Interspeech 2022
18-22 September 2022, Incheon, Korea

Cross-Layer Similarity Knowledge Distillation for Speech Enhancement

Jiaming Cheng[1], *Ruiyu Liang*[1,2], *Yue Xie*[2], *Li Zhao*[1], *Björn W. Schuller*[3,4], *Jie Jia*[5], *Yiyuan Peng*[5]

[1]School of Information Science and Engineering, Southeast University, China
[2]School of Information and Communication Engineering, Nanjing Institute of Technology, China
[3]Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[4]GLAM – Group on Language, Audio, & Music, Imperial College London, UK
[5]vivo Mobile Commun co Ltd, China

230198469@seu.edu.cn, liangry@njit.edu.cn, xieyue0109@njit.edu.cn, zhaoli@seu.edu.cn, schuller@tum.de, jie.jia@vivo.com, pengyiyuan@vivo.com

Abstract

Speech enhancement (SE) algorithms based on deep neural networks (DNNs) often encounter challenges of limited hardware resources or strict latency requirements when deployed in real-world scenarios. However, a strong enhancement effect typically requires a large DNN. In this paper, a knowledge distillation framework for SE is proposed to compress the DNN model. We study the strategy of cross-layer connection paths, which fuses multi-level information from the teacher and transfers it to the student. To adapt to the SE task, we propose a frame-level similarity distillation loss. We apply this method to the deep complex convolution recurrent network (DCCRN) and make targeted adjustments. Experimental results show that the proposed method considerably improves the enhancement effect of the compressed DNN and outperforms other distillation methods.

Index Terms: speech enhancement, knowledge distillation, cross-layer connection, pairwise similarity

1. Introduction

Speech enhancement (SE) has been a hot topic in the field of speech for decades. This paper focuses on the monaural SE task. The traditional SE methods are mainly based on statistical signal processing that has a low requirement of computation and hardware, thus having good real-time performance. Typical algorithms include spectral subtraction [1], Wiener filtering [2], minimum mean square error (MMSE) methods [3], or non-negative matrix factorization-based approaches [4]. However, their implementations are often based on assumptions that are unreasonable in real-world scenarios (such as the stationarity of noise), which limits their performance.

Recent developments in SE methods based on deep neural networks (DNNs) have shown superior performance compared with traditional machine learning and signal processing methods [5, 6, 7]. Many deep learning-based SE models have reported excellent performance on real-time and non-real-time tracks in the recent deep noise suppression challenge (DNS) series [8, 9]. However, a large DNN is generally required to achieve ideal performance, which is both computationally intensive and memory-consuming. Even if the real-time requirements of the DNS Challenge are met, deployment difficulties will occur in latency-sensitive applications or on resource-constrained devices (e. g., headsets). Therefore, reducing the size of the DNN has become increasingly important in deep learning-based SE systems.

The mainstream model compression techniques, such as pruning, quantization, and knowledge distillation, all have certain effects in reducing the complexity of the model [10]. This paper mainly focuses on the knowledge distillation mechanism. Its main idea is to shift knowledge from a large teacher model into a small one. The research of knowledge distillation started from the work of Hinton et al. [11] and has been further developed in recent years. PKT [12] performed knowledge transfer by matching the probability distribution of the data in the feature space. SPKD [13] modeled the knowledge as pairwise similarities. The knowledge review framework [14] studied the cross-stage connection paths of the teacher-student model. All these solutions focused on the transformation of the intermediate representation.

However, the existing knowledge distillation methods are mostly applied to classification tasks, and the related work on regression tasks such as SE is rare. A low-latency online extension of wave-U-net was proposed in [15], which directly reduces the difference between the teacher and student output. Teacher-student learning was used in [16] to train a general sub-band enhancement model. However, these methods did not study the intermediate representation of the DNN model. In this paper, we propose a cross-layer knowledge distillation framework for SE tasks. Inspired by the knowledge review method [14], multi-layer feature representations are fused to guide the single layer of the student network. The difference is that we use the frame-level pairwise similarities distance as the distillation loss instead of the hierarchical context loss (HCL) [14]. We apply this strategy to the DCCRN model [17] which ranked first on the real-time track of the DNS Challenge. The experimental results show that the proposed method achieves better performance when compared to other distillation methods.

2. Methodology

2.1. System Overview

The state-of-the-art SE model DCCRN is chosen as the baseline model to perform the knowledge distillation method. Although it meets the real-time requirements of the challenge, it still has 3.7M parameters. Compared with the RNNoise model [19] (only 0.06M parameters) designed for real-time applications, there is still a big gap. Therefore, it is necessary to perform further compression of the DCCRN model. The encoder and decoder are composed of complex convolution/deconvolution layers and complex long-short-term memory (LSTM) layers are inserted between the encoder and decoder to model the temporal dependencies. According to the symmetrical structure, we set the position of distillation in the encoder, decoder and the middle complex LSTM layers respectively. The overall framework of

10.21437/Interspeech.2022-429

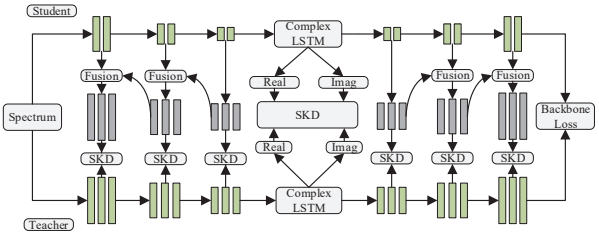

Figure 1: *The overall framework of the proposed method. The intermediate representation features of the student model are recursively integrated through the feature fusion method introduced in Section 2.2, so that the single layer of the student model can learn the effective information from multiple teacher layers. The feature fusion is expanded to both sides with the complex LSTM layer as the center, and the arrows in Figure 1 represent the flow of features. SKD represents the frame-level similarity knowledge distillation loss introduced in Section 2.3. The backbone loss of the teacher and student networks is the multi-resolution STFT (MRSTFT) loss [18] between the enhanced waveform and the clean waveform.*

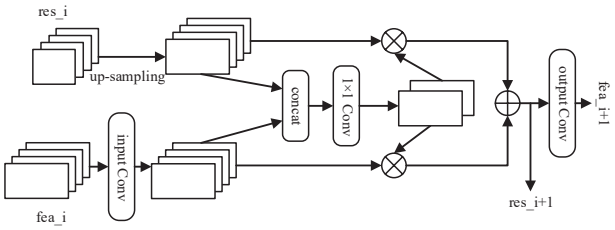

Figure 2: *The architecture of the recursive fusion $\mathcal{U}'(\cdot,\cdot)$ from the i-th layer to the $i+1$-th layer. res_i and fea_i denote the cumulative recursive features and the current layer features of the i-th layer, respectively. The $concat$ operation is performed on the channel dimension. The input and output convolutions are both 2D convolutions. The former is used to unify the channel dimension of the recursive features and the current layer features, and the latter is used to restore the channel dimension of the current layer features.*

the proposed method is shown in Figure 1. For the distillation of the encoder and decoder, the residual learning strategy in [14] is adopted to progressively extract useful information through the fusion of multiple-layer teacher output. For the middle complex LSTM layers, frame-level similarity distillation is performed on the output of the real and imaginary parts, respectively.

2.2. Feature Fusion

The feature fusion is set at the output of each complex convolutional layer of the encoder and decoder. In the design of the original DCCRN model, the real and imaginary parts of the complex convolutional layer output will be concatenated and processed by complex batch normalization. Therefore, the feature fusion is performed on the concatenated output without additional diversion. Given the intermediate features of the n stages to be distilled in the student model $\left(\mathbf{F}_S^1, \mathbf{F}_S^2, \ldots, \mathbf{F}_S^n\right)$, the features of the i-th teacher layer can be similarly defined as $\mathbf{F}_T^i$. To preserve the effective information contained in the teacher model, we only transform the student features. The one-to-one knowledge distillation loss can be expressed as

$$\mathcal{L}_{single} = \sum_{i \in \mathbf{I}} \mathcal{D}\left(\mathcal{M}_S^i\left(\mathbf{F}_S^i\right), \mathbf{F}_T^i\right), \tag{1}$$

where $\mathbf{I}$ denotes the intermediate representation feature set of multiple layers to be distilled in the student model. $\mathcal{M}$ represents the feature transformation; here, the up-sampling method of nearest interpolation is used to execute the transformation. $\mathcal{D}$ describes the distance of the intermediate features. Similarly, the one-to-many distillation loss is written as

$$\mathcal{L}_{multiple} = \sum_{i \in \mathbf{I}} \left(\sum_{j=1}^{i} \mathcal{D}\left(\mathcal{M}_S^{i,j}\left(\mathbf{F}_S^i\right), \mathbf{F}_T^j\right)\right). \tag{2}$$

For the encoder, the teacher's first i levels of features are used to guide the current layer of the student model, while for the decoder, we use the last i layers. In order to avoid the mutual interference of multi-layer features and reduce redundant calculations, we adopt the residual learning strategy [14]. Assuming that n stages of features are distilled, we can switch the order of the two summations of i and j in (2) as

$$\mathcal{L}_{multiple} = \sum_{j=1}^{n} \left(\sum_{i=j}^{n} \mathcal{D}\left(\mathbf{T}_S^{i,j}, \mathbf{F}_T^j\right)\right), \tag{3}$$

where the transformed student features $\mathcal{M}_S^{i,j}\left(\mathbf{F}_S^i\right)$ are abbreviated as $\mathbf{T}_S^{i,j}$. However, there are still $n(n+1)/2$ pairs of features that need to be calculated. The sum of multiple distance pairs is approximated as the distance of the fused features using the n-dimensional operator $\mathcal{U}$ [14]:

$$\sum_{i=j}^{n} \mathcal{D}\left(\mathbf{T}_S^{i,j}, \mathbf{F}_T^j\right) \approx \mathcal{D}\left(\mathcal{U}\left(\mathbf{T}_S^{i,j}, \ldots, \mathbf{T}_S^{i,n}\right), \mathbf{F}_T^j\right). \tag{4}$$

The process of feature fusion can be carried out in a progressive manner. We redefine the fusion operation as a recursive function $\mathcal{U}'(\cdot,\cdot)$, and the fusion distillation loss can be calculated as

$$\mathcal{L}_{fusion} = \mathcal{D}\left(\mathbf{T}_S^n, \mathbf{F}_T^n\right) + \sum_{j=n-1}^{1} \mathcal{D}\left(\mathcal{U}'\left(\mathbf{T}_S^j, \mathbf{T}_S^{j+1}\right), \mathbf{F}_T^j\right), \tag{5}$$

where $\mathbf{T}_S^{j+1}$ denotes the fusion from $\mathbf{T}_S^n$ to $\mathbf{T}_S^{j+1}$. Here, the loop starts from the $n-1$-th in reverse order and there is only a dimensional transformation in the n-th layer. The recursive fusion process from the i-th layer to the $i+1$-th layer is illustrated in Figure 2. First, the up-sampling method is used to resize the high-level features to the same shape as the lower-level features. Then, the current layer features and the recursive features are weighted by attention using 1×1 convolution and are added to update. The entire process is shown in Figure 1, the complex LSTM layers are used as the center of symmetry to perform a recursive fusion on both sides of the encoder and decoder. We believe that due to the symmetry of the DCCRN model, the student high-level stage close to the middle has a strong ability to learn effective information from the low-level features near the two ends of the teacher.

2.3. Similarity Knowledge Distillation (SKD)

After the feature fusion process, the feature dimensions of the teacher and student models become the same in the encoder

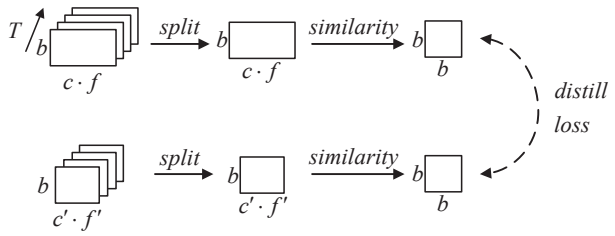

Figure 3: *The calculation of similarity distillation loss. Given an input minibatch with dimensions* (b, c, t, f)*, first, the input is sliced along the time dimension* T *into feature maps* (b, c, f) *of* t *groups. Then, we derive* $b \times b$ *pairwise similarity matrices from the feature maps, and the similarity distance is calculated on the matrices produced by the student and the teacher.*

and decoder stages, but there are still differences in the middle complex LSTM layers. For consistency, a distance calculation method that is not restricted by feature dimensions is needed. Inspired by [13], we derive the pairwise similarity matrices from the intermediate feature representations of the teacher and student models. Such similarity distillation loss can simultaneously achieve dimensional compression and similarity information transmission. Given a mini-batch input, the feature map of the complex convolutional layer is $O_T \in \mathbf{R}^{b\times c\times t\times f}$, and the output size of the complex LSTM layer is $(b \times t \times f)$, where b is the batch size, c is the number of output channels, t the number of speech frames, and f is the dimension of the feature space. Unlike [13], we independently calculate the similarity matrix of each frame because the information of different frames may interfere with each other, and we hope that each frame has a unique contribution to the distillation, while direct flattening will smooth out the differences between frames. The specific process of similarity distillation is shown in Figure 3. We first perform frame-level segmentation on the feature map of the l-th layer, and then flatten the features into two dimensions. Let the transformed feature of the j-th frame be $Q_T^{(l,j)} \in \mathbf{R}^{b\times f'}$, the similarity matrix of the teacher and the student are calculated separately, and then, L2 normalization is applied to each row of the matrix:

$$\begin{aligned}\tilde{Q}_T^{(l,j)} &= Q_T^{(l,j)} \cdot Q_T^{(l,j)\,\mathrm{T}}; G_{T[i,:]}^{(l,j)} = \tilde{Q}_{T[i,:]}^{(l,j)} \Big/ \left\| \tilde{Q}_{T[i,:]}^{(l,j)} \right\|_2 \\ \tilde{Q}_S^{(l,j)} &= Q_S^{(l,j)} \cdot Q_S^{(l,j)\,\mathrm{T}}; G_{S[i,:]}^{(l,j)} = \tilde{Q}_{S[i,:]}^{(l,j)} \Big/ \left\| \tilde{Q}_{S[i,:]}^{(l,j)} \right\|_2\end{aligned} \tag{6}$$

where $[i,\ :]$ denotes the i-th row in a matrix. The dimension of the similarity matrix $G_T^{(l,j)}, G_S^{(l,j)}$ calculated for each frame is $b \times b$. Finally, the distillation loss of the l-th layer is defined as the accumulation of all frames' similarity distances:

$$\mathcal{L}_{SKD}^{l} = \frac{1}{b^2} \sum_{j=1}^{t} \left\| G_T^{(l,j)} - G_S^{(l,j)} \right\|_F^2, \tag{7}$$

where $\|\cdot\|_F$ is the Frobenius norm.

2.4. Training Procedure

This section describes the entire training procedure shown in Figure 1. The multi-resolution STFT (MRSTFT) loss [18] is used as the backbone loss to train the teacher model, and then, it is frozen during the training of the student model, and knowledge distillation is performed simultaneously with the training of the student network. During the forward inference of the student and teacher networks, the output of each layer is saved for the calculation of the knowledge distillation. The distillation is set in the encoder, decoder, and the middle complex LSTM layers, respectively. For the encoder and decoder, the features of the student model are first transformed using the feature fusion method in Section 2.2. Given the transformed features of the student l-th layer $\mathbf{T}_{S|x}^{l}$ and the corresponding teacher model features $\mathbf{F}_{T|x}^{l}$, where x represents the encoder or the decoder, the distillation loss of the encoder $\mathcal{L}_{distill}^{enc}$ and that of the decoder $\mathcal{L}_{distill}^{dec}$ is calculated using the similarity distillation method in Section 2.3:

$$\begin{aligned}\mathcal{L}_{distill}^{enc} &= \sum_{l=1}^{M} \mathcal{L}_{SKD}^{l}(\mathbf{F}_{T|enc}^{l}, \mathbf{T}_{S|enc}^{l}), \\ \mathcal{L}_{distill}^{dec} &= \sum_{l=1}^{N} \mathcal{L}_{SKD}^{l}(\mathbf{F}_{T|dec}^{l}, \mathbf{T}_{S|dec}^{l}),\end{aligned} \tag{8}$$

where M and N represent the number of layers of the encoder and decoder, respectively. For the middle complex LSTM layers, distillation is performed on the output of the real and imaginary parts, respectively. Given the features of the student l-th complex LSTM layer $\mathbf{F}_{S|y}^{l}$ and the corresponding teacher model features $\mathbf{F}_{T|y}^{l}$, where y represents the real part or the imaginary part, the distillation loss of complex LSTM layers $\mathcal{L}_{distill}^{CLSTM}$ is:

$$\begin{aligned}\mathcal{L}_{distill}^{CLSTM} = &\sum_{l=1}^{K} \mathcal{L}_{SKD}^{l}(\mathbf{F}_{T|real}^{l}, \mathbf{F}_{S|real}^{l}) \\ &+ \sum_{l=1}^{K} \mathcal{L}_{SKD}^{l}(\mathbf{F}_{T|imag}^{l}, \mathbf{F}_{S|imag}^{l}),\end{aligned} \tag{9}$$

where K denotes the number of complex layers. The overall loss of the student model $\mathcal{L}_{Stu}$ is the combination of the backbone loss $\mathcal{L}_{MRSTFT}$ and the distillation losses:

$$\mathcal{L}_{Stu} = \mathcal{L}_{MRSTFT} + \mathcal{L}_{distill}^{enc} + \mathcal{L}_{distill}^{dec} + \mathcal{L}_{distill}^{CLSTM}. \tag{10}$$

3. Experiments and Analysis

3.1. Dataset

The Interspeech 2020 DNS Challenge dataset [8] is used to prepare the training and test sets. The DNS dataset contains 500 hours of clean clips from 2150 speakers and 65,000 noise clips in a total of 180 hours. We randomly split the corpus into 60,000 and 1,000 utterances each in the training set and the validation set. The noisy utterances are generated by mixing randomly selected speech and noise at random SNR between -5 and 15 dB using official scripts provided by DNS Challenge [8]. The official non-reverb test set is used for objective scoring comparison.

3.2. Implementation Details

The DCCRN-CL model [17] is chosen as the baseline model. The kernel size and stride of the teacher and the student model are both set to $(5, 2)$ and $(2, 1)$ in the frequency and time axes. The number of channels for the teacher is $\{32, 64, 128, 256, 256, 256\}$, while the student is $\{8, 16, 32, 64, 64, 64\}$. The teacher model uses the complex LSTM with 128 units for the real part and the imaginary part, respectively, while the student model has 32 units. The compressed student model has only 0.23M parameters, which is 6 %

Table 1: *Comparison of objective speech indicators between distilled and undistilled models in the non-reverb test set*

Distillation Mechanism	Model	Param.(M)	WB-PESQ	STOI(%)
-	Noisy	-	1.582	91.52
None	NSNet [20]	1.26	2.145	94.47
None	RNNoise [19]	0.06	1.973	-
None	DTLN [21]	0.99	-	94.76
None	DCCRN-T [17]	3.67	2.803	96.43
None	DCCRN-S	0.23	2.396	94.98
Diff [15]	DCCRN-S	0.23	2.429	95.28
PKT [12]	DCCRN-S	0.23	2.425	**95.30**
ReviewKD [14]	DCCRN-S	0.23	2.404	94.94
SPKD [13]	DCCRN-S	0.23	2.464	95.13
SKD	DCCRN-S	0.23	2.500	95.19
CLSKD	DCCRN-S	0.23	**2.518**	95.29

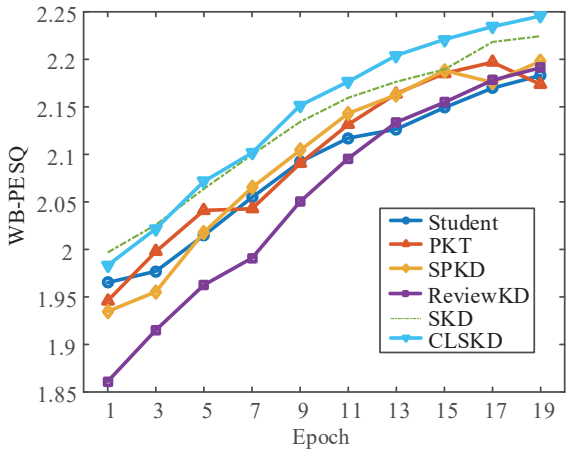

Figure 4: *The WB-PESQ score curve of the validation set*

of the teacher model (3.7M). For the feature fusion, convolutional layers with a kernel size of $(5, 1)$ are used as input and output convolutions.

All the utterances are sampled at 16 kHz and chunked to 2 seconds. The window length and hop size are 32 ms and 16 ms, and the FFT length is 512. The MRSTFT loss [18] is used as the loss function of the baseline model instead of the SI-SNR loss used by the original DCCRN, because its variation range is more suitable for distillation tasks. We use Pytorch to implement the method. All models are optimized using Adam with a learning rate of 0.0006, a batch size of 32, and an epoch number of 20.

3.3. Experimental Results and Discussion

Since this paper focuses on real-time SE applications, the low-complexity methods NSNet [20], RNNoise [19] (data from [22]), and DTLN [21] which have objective scores reported on the DNS Challenge 2020 are chosen as the undistilled models. We retrain the original DCCRN as the teacher model DCCRN-T, and use the compressed one as the student model DCCRN-S. Regarding the comparison of distillation methods, the method that reduces the difference of model output (Diff) [15] is first chosen for comparison. Since the distillation for the intermediate representation of the model is rare to find in the SE field, we select the mainstream distillation methods in the image processing field for comparison. The reviewKD [14] framework that uses the hierarchical context loss (HCL) as the distillation loss and the SPKD [13] method that calculates similarities between the same level's features are selected as the comparison algorithms to prove the effectiveness of the proposed cross-layer similarity knowledge distillation (CLSKD) method that combines the two strategies. PKT [12] uses cosine similarity to compress the intermediate representation of the model, which is similar to the idea of this paper, so it is also used for comparison. Wideband PESQ (WB-PESQ) [23] and STOI [24] are used as objective indicators for speech quality assessment.

Figure 4 compares the WB-PESQ indicators of each distillation method for the intermediate representation on the validation set. It is worth noting that the frame-level similarity knowledge distillation method (SKD) has achieved a more stable and effective improvement than the original SPKD algorithm. And the proposed CLSKD method has the largest improvement.

Regarding the indicators shown in Table 1, compared with the Diff method that directly narrows the output distance, the proposed CLSKD has an improvement on WB-PESQ, which is brought by the distillation of intermediate representation. Among feature distillation methods, the PKT method shows advantages in STOI indicators, but its improvement on WB-PESQ is limited. The SPKD's comprehensive performance on the test set and the validation set is slightly better than the one of PKT, which shows that compared to the probability distribution, the use of pairwise similarity to model knowledge from the teacher is more suitable for the SE field. Compared with SPKD, the frame-level similarity distillation method (SKD) proposed in this paper has a further improvement in each indicator. The ReviewKD method, which uses HCL as the distillation loss, fails to achieve advantages in both indicators. This may be due to the loss of frame-level effective information by the downsampling operation of HCL. The proposed CLSKD ranks first on WB-PESQ and is equivalent to PKT on STOI, reflecting that the introduction of cross-layer information on the basis of SKD can achieve further improvements. Compared with other low-complexity undistilled models, the distilled student model using the CLSKD method maintains a competitive enhancement effect at a low parameter level (0.23M). Moreover, the proposed distillation method is completely cost-free at test time, because the student model remains the same during inference.

4. Conclusions

In this paper, we proposed a new knowledge distillation framework for SE. Intermediate features of multiple layers in the teacher are used to guide one layer in the student, and a frame-level pairwise similarity distance is calculated as the distillation loss. To the best of our knowledge, this is the first time that the intermediate representations of the network were used to distill SE models. Experimental results show that the proposed cross-layer similarity distillation method can considerably improve the enhancement effect of the student model, and outperforms other distillation methods. For future work, we hope to apply our distillation method to more structures.

5. Acknowledgements

The work was supported in part by the National Key Research and Development Program of China under Grant Nos. 2020YFC2004002 and 2020YFC2004003, and the National Natural Science Foundation of China under Grant No. 62001215.

6. References

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[2] P. Scalart and J. Filho, "Speech enhancement based on a priori signal to noise estimation," in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, vol. 2, 1996, pp. 629–632 vol. 2.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[4] F. Weninger and B. Schuller, "Optimization and Parallelization of Monaural Source Separation Algorithms in the openBliSSART Toolkit," *Journal of Signal Processing Systems*, vol. 69, no. 3, pp. 267–277, 2012.

[5] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[6] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, 2014.

[7] S. Liu, G. Keren, E. Parada-Cabaleiro, and B. W. Schuller, "N-HANS: A neural network-based toolkit for in-the-wild audio enhancement," *Multimedia Tools and Applications*, pp. 1–25, 6 2021.

[8] C. K. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, S. Matusevych, R. Aichner, A. Aazami, S. Braun, P. Rana, S. Srinivasan, and J. Gehrke, "The INTERSPEECH 2020 Deep Noise Suppression Challenge: Datasets, Subjective Testing Framework, and Challenge Results," in *Proc. Interspeech 2020*, 2020, pp. 2492–2496. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-3038

[9] C. K. A. Reddy, H. Dubey, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "Icassp 2021 deep noise suppression challenge," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6623–6627.

[10] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural network with pruning, trained quantization and huffman coding," in *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. [Online]. Available: http://arxiv.org/abs/1510.00149

[11] G. E. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *CoRR*, vol. abs/1503.02531, 2015. [Online]. Available: http://arxiv.org/abs/1503.02531

[12] N. Passalis, M. Tzelepi, and A. Tefas, "Probabilistic knowledge transfer for lightweight deep representation learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 5, pp. 2030–2039, 2021.

[13] F. Tung and G. Mori, "Similarity-preserving knowledge distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1365–1374.

[14] P. Chen, S. Liu, H. Zhao, and J. Jia, "Distilling knowledge via knowledge review," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5008–5017.

[15] S. Nakaoka, L. Li, S. Inoue, and S. Makino, "Teacher-student learning for low-latency online speech enhancement using wave-u-net," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 661–665.

[16] X. Hao, S. Wen, X. Su, Y. Liu, G. Gao, and X. Li, "Sub-Band Knowledge Distillation Framework for Speech Enhancement," in *Proc. Interspeech 2020*, 2020, pp. 2687–2691. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-1539

[17] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement," in *Proc. Interspeech 2020*, 2020, pp. 2472–2476. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-2537

[18] A. Dfossez, G. Synnaeve, and Y. Adi, "Real Time Speech Enhancement in the Waveform Domain," in *Proc. Interspeech 2020*, 2020, pp. 3291–3295. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-2409

[19] J.-M. Valin, "A hybrid dsp/deep learning approach to real-time full-band speech enhancement," in *2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP)*, 2018, pp. 1–5.

[20] Y. Xia, S. Braun, C. K. A. Reddy, H. Dubey, R. Cutler, and I. Tashev, "Weighted speech distortion losses for neural-network-based real-time speech enhancement," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 871–875.

[21] N. L. Westhausen and B. T. Meyer, "Dual-Signal Transformation LSTM Network for Real-Time Noise Suppression," in *Proc. Interspeech 2020*, 2020, pp. 2477–2481. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-2631

[22] U. Isik, R. Giri, N. Phansalkar, J.-M. Valin, K. Helwani, and A. Krishnaswamy, "PoCoNet: Better Speech Enhancement with Frequency-Positional Embeddings, Semi-Supervised Conversational Data, and Biased Loss," in *Proc. Interspeech 2020*, 2020, pp. 2487–2491. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-3027

[23] A. Takahashi, A. Kurashima, C. Morioka, and H. Yoshino, "Objective quality assessment of wideband speech by an extension of itu-t recommendation p. 862," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[24] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.