A low-computational DNN-based speech enhancement for hearing aids based on element selection

Chiho Haruta and Nobutaka Ono
Department of Computer Science, Graduate School of Systems Design
Tokyo Metropolitan University
Tokyo, Japan
haruta-chiho@ed.tmu.ac.jp, onono@tmu.ac.jp

***Abstract*—In this study, we propose a low-computational deep neural network (DNN)-based speech enhancement scheme for hearing aids. Since the computational resources in the digital signal processor embedded in a hearing aid are very limited, we reduce the input feature dimension for the DNN. To achieve low computational processing, we consider a dimensionality reduction by element selection. The elements are selected by minimizing the reconstruction error of a linear autoencoder. Because it is a selection, the proposed dimensionality reduction does not need any multiplications, unlike other dimensionality reduction algorithms such as principal component analysis. Therefore, our algorithm can reduce computational cost with little degradation of the speech enhancement performance. We evaluate the performance and the computational cost of the proposed algorithm compared with conventional algorithms.**

***Index Terms*—speech enhancement, noise reduction, hearing aids, dimensionality reduction, low computational cost.**

I. Introduction

Hearing-impaired people have great difficulty understanding speech in noisy environments such as parties or crowded restaurants. To solve this problem, several types of single-channel speech enhancement algorithm based on spectral subtraction techniques have been proposed and implemented in hearing aid systems [1]. Although most of these algorithms improve listening comfort of hearing-impaired people in noisy situations but rarely improve speech intelligibility [2]–[4]. A deep neural network (DNN)-based speech enhancement algorithm with time-frequency masking in the short-time Fourier transform (STFT) domain has been proposed and shown to exhibit a significant performance gain for speech intelligibility in noisy environments for hearing-impaired people [5]–[8]. However, sound processing in hearing aids is required to work with low computational cost. In such systems, a large number of multiplications can greatly increase the computational burden. Several algorithms have been proposed to reduce the computational cost of DNNs. For example, teacher-student learning can be used to design a small and compact network with high accuracy based on larger models [9]–[11]. Pruning [12] removes inessential parameters without incurring accuracy loss. Quantization [13]–[15] reduces the computational precision of weight coefficients and activation functions.

We focus on reducing the dimensionality of input features to reduce computational cost. Dimensionality reduction is an essential preprocessing step improving the computational efficiency and accuracy of machine learning [16]. There are various dimensionality reduction methods such as principal component analysis (PCA), linear discriminant analysis, and multidimensional scaling. However, a linear transformation such as PCA requires matrix multiplication, which increases the number of multiplications and thus imposes a large computational burden on the hearing aid.

In order to realize dimensionality reduction with low computational cost, we propose a new approach of reducing dimensionality by the element selection proposed in [17], rather than a linear transformation. For an appropriate selection, the elements are selected to minimize the reconstruction error of the linear autoencoder of the original data. After dimensionality reduction, the selected elements are fed to a DNN as input features. If the selected elements retain the features of the original data, the DNN is expected to be trained to achieve the same accuracy as one trained with original data, even if the number of multiplications at the input layer is reduced. Furthermore, the proposed algorithm has the possibility of being used as a preprocessing step in other algorithms to improve their performance to computational cost ratio.

We apply this scheme to DNN-based speech enhancement. We compare the performance of the proposed low-computational DNN-based speech enhancement and that of a conventional approach.

II. DNN-based speech enhancement using time-frequency masking

A. DNN-based time-frequency masking

We apply DNN-based time-frequency masking to speech enhancement. Let $X(\omega, \tau)$ be an STFT of a mixture of the target clean speech and a noise signal. In time-frequency masking techniques, the target clean signal is estimated by

applying the time-frequency mask $M(\omega,\tau)$ to $X(\omega,\tau)$ in the STFT domain.

$$\hat{S}(\omega,\tau) = M(\omega,\tau)X(\omega,\tau) \tag{1}$$

Then, the estimated clean speech in the time domain can be obtained by the inverse discrete Fourier transform and overlap-add. There are various masking targets, such as ideal binary mask [18], [19], ideal ratio mask (IRM) [20], and spectral magnitude mask [8]. In this work, we choose the IRM since it has been reported that it improves the speech intelligibility of hearing-impaired people in noisy environments [5], [6]. The IRM is given by

$$M(\omega,\tau) = \frac{|S(\omega,\tau)|^2}{|X(\omega,\tau)|^2}, \tag{2}$$

where $S(\omega,\tau)$ is an STFT of the target clean signal. Since we do not know $S(\omega,\tau)$, we use a DNN to estimate $M(\omega,\tau)$ from the noisy speech $X(\omega,\tau)$.

B. Low-computational DNN-based speech enhancement

The number of multiplications often affects the power consumption when considering applications to small devices such as hearing aids. Therefore, in the rest of this paper, we define the computational cost by the number of multiplications. DNN-based time-frequency masking often requires many multiplications, especially in input layers.

Several methods can be used to reduce the computational cost of DNN inference. Pruning [12] removes connections or units that are unnecessary or redundant. Quantization reduces the precision of the parameters or activations to lower bit representations [13]–[15]. Teacher-student learning transfers a trained model to a smaller model [9]–[11]. As one of these techniques, we focus on dimensionality reduction which is widely used to reduce computational cost and increase accuracy. If we can reduce the dimensionality of the input vector to the DNN without losing useful information from the original data, we can develop an algorithm that reduces the computaional cost on its own or when combined with conventional algorithms.

III. Dimensionality reduction based on element selection

A. Linear dimensionality reduction

Dimensionality reduction is a preprocessing step for removing redundant features to improve the accuracy and reduce the computational cost [16]. We consider obtaining M-dimensional real-valued vector $\boldsymbol{y}$ from K-dimensional real-valued vector $\boldsymbol{x}$ ($M < K$). One of the most widely used methods for dimensionality reduction is PCA [16]. In PCA, we obtain $\boldsymbol{y}$ by multiplying matrix $P \in \mathbb{R}^{M\times K}$ with $\boldsymbol{x}$ as

$$\boldsymbol{y} = P\boldsymbol{x}. \tag{3}$$

P is a projection matrix that transforms vectors from the original space to a new space whose basis is the M components of the original vectors. However, this implementation requires $M \times K$ multiplications in (3), which is a significant problem for hearing aids with limited computational resources.

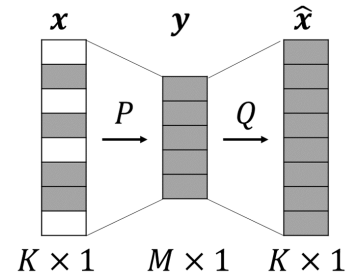

Fig. 1. Dimensionality reduction by element selection.

B. Dimensionality reduction based on element selection [17]

To avoid additional multiplications due to dimensionality reduction, we apply element selection proposed in [17], where we obtain an M-dimensional vector by selecting M elements from a K-dimensional vector. Element selection requires no multiplication and it is equivalent to (3) when the (i,j)-th element of P satisfies

$$p_{ij} = \begin{cases} 1 & (j = \sigma(i)) \\ 0 & (\text{otherwise}), \end{cases} \tag{4}$$

where $\{\sigma(i)|i = 1,2,\cdots,M\}$ indicates M selected indices from K indices $\{1,2,\cdots,K\}$ without duplication. $\sigma(i)$ is selected by minimizing the reconstruction error of a linear autoencoder of $\boldsymbol{x}$. In other words, when $\boldsymbol{x}$ is reconstructed with the reconstruction matrix Q to $\boldsymbol{y}$, we set P that minimizes the mean squared error of reconstructed vector $\hat{\boldsymbol{x}} = Q\boldsymbol{y}$ and original vector $\boldsymbol{x}$, where $Q \in \mathbb{R}^{K\times M}$ is optimized. P is derived as

$$P = \underset{P}{\operatorname{argmin}}\ \epsilon(P), \tag{5}$$

where

$$\epsilon(P) = \min_{Q} \mathrm{E}(|Q\boldsymbol{y} - \boldsymbol{x}|^2) = \min_{Q} \mathrm{E}(|QP\boldsymbol{x} - \boldsymbol{x}|^2). \tag{6}$$

Fig. 1 presents an image of dimensionality reduction by element selection and reconstruction.

We optimize $\sigma(i)(i = 1,2,\cdots,M)$ by iteratively swapping a selected index and a non selected index. Let Δ_{hj} be $\epsilon(PR_{hj}) - \epsilon(P)$, where h is a selected index (one of $\sigma(i)$), j is a non selected index, and R_{hj} is an identity matrix except for the four entries: $r_{hh} = r_{jj} = 0$ and $r_{hj} = r_{jh} = 1$ where r_{hj} is the (h,j)-th element of R_{hj}, and others as well. R_{hj} works for swapping the index h and j. Then, the optimization algorithm can be written as follows.

1. Initialize $\sigma(i)(i = 1,2,\cdots,M)$ by selecting indices randomly.
2. For each of $i = 1,2,\cdots,M$, evaluate $\Delta_{\sigma(i)j}$ for all non selected j and find its minimum in terms of j. If the minimum is less than 0, swap $\sigma(i)$ and j^* achieving the minimum, and update P. If not, just move to the next i.
3. If $\Delta_{\sigma(i)j}$ is equal to or larger than 0 for all $i = 1,2,\cdots,M$, the algorithm is converged.

This algorithm is greedy and the reconstruction error in (6) does not increase while the above algorithm is running. The

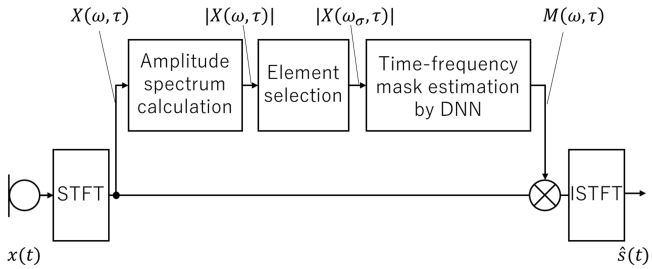

Fig. 2. Block diagram of the proposed algorithm.

fast computation for this algorithm was preliminarily presented in [17] and will be presented in another paper.

IV. Low-computational DNN-based speech enhancement based on element selection

Our proposed algorithm reduces the computational cost of DNN-based speech enhancement using dimensionality reduction based on element selection, with the ultimate goal of application to hearing aids. Fig. 2 presents a block diagram of the proposed algorithm. Note that the element selection does not need any multiplications because it is a processing of just selecting pre-determined indexed elements of $\boldsymbol{x}$.

We conducted experiments to confirm the effectiveness of the dimensionality reduction method by element selection. The performance of the proposed algorithm is compared with that of a conventional algorithm. To estimate the IRM, a fully-connected DNN, which is the most basic and simplest speech enhancement architecture [21], is used. We prepare several conditions that require various numbers of multiplications for both the conventional and proposed algorithms by changing the number of units in hidden layers.

V. Experiments

A. Dataset

We used clean speech from the Japanese Newspaper Article Sentences corpus [22] and noise from the TUT dataset [23], [24]. To generate training data, 12 hours of speech (50 male and female speakers each) of the corpus were combined with noise recorded in 15 types of environment at signal-to-noise ratios (SNRs) of 0, 5, and 10 dB. Evaluation data were obtained from 20 min of speech (five male and five female speakers different from the speakers used for the training data) and combined with 2 kinds of noise ('Traffic' and 'Cafe/restaurant' noise from different recordings used for the training) at SNRs of 0, 5, and 10 dB.

B. Setup

The sampling frequency is 16 kHz, the frame length is 1024 samples, and the frame shift length is 512 samples. A Hamming window is used for frame analysis. The input of the DNN is the amplitude spectrum for the current frame of noisy speech, and the target of the DNN is the IRM at the same frame. Therefore, the DNN has 513 input and 513 output dimensions when without dimensionality reduction. When the dimensionality reduction is adopted, the DNN has fewer input dimensions than 513. DNN has 3 hidden layers under each condition. Table I and Table II show the conditions used in the experiment. The conventional method without dimensionality reduction is denoted as All_n, where n indicates the number of units in the hidden layers. Similarly, the proposed method with dimensionality reduction by element selection to minimize the mean reconstruction error is denoted as Sel_MMRE_n. Also, the method with dimensionality reduction by random element selection is denoted as Sel_Rand_n. The condition All_n has 513 input dimensions because there is no dimensionality reduction. Sel_MMRE_n and Sel_Rand_n reduce the input dimensions by half by element selection.

TABLE I
Conditions without dimensionality reduction

Condition	Network structure	Multiplications
All_{512}	513-512-512-512-513	1049600
All_{256}	513-256-256-256-513	393728
All_{102}	513-102-102-102-513	125460
All_{51}	513-51-51-51-513	57528
All_{25}	513-25-25-25-513	26900

TABLE II
Conditions for dimensionality reduction by element selection

Condition	Network structure	Multiplications
$\text{Sel\_MMRE}_{512}/\text{Rand}_{512}$	256-512-512-512-513	918016
$\text{Sel\_MMRE}_{256}/\text{Rand}_{256}$	256-256-256-256-513	327936
$\text{Sel\_MMRE}_{102}/\text{Rand}_{102}$	256-102-102-102-513	99246
$\text{Sel\_MMRE}_{51}/\text{Rand}_{51}$	256-51-51-51-513	44421
$\text{Sel\_MMRE}_{25}/\text{Rand}_{25}$	256-25-25-25-513	20475

For training, the learning rate is set to 0.01 and Adam [25] is used as the optimization algorithm with decay rates $\beta_1 = 0.9$ and $\beta_2 = 0.999$. ReLU [26] is used as the activation function for all layers, except for the output layer, which uses a sigmoid function. Each model is trained for 400 epochs and then evaluated. The DNN is trained to minimize the following loss function:

$$J = \sum_{\omega=1}^{\omega_{all}} |M(\omega,\tau)X(\omega,\tau) - S(\omega,\tau)|^2, \tag{7}$$

where $M(\omega,\tau)$ is the predicted IRM, $S(\omega,\tau)$ is the STFT of the target clean signal, and ω_{all} is the total number of the frequency bins. The scale-invariant source-to-distortion ratio (SI-SDR) [27] is used to evaluate the speech enhancement performance. Fig. 3 shows the selected element indices for the Sel_MMRE_n and Sel_Rand_n. It shows that many low-frequency components are selected in Sel_MMRE_n.

C. Results

Fig. 4 shows the relationship between the number of multiplications and the speech enhancement performance. The average improvement of the SI-SDR for each method compared with that of the non-processed noisy signal. For the conditions that have approximately the same numbers of multiplications, in the other words, the difference between the number of multiplications is less than twofold, the proposed method Sel_MMRE_n outperformed the conventional method

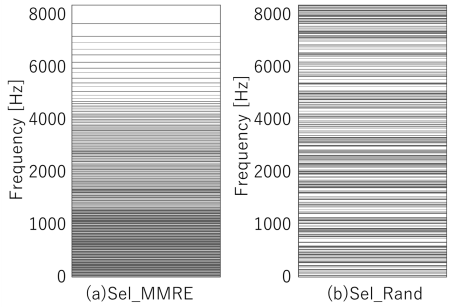

Fig. 3. Selected frequencies in (a) Sel_MMRE and (b) Sel_Rand.

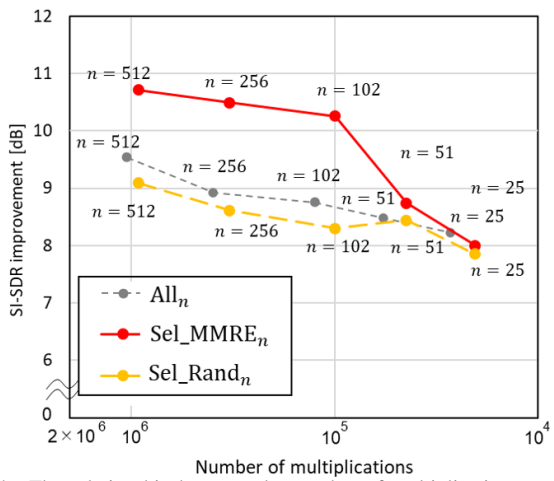

Fig. 4. The relationship between the number of multiplications and speech enhancement performance. n represents the number of units in hidden layers.

All$_n$, except for the cases of $n = 25$. The average SI-SDR improvements of the proposed methods Sel_MMRE$_{512}$, Sel_MMRE$_{256}$, and Sel_MMRE$_{102}$ were 10.72, 10.50, and 10.26 dB, which were higher than those of the conventional methods All$_{512}$, All$_{256}$, and All$_{102}$, respectively, even though the number of multiplications was slightly reduced in each case. In addition, compared with the pair of All$_{512}$ and Sel_MMRE$_{256}$ and the pair of All$_{256}$ and Sel_MMRE$_{102}$, the proposed methods obtained better performance than the conventional method with less than 1/3 the number of multiplications of the conventional methods. Additionally, it is not shown in figures here, but the training loss in ALL$_n$ was smaller than that in Sel_MMRE$_n$ for all n while the validation loss in ALL$_n$ was almost larger than Sel_MMRE$_n$ except for the case of $n = 51$. It shows that the proposed method may enable the model to be generalized more to unseen conditions. On the other hand, the conventional method All$_n$ outperformed Sel_Rand$_n$ in almost all cases except for Sel_Rand$_{51}$. In summary, these results indicate that reducing the input dimensions by appropriate element selection achieved a better speech enhancement performance than did the conventional method even with the same or smaller number of multiplications.

Fig. 5 presents some spectrograms of clean speech and processed signals as examples. We can also see that Sel_MMRE$_{256}$ was denoised more than All$_{256}$ and Sel_Rand$_{256}$ from the spectrograms.

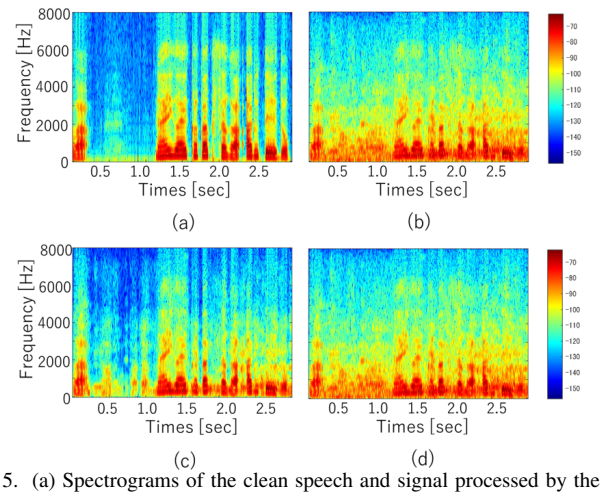

Fig. 5. (a) Spectrograms of the clean speech and signal processed by the (b) conventional method All$_{256}$, (c) proposed method Sel_MMRE$_{256}$, and (d) Sel_Rand$_{256}$ for the mixture of speech and 'cafe/restaurant' noise at SNR 5 dB.

To further investigate charasteristics of Sel_MMRE and Sel_Rand, we compared frequency-wise SNRs defined as

$$\mathrm{SNR}(\omega) = 10\log_{10}\frac{\sum_\tau |S(\omega,\tau)|^2}{\sum_\tau |M(\omega,\tau)X(\omega,\tau) - S(\omega,\tau)|^2}. \quad (8)$$

Fig. 6 and Fig. 7 show the SNR improvements in Sel_MMRE$_{512}$ and Sel_Rand$_{512}$ for all the evaluation data, respectively. The SNR improvements in Sel_MMRE$_{512}$ in Fig. 6 are larger than those in Sel_Rand$_{512}$ in Fig. 7 at most of the frequencies except those around 6000 Hz and those higher than 7000 Hz. As in Fig. 6, regardless of whether the frequency is selected or not, the plotted points were almost on a smooth curve. Although fewer points are selected at higher frequencies, there is not a significant degradation compared with those at lower frequencies. We can see that the proposed algorithm estimated IRM with high accuracy only from information at the selected frequencies.

The results indicate that the element selection method proposed and demonstrated above is suitable for DNN-based speech enhancement and can reduce the input dimensions of the DNN effectively and efficiently. Further studies on element selection will be needed to improve the performance and work for extended situations such as when signals from several frames are used as input features of the DNN.

VI. Conclusion

In this paper, we proposed a new method for low-computational DNN-based speech enhancement. In the proposed method, the input feature dimensions were reduced by element selection. The elements that minimize the reconstruction error of the linear autoencoder of the original input signal were selected. The proposed method achieved better performance than the conventional method even when fewer numbers of multiplications were required. As a future task, setting a more suitable element selection criterion in accordance with the applied problem situation or the postprocessing of the DNN may further improve the performance.

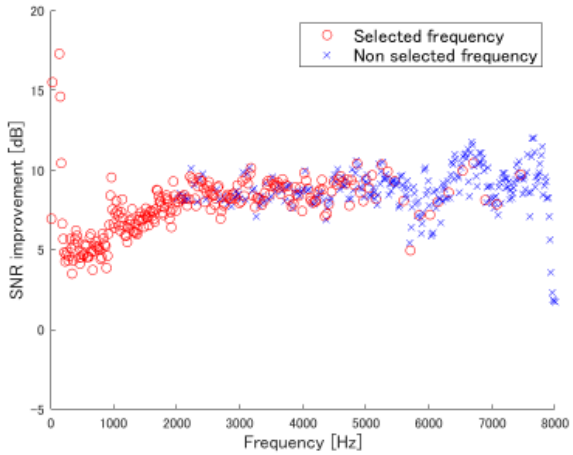

Fig. 6. SNR improvement under the condition Sel_MMRE$_{512}$. The red circles represent the SNR improvement at selected frequencies and the blue crosses represent that at non selected frequencies.

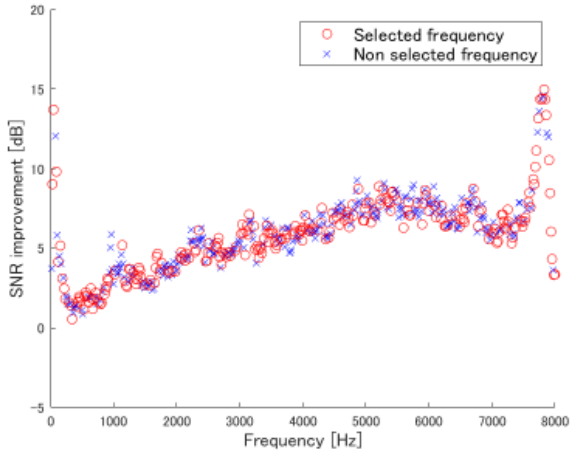

Fig. 7. SNR improvement under the condition Sel_Rand$_{512}$. The meanings of the symbols are the same as in Fig. 6

ACKNOWLEDGMENT

This work was supported by JST CREST Grant Number JPMJCR19A3, Japan.

REFERENCES

[1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error short-time spectral amplitude estimator," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[2] H. Luts, K. Eneman, J. Wouters, M. Schulte, M. Vormann, M. Buechler, N. Dillier, R. Houben, W. A. Dreschler, M. Froehlich, H. Puder, G. Grimm, V. Hohmann, A. Leijon, A. Lombard, D. Mauler, and A. Spriet, "Multicenter evaluation of signal enhancement algorithms for hearing aids," *Journal of the Acoustical Society of America*, vol. 127, no. 3, pp. 1491–1505, 2010.

[3] F. Dubbelboer and T. Houtgast, "The concept of signal-to-noise ratio in the modulation domain and speech intelligibility," *Journal of the Acoustical Society of America*, vol. 124, no. 6, pp. 3937–3946, 2008.

[4] Y. Hu and P. C. Loizou, "A comparative intelligibility study of single-microphone noise reduction algorithms," *Journal of the Acoustical Society of America*, vol. 122, no. 3, pp. 1777–1786, 2007.

[5] D. L. 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] M. Aubreville, K. Ehrensperger, A. Maier, T. Rosenkranz, B. Graf, and H. Puder, "Deep denoising for hearing aid applications," in *Proceedings of International Workshop on Acoustic Signal Enhancement* , pp. 361–365, 2018.

[7] G. S. Bhat, N. Shankar, C. K. A. Reddy, and I. M. S. Panahi, "A real-time convolutional neural network based speech enhancement for hearing impaired listeners using smartphone," in IEEE Access, vol. 7, pp. 78421–78433, 2019.

[8] Y. Wang, A. Narayanan, and D. L. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[9] A. S. Subramanian, S. Chen, S. Watanabe, "Student-Teacher Learning for BLSTM Mask-based Speech Enhancement," in *Proceedings of Interspeech*, pp. 3249–3253, 2018.

[10] R. Aihara, T. Hanazawa, Y. Okato, G. Wichern, and J. L. Roux, "Teacher-student deep clustering for low-delay single channel speech separation," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 690–694, 2019.

[11] Yan-Hui Tu, Jun Du, and Chin-Hui Lee, "Speech enhancement based on teacher–student deep learning using improved speech presence probability for noise-robust speech recognition," *IEEE/ACM Transactions on. Audio, Speech and Language Processing*, vol. 27, no. 12, pp. 2080–2091, 2019.

[12] S. Han, H. Mao, and W. Dally. "Deep compression: compressing deep neural network with pruning, trained quantization and Huffman coding," in *The International Conference on Learning Representations*, 2016.

[13] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, "Quantized neural networks: training neural networks with low precision weights and activations," Journal of Machine Learning Research, vol. 18, no. 1, pp.6869–6898, 2018.

[14] Y. Choukroun, E. Kravchik, F. Yang, and P. Kisilev, "Low-bit Quantization of Neural Networks for Efficient Inference," in *Proceedings of International Conference on Computer Vision Workshop*, pp. 3009–3018, 2019.

[15] S. Abdullah, M. Zamani, and A. Demosthenous, "Towards more efficient DNN-based speech enhancement using quantized correlation mask," in IEEE Access, vol. 9, pp. 24350–24362, 2021.

[16] S. Velliangiri, S. Alagumuthukrishnan, S. I. Thankumar joseph, "A review of dimensionality reduction techniques for efficient computation," *Procedia Computer Science*, vol. 165, pp. 104–111, 2019.

[17] N. Ono, "Dimension reduction without multiplication in machine learning", in *The Technical Report of The Proceeding of The Institute of Electronics, Information and communication Engineers*, vol. 119, no. 439, pp. 21–26, 2020 (in Japanese).

[18] D. L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," Speech Separation by Humans and Machines, Springer, pp. 181–197, 2005.

[19] S. Srinivasan, N. Roman, and D. L. Wang, "Binary and ratio time-frequency masks for robust speech recognition," Speech Communication, vol. 48, pp. 1486–1501, 2006.

[20] D. S. Williamson, Y. Wang, and D. L. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no.3, pp. 483–492, 2016.

[21] S. A. Nossier, J. Wall, M. Moniri, C. Glackin, and N. Cannings, "An experimental analysis of deep learning architectures for supervised speech enhancement," *Electronics*, vol. 10, no. 17, 2021.

[22] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "Jnas: Japanese speech corpus for large vocabulary continuous speech recognition research," vol. 20, no. 3, pp. 199–206, 1999.

[23] A. Mesaros, T. Heittola, and T. Virtanen. TUT Acoustic Scenes 2017. https://zenodo.org/record/400515

[24] A. Mesaros, T. Heittola and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," in *Proceedings of European Signal Processing Conference*, pp. 1128–1132, 2016.

[25] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of International Conference for Learning Representations*, 2015.

[26] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of International Conference on Artificial Intelligence and Statistics*, vol. 15, pp. 315–323, 2011.

[27] J. L. Roux, S. Wisdom, H. Erdogan and J. R. Hershey, "SDR – half-baked or well done?," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 626–630, 2019.