INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Sparse Mixture of Local Experts for Efficient Speech Enhancement

Aswin Sivaraman, Minje Kim

Indiana University, Department of Intelligent Systems Engineering, USA
asivara@indiana.edu, minje@indiana.edu

Abstract

This work proposes a novel approach for reducing the computational complexity of speech denoising neural networks by using a sparsely active ensemble topology. In our ensemble networks, a gating module classifies an input noisy speech signal either by identifying speaker gender or by estimating signal degradation, and exclusively assigns it to a best-case specialist module, optimized to denoise a particular subset of the training data. This approach extends the hypothesis that speech denoising can be simplified if it is split into non-overlapping subproblems, contrasting earlier approaches that train large generalist neural networks to address a wide range of noisy speech data. We compare a baseline recurrent network against an ensemble of similarly designed, but smaller networks. Each network module is trained independently and combined to form a naïve ensemble. This can be further fine-tuned using a sparsity parameter to improve performance. Our experiments on noisy speech data–generated by mixing LibriSpeech and MUSAN datasets–demonstrate that a fine-tuned sparsely active ensemble can outperform a generalist using significantly fewer calculations. The key insight of this paper, leveraging model selection as a form of network compression, may be used to supplement already-existing deep learning methods for speech denoising.

Index Terms: speech enhancement, adaptive mixture of local experts, model selection, neural network compression

1. Introduction

Speech denoising is a highly-studied research problem aimed at improving speech quality and intelligibility within noisy recordings. Denoising methods are often assessed by the removal of non-speech components and the minimization of any artifacts introduced by the enhancement process. This work addresses the specific scenario of removing non-stationary uncorrelated background noise from a monophonic recording of a single English speaker. While there are well-established algorithms for speech denoising—such as Wiener filtering [1], spectral subtraction [2], and the short-time spectral amplitude method [3]—recent advances in deep learning technology have significantly improved performance by reformulating speech denoising as a supervised learning task.

Time-frequency (TF) mask estimation is a prevailing supervised learning strategy for speech denoising using deep neural networks (DNN), in which each TF bin is classified into two categories: speech and noise [4, 5]. In order to capture the temporal structure inherent in speech signals, complex deep learning models incorporating recurrent neural networks (RNN) have seen greater usage [6, 7]. TF masking algorithms can also work for speech separation problems. A recent DNN-based unsupervised learning approach known as “deep clustering” can perform speaker-independent separation on an arbitrary number of sound sources [8]. Deep clustering has been shown to benefit from jointly estimating time-frequency masks [9] or through incorporating spatial audio features such as phase difference [10].

The growing number of DNN-based speech denoising methods is a consequence of the ubiquitous increase in computing power, in part due to accelerated matrix multiplications on graphics processing units (GPUs). In order to learn highly nonlinear mapping functions, modern-day neural networks now perform millions of calculations on millions of learnable parameters. This trend of larger DNNs with higher computational and spatial complexity is at odds with the commercial need for robust low-power models to work in resource-limited devices. The fundamental trade-off between model performance and model complexity is the subject of ongoing deep learning research. Popular domain-agnostic techniques for network compression include pruning weights and filters [11, 12, 13] and quantizing network parameters [14, 15, 16]. With regards to speech denoising, bit-depth reduction has been successfully shown to compress fully-connected and recurrent models [17, 18].

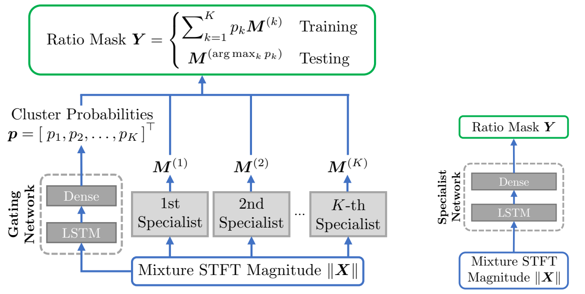
This paper examines an approach to network compression by leveraging a neural network design philosophy popularly known as “mixture of local experts” (MLE) [19]. In the MLE procedure, an ensemble of specialist networks are independently trained to handle a subset of all training cases. A classifier submodule, referred to as the gating network, is trained to either select the best-suited specialist for a given input sample or calculate a weighted sum of all the specialist inferences. Recent literature in speech separation and speech enhancement has explored different uses for ensemble architectures–either for multicontext feature extraction [20], for phoneme-specific enhancement [21], or for model selection [22, 23]–however, our work is the first to use the MLE design as a means for reducing the number of inference-time calculations (i.e. arithmetic complexity) of speech denoising neural networks.

Prior work has shown that the speech denoising problem can be decomposed into independent subproblems constituting the various dimensions along which noisy speech signals may vary [24], while their complexity-related issues were not discussed. Our main contribution is that we constrain the MLE model to be sparsely active, so that only one specialist is used at a given time. This approach can reduce the total number of calculations during denoising inference assuming that the sum of network parameters (from the one specialist and the gating network) are lesser than an equally-performing general-purpose speech denoising network. Our experiments comprehensively evaluate the sparsely active ensemble of specialists topology and show that the savings in arithmetic complexity do not result in compromised speech denoising performance.

2. The Proposed Method

Given that the speech denoising task can be divided into mutually exclusive subproblems, we propose that it must be possible to split a complete noisy speech dataset along some latent

http://dx.doi.org/10.21437/Interspeech.2020-2989



(a) (b)

Figure 1: *Comparison between (a) the proposed ensemble of specialists model and (b) the baseline model.*

dimension in order to form non-overlapping subsets (i.e. clusters). Although the MLE procedure is theoretically capable of learning latent clusters in an unsupervised fashion, in this work, we utilize prior knowledge about the problem domain to designate two latent spaces: (1) different speech degradation levels and (2) speaker gender.

The proposed model, shown in Figure 1a, is an ensemble of specialist networks regulated by a gating network. While it is fundamentally possible to utilize the inferences of multiple specialists, we propose using only a single specialist in order to bring computational complexity during inference to a minimum. We assume that the noisy speech data can be split into distinct subsets. Consequently, we pre-train each specialist network to individually address one subproblem. Our experiments compare the proposed ensemble model against a baseline model, shown in Figure 1b, which is architecturally equivalent to a specialist network but is trained using the entire noisy speech training set. In the following subsections of this paper, we define the specialist and gating modules more formally.

2.1. Specialist Networks

A monaural time-domain mixture signal $\boldsymbol{x}$ is defined as the sum of a clean speech signal $\boldsymbol{s}$ and an additive background noise signal $\boldsymbol{n}$: $\boldsymbol{x} = \boldsymbol{s} + \boldsymbol{n}$. The goal of speech denoising is to learn a mapping function g which produces an estimated signal $\hat{\boldsymbol{s}}$ such that $g(\boldsymbol{x}) = \hat{\boldsymbol{s}} \approx \boldsymbol{s}$.

A well-known objective metric for this single-channel denoising task is the signal-to-distortion ratio (SDR). Implemented as part of the *BSS_eval* toolkit [25], SDR expresses the ratio of energy between a source signal and an estimate signal. A more robust modification of SDR, known as scale-invariant SDR (SI-SDR), uses a scaling factor α to ensure that the residual vector (between source $\boldsymbol{s}$ and estimate $\hat{\boldsymbol{s}}$) maintains orthogonality to the source [26] as follows:

$$\text{SI-SDR}(\boldsymbol{s}, \hat{\boldsymbol{s}}) = 10 \log_{10} \left[\frac{\sum_t (\alpha s_t)^2}{\sum_t (\alpha s_t - \hat{s}_t)^2} \right]. \tag{1}$$

For standard SDR, $\alpha = 1$; for SI-SDR, $\alpha = \frac{\hat{\boldsymbol{s}}^\top \boldsymbol{s}}{\boldsymbol{s}^\top \boldsymbol{s}}$. Both specialist and baseline networks are trained to maximize this metric between the recovered estimate speech $\hat{\boldsymbol{s}}$ and the reference clean speech $\boldsymbol{s}$.

There are many possible ways to produce $\hat{\boldsymbol{s}}$ given only $\boldsymbol{x}$. One established approach is known as time-frequency (TF) masking, in which models estimate a TF mask matrix $\boldsymbol{Y}$ such that $\hat{\boldsymbol{S}} = \boldsymbol{Y} \odot \boldsymbol{X}$, where $\odot$ denotes Hadamard product and $\hat{\boldsymbol{S}}$ and $\boldsymbol{X}$ are the discrete short-time Fourier transforms (STFT) of the estimate signal and the noisy mixture signal respectively. The mask matrix is a ratio at each TF-point in the mixture signal belonging to either noise or speech, with values between 0 and 1 respectively. The inverse STFT transforms $\hat{\boldsymbol{S}}$ from the time-frequency domain back to the time domain $\hat{\boldsymbol{s}}$. To estimate $\boldsymbol{Y}$ through supervised learning, both specialist and baseline models target the ideal ratio mask (IRM) [27], which is defined as:

$$\text{IRM} = \sqrt{\frac{|\boldsymbol{S}|^2}{|\boldsymbol{S}|^2 + |\boldsymbol{N}|^2}} \tag{2}$$

$|\boldsymbol{S}|$ and $|\boldsymbol{N}|$ denote the magnitude STFT of speech and noise respectively. IRM has been shown to work well as a masking target assuming that the interfering noise signal $\boldsymbol{n}$ is uncorrelated with target speech signal $\boldsymbol{s}$ [28, 29].

To focus our attention on the benefits of the ensemble philosophy, with consideration for the constraints of resource-limited environments, we design our specialist network with unidirectional recurrent layers followed by a feed-forward dense layer. The recurrent layers are made up of long short-term memory (LSTM) cells [30]. The number of recurrent layers as well as the number of hidden units per layer are adjustable experiment parameters which affect the overall complexity of the model. The specialist network takes the noisy speech magnitude STFT $|\boldsymbol{X}|$ as input and predicts a ratio mask matrix $\boldsymbol{Y}$. Subsequently, inv-STFT $(\boldsymbol{Y} \odot \boldsymbol{X})$ yields the denoised speech estimate $\hat{\boldsymbol{s}}$.

We note that convolutional neural networks (CNN) on time-domain signals currently achieve the state-of-the-art performance in source separation [31]. Despite their low model complexity, convolutional architectures are able learn the sequence-to-sequence mapping. We leave general application of our proposed ensemble model to different architectures for future work.

2.2. Gating Network

The gating network is responsible for assigning an input signal to the appropriate specialist. It introduces a classification subtask as overhead to the overarching denoising task, splitting the full training dataset into some number of latent clusters.

Identifying latent clusters in a noisy speech corpus is non-trivial. Prior works using ensemble models for speech enhancement have shown that specialists may be trained to denoise a particular phoneme [21]. This approach, which requires training data to be phoneme-labeled, is naturally language-dependent but also non-sparse, as multiple specialists may actively perform some computations due to the high variance of phonemes in speech. To ensure a sparse activation of specialists (ideally one specialist per input signal), a more generalized latent clustering is preferred. For this reason, we design two types of gating networks to classify inputs based on either *speech degradation level* or *speaker gender*.

Similar to the specialist architecture, our gating networks are also designed with multiple recurrent layers and a single dense layer. However, in our current proposed model the gating network does not make predictions frame-by-frame; after processing the entire input sequence, the network produces a single softmax vector $\boldsymbol{p}$, with K elements corresponding to the number of clusters (i.e. the number of specialists). The index of the maximum value in $\boldsymbol{p}$ should correspond to the index of the best-suited specialist.

2.3. Ensemble Network

The proposed ensemble model combines K specialist networks together with a gating network. First, all of the sub-networks are independently trained. The combination of these pre-trained modules forms a primitive ensemble, as the gating network can already assign an incoming test example to one of the specialists. The output mask $\boldsymbol{Y}$ is chosen from the specialist which corresponds to the maximum value of gating network softmax vector $\boldsymbol{p}$. The "hard" gating mechanism is formulated as:

$$\boldsymbol{Y} = \boldsymbol{M}^{(k^*)}, \quad k^* = \arg\max_k p_k, \tag{3}$$

where $\boldsymbol{M}^{(k)}$ denotes the predicted ratio mask matrix from the k-th specialist.

However, this naïve ensemble is sub-optimal as it lacks the potential co-adaptation between gating and specialist networks. For example, given the fact that the gating network cannot classify mixtures with 100% accuracy, the specialists should adapt to the situation where it processes a misclassified sample (e.g., a male speech sample falls in the female speaker's specialist). Knowing this, we can further train the submodules in unison. During this fine-tuning phase, the ensemble model estimates the output ratio mask $\boldsymbol{Y}$ by performing a normalized sum over the individual masks $\boldsymbol{M}^{(k)}$ produced by all specialists weighted by the gating network softmax vector $\boldsymbol{p}$. This "soft" gating mechanism ensures that the ratio mask calculation is differentiable, and is formulated as:

$$\boldsymbol{Y} = \sum_k p_k \boldsymbol{M}^{(k)}. \tag{4}$$

During the test phase, the weighted sum is replaced by the hard-decision shown in Eq. 3. This difference between training-time and evaluation-time computation in the ensemble architecture is the crux of its efficiency; only one out of all the specialists is used to process the entire mixture spectrogram $|\boldsymbol{X}|$, making the total used network parameters a fraction of the total learned. We reduce the discrepancy between the hard and soft gating mechanisms, used during testing and fine-tuning respectively, by introducing a scaling parameter λ to the softmax gating network output:

$$p_k = \frac{\exp(\lambda \cdot o_k)}{\sum_{j=1}^{K} \exp(\lambda \cdot o_j)}. \tag{5}$$

Each element of the gating network output cluster probability vector (p_k) is dependent on the corresponding element of dense layer output (o_k) normalized by the sum of all dense layer output elements. While the traditional softmax function can be calculated using $\lambda = 1$, we elevate the sparsity of $\boldsymbol{p}$ by setting $\lambda = 10$. This saturates $\boldsymbol{p}$ to be near-1 at a single index and near-0 at every other index, making the weighted sum for ratio mask $\boldsymbol{Y}$ (Eq. 4) effectively select the best-case specialist mask. This modification of the softmax function has been successfully used for quantizing vectors with image compression [32].

3. Experiment Setup

All models (specialist, gating, baseline, and ensemble) are trained using a stochastic data sampling strategy which dynamically mixes clean speech recordings from the LibriSpeech[1] corpus [33] with noise recordings from the MUSAN[2] corpus [34]. This exposes the models to up to 251 unique speakers[3] and 843 unique noise types[4] during training. 40 unseen speakers[5] and 87 unseen noise types[6] are used to test the models. 5% of the training utterances and noises are set aside for validation to help determine training convergence.

All experiment audio files use a sampling rate of 16000 Hz. Spectrograms are generated using the STFT with a frame size of 1024 samples and a hop size of 256 samples. Per epoch, for each example in the training batch, the sampler mixes a normalized 1-second snippet of a random training speaker's utterance with a normalized 1-second snippet from a random training noise, chosen with uniform probability. There are 100 mixture signals in a batch. Unlike the training mixtures, test mixtures vary in duration; this gives our models an effective RNN lookback size of 1-second.

We assess the proposed ensemble of specialists methodology across two latent spaces. For the *signal degradation* latent space, we instantiate $K = 4$ specialists and generate noisy speech mixtures with specific signal-to-noise (SNR) levels—either -5, 0, 5, or 10 dB—for each of the four specialists. Similarly for the *speaker gender* experiment, there are $K = 2$ specialists which see a gender-filtered subset of the training data with uniformly varying input SNR values out of the four above listed. In contrast, the baseline model must generalize to all levels of signal degradation and all speaker genders; its training batches consist of 100 mixed gender 1-second-long mixtures with input SNR uniformly distributed between the four values.

All networks are optimized using the Adam optimizer [35] with an initial learning rate of $\eta = 0.001$. The specialist network uses the additive inverse of the SI-SDR metric (Eq. 1) between $\hat{\boldsymbol{s}}$ and $\boldsymbol{s}$ as the loss function, whereas the gating network minimizes the binary cross entropy (BCE) metric between its output, softmax vector $\boldsymbol{p}$, and a ground-truth one-hot vector representing the index of the best-suited specialist. Each network variant is trained for approximately three hours on a NVIDIA Titan Xp GPU, after which the validation metric is considered to have converged.

4. Experiment Results

We report the denoised signal SI-SDR improvement for all models averaged across 1000 test set mixtures. Figure 2a compares the test signal speech denoising performance between the four *signal degradation*-based specialists and the one baseline model. It is evident that, at all mixture SNR levels, a neural network specifically trained to denoise those mixtures can outperform a generalist network. This gap in performance is most prominent with the extrema mixture levels (i.e. the -5dB and +10dB mixture SNR cases). As the number of RNN hidden units and layers increases, the performance gap between specialists and baseline model diminishes. With larger network complexity, the generalist's performance eventually matches the specialist's, which saturates after a particular network size.

The specialist curves in Figure 2a, 2c, and 2d set a theoretical upper bound to the naïve ensemble model: even with a perfect gating network, the naïve ensemble cannot outperform the sum of its parts. The superior performance of the naïve ensemble model to the baseline comes from the fact that each specialist focuses on a smaller subset of the original prob-

[1] Available for download at `http://www.openslr.org/12/`.
[2] Available for download at `http://www.openslr.org/17/`.
[3] From the `librispeech/train-clean-100` folder.
[4] From the `musan/noise/free-sound` folder.
[5] From the `librispeech/test-clean` folder.
[6] From the `musan/noise/sound-bible` folder.

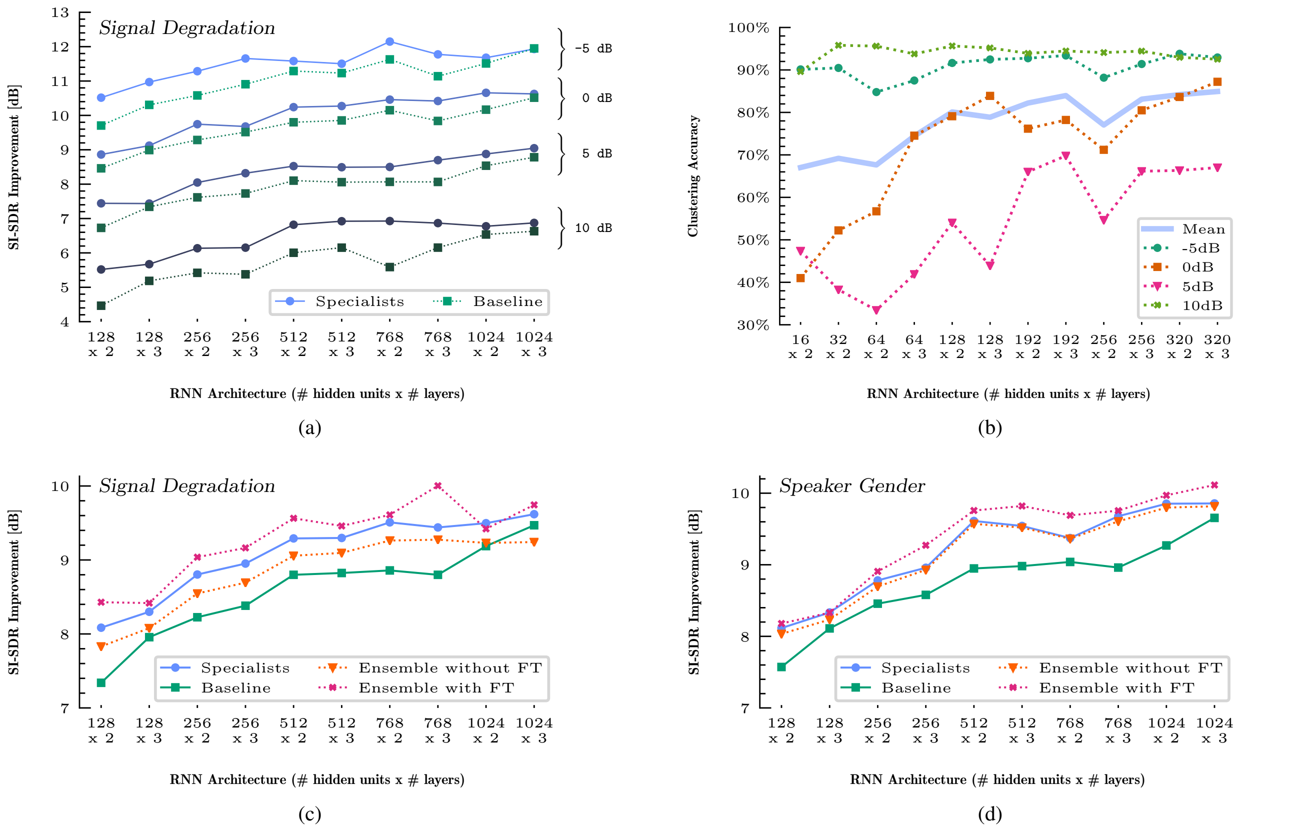


Figure 2: *Results from the signal degradation and speaker gender experiments. The LSTM component of the specialist network increases in computational complexity going across the x-axis on all subplots.*

lem with the same model capacity. In this hypothetical context where the best-suited specialist is always selected, an ensemble of smaller specialist networks will consistently outperform the baseline generalists.

Therefore, the gating network's classification accuracy matters. As shown in Figure 2b, *signal degradation*-based gating networks with a smaller RNN architecture are only able to distinguish the extrema mixture levels with high confidence. Increasing the number of hidden units and layers brings up the classification accuracy of the non-extrema mixture levels (i.e. 0dB and +5dB mixture SNR). Based on these results, we chose the 128×2 gating network architecture to be used for the subsequent ensemble experiments, as it adequately clusters test mixtures (with $\approx 80\%$ accuracy on average) while only incurring a small computational overhead.

Figure 2c compares the averaged denoising performance of the individual specialists, the baseline, and the ensemble models (with and without fine-tuning) across all four mixture SNR cases. We can see that the naïve ensemble improves upon the baseline with a significant margin, but cannot pass the theoretical upper bound set by the oracle choice of specialist. Still, the naïve ensemble model can compete as an efficient inference model with the high-complexity baseline model of size 1024×2 with a simpler architectural choice, 512×2.

Figure 2c also shows that the fine-tuning step greatly improves our ensemble model, surpassing the oracle specialist upper bound. This suggests that through fine-tuning, the specialists learn to compensate for imperfect classification results from the gating module. We can see that a fine-tuned ensemble with a smaller specialist RNN architecture, 512×2, outperforms the most complex baseline model of size 1024×3. This is a significant amount of computation reduced during the test time, even considering the overhead cost of the 128×2 gating network.

A similar trend is present in the *speaker gender* experiment, summarized in Figure 2d. Since this setup consists of only two specialists, the gating network's job is an easier binary classification. A 16×2 RNN architecture sufficiently classifies speaker gender with 90% classification accuracy. Using that, the naïve ensemble achieves near-optimal performance, reaching the upper bound in nearly every architecture. The fine-tuning process lifts the performance even further.

5. Conclusion

This work shows that speech denoising neural networks can benefit from the MLE design philosophy, boosting performance while reducing arithmetic complexity. Our specialist networks are trained on specific partitions of a large noisy speech corpus across two latent spaces: *signal degradation* and *speaker gender*. Despite the small overhead cost of a gating network, a naïve ensemble network is shown to match the performance of generalist denoising networks with fewer parameters i.e. fewer inference-time calculations. Furthermore, fine-tuning the ensemble with the inclusion of a sparsity parameter helps the model exceed the theoretical upper bound of the oracle specialist. Denoised speech examples and source code for this project are available online at `http://saige.sice.indiana.edu/research-projects/sparse-mle`.

6. References

[1] N. Wiener, *Extrapolation, interpolation, and smoothing of stationary time series*. MIT Press, 1964.

[2] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, pp. 113–120, 1979.

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

[4] Y. Wang and D. L. Wang, "Towards scaling up classification-based speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1381–1390, July 2013.

[5] 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.

[6] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr," in *Proceedings of the International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, Aug. 2015.

[7] L. Sun, J. Du, L.-R. Dai, and C.-H. Lee, "Multiple-target deep learning for LSTM-RNN based speech enhancement," in *2017 Hands-free Speech Communications and Microphone Arrays (HSCMA)*. IEEE, 2017, pp. 136–140.

[8] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Mar. 2016.

[9] Y. Luo, Z. Chen, J. R. Hershey, J. Le Roux, and N. Mesgarani, "Deep clustering and conventional networks for music separation: Stronger together," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, March 2017, pp. 61–65.

[10] E. Tzinis, S. Venkataramani, and P. Smaragdis, "Unsupervised deep clustering for source separation: Direct learning from mixtures using spatial information," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019, pp. 81–85.

[11] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016.

[12] H. Li, A. Kadav, I. Durdanovic, H. Samet, and H. P. Graf, "Pruning Filters for Efficient ConvNets," *arXiv preprint arXiv:1608.08710*, 2016.

[13] J.-H. Luo, J. Wu, and W. Lin, "Thinet: A filter level pruning method for deep neural network compression," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2017, pp. 5058–5066.

[14] D. Soudry, I. Hubara, and R. Meir, "Expectation backpropagation: Parameter-free training of multilayer neural networks with continuous or discrete weights," in *Advances in Neural Information Processing Systems (NIPS)*, 2014, pp. 963–971.

[15] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, "Xnor-net: Imagenet classification using binary convolutional neural networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016, pp. 525–542.

[16] M. Kim and P. Smaragdis, "Bitwise neural networks," in *International Conference on Machine Learning (ICML) Workshop on Resource-Efficient Machine Learning*, Jul 2015.

[17] ——, "Bitwise neural networks for efficient single-channel source separation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2018.

[18] S. Kim, M. Maity, and M. Kim, "Incremental binarization on recurrent neural networks for single-channel source separation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019.

[19] R. A. Jacobs and M. I. Jordan, "A competitive modular connectionist architecture," in *Advances in Neural Information Processing Systems (NIPS)*, 1990, pp. 767–773.

[20] X.-L. Zhang and D. Wang, "A deep ensemble learning method for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 5, pp. 967–977, 2016.

[21] S. Chazan, J. Goldberger, and S. Gannot, "Deep recurrent mixture of experts for speech enhancement," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2017, pp. 359–363.

[22] M. Kim, "Collaborative deep learning for speech enhancement: A run-time model selection method using autoencoders," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2017.

[23] A. Liu, P. Seetharaman, and B. Pardo, "Model selection for deep audio source separation via clustering analysis," *arXiv preprint arXiv:1910.12626*, 2019.

[24] M. Kolbæk, Z. H. Tan, and J. Jensen, "Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 153–167, Jan 2017.

[25] E. Vincent, C. Fevotte, and R. Gribonval, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[26] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR - half-baked or well done?" *arXiv preprint arXiv:1811.02508*, 2018.

[27] S. Srinivasan, N. Roman, and D. L. Wang, "Binary and ratio time-frequency masks for robust speech recognition," *Speech Communication*, vol. 48, no. 11, pp. 1486–1501, 2006.

[28] A. Narayanan and D. L. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2013, pp. 7092–7096.

[29] 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.

[30] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[31] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[32] E. Agustsson, F. Mentzer, M. Tschannen, L. Cavigelli, R. Timofte, L. Benini, and L. V. Gool, "Soft-to-hard vector quantization for end-to-end learning compressible representations," in *Advances in Neural Information Processing Systems (NIPS)*, 2017, pp. 1141–1151.

[33] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[34] D. Snyder, G. Chen, and D. Povey, "MUSAN: A Music, Speech, and Noise Corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[35] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.