (MVDR) + MatchboxNet (Single-Ch) [4]	140	0.121	0.048	0.169	0.117	0.103	0.220
90° Beamform (MVDR) + KWT-1 (Single-Ch) [23]	607	0.135	0.060	0.195	0.054	0.153	0.207
90° Beamform (MVDR) + ConvMixer (Single-Ch) [13]	124	0.056	0.088	0.144	0.048	0.121	0.169
Multi-channel 3D-ResNet (DA) [24]	13,619[1]	0.053	0.082	0.135	N.A.	N.A.	0.158
Multi-channel ConvMixer (6-Channel) [Ours]	415	0.050	0.074	**0.124**	0.043	0.118	0.161
Centroid Aware Multi-channel ConvMixer (6-Channel) [Ours]	622	0.034	0.091	0.125	0.044	0.107	**0.152**

Table 2: *Performance on eval set with front-end processing*

Models	Evaluation Set			
	Params (K)	FAR (↓)	FRR (↓)	Score (↓)
3-look BF	473	0.047	0.090	0.137
6-channel + WPE	622	0.040	0.136	0.176
WPE + 3-look BF [Ours]	**473**	0.054	0.072	**0.126**
3D-ResNet (DA + BF) [24]	13,619[1]	N.A.	N.A.	0.122
A-transformer (BF) [26]	15,417[1]	N.A.	N.A.	**0.106**
A-conformer (BF) [26]	27,147[1]	N.A.	N.A.	0.116

3.3. Multi-look Beamformer and WPE

In this section, we combine front-end enhancement as in [16, 25] to maximize the potential of our framework by modifying it into a multi-look beamforming KWS. We replace the raw microphone array with a set of beamformed (BF) signals aimed at 10°, 90° and 170°, respectively. We also include a reference wav signal from channel-0 raw to preserve the information of the original utterance. Additionally, we have considered dereverberation with WPE [2]. Altogether, we obtain an enhanced multi-look (3-look BF + ch0) KWS ConvMixer model with centroid awareness.

The results presented in Table 2 show that the 3-look beamformed signals of our model obtain a score of 0.137 and a gain of 10%. However, dereverberation with WPE alone does not help to build a better system, likely due to the generated artifacts that instil adverse distortion. Furthermore, dereverberation seems to lower our FAR performance. By combining the two, our proposed model has attained the score of **0.126** with a relative gain of 63% compared to the baseline, despite incurring only a small fractional cost in latency for the front-end enhancement. Furthermore, we compare our final system against the latest SOTA performance for this competition dataset. Our model fares well against the 3D-ResNet with similar performance but is only 3.4% of its size. However, we did not outperform the transformer and conformer models. Despite that, our model is only 3.0% of the size of transformer and 1.7% of conformer. Moreover, all of the reported SOTA models (3D-ResNet, transformer, conformer) were trained with augmented noise and speed perturbation that increased the training data by more than 3 times. In contrast, we did not use those technique as we aimed to analyze and evaluate our proposed architecture with minimal benefits from increasing data volume.

3.4. Effect of the proposed Centroid Based Awareness

To demonstrate the effectiveness of our proposed Centroid Based Awareness method, we use a t-SNE plot as Figure 2 to show the difference between our centroid method and the vanilla method for the latent space embedding approach using the Google Command dataset [27] mixed with -5dB FreeSound [28] noise. This helps to illustrate the effect of our method with a controlled noise level at -5dB. As shown in the plot, our centroid method results in well-separated clusters without multiple sub-clusters and a smaller area of discrepancy, indicating its higher confidence and lower model error compared to the vanilla method.

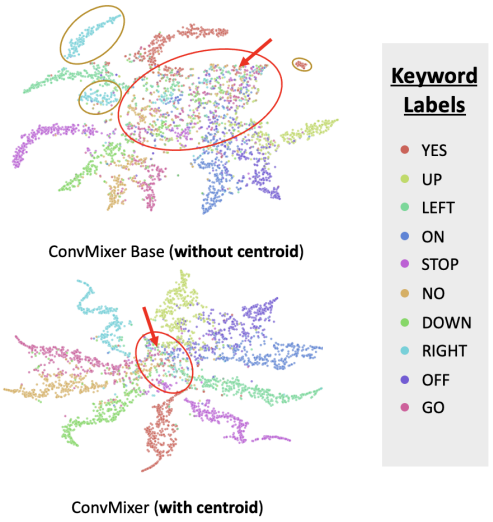

Figure 2: *The t-SNE plot to show the effectiveness of the centroid method from the latent space embedding approach using Google Command dataset [27] and mixing with -5dB FreeSound [28] noise.*

4. Conclusions

To conclude, we proposed a novel small-footprint model for multi-channel KWS with a ConvMixer module and centroid-based awareness. Our model achieves a compelling gain with a score of 0.126 and a 63% boost against the official baseline in noisy and far-field environments. Additionally, our best model framework outperforms recent SOTA small footprint KWS models with 473K parameters and is competitive against SOTA models for this dataset but with smaller parameters (3% in size). Overall, our model demonstrates better robustness in noisy and far-field environments.

5. Acknowledgements

This work was supported by Alibaba Group through Alibaba Innovative Research (AIR) Program and Alibaba-NTU Singapore Joint Research Institute (JRI), Nanyang Technological University, Singapore

[2] https://github.com/fgnt/nara_wpe

6. References

[1] Y. Zhang, N. Suda, L. Lai, and V. Chandra, "Hello edge: Keyword spotting on microcontrollers," *arXiv preprint:1711.07128*, 2017.

[2] 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," *Proc. INTERSPEECH 2019 – 20th Annual Conference of the International Speech Communication Association*, 2019.

[3] O. Rybakov, N. Kononenko, N. Subrahmanya, M. Visontai, and S. Laurenzo, "Streaming Keyword Spotting on Mobile Devices," in *Proc. INTERSPEECH 2020 – 21st Annual Conference of the International Speech Communication Association*, 2020, pp. 2277–2281.

[4] S. Majumdar and B. Ginsburg, "MatchboxNet: 1d time-channel separable convolutional neural network architecture for speech commands recognition," *Proc. INTERSPEECH 2020 – 21st Annual Conference of the International Speech Communication Association*, 2020.

[5] I. López-Espejo, Z.-H. Tan, J. H. Hansen, and J. Jensen, "Deep spoken keyword spotting: An overview," *IEEE Access*, vol. 10, pp. 4169–4199, 2021.

[6] D. Ng, R. Zhang, J. Q. Yip, C. Zhang, Y. Ma, T. H. Nguyen, C. Ni, E. S. Chng, and B. Ma, "Contrastive speech mixup for low-resource keyword spotting," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[7] A. Zhang, H. Wang, P. Guo, Y. Fu, L. Xie, Y. Gao, S. Zhang, and J. Feng, "Ve-kws: Visual modality enhanced end-to-end keyword spotting," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023.

[8] Z. Yang, D. Ng, X. Li, C. Zhang, R. Jiang, W. Xi, Y. Ma, C. Ni, J. Zhao, B. Ma, and E. S. Chng, "Dual-memory multi-modal learning for continual spoken keyword spotting with confidence selection and diversity enhancement," *Proc. INTERSPEECH 2023 – 24th Annual Conference of the International Speech Communication Association*, 2023.

[9] Y. Wang, P. Getreuer, T. Hughes, R. F. Lyon, and R. A. Saurous, "Trainable frontend for robust and far-field keyword spotting," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5670–5674.

[10] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.

[11] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 246–250.

[12] D. Ng, J. Q. Yip, T. Surana, Z. Yang, C. Zhang, Y. Ma, C. Ni, E. S. Chng, and B. Ma, "I2cr: Improving noise robustness on keyword spotting using inter-intra contrastive regularization," *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 605–611, 2022.

[13] 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 *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 3603–3607.

[14] B. Li, T. N. Sainath, R. J. Weiss, K. W. Wilson, and M. Bacchiani, "Neural network adaptive beamforming for robust multichannel speech recognition," *Proc. INTERSPEECH 2016 – 17th Annual Conference of the International Speech Communication Association*, pp. 1976–1980, 2016.

[15] P. C. Yong and S. Nordholm, "Real time noise suppression in social settings comprising a mixture of non-stationary anc transient noise," in *Proc. 25th European Signal Processing Conference (EUSIPCO)*. IEEE, 2017, pp. 588–592.

[16] T. N. Sainath, R. J. Weiss, K. W. Wilson, B. Li, A. Narayanan, E. Variani, M. Bacchiani, I. Shafran, A. Senior, K. Chin *et al.*, "Multichannel signal processing with deep neural networks for automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 965–979, 2017.

[17] T. Ochiai, S. Watanabe, T. Hori, J. R. Hershey, and X. Xiao, "Unified architecture for multichannel end-to-end speech recognition with neural beamforming," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1274–1288, 2017.

[18] J. Wu, Y. Huang, H.-J. Park, N. Subrahmanya, and P. Violette, "Small footprint multi-channel keyword spotting," in *Proc. Odyssey 2020 The Speaker and Language Recognition Workshop*, 2020, pp. 391–395.

[19] I. O. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit *et al.*, "Mlp-mixer: An all-mlp architecture for vision," *Advances in Neural Information Processing Systems*, vol. 34, pp. 24 261–24 272, 2021.

[20] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.

[21] D. Pfau, "A generalized bias-variance decomposition for bregman divergences," *Unpublished Manuscript*, 2013.

[22] H. Chen, H. Zhou, J. Du, C.-H. Lee, J. Chen, S. Watanabe, S. M. Siniscalchi, O. Scharenborg, D.-Y. Liu, B.-C. Yin, J. Pan, J.-Q. Gao, and C. Liu, "The first multimodal information based speech processing (misp) challenge: Data, tasks, baselines and results," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[23] A. Berg, M. O'Connor, and M. T. Cruz, "Keyword Transformer: A Self-Attention Model for Keyword Spotting," in *Proc. INTERSPEECH 2021 – 22nd Annual Conference of the International Speech Communication Association*, 2021, pp. 4249–4253.

[24] M. Cheng, H. Wang, Y. Wang, and M. Li, "The dku audio-visual wake word spotting system for the 2021 misp challenge," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 9256–9260.

[25] X. Ji, M. Yu, J. Chen, J. Zheng, D. Su, and D. Yu, "Integration of multi-look beamformers for multi-channel keyword spotting," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7464–7468.

[26] Y. Xu, J. Sun, Y. Han, S. Zhao, C. Mei, T. Guo, S. Zhou, C. Xie, W. Zou, and X. Li, "Audio-visual wake word spotting system for misp challenge 2021," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 9246–9250.

[27] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," *arXiv preprint:1804.03209*, 2018.

[28] E. Fonseca, J. Pons Puig, X. Favory, F. Font Corbera, D. Bogdanov, A. Ferraro, S. Oramas, A. Porter, and X. Serra, "Freesound datasets: a platform for the creation of open audio datasets," in *Proc. the 18th ISMIR (International Society for Music Information Retrieval) Conference*, 2017.