IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 20 August 2023, accepted 4 September 2023, date of publication 11 September 2023,
date of current version 20 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3314191

RESEARCH ARTICLE

Joint Framework of Curriculum Learning and Knowledge Distillation for Noise-Robust and Small-Footprint Keyword Spotting

JAEBONG LIM AND YUNJU BAEK
School of Computer Science and Engineering, Pusan National University, Busan 46241, South Korea
Corresponding author: Yunju Baek (yunju@pusan.ac.kr)

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2023-RS-2023-00260098) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation).

ABSTRACT Spoken keyword spotting, which is characterized by simplicity and low latency, has been widely used in consumer electronics to facilitate always-on voice interfaces. Small-footprint keyword spotting based on tiny convolutional neural networks can be implemented on resource-constrained, yet energy-efficient, microcontrollers in real time. However, it is difficult for tiny neural networks to learn the noise-robustness properties essential for successful voice interfaces. To overcome this problem, this study proposes a joint framework of curriculum learning and knowledge distillation for noise-robust small-footprint keyword spotting. The proposed joint framework applies noise-mixture curriculum learning to a network that is sufficiently large, to learn various noise situations. Subsequently, knowledge distillation is applied to compress the large network into a sufficiently small network for use in an onboard microcontroller. To enhance the effectiveness of the joint framework, a curriculum learning approach is proposed with a new noise mixture strategy along with knowledge distillation that employs an effective ensemble of neural network snapshots for each curriculum stage. The proposed methods enable large networks to effectively learn noisy situations, thereby transferring noise robustness to small networks. The effectiveness of the joint framework was illustrated on the Google Speech Commands dataset with noise mixtures incorporated from various public noise datasets. The performance of the joint framework was superior in noisy situations compared to that of state-of-the-art noise-robust keyword-spotting methods. Consequently, the proposed framework significantly improves the usability of voice interfaces in consumer electronics.

INDEX TERMS Curriculum learning, data augmentation, joint framework, knowledge distillation, neural network compression, noise-robust keyword spotting, small-footprint keyword spotting.

I. INTRODUCTION

With the widespread use of intelligent consumer electronics in daily life, voice interfaces have become prevalent in natural and intuitive interactions. Spoken keyword spotting, owing to its simplicity and low latency, is widely used in consumer electronics to enable an always-on voice interface. Deep neural networks are commonly used for keyword spotting because of their superior accuracy [1]. A neural network that is trained on spoken keyword samples for predefined keyword identification recognizes the trained keywords and distinguishes them from all the other words or phrases in an audio stream. A keyword can involve a specific word or multiword phrase that functions as a wake-up word or command. In artificial intelligence assistance, wake-up words can be "OK Google," "Alexa," and "Hey Siri." Keyword spotting is also used to recognize simple commands, such as "Turn on," "Turn off," and "Play," in various consumer electronics applications.

Unlike other acoustic models, such as automatic speech recognition that is based on large neural networks with huge computational overhead, keyword spotting can be

The associate editor coordinating the review of this manuscript and approving it for publication was Wei-Wen Hu.

implemented using relatively small neural networks. Tiny convolutional neural networks have been actively studied to improve the performance of keyword spotting and reduce the computational load [2], [3], [4]. These networks convert input audio into image-like spectrograms and process them efficiently using only a few network parameters. In addition, studies have proposed various model compression techniques for deep neural networks [5], [6], [7]. One of the effective techniques is knowledge distillation, which transfers the superior performance of a large teacher network to a small student network [8]. In previous studies, keyword-spotting models based on tiny convolutional networks were implemented on resource-constrained but energy-efficient microcontrollers in real time [9], [10].

Noise robustness of acoustic models is crucial for a successful voice interface. Performance degradation of the acoustic model can arise from ambient noise which can significantly damage the usability of the voice interface. Speech enhancement [11], [12] and data augmentation have been proposed for noise robustness of acoustic models. A promising data augmentation approach for improving noise robustness is noise-mixture training, in which background noise is injected into clean samples in the training data [13], [14]. Noise-mixture training artificially introduces difficult training samples into the keyword-spotting model, rendering the model more robust against noise. Neural networks have difficulty in learning loud noise situations. Thus, curriculum learning for acoustic models has been studied [15], [16], as a learning method that increases the difficulty of training samples, whereby after the initial training on easy, clean samples, training is continued on difficult samples with loud noise, which is similar to a person studying according to a curriculum [17]. Curriculum learning is widely used for noise-robust keyword spotting [18].

Recently, converging voice interfaces with wearable devices, such as hearables and IoT devices for smart homes, has been actively investigated [19], [20]. These tiny consumer electronics utilize spoken keyword spotting to create a simple but valuable hand-free voice interface. Unlike smart speakers with high-performance processors [21], wearables and IoT devices utilize low-performance and low-power microcontrollers [22] for spoken keyword spotting. Although smart speakers exhibit good performance when using multiple microphones [23], mounting multiple microphones is difficult because of the nature of these small devices [24]. Therefore, the development of noise-robust small-footprint keyword spotting is required for devices with minimal resources and single microphones.

Although the effectiveness of curriculum learning has been demonstrated in keyword spotting, it is challenging to develop noise-robust and small-footprint keyword spotting because it is difficult to thoroughly learn loud noise situations [25]. [26], [27]. When there are several difficult samples, as in loud noise scenarios, a joint framework for robust training and compression is a popular approach in other domains. Existing studies in the fields of computer vision and natural language processing have proposed joint frameworks for combining curriculum learning and knowledge distillation methods [28], [29]. However, studies on joint optimization methods for noise robustness are scarce in the field of acoustic modeling.

In this study, we propose a joint framework for curriculum learning and knowledge distillation for noise-robust small-footprint keyword spotting. The joint framework comprises a two-phase algorithm. First, curriculum learning is applied to a sufficiently large network to learn various noisy situations. Subsequently, knowledge distillation is applied to compress the large network into a network that is sufficiently small for use as an onboard microcontroller. We found that distilling the small network after applying curriculum learning to the large teacher network is superior to directly applying curriculum learning to the small network. To enhance the effectiveness of the joint framework, a curriculum learning based on a new noise mixture strategy is proposed along with knowledge distillation that employs an effective ensemble of neural network snapshots for each curriculum stage. Using these methods, large networks can effectively learn to handle loud noise situations, and the resulting noise robustness can be transferred to the smaller networks. The main contributions of this study are summarized as follows.

1) To the best of our knowledge, this is the first study employing a joint framework of curriculum learning and knowledge distillation for noise-robust small-footprint keyword spotting.

2) Novel curriculum learning and knowledge distillation methods are proposed to enhance the effectiveness of the joint framework.

3) The proposed framework exhibited state-of-the-art performance on four well-known noise datasets comparable to that of current state-of-the-art methods in terms of noise robustness.

The remainder of this paper is organized as follows: In Section II, we review related studies on small-footprint keyword spotting and noise-robust keyword spotting. In Section III, the proposed joint framework for noise-robust small-footprint keyword spotting is described. In Section IV, we discuss the experimental results obtained using public datasets. Finally, in Section V, conclusions and future work are presented.

II. RELATED WORK

A. SMALL-FOOTPRINT KEYWORD SPOTTING

With the advances in keyword spotting in industry and academia, smart consumer electronics with voice interfaces have become widespread. These devices utilize high-performance processors and multiple microphones for improved accuracy under various conditions such as noise. The Google Smart Speaker series have a quad-core application processor and more than two microphones [21]. For such smart speakers, Yu et al. proposed a noise-robust keyword-spotting model with noise cancellation, using six

microphones [23]. The number of parameters in the model was over 5.1 M.

Wearables and IoT devices with always-on voice interfaces have become increasingly attractive for hands-free control and audio scene understanding [30]. For example, hearables can benefit from built-in spoken keyword spotting for hands-free playback and volume control. In smart homes, IoT devices can be controlled using natural voice commands. These devices use low-power microcontrollers with only hundreds of KB of flash memory and tens of KB of RAM. Because of the nature of these small devices, it is difficult to mount multiple microphones. Fernandez-Marqueset et al. investigated tiny keyword spotting models for various microcontrollers [22]. Kim proposed a dedicated chip for wearable IoT devices for keyword-spotting models [24]. However, these studies on small-footprint keyword spotting do not consider noise robustness, which is crucial for successful voice interfaces.

Tiny convolutional neural networks (CNNs) have been widely used for small-footprint keyword spotting because they can efficiently learn spatial information from audio sequences. State-of-the-art networks use efficient structures composed of repeated blocks based on residual and depth-wise separable convolutions [2]. For efficiency, these convolutional networks use one-dimensional (1D) or partial two-dimensional (2D) convolutions. Utilizing temporal 1D convolutions, TC-ResNet requires fewer computations than 2D approaches [3]. BC-ResNet employs broadcast residual learning to address the inefficiency of 2D convolution and inferior performance of 1D convolution [4]. The authors apply frequency-wise 1D convolution to 2D audio features and then average the 2D features over frequency to obtain temporal features. After temporal operations, the 2D features are subjected to residual mapping, whereby the 1D residual information is broadcast.

Neural network compression has also received considerable attention in small-footprint keyword spotting and is categorized into three approaches: pruning, quantization, and knowledge distillation. Pruning reduces model size and number of operations by eliminating unimportant parameters that do not degrade performance [31]. The quantization method approximates floating-point values by a set of integers and scaling factors, to achieve a smaller size for more efficient computations at the expense of lower bit-width representation [32]. Knowledge distillation is a learning framework that utilizes a teacher–student network, whereby the teacher network transfers its knowledge to enhance the performance of the student network [8]. Ensemble distillation improves the performance of a distilled student by extracting knowledge from the ensemble of multiple teachers, thereby encoding it for the student [33].

Song et al. proposed a knowledge distillation technique for a lightweight encoder as a replacement for complex speech front ends, such as mel-frequency cepstral coefficient (MFCC) or convolutions, without sacrificing the performance of small-footprint keyword spotting [5]. Tucker et al. investigated two methods to improve the efficiency of a small-footprint keyword-spotting model, namely, knowledge distillation and quantization [6]. Kim et al. introduced a method that simultaneously applies pruning, quantization, and knowledge distillation to small-footprint keyword spotting [7]. However, these studies on small-footprint keyword spotting do not consider noise robustness.

TABLE 1. Related work for noise-robust keyword spotting.

	Model Params #	Mic. #	**Curriculum Learning**	**Knowledge Distillation**	Public Dataset	Test SNR (dB)	Test Accuracy (%)
Larger model [26]	24000k	1	✔	✔	O	0 ~ 20	91.9
Multi-microphone [23]	5100k	6			X	-12 ~ 6	94.0
ImportantAug, 2022 [26]	100k	1			O	−12.5	56.5
[18]	-	1	✔		X	-	-
ConvMixer, 2022 [25]	119k	1	✔		O	-10	71.9
PKD, 2022 [27]	102k	1		✔	O	0	82.3
Proposed Joint Framework	27k	1	✔	✔	O	−12.5	78.6

B. NOISE-ROBUST KEYWORD SPOTTING

Data augmentation has been widely applied to acoustic models to improve neural network performance by adding diversity to the training data. For a successful voice interface, acoustic models must effectively distinguish speech from background noise. A promising augmentation approach for improving noise robustness is noise-mixture training, in which environmental noise is added to clean samples in the training data [13], [14]. Noise mixture training introduces difficult training samples into the model, thereby making the model more robust against noise. Trinh et al. proposed ImportantAug, which augments training data by injecting noise into unimportant speech regions, and a data augmentation agent trained to optimize noise addition and minimize its effect on performance, predicts the importance level of the speech [26].

Noise mixture training is a promising technique that achieves noise robustness in keyword spotting. However, it is difficult for a small-footprint keyword-spotting model to learn loud noise situations; thus, the model becomes incompetent when confronted by louder noise. Curriculum learning, which involves progressive training, has been studied to overcome this problem of acoustic models [15], [16]. Curriculum learning divides the learning process into several stages and gradually increases the difficulty of training samples, training first on easy samples such as clean speech and then on difficult samples with loud noise [17]. Curriculum learning is more effective than conventional training in obtaining noise robustness, similar to how a person studies according to a curriculum. Therefore, research on keyword spotting widely employs curriculum learning for noise robustness.

As shown in Table 1, although previous studies on small-footprint keyword spotting have improved noise

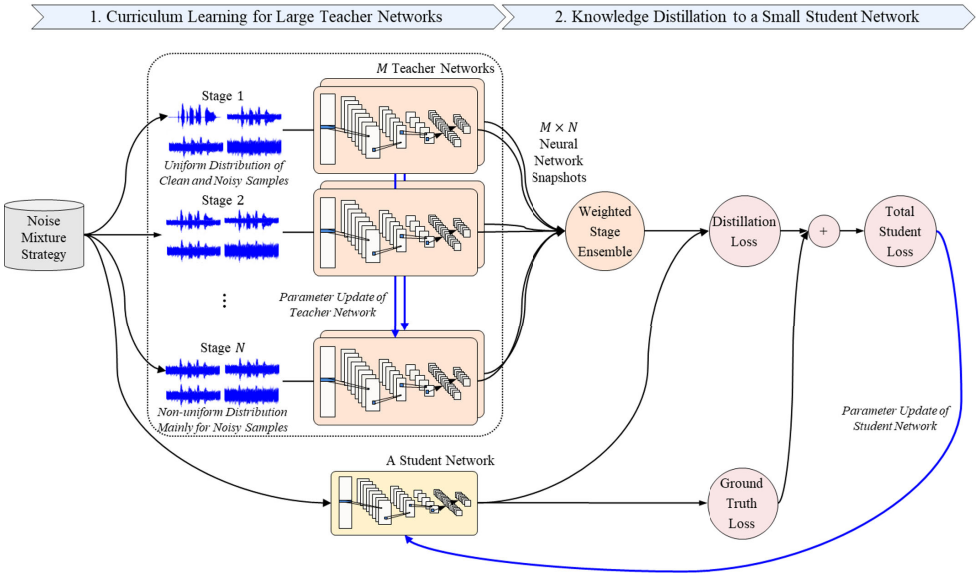

FIGURE 1. Overview of the proposed joint framework of curriculum learning and knowledge distillation. The proposed joint framework applies the proposed noise mixture curriculum learning to a large model, whereby the large model is compressed into a small one using weighted ensemble knowledge distillation.

robustness, their performance still needs to be enhanced when the noise is intense. Larger models or multimicrophone approaches display better robustness but are unsuitable for wearables and IoT devices [23], [26]. Developing robust and small-footprint keyword spotting is challenging because it is difficult for tiny neural networks to learn noise robustness. To address this issue, a joint framework of curriculum learning and compression techniques, such as knowledge distillation, is required to improve noise robustness in small-footprint keyword spotting. Although Guimaraes et al. utilized both curriculum learning and knowledge distillation for keyword spotting, they used distillation to exploit a massive self-supervised model, not to compress the keyword spotting model [26].

Curriculum learning and knowledge distillation are widely used in small-footprint keyword spotting. Higuchi et al. proposed dynamic curriculum learning to train a keyword-spotting model on clean and noisy samples according to the data parameters [18]. These parameters automatically learn the difficulty of the classes and instances using gradient optimization. Ng et al. proposed a keyword-spotting model called ConvMixer, which along with a traditional curriculum learning method enhances noise robustness [25]. The training process is divided into five progressively harder stages, and the model is initially trained on clean speech samples without noise. A knowledge distillation method for noise-robust small-footprint keyword spotting, referred to as prototypical knowledge distillation (PKD), has been proposed to address the issue of network bias toward samples that are easy to train [27]. To effectively train hard samples, the PKD method utilizes a prototype distribution based on the distance between the class centroids and each embedding vector for knowledge distillation. However, further research that jointly utilizes curriculum learning and knowledge distillation is required for small-footprint keyword spotting.

Studies on joint frameworks have been actively conducted in other fields, such as computer vision and natural language processing. Zhu et al. proposed a combination of curriculum learning and knowledge distillation methods to solve the long-sentence training problem in natural language processing [27]. Panagiotatos et al. proposed a combination of curriculum learning and knowledge distillation methods for computer vision [28]. However, because such studies are scarce in the field of acoustic modeling, joint frameworks for noise robustness should be studied.

III. PROPOSED JOINT FRAMEWORK

To overcome the difficulty of learning the noise-robustness property in small-footprint keyword spotting, we propose a joint framework that first learns a large teacher network that is robust to loud noise and subsequently compresses it into a small student network. To maximize the effectiveness of the joint framework, we propose and combine novel curriculum learning and ensemble distillation methods. First, by applying the proposed noise mixture curriculum learning to a large model, noise robustness under loud noise scenarios is maximized. Subsequently, the noise-robust large network is effectively compressed into a small network using the proposed ensemble distillation.

Fig. 1 shows the characteristics of the proposed joint framework. Unlike traditional curriculum learning, in which the range of noise strength is gradually widened, the proposed curriculum adopts a strategy of gradually increasing the ratio of noisy samples while learning the noisy samples from an early stage for overall effective learning. This demonstrates better performance for loud noise scenarios. The proposed

Algorithm 1 Proposed Curriculum Learning

Input: Speech dataset D_{Speech}, Noise dataset D_{Noise}, SNR distribution R_n

Output: N Teacher Network Snapshots

For n **in** 1: N **do**
  For e **in** 1: E_n **do**
    For d_{Speech} **in** D_{Speech} **do**
      Randomly select a d_{Noise} from D_{Noise}
      Select a SNR r in SNR distribution R_n
      Calculate weight factor ω to get SNR r
      $d_{Mix} \leftarrow d_{Speech} + \omega d_{Noise}$
    End for
    Train teacher network θ^T with D_{Mix}
  End for
  Get teacher network snapshot θ^{T_n} for stage n
End for

ensemble distillation method utilizes neural network snapshots at each stage of the curriculum learning process. Thus, the knowledge of the large teacher network is effectively transferred to the small student network, and the performance of the compressed network is more effectively preserved.

The proposed curriculum learning and knowledge distillation work synergistically, as follows: the proposed curriculum learning induces better performance under loud noise scenarios even though the performance is slightly lower when the noise is weak. The performance degradation under weak noise scenarios is mitigated by the proposed knowledge distillation, which ensembles all neural network snapshots generated during curriculum learning. Therefore, a distilled network achieves good performance both in early stage network snapshots with weak noise and later-stage network snapshots with loud noise.

A. PROPOSED CURRICULUM LEARNING

Unlike the gradual widening of the loudness range of noisy samples in traditional curriculum learning, the proposed curriculum learning jointly trains on clean and noisy samples in the initial stage, thereafter gradually increasing the proportion of noisy samples in each subsequent stage. As a result of the proposed curriculum learning, the performance of the neural network trained on the initial stage is improved under weak-noise scenarios, and the neural network trained up to the final stage has improved performance under loud-noise scenarios. In subsequent noise distillations, these neural network snapshots efficiently provide information on various noise intensities.

The key to successful curriculum learning for noise-robust keyword spotting, is an effective noise mixture strategy for each curriculum stage. In the curriculum learning process, the noise mixture strategy determines how to mix noise with clean speech at each stage to create training samples for the neural network. The mixture strategy dictates the characteristics of the final neural network as the stages progress. For example, with traditional curriculum training, in which

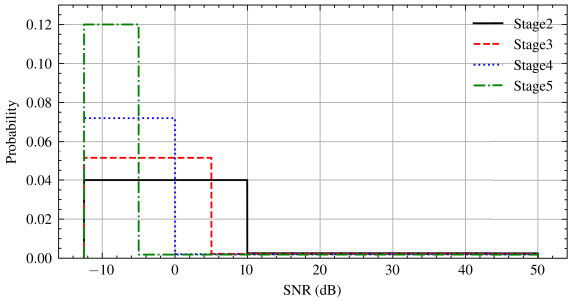

FIGURE 2. SNR distribution R_n for each stage n. The probability of selecting low-SNR samples for each subsequent stage is increased.

the signal-to-noise ratio (SNR) range of noise is gradually widened, neural networks eventually perform well under various degrees of noise. However, traditional curriculum learning is less effective because it is difficult for small neural networks to learn loud noises. Therefore, the proposed curriculum learning focuses on loud noise situations as the stages progress, which improves the performance under loud noise despite a slight drop in performance when noise is weak.

To enhance noise robustness, SNR was employed as the main metric for the noise mixture strategy. The SNR of an audio sample is defined as the ratio of the power of speech to the power of background noise, as in (1).

$$SNR = 10log_{10}(\frac{P_{speech}}{P_{noise}}), \tag{1}$$

where P denotes the average power. An SNR of less than 0 dB indicates that the background noise is louder than the noise in speech. An SNR of -10 dB indicates that the power of the noise is 10 times greater than that of a speech sample.

Algorithm 1 presents the training process for the proposed curriculum learning. The training process is divided into N progressively more difficult stages. For each stage, the teacher network is trained on a noise-mixed dataset D_{Mix} for e epochs. Samples from the dataset D_{Mix} are newly synthesized for each epoch according to the SNR distribution R_n of the corresponding stage. A clean sample d_{Speech} is mixed with randomly selected noise d_{Noise} of SNR r. The parameters θ^T of the teacher network f are trained on D_{Mix}. Consequently, N neural network snapshots θ^{T_n} of the teacher network are obtained.

The SNR distribution R_n comprises the sampling and main ranges. The SNR of noise samples is selected from the main range with probability ρ and from the sampling range, excluding the main range, with probability $1 - \rho$. In the first stage, the SNR is uniformly selected from the sampling range since the sampling and main ranges are the same. In subsequent stages, the main range is gradually narrowed to increase the proportion of noisy samples. Consequently, the SNR distribution R_n gradually focuses on loud noise situations as the stages progress. The selected SNR distribution R_n for each stage n is shown in Fig. 2.

The settings of the parameters for the proposed curriculum learning are summarized as follows:

R_n: R_n is a factor that adjusts the focus on loud situations as the stages progress. The sampling range has a lower bound of −15 dB for samples with the loudest noise and an upper bound of 50 dB for almost clean samples. The main ranges for each stage are [−15, 50] dB, [−15, 10] dB, [−15, 5] dB, [−15, 0] dB, and [−15, −5] dB. We set ρ to 0.9. The larger ρ is, the higher is the probability that the SNR of the main range is sampled. In other words, the larger ρ is, the higher is the concentration of loud noises. These choices are validated in Section IV.

N: number of stages. We set the number of stages to five, as in other curriculum learning studies [18], [25].

Speech augmentation, which is commonly used in the training of acoustic models, was applied under the same conditions at all stages [35].

Volume change: The volume of the clean samples was randomly varied in the range of 40–180%.

Shifting: The speech in the sample was shifted by 25–125% to change the central position.

Resampling: Samples were randomly resampled in the range of 90–110% to change the speed or pitch.

SpecAugment [36]: Samples in the range of 0–5 were masked in the frequency and time domains.

B. PROPOSED ENSEMBLE DISTILLATION

Knowledge distillation was used to compress curriculum-based teacher networks into smaller student networks. The proposed ensemble distillation method considers noise robustness to maximize the effect of distilling the student network based on the curriculum of teacher networks. The proposed ensemble distillation effectively uses all the neural network snapshots for each curriculum stage.

The knowledge of teacher networks is reflected when calculating the loss function while distilling the student network. It is necessary to efficiently determine which of the numerous large networks to receive knowledge from. In the proposed ensemble distillation, a weighted-stage ensemble distillation mechanism is adopted, to adjust the weight of the teacher's knowledge that should be prioritized according to the training data when distilling students. Teachers, which are neural network snapshots of the previous curriculum learning phase, focus more on loud situations as the stages progress. Therefore, the noisier the data to be trained, the more knowledge is passed on from the snapshots in the later stages, and the cleaner the data, the more knowledge is passed on from the snapshots in the earlier stages, as shown in Fig. 3.

We first explain how traditional knowledge distillation and ensemble distillation work in terms of the loss function, prior to describing ensemble distillation using all the snapshots for each curriculum stage. Finally, the proposed weighted-stage curriculum learning method is presented along with the loss function and weight-adjustment method.

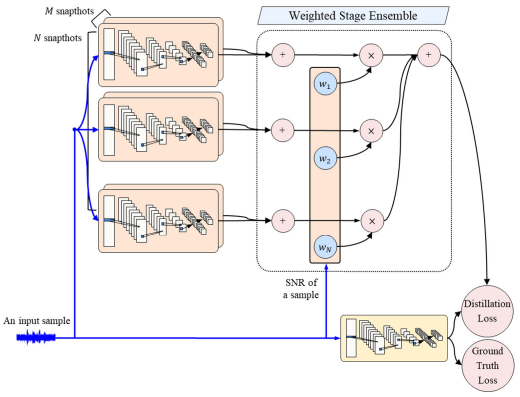

FIGURE 3. Overview of proposed ensemble distillation. The SNR of a sample determines the snapshots that transfer the most knowledge.

In traditional knowledge distillation, a teacher network is used to distill the student network. The teacher logit output can be represented as $z^T = f(\theta^T, x)$, where x is the input sample, and θ^T refers to the trained parameters of the teacher network f; $z^S = g(\theta^S, x)$ denotes the student logit output, where θ^S represents the parameters of the student network g. To determine the best parameters θ^S, the student is trained to minimize the cross entropy $\mathcal{H}$ between the ground truth y and the student probability output p^S after applying the SoftMax function σ, as shown in (2) and (3).

$$p^S = \sigma(z^S) = \frac{exp(z_k^T)}{\sum_{j=1}^{K} exp(z_j^T)} \text{for all classes } k, \quad (2)$$

$$L_{CE}^S = \mathcal{H}\left(y, p^S\right) = \sum_{k=1}^{K} -y_k \, log\left(p_k^S\right). \quad (3)$$

Along with the cross-entropy loss L_{CE}^S, the Kullback–Leibler (KL) divergence between the softened probability outputs of the teacher and student are considered in transferring the knowledge of the teacher network, as shown in (4) and (5). The softened probability outputs p^T and p^S are smoothed by a temperature τ, as in other distillation studies. Consequently, the total loss function is the weighted sum of the cross-entropy loss L_{CE}^S and the knowledge distillation loss L_{KD} with the loss weight λ, as shown in (6).

$$p^T = \sigma(z^T, \tau) = \frac{exp(z_k^T/\tau)}{\sum_{j=1}^{K} exp(z_j^T/\tau)} \text{ for all classes } k, \quad (4)$$

$$L_{KD} = \tau^2 KL\left(p^T, p^S\right) = \tau^2 \sum_{k=1}^{K} p_k^T \, log\left(p_k^T/p_k^S\right), \quad (5)$$

$$L^S = (1-\lambda) L_{CE}^S + \lambda L_{KD}. \quad (6)$$

In traditional ensemble distillation, multiple teacher networks are used to distill the student network. Let $z^{T_m} = f(\theta^{T_m}, x)$ denote the logit output of each teacher. Some independently trained neural networks with the same architecture f, same training algorithm on the same data distribution, and different initializations by random seeds result in teachers with different parameters θ^{T_m}. As shown in (7), the mean softened output P^E of N teachers improves performance. Thus,

Algorithm 2 Proposed Ensemble Distillation

Input: $M \times N$ teacher network snapshots
Output: A student network

```
For e in 1: E_n do
  For d_Mix in D_Mix do
    For n in 1: N do
      w_n ← β
      If SNR of d_Mix is main range of R_n
        w_n ← α
      End if
    End for
    Get L̂^S and P̂^E with w_n Σ_{n=1}^{N} z^{T_{m,n}}
  End for
  Train student network with L̂^S
End for
```

the KL divergence between the mean softened probability output P^E of the teachers and the probability output p^S of the student is used for ensemble distillation, as shown in (8). The ensemble distillation loss L_{ED} replaces the knowledge distillation loss L_{KD} with the total student loss, expressed as in (9).

$$P^E = \sigma(\frac{1}{M}\sum_{m=1}^{M} z^{T_m}, \tau), \tag{7}$$

$$L_{ED} = \tau^2 KL\left(P^E, p^S\right), \tag{8}$$

$$\widehat{L}^S = (1-\lambda)\, L_{CE}^S + \lambda L_{ED}. \tag{9}$$

The proposed ensemble distillation method uses the neural network snapshots learned up to the n-th stage to improve ensemble efficiency. The network parameters $\theta^{T_{m,n}}$ of the neural network snapshots can be obtained from as many as n curriculum stages for the m-th network parameter θ^{T_m}. Let $z^{T_{m,n}} = f(\theta^{T_{m,n}}, x)$ denote the logit output of each teacher. With N snapshots for each M independently trained neural networks, a total of $M \times N$ snapshots are used by the ensemble. When assembling a large number of networks, it is important to induce different networks to create a synergistic effect without conflicts. However, simply assembling them all, as in (10), can often cause conflict.

$$\dot{P}^E = \sigma(\frac{1}{M \times N}\sum_{m=1}^{M}\sum_{n=1}^{N} z^{T_{m,n}}, \tau), \tag{10}$$

In the proposed ensemble distillation, the ensemble differs according to the samples used to train the neural network snapshots $\theta^{T_{m,n}}$. Unlike all the snapshots contributing equally to the ensemble, the proposed strategy is to vary the contribution of each snapshot to the ensemble according to the training sample. The proposed strategy for adjusting the weight w_n according to the training sample is shown in Algorithm 2. The weight w_n emphasizes the snapshots for the n-th curriculum stage, as expressed in (11):

$$\widehat{P}^E = \sigma(\frac{1}{M \times N}\sum_{m=1}^{M} w_n \sum_{n=1}^{N} z^{T_{m,n}}, \tau), \tag{11}$$

The $\sum_{n=1}^{N} z^{T_{m,n}}$ denotes the subset ensemble with only snapshots of the nth stage. Thus, the subset ensemble is a subset of snapshots $\theta^{T_{m,n}}$. If the SNR of a training sample x is included in the main range of the SNR distribution R_n of the nth stage, the weight w_n should be increased to emphasize the subset ensemble of the nth stage. The weight w_s is assigned the constant α, if the SNR is included in the main range of the SNR distribution R_n, or the constant β if it is not included, as shown in (12). The constant α is greater than the constant β, and these constants are set to appropriate values through a grid search. Consequently, the proposed ensemble distillation method is adaptive in emphasizing the snapshots for the nth curriculum stage according to the training sample x.

$$w_n = \begin{cases} \alpha & if\ x \in main\ range\ of\ R_n \\ \beta & else \end{cases} \tag{12}$$

The settings of parameters for the proposed ensemble distillation are summarized as follows:

τ: The temperature τ of the knowledge distillation loss. The larger τ is, a smoother teacher's logit is transmitted to the student. For the grid search, the temperature τ was set to five, which is a commonly used value for knowledge distillation.

λ: The weight λ of the knowledge distillation loss. The larger λ is, the larger is the reflection ratio of distillation loss. For the grid search, the loss weight λ was set to 0.1, a commonly selected value for knowledge distillation.

N: number of stages. N was set to five.

M: number of teacher networks. As M increases, the performance improves, but the learning cost also increases.

w_n: a factor that adjusts the focus on snapshots for loud scenarios. The w_n consists of α and β. As α increases, the weight of the snapshots for loud situations increases. As β increases, the snapshots are reflected evenly. We set α to 1 and β to 0. These choices are validated in Section IV.

IV. EXPERIMENTS

In this section, we present the experiments conducted to evaluate the effectiveness of the proposed joint framework for noise-robust small-footprint keyword spotting. The experiments examine the performance and model size of keyword-spotting models with respect to the methods employed. To evaluate noise robustness, we compared the accuracy for each SNR level by adjusting the noise volume in an unseen noise environment setting, using four public noise datasets: MUSAN [39], QUT [40], UrbanSound8K [41], and WHAM [42] for a quantitative comparison.

The experiment consisted of the following five steps: The proposed joint framework was compared to three state-of-the-art methods: ConvMixer, ImportantAugust, and PKD. The efficacy of the proposed curriculum learning was verified and compared to the baseline curriculum learning methods without curriculum learning and traditional curriculum learning methods. The efficacy of the proposed weighted-stage ensemble distillation was further verified through comparisons with baseline ensemble distillation, that is, ensemble

distillation without curriculum learning, with curriculum learning, and stage ensemble distillation. The hyperparameters of the proposed curriculum-learning and knowledge distillation methods were also explored. Finally, we extended the SNR of the public datasets and evaluated accuracy based on the SNR in loud noise situations.

A. EXPERIMENTAL SETTINGS

The experimental setup is described in detail to ensure the reproducibility of this study. To evaluate the effectiveness of the proposed joint framework, we selected the experimental settings, including the neural network architecture, data preprocessing method, and training hyperparameters.

For small-footprint keyword spotting, we used BC-ResNet, a state-of-the-art neural network architecture [4]. The BC-ResNet-8 architecture with 321k parameters was used as the teacher network, and the BC-ResNet-2 architecture with 27.3k parameters was used as the student network. To generate input data using BC-ResNet, 1 s of raw audio was preprocessed into 2D features of size 49×40 using MFCC conversion. BC-ResNet uses these inputs to output the classification results for 11 classes consisting of ten predefined keywords and one negative class. The training samples for the negative class included words other than the predefined keywords or background audio.

When training the teacher or student network from scratch, the experimental settings were as follows: all code was implemented in TensorFlow; the Adam optimizer was used with a default learning rate of 0.001 and a common mini-batch size of 256. When applying curriculum learning, the neural networks were trained for 4000 epochs, with 2000, 500, 500, 500, and 500 epochs in five stages. The neural networks were trained for 4000 epochs even when curriculum learning was not applied. As the number of training samples for the negative class was more than 10 times greater than that for one predefined keyword, the class weight of the negative class was set to 0.1 to prevent bias against the negative class.

The experimental settings specified to distinguish a student from the teacher were as follows: the Adam optimizer was used with a learning rate of 0.001 and a mini-batch size of 256, similar to scratch learning. The students were trained for 4000 epochs, regardless of the distillation method.

B. DATASETS

The Google Speech Commands (GSC) dataset version 2 was used in all the experiments [37]. This dataset included approximately 100k single-word audio clips of 35 unique words. The neural networks were trained using the training set and evaluated using the test set. Randomly selected noisy and clean samples were mixed at different SNR levels to create noisy environments. We used the FSK50k [38] dataset for training and the MUSAN [39], QUT [40], UrbanSound8K [41], and WHAM [42] datasets to evaluate the performance in unseen noise environments.

1) MUSAN

The public GSC-MUSAN dataset was used in a related study for noise-robust keyword spotting [26]. We used GSC-MUSAN mixed with noise from the MUSAN dataset to augment the clean samples from the GSC dataset. Because the recordings in MUSAN have variable lengths, GSC-MUSAN only uses the initial 1 s of each recording and discards shorter recordings, as the samples are limited to a maximum of 1 s. After removing the short samples, GSC-MUSAN randomly selects 175 noisy audio files and combines them with samples from the test set.

2) QUT

This dataset contains an audio file named ''HOME-LIVINGB-1.wav,'' which comprises 40 min of background noise recorded in a living room setting. The audio file was used to construct the test data in a manner similar to that of a related study on noise-robust keyword spotting [26]. Segments were randomly selected from this noisy audio file and blended with clean samples from the test set. Subsequently, the file sampling rate was modified from 48 to 16 kHz to match the GSC dataset.

3) UrbanSound8K

This dataset contains 8732 audio files of urban sounds that are shorter than 4 s from 10 classes: air conditioner, car horn, children playing, dog bark, drilling, engine idling, gunshot, jackhammer, siren, and street music. All excerpts were obtained from field recordings uploaded at www.freesound.org. This dataset was used in a manner similar to the QUT dataset.

4) WHAM

The WHAM or WSJ0 hipster ambient mixtures dataset is a compilation of two-speaker mixtures derived from the wsj0-2 mix dataset merged with authentic ambient noise samples. These samples were gathered from various public spaces such as coffee shops, restaurants, and bars within the San Francisco Bay Area. This dataset was used in a manner similar to the QUT dataset.

C. STATE-OF-THE-ART METHODS

The keyword-spotting performance of the joint framework was evaluated and compared with that of the state-of-the-art methods for noise-robust keyword spotting.

1) ConvMixer

Ng et al. [25] proposed ConvMixer, a new convolutional neural network architecture that adds a mixer layer to mix frequency and temporal domain features with 1D temporal, 2D frequency $\times$ temporary, and partial 2D convolutions. ConvMixer applies traditional noise mixture curriculum learning as a training strategy to enhance noise robustness. The training process is divided into five progressively more challenging stages. Initially, ConvMixer is trained on clean speech

TABLE 2. Comparison with state-of-the-art noise-robust and small-footprint keyword spotting methods.

	Model Parms #	Accuracy on Clean (%)	Accuracy on MUSAN (%)							Accuracy on QUT (%)						
			40 dB	30 dB	20 dB	10 dB	0 dB	-10 dB	-12.5 dB	40 dB	30 dB	20 dB	10 dB	0 dB	-10 dB	-12.5 dB
Curriculum Learning	**27k**	93.1	93.1	92.9	92.3	90.7	86.1	74.5	70.7	93.0	92.9	92.0	89.2	80.1	56.8	49.6
ConvMixer, 2022 [25]	119k	93.2	-	-	90.8	-	83.0	71.9	-	-	-	-	-	-	-	-
ImportantAug, 2022 [26]	100k	95.0	94.9	94.8	94.3	92.6	86.7	65.0	56.5	95.2	94.9	94.2	91.1	76.5	38.7	28.0
Proposed Joint Framework	**27k**	97.1	**97.0**	**96.8**	**96.4**	**95.2**	**91.1**	**82.3**	**78.6**	**97.3**	**97.0**	**96.6**	**94.3**	**87.8**	**68.5**	**64.1**

	Model Parms #	Accuracy on Clean (%)	Accuracy on UrbanSound8k (%)					Accuracy on WHAM (%)				
			20 dB	15 dB	10 dB	5 dB	0 dB	20 dB	15 dB	10 dB	5 dB	0 dB
Curriculum Learning	**27k**	93.1	92.5	91.7	90.3	87.7	82.5	92.8	92.1	91.0	88.6	84.0
PKD, 2022 [27]	102k	**97.7**	**96.7**	95.6	93.8	89.2	82.3	96.6	95.9	94.2	89.4	79.4
Proposed	**27k**	97.1	96.4	**95.8**	**94.9**	**93.1**	**89.2**	**96.9**	**96.4**	**96.1**	**94.5**	**91.1**

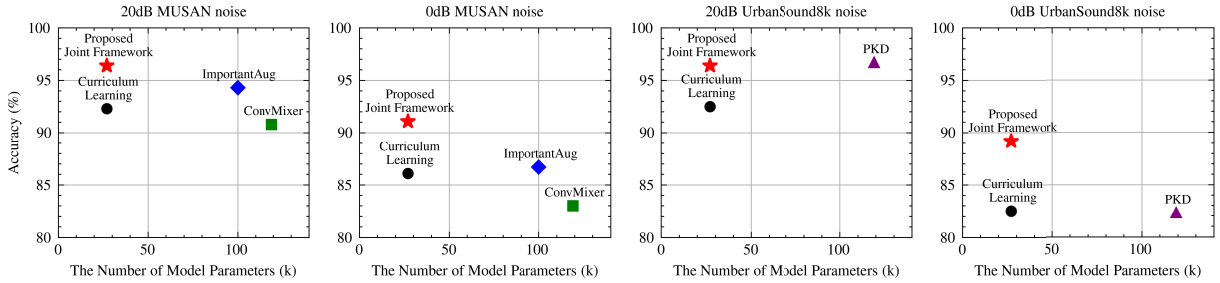

FIGURE 4. Efficacy of the proposed joint framework compared with curriculum learning and state-of-the-art methods on the MUSAN and UrbanSound8k datasets. The proposed joint framework showed accuracy improvement and model size reduction compared to the state-of-the-art.

samples without noise. In the subsequent three stages, noisy samples are introduced in increments of −5 dB as follows: {clean, 0 dB}, {clean, 0 dB, −5 dB}, {clean, 0 dB, −5 dB, −10 dB}.

2) IMPORTANT AUG

Trinh et al. [26] proposed Important Aug, a method that enhances the quality of noise samples for training acoustic models. This approach involves injecting noise exclusively into unimportant speech areas while leaving important regions untouched. The significance level for each speech is estimated by the augmentation agent trained to optimize the amount of noise introduced while minimizing its effect on accuracy.

3) PKD

Kim et al. [27] presented a robust knowledge distillation and feature extraction multilayer method. They proposed three different distance metrics for knowledge distillation and a novel feature extraction method that uses the distribution between class centroids and embedding vectors. Background noise is randomly injected, and random time shifting is applied.

D. EXPERIMENTAL RESULTS

1) PROPOSED JOINT FRAMEWORK RESULTS

Table 2 presents the results of the comparison between the proposed joint framework and recent state-of-the-art methods. The state-of-the-art ConvMixer, ImportantAug, and PKD methods were compared in terms of accuracy and model size. Keyword-spotting accuracy was compared at different SNRs for four noise datasets: MUSAN, QUT, UrbanSound8k, and WHAM. The performance of ImportantAug on the MUSAN dataset constitutes the experimental results under observed noise conditions, and the performances of the other methods, including the proposed method, are reported under the experimental results for unseen noise conditions. The proposed joint framework exhibits better keyword-spotting performance than all the state-of-the-art methods under all noise conditions. In particular, the proposed joint framework displays a significant performance improvement under loud noise condition, where the power of the noise audio is greater than or equal to that of the speech audio.

In Fig. 4, the keyword-spotting accuracy of the proposed joint framework is compared with that of the existing curriculum learning method. The black point shows the results of applying curriculum learning directly to the small network, and the red star denotes the results of the proposed method of distilling the small network after applying curriculum learning to the large teacher network. The proposed joint framework shows an average 8% improvement in accuracy and an average four-fold reduction in model size compared to state-of-the-art methods. The proposed joint framework has an accuracy of 96.4% at 20 dB MUSAN, 91.1% at 0 dB MUSAN, 96.4% at 20 dB UrbanSound8k, and 89.2% at 0 dB UrbanSound8k, compared with the accuracies of curriculum learning of 92.3%, 86.1%, 92.5%, and 82.5%, respectively.

TABLE 3. Evaluation of small and large networks according to curriculum methods.

	Model Size (Parms #)	Accuracy (%) on Clean (F1-score)	Averaged Accuracy (%) of Noises (F1-score)	Accuracy (%) on MUSAN (F1-score)							Accuracy (%) on QUT (F1-score)							Accuracy (%) on UrbanSound8k (F1-score)					Accuracy (%) on WHAM (F1-score)				
				40 dB	30 dB	20 dB	10 dB	0 dB	-10 dB	-12.5 dB	40 dB	30 dB	20 dB	10 dB	0 dB	-10 dB	-12.5 dB	20 dB	15 dB	10 dB	5 dB	0 dB	20 dB	15 dB	10 dB	5 dB	0 dB
No Curriculum	Small (27k)	95.7 (0.94)	86.5 (0.82)	95.6 (0.94)	95.4 (0.93)	94.8 (0.92)	93.1 (0.90)	87.1 (0.82)	73.0 (0.65)	68.9 (0.60)	95.5 (0.93)	95.3 (0.93)	94.5 (0.92)	92.1 (0.88)	81.5 (0.75)	53.2 (0.44)	45.8 (0.36)	94.6 (0.92)	93.7 (0.91)	92.1 (0.89)	88.9 (0.85)	82.8 (0.78)	95.0 (0.93)	94.5 (0.92)	93.1 (0.90)	90.4 (0.87)	84.8 (0.80)
	Large (313k)	97.9 (0.97)	91.4 (0.88)	97.9 (0.97)	97.8 (0.96)	97.4 (0.96)	96.4 (0.94)	93.4 (0.90)	82.2 (0.75)	77.4 (0.69)	97.9 (0.97)	97.7 (0.96)	97.2 (0.95)	95.8 (0.93)	89.7 (0.85)	64.3 (0.55)	55.1 (0.45)	97.4 (0.96)	96.9 (0.95)	95.9 (0.94)	94.2 (0.91)	90.8 (0.86)	97.6 (0.96)	97.2 (0.95)	96.4 (0.94)	94.9 (0.92)	91.7 (0.88)
Traditional Curriculum	Small (27k)	91.7 (0.89)	84.6 (0.80)	91.7 (0.89)	91.5 (0.88)	91.2 (0.88)	89.7 (0.86)	85.2 (0.80)	73.4 (0.66)	69.4 (0.61)	91.7 (0.89)	91.4 (0.88)	90.6 (0.87)	88.3 (0.84)	80.5 (0.75)	58.4 (0.49)	51.1 (0.41)	91.6 (0.88)	90.8 (0.87)	89.4 (0.86)	86.9 (0.83)	81.9 (0.77)	91.7 (0.89)	91.1 (0.88)	90.1 (0.86)	88.1 (0.84)	84.0 (0.79)
	Large (313k)	98.0 (0.97)	91.3 (0.88)	98.0 (0.97)	97.8 (0.96)	97.5 (0.96)	96.5 (0.95)	93.6 (0.90)	82.0 (0.75)	77.4 (0.69)	98.0 (0.97)	97.8 (0.96)	97.4 (0.96)	95.9 (0.94)	89.7 (0.85)	62.9 (0.54)	53.7 (0.44)	97.4 (0.96)	96.8 (0.95)	95.9 (0.94)	94.2 (0.91)	90.5 (0.86)	97.6 (0.96)	97.2 (0.95)	96.5 (0.94)	95.0 (0.92)	90.9 (0.87)
Proposed Curriculum	Small (27k)	93.1 (0.91)	85.3 (0.81)	93.1 (0.91)	92.9 (0.90)	92.3 (0.89)	90.7 (0.87)	86.1 (0.81)	74.5 (0.67)	70.7 (0.62)	93.0 (0.90)	92.9 (0.90)	92.0 (0.89)	89.2 (0.85)	80.1 (0.74)	56.8 (0.48)	49.6 (0.40)	92.5 (0.90)	91.7 (0.89)	90.3 (0.87)	87.7 (0.84)	82.5 (0.77)	92.8 (0.90)	92.1 (0.89)	91.0 (0.88)	88.6 (0.85)	84.0 (0.79)
	Large (313k)	97.4 (0.96)	91.4 (0.88)	97.3 (0.96)	97.1 (0.96)	96.8 (0.95)	96.0 (0.94)	93.4 (0.90)	83.9 (0.77)	79.7 (0.72)	97.3 (0.96)	97.1 (0.95)	96.5 (0.94)	94.9 (0.92)	89.3 (0.84)	66.7 (0.58)	58.3 (0.49)	96.8 (0.95)	96.4 (0.94)	95.5 (0.93)	94.0 (0.91)	91.2 (0.87)	97.1 (0.95)	96.7 (0.95)	96.0 (0.94)	94.6 (0.92)	91.8 (0.88)

The performance improvement is significant for loud noise samples when the joint framework is applied.

2) PROPOSED CURRICULUM LEARNING RESULTS

In this experiment, the performance of the proposed curriculum learning method of our joint framework was assessed. We verified the efficacy of the proposed curriculum learning method compared to baseline curriculum learning methods without curriculum learning and traditional curriculum learning, similar to that used in ConvMixer.

Table 3 presents the accuracy of keyword spotting on the MUSAN noise dataset resulting from the application of curriculum learning to a small network (BC-ResNet-2). "No curriculum" denotes the result of applying a random noise mixture, such as ImportantAug or PKD. "Traditional curriculum" denotes the result of applying a general curriculum learning approach, such as ConvMixer. In the experimental results, the proposed noise mixture curriculum exhibits the best performance.

Fig. 5 presents a comparison of keyword-spotting accuracy for the application of curriculum learning to a large network (BC-ResNet-8). The traditional noise mixture curriculum learning applied in ConvMixer displays no performance improvement compared with the random noise mixture used in ImportantAug and PKD. With the proposed curriculum learning, the performance is improved to 83.9% at −10 dB and 79.7% at −12.5 dB, compared with 82.2% and 77.4%, respectively, under loud noise scenarios. In addition, compared with the performance of the small network in Fig. 5, that is, 74.6% at −10 dB and 70.7% at −12.5 dB, a significant performance improvement is observed. Therefore, it is difficult to cope with loud noises using a small network, even when curriculum learning is applied.

Fig. 6 shows the improvement trend of keyword spotting accuracy at −12 dB on the MUSAN noise dataset in terms of epochs as learning progresses according to the applied curriculum learning method. When a random noise mixture is applied without curriculum learning, the performance converges to 77.4% at 3000 epochs. In addition, when traditional curriculum learning is applied, performance improvement

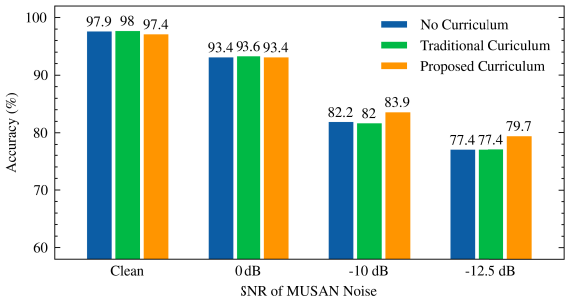

FIGURE 5. Accuracy comparison of large networks for the curriculum methods on the MUSAN dataset.

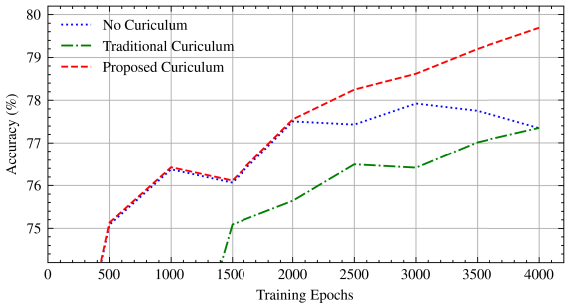

FIGURE 6. Accuracy of the large networks according to curriculum methods for different training epochs.

is slower than without traditional curriculum, and the highest performances are similar. When the proposed method is applied, performance improves.

3) PROPOSED ENSEMBLE DISTILLATION RESULTS

In this experiment, we demonstrated the performance of the proposed ensemble distillation in our joint framework. The efficacy of the proposed weighted-stage ensemble distillation was verified compared to three baseline ensemble distillations: ensemble distillation without curriculum learning, with curriculum learning, and stage ensemble distillation.

Table 4 shows the comparison results of keyword-spotting accuracy for the distillation methods. The performances of

TABLE 4. Evaluation of teacher ensemble and distilled students according to ensemble distillation methods.

	Before or After of Distillation (Model Parms #)	Accuracy (%) on Clean (F1-score)	Averaged Accuracy (%) of Noises (F1-score)	Accuracy (%) on MUSAN (F1-score)							Accuracy (%) on QUT (F1-score)							Accuracy (%) on UrbanSound8k (F1-score)					Accuracy (%) on WHAM (F1-score)				
				40 dB	30 dB	20 dB	10 dB	0 dB	-10 dB	-12.5 dB	40 dB	30 dB	20 dB	10 dB	0 dB	-10 dB	-12.5 dB	20 dB	15 dB	10 dB	5 dB	0 dB	20 dB	15 dB	10 dB	5 dB	0 dB
Ensemble Distillation without Curriculum	Before (1.5M)	98.5 (0.98)	92.5 (0.89)	98.5 (0.98)	98.3 (0.97)	98.0 (0.97)	97.0 (0.95)	94.4 (0.91)	84.5 (0.78)	80.1 (0.72)	98.4 (0.97)	98.2 (0.97)	97.8 (0.96)	96.7 (0.95)	91.5 (0.87)	67.8 (0.58)	59.2 (0.48)	97.9 (0.97)	97.3 (0.96)	96.4 (0.94)	95.1 (0.92)	91.8 (0.88)	98.0 (0.97)	97.7 (0.96)	97.1 (0.95)	95.7 (0.93)	92.9 (0.89)
	After (27k)	97.1 (0.95)	90.4 (0.85)	97.0 (0.95)	96.8 (0.95)	96.3 (0.94)	95.1 (0.92)	90.9 (0.85)	79.6 (0.69)	76.1 (0.64)	97.3 (0.95)	97.0 (0.95)	96.6 (0.94)	94.5 (0.91)	87.2 (0.80)	66.3 (0.49)	60.2 (0.41)	96.3 (0.94)	95.7 (0.93)	94.3 (0.91)	92.2 (0.88)	87.9 (0.82)	96.9 (0.95)	96.5 (0.94)	96.1 (0.92)	93.2 (0.89)	89.3 (0.83)
Ensemble Distillation with Curriculum	Before (1.5M)	98.1 (0.97)	92.8 (0.90)	98.0 (0.97)	97.9 (0.97)	97.6 (0.96)	96.8 (0.95)	94.5 (0.92)	86.0 (0.80)	82.1 (0.74)	98.1 (0.97)	97.9 (0.97)	97.3 (0.96)	96.1 (0.94)	91.3 (0.87)	70.7 (0.62)	62.5 (0.52)	97.5 (0.96)	97.1 (0.95)	96.3 (0.94)	95.2 (0.93)	92.6 (0.89)	97.8 (0.96)	97.3 (0.96)	96.7 (0.95)	95.5 (0.93)	93.2 (0.90)
	After (27k)	96.9 (0.95)	90.9 (0.85)	96.8 (0.94)	96.2 (0.94)	96.2 (0.94)	94.6 (0.92)	90.9 (0.85)	82.3 (0.72)	78.7 (0.66)	97.0 (0.95)	96.8 (0.95)	96.2 (0.94)	94.1 (0.91)	87.8 (0.81)	68.6 (0.50)	64.1 (0.43)	96.2 (0.94)	95.6 (0.93)	94.4 (0.91)	93.1 (0.89)	89.0 (0.83)	96.6 (0.94)	96.0 (0.93)	95.4 (0.92)	94.4 (0.90)	91.0 (0.85)
Stage Ensemble Distillation	Before (7.5M)	98.1 (0.97)	92.3 (0.89)	98.1 (0.97)	98.0 (0.97)	97.6 (0.96)	96.7 (0.95)	94.2 (0.91)	85.0 (0.78)	80.8 (0.72)	98.0 (0.97)	97.8 (0.96)	97.3 (0.96)	96.1 (0.94)	90.7 (0.86)	67.9 (0.58)	59.1 (0.48)	97.5 (0.96)	97.1 (0.95)	96.3 (0.94)	94.9 (0.92)	92.1 (0.88)	97.7 (0.96)	97.5 (0.96)	96.8 (0.95)	95.6 (0.93)	92.7 (0.89)
	After (27k)	97.0 (0.95)	90.5 (0.85)	97.0 (0.95)	96.8 (0.95)	96.2 (0.94)	94.6 (0.92)	90.7 (0.85)	80.7 (0.70)	76.9 (0.64)	97.0 (0.95)	96.8 (0.95)	96.2 (0.94)	94.1 (0.91)	87.0 (0.80)	67.9 (0.49)	62.7 (0.41)	96.2 (0.94)	95.6 (0.93)	94.4 (0.91)	92.3 (0.88)	88.3 (0.83)	96.6 (0.94)	96.0 (0.93)	95.4 (0.92)	93.6 (0.89)	90.0 (0.83)
Weighted Stage Ensemble Distillation	Before (7.5M)	98.4 (0.97)	92.9 (0.90)	98.3 (0.97)	98.2 (0.97)	97.9 (0.97)	96.9 (0.95)	94.8 (0.92)	85.8 (0.80)	82.0 (0.74)	98.3 (0.97)	98.1 (0.97)	97.7 (0.96)	96.6 (0.95)	91.7 (0.88)	70.3 (0.61)	61.9 (0.51)	97.8 (0.97)	97.3 (0.96)	96.5 (0.95)	95.3 (0.93)	92.6 (0.89)	98.0 (0.97)	97.6 (0.96)	97.0 (0.95)	95.8 (0.93)	93.3 (0.90)
	After (27k)	97.1 (0.95)	91.1 (0.86)	97.0 (0.95)	96.8 (0.95)	96.4 (0.94)	95.2 (0.92)	91.1 (0.86)	82.3 (0.72)	78.6 (0.66)	97.3 (0.95)	97.0 (0.95)	96.6 (0.94)	94.3 (0.91)	87.8 (0.81)	68.5 (0.50)	64.1 (0.43)	96.4 (0.94)	95.8 (0.93)	94.9 (0.92)	93.1 (0.89)	89.2 (0.84)	96.9 (0.95)	96.4 (0.94)	96.1 (0.92)	94.5 (0.90)	91.1 (0.85)

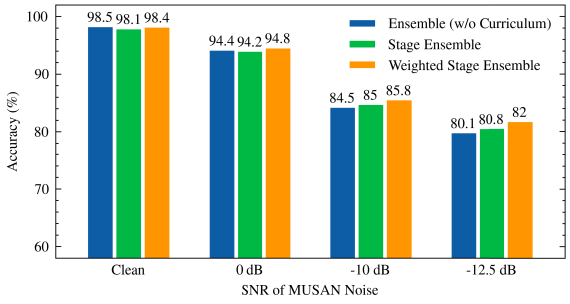

FIGURE 7. Ensemble accuracy of teacher networks according to ensemble methods on the MUSAN noise dataset.

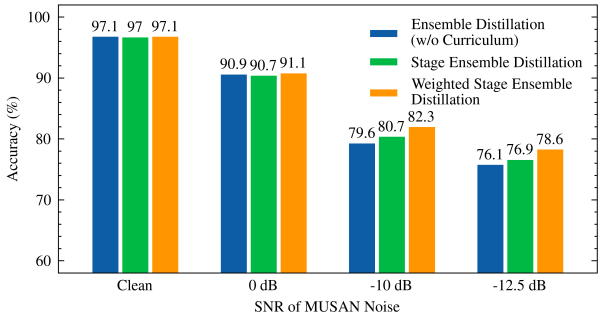

FIGURE 8. Accuracy comparison of the distilled students according to ensemble methods on the MUSAN noise dataset.

the teacher ensembles and distilled students were compared with respect to the distillation method. Compared with learning from scratch using a small network without knowledge distillation, the performance is significantly improved when students are extracted through any of the distillation methods. In addition, ensemble distillation with the proposed curriculum learning improves performance compared to those without curriculum learning. The comparison confirms that the proposed weighted-stage ensemble distillation exhibits the best performance under all noise scenarios.

Fig. 7 shows the keyword spotting accuracy on the MUSAN dataset for the teacher ensemble methods. The ensemble using only the final independently trained teachers, as shown in (7), displays an accuracy of 80.1% at −12.5 dB. The stage ensemble, which uses not only the final independently trained teachers but also the intermediate teachers at each stage, as shown in (10), displays an accuracy of 80.8% at −12.5 dB. Finally, the proposed weighted-stage ensemble, which provides weights for the intermediate teachers at each stage according to the SNR of the training sample, as shown in (11), displays an accuracy of 82% at −12.5 dB. This is the best performance compared with the baseline.

Fig. 8 shows the accuracy of the distilled student networks on the MUSAN dataset for the ensemble distillation methods. The ensemble distillation using only the final independently trained teachers without curriculum learning, as shown in (7), displays an accuracy of 76.1% at −12.5 dB. The stage ensemble, which uses not only the final independently trained teachers but also the intermediate teachers for each stage, as shown in (10), displays an accuracy of 76.9% at −12.5 dB. Finally, the proposed weighted-stage ensemble, which provides weights for the intermediate teachers at each stage according to the SNR of the training sample, as shown in (11), displays an accuracy of 78.6% at −12.5 dB. This is the best performance compared with the baseline.

E. JOINT FRAMEWORK HYPERPARAMETERS

1) PROPOSED CURRICULUM LEARNING

As described in Section III, the proposed curriculum learning method varies the ratio of loud noise samples at each stage with speech augmentation applied simultaneously. In this experiment, we compared keyword-spotting accuracy according to the noise mixture strategy for each stage and considered whether speech augmentation was applied. The MUSAN dataset was used for the comparisons.

Fig. 9 compares the accuracy of the teacher network in terms of the probability ρ for the SNR distribution R_n. When

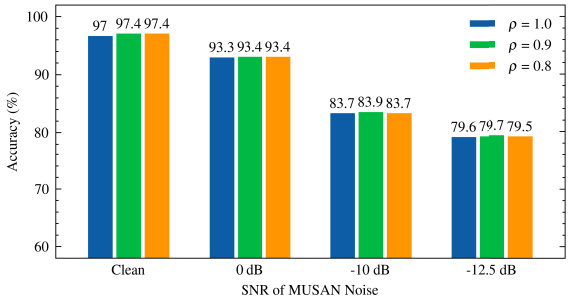

FIGURE 9. Comparison of the noise mixture strategy according to probability ρ of the main range on the MUSAN dataset.

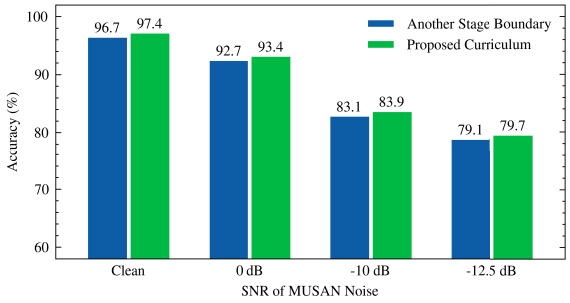

FIGURE 10. Comparison of the noise mixture strategy between a stage boundary on the MUSAN dataset.

the selected probability ρ is 0.9, the accuracy improves slightly from 79.5% or 79.6% to 79.7% at −12.5 dB compared to other values of probability ρ.

Fig. 10 shows a comparison of the accuracy of the teacher network in terms of the loud–noise ratio at each stage. The selected curriculum increases the main range for each stage as follows: [−15, 50] dB, [−15, 10] dB, [−15, 5] dB, [−15, 0] dB, and [−15, −5] dB, displaying better performance than other curriculums. In particular, the selected curriculum shows slightly better performance, 79.7% versus 79.1% at −12.5 dB, at the main ranges of [−15, 50] dB, [−15, 5] dB, [−15, 0] dB, [−15, −5] dB, and [−15, −10] dB for each stage.

Fig. 11 compares the accuracy of the teacher based on whether speech augmentation was applied. The accuracy improves slightly from 79.4% to 79.7% at −12.5 dB, when speech augmentation is applied.

2) PROPOSED ENSEMBLE DISTILLATION

As described in Section III, the proposed ensemble distillation method distills a student using neural network snapshots for each curriculum stage. In this experiment, we explored the hyperparameters for the proposed ensemble distillation, including the number of teachers used and the size of the teacher network. The MUSAN dataset was used for the exploration.

Fig. 12 shows a comparison of the accuracy with respect to the constants α and β. The constants α and β were set to 1 and 0, respectively. The performance of the proposed ensemble for the selected constant s, as shown in (11), was

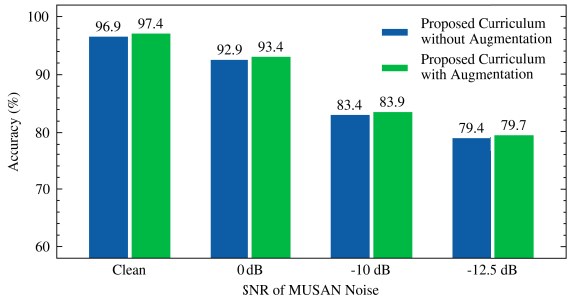

FIGURE 11. Comparison of the noise mixture strategy depending on whether augmentation was applied on the MUSAN dataset.

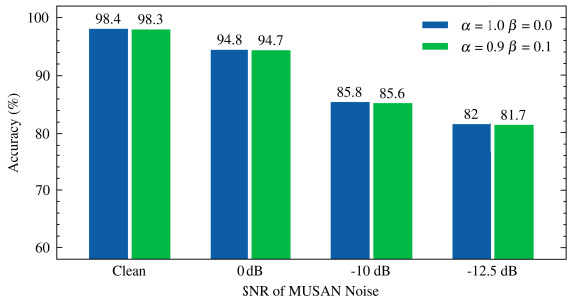

FIGURE 12. Comparison of the proposed ensemble according to α and β of the weight factor w_n on the MUSAN dataset.

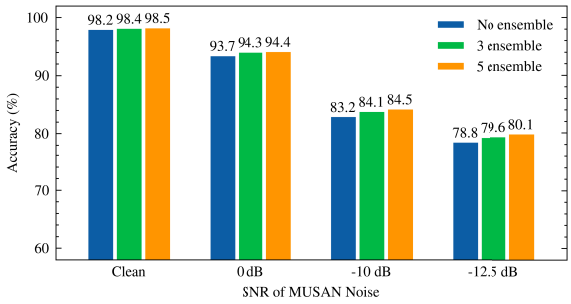

FIGURE 13. Comparison of the ensemble depending on the number of teachers on the MUSAN dataset.

better than those of other choices, with the accuracy improving slightly from 81.7% to 82% at −12.5 dB.

Fig. 13 shows a comparison of the accuracy in terms of the number of teachers for the ensemble, as shown in (7). The performance of the ensembles is better than that of the teachers. Accuracy improves as the number of teachers increases, 78.8% with no ensemble at −12.5 dB to 80.1% with an ensemble of five teachers.

Fig. 14 shows a comparison of the accuracy of the teacher network with respect to the number of network parameters. The larger the size of the teacher, the higher is the accuracy. In the case of a teacher with 120k parameters, the accuracy is 75.1%, as opposed to 79.7% when the number of parameters is increased to 350k and 80.9% when the number of parameters is increased to 700k.

The ensemble using a larger size and number of teachers displays better accuracy. An appropriate size or number of

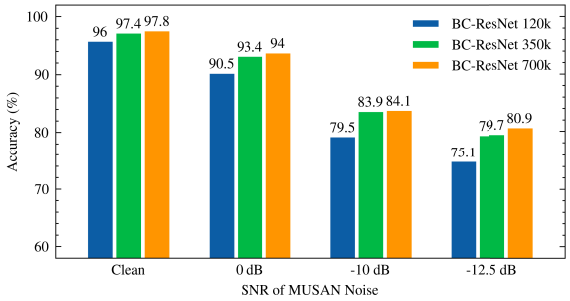

FIGURE 14. Comparison of teachers according to the number of network parameters on the MUSAN dataset.

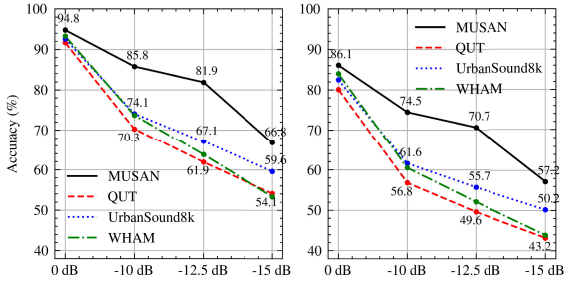

FIGURE 15. Accuracy of the (a) proposed joint framework and (b) curriculum learning according to the SNR on four noise datasets.

teachers can be selected considering the limitations of computing resources.

F. RESULTS ON EXTENDED SNR OF NOISE DATASETS

Thus far, the experiments were conducted using public datasets to compare three state-of-the-art methods. We extended the SNR of public datasets and evaluated their accuracy according to the SNR in loud noise situations. Fig. 15 illustrates the keyword-spotting accuracy of the proposed joint framework in terms of SNR on four public noise datasets. The proposed joint framework showed an accuracy of 66.8% at −15 dB MUSAN, 54.1% at −15 dB QUT, 59.6% at −15 dB UrbanSound8k, and 53.3% at −15 dB WHAM, compared with the accuracies of curriculum learning of 57.2%, 43.2%, 50.2%, and 43.9%, respectively. The performance improvement is significant for loud noise samples when the joint framework is applied.

V. CONCLUSION

This paper presents the first study on a joint framework of curriculum learning and knowledge distillation for noise-robust and small-footprint keyword spotting. The main finding is that distilling a small network after applying curriculum learning to the large teacher network is superior to directly applying curriculum learning to the small network. Keyword spotting is a voice interface technology that is already widespread in consumer electronics, such as smart speakers. However, keyword-spotting models must be more robust and lighter for battery-powered and resource-constrained devices. We believe that the proposed joint framework will promote keyword-spotting-based voice interfaces for wearable devices, such as hearables and IoT devices for smart homes.

This study proposes a joint framework that benefits from curriculum learning and knowledge distillation. We propose curriculum learning with a new noise mixture strategy and knowledge distillation with an effective ensemble of neural network snapshots for each curriculum stage to enhance the effectiveness of the joint framework. In particular, the proposed joint framework applies the proposed curriculum learning to a network that is sufficiently large to learn various noise situations. Subsequently, the proposed ensemble distillation is applied to compress the large network into a sufficiently small network for onboard microcontrollers. The proposed joint framework achieved superior accuracy in noisy situations compared to state-of-the-art methods. In particular, the proposed joint framework achieved small-footprint keyword spotting with an accuracy of 79.1% at an SNR of −12.5 dB on the MUSAN dataset. Therefore, the proposed framework can significantly improve the usability of voice interfaces in consumer electronics.

The limitations of our study and potential directions for future research are summarized as follows:

1) Our research targeted low performance, low power, and tiny consumer devices. A limitation is that more research is required on the adaptive operation of devices with different numbers of microphones and performances. Therefore, our future work will focus on a joint framework that considers hardware-specific conditions, similar to the research field of hardware-aware neural architecture searches [43].

2) We treated all sounds except voice commands as noise, but audio-based consumer technology can advance scene understanding through comprehensive noise analysis [30]. Recognizing the context of ambient noise can improve voice recognition or situational awareness.

3) We demonstrated the effect of the joint framework and presented a knowledge distillation method to enhance it; however, there are additional considerations. Studies on new loss functions or knowledge transfer methods [44] have been actively conducted in knowledge distillation to boost teacher-student matching. Therefore, in future work, we will apply advanced loss functions and transfer methods to the joint framework to improve the performance under loud noise scenarios.

REFERENCES

[1] P. Busia, G. Deriu, L. Rinelli, C. Chesta, L. Raffo, and P. Meloni, "Target-aware neural architecture search and deployment for keyword spotting," *IEEE Access*, vol. 10, pp. 40687–40700, 2022.
[2] R. Tang and J. Lin, "Deep residual learning for small-footprint keyword spotting," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5484–5488.
[3] 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 *Proc. Interspeech*, Sep. 2019, pp. 3372–3376.
[4] B. Kim, S. Chang, J. Lee, and D. Sung, "Broadcasted residual learning for efficient keyword spotting," in *Proc. Interspeech*, Aug. 2021, pp. 1–5.
[5] Z. Song, Q. Liu, Q. Yang, and H. Li, "Knowledge distillation for in-memory keyword spotting model," in *Proc. Interspeech*, Sep. 2022, pp. 4128–4132.

[6] G. Tucker, M. Wu, M. Sun, S. Panchapagesan, G. Fu, and S. Vitaladevuni, "Model compression applied to small-footprint keyword spotting," in *Proc. Interspeech*, Sep. 2016, pp. 1878–1882.

[7] J. Kim, S. Chang, and N. Kwak, "PQK: Model compression via pruning, quantization, and knowledge distillation," in *Proc. Interspeech*, Aug. 2021, pp. 1–5.

[8] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.

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

[10] E. Liberis, Ł. Dudziak, and N. D. Lane, "μNAS: Constrained neural architecture search for microcontrollers," in *Proc. 1st Workshop Mach. Learn. Syst.*, Apr. 2021, pp. 70–79.

[11] T. O'Malley, A. Narayanan, Q. Wang, A. Park, J. Walker, and N. Howard, "A conformer-based ASR frontend for joint acoustic echo cancellation, speech enhancement and speech separation," in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Dec. 2021, pp. 304–311.

[12] Y. A. Huang, T. Z. Shabestary, and A. Gruenstein, "Hotword cleaner: Dual-microphone adaptive noise cancellation with deferred filter coefficients for robust keyword spotting," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 6346–6350.

[13] M. L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 7398–7402.

[14] A. Pervaiz, F. Hussain, H. Israr, M. A. Tahir, F. R. Raja, N. K. Baloch, F. Ishmanov, and Y. B. Zikria, "Incorporating noise robustness in speech command recognition by noise augmentation of training data," *Sensors*, vol. 20, no. 8, p. 2326, Apr. 2020.

[15] S. Braun, D. Neil, and S.-C. Liu, "A curriculum learning method for improved noise robustness in automatic speech recognition," in *Proc. 25th Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2017, pp. 548–552.

[16] S. Ranjan and J. H. L. Hansen, "Curriculum learning based approaches for noise robust speaker recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 1, pp. 197–210, Jan. 2018.

[17] A. Danyluk, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. Annu. Int. Conf. Mach. Learn. (ICML)*, 2009, pp. 41–48.

[18] T. Higuchi, S. Saxena, M. Souden, T. D. Tran, M. Delfarah, and C. Dhir, "Dynamic curriculum learning via data parameters for noise robust keyword spotting," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021.

[19] F. Kawsar, C. Min, A. Mathur, A. Montanari, U. G. Acer, and M. Van den Broeck, "eSense: Open earable platform for human sensing," in *Proc. 16th ACM Conf. Embedded Netw. Sensor Syst.*, Nov. 2018, pp. 371–372.

[20] L. Turchet, G. Fazekas, M. Lagrange, H. S. Ghadikolaei, and C. Fischione, "The Internet of Audio Things: State of the art, vision, and challenges," *IEEE Internet Things J.*, vol. 7, no. 10, pp. 10233–10249, Oct. 2020.

[21] Google. *Google Nest and Home Device Specifications*. Accessed: Jul. 17, 2023. [Online]. Available: https://support.google.com/googlenest/answer/7072284

[22] J. Fernandez-Marques, V. Tseng, S. Bhattachara, and N. D. Lane, "BinaryCmd: Keyword spotting with deterministic binary basis," in *Proc. Conf. Mach. Learn. Syst. (MLSys)*, 2018, pp. 1–3.

[23] M. Yu, X. Ji, B. Wu, D. Su, and D. Yu, "End-to-end multi-look keyword spotting," in *Proc. Interspeech*, Oct. 2020, pp. 66–70.

[24] K. Kim, C. Gao, R. Graça, I. Kiselev, H.-J. Yoo, T. Delbruck, and S.-C. Liu, "A 23-μW keyword spotting IC with ring-oscillator-based time-domain feature extraction," *IEEE J. Solid-State Circuits*, vol. 57, no. 11, pp. 3298–3311, Nov. 2022.

[25] 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 Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 3603–3607.

[26] V. A. Trinh, H. S. Kavaki, and M. I. Mandel, "Importantaug: A data augmentation agent for speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 8592–8596.

[27] D. Kim, G. Kim, B. Lee, and H. Ko, "Prototypical knowledge distillation for noise robust keyword spotting," *IEEE Signal Process. Lett.*, vol. 29, pp. 2298–2302, 2022.

[28] Q. Zhu, X. Chen, P. Wu, J. Liu, and D. Zhao, "Combining curriculum learning and knowledge distillation for dialogue generation," in *Proc. Findings Assoc. Comput. Linguistics (EMNLP)*, 2021, pp. 1284–1295.

[29] G. Panagiotatos, N. Passalis, A. Iosifidis, M. Gabbouj, and A. Tefas, "Curriculum-based teacher ensemble for robust neural network distillation," in *Proc. 27th Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2019, pp. 1–5.

[30] X. Chang, P. Ren, P. Xu, Z. Li, X. Chen, and A. Hauptmann, "A comprehensive survey of scene graphs: Generation and application," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 1–26, Jan. 2023.

[31] S. Han, J. Pool, J. Tran, and W. Dally, "Learning both weights and connections for efficient neural network," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1–9.

[32] J. Wu, C. Leng, Y. Wang, Q. Hu, and J. Cheng, "Quantized convolutional neural networks for mobile devices," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4820–4828.

[33] Z. Allen-Zhu and Y. Li, "Towards understanding ensemble, knowledge distillation and self-distillation in deep learning," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2023, pp. 1–73.

[34] H. R. Guimaraes, A. Pimentel, A. R. Avila, M. Rezagholizadeh, and T. H. Falk, "Improving the robustness of DistilHuBERT to unseen noisy conditions via data augmentation, curriculum learning, and multi-task enhancement," in *Proc. ENLSP-II NeurIPS Workshop*, 2022, pp. 1–6.

[35] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proc. Interspeech*, Sep. 2015, pp. 1–4.

[36] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech*, Sep. 2019, pp. 2613–2617.

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

[38] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50K: An open dataset of human-labeled sound events," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 30, pp. 829–852, 2022.

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

[40] D. Dean, S. Sridharan, R. Vogt, and M. Mason, "The QUT-NOISE-TIMIT corpus for the evaluation of voice activity detection algorithms," in *Proc. Interspeech*, Sep. 2010, pp. 3110–3113.

[41] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proc. 22nd ACM Int. Conf. Multimedia*, Nov. 2014, pp. 1041–1044.

[42] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. L. Roux, "WHAM! Extending speech separation to noisy environments," in *Proc. Interspeech*, Sep. 2019, pp. 1368–1372.

[43] H. Benmeziane, K. El Maghraoui, H. Ouarnoughi, S. Niar, M. Wistuba, and N. Wang, "Hardware-aware neural architecture search: Survey and taxonomy," in *Proc. IJCAI*, Aug. 2021, pp. 4322–4329.

[44] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *Int. J. Comput. Vis.*, vol. 129, no. 6, pp. 1789–1819, Jun. 2021.

JAEBONG LIM was born in 1992. He received the B.S. and M.S. degrees from Pusan National University, Busan, Republic of Korea, in 2016 and 2018, respectively, where he is currently pursuing the Ph.D. degree. His research interests include embedded systems, low-power devices, embedded AI, TinyML, and driver behavior analysis.

YUNJU BAEK was born in 1967. He received the Ph.D. degree in computer science from KAIST, Republic of Korea, in 1997. He was an invited Professor with KAIST, the CTO of Naver Corporation, and an Assistant Professor with Sookmyung Women's University. He is currently a Professor with the School of Computer Science and Engineering, Pusan National University. His research interests include embedded systems, RTLS systems, wireless sensor networks, embedded AI, TinyML, and driver behavior analysis.