# Speech Enhancement Based on Deep Mixture of Distinguishing Experts

Xupeng Jia and Dongmei Li
Department of Electronic Engineering, Tsinghua University, Beijing, China
E-mail: jxp12@mails.tsinghua.edu.cn Tel/Fax: +86-10-62782693

Abstract—In this work, we propose a new strategy for deep mixture of experts (DMoE) based speech enhancement. DMoE system is difficult to train due to the specific network structure and the necessity of carefully designed pre-training methods to guarantee good performance. We propose using distinguishing deep neural networks (DNNs) as experts, dealing with magnitude spectrogram and log-magnitude spectrogram respectively. The proposed method is compared with the state-of-art DMoE system utilizing hard expectation maximization (HEM) pre-training method. Speech enhancement experiments in 30 (5*6) noise and SNR conditions show the superiority of the proposed method over the baseline method. The average improvements obtained for matched conditions are 0.076 in perceptual evaluation of speech quality (PESQ), 1.824dB in segmental signal to noise ratio (segSNR) and 0.043 in short time objective intelligibility (STOI).

## I. Introduction

Speech enhancement is widely used in many real world applications such as speech communication, automatic speech recognition, hearing aids, etc.[1] The main task of speech enhancement is to suppress the noise in a noisy speech recording while keeping the distortion level as low as possible. Monaural speech enhancement is very challenging due to the nonstationary property of both the speech and noise signals, while lacking of other useful clues such as spatial information.

DNN is a powerful tool for complex classification and regression tasks. It uses a cascade of nonlinear processing units to model the relationship between inputs and outputs, and learns the model parameter from data. During the last decades, DNN has shown great advantage in speech enhancement area. DNN based monaural speech enhancement method can be divided into two categories according to the training target, i.e., mask based ones and mapping based ones. Ideal binary mask (IBM) [2] is the first training target used in DNN based speech enhancement methods. It uses time-frequency analysis techniques such as short time frequency transform (STFT) to get the spectrogram of the noisy signal. A mask, 0 or 1 for each time-frequency unit in IBM, is generated according to the local signal-to-noise ratio (SNR) of the unit. The mask is then applied to the spectrogram and an iSTFT step is carried out to get the enhanced signal. Following IBM, some other masking strategies are proposed, for example, ideal ratio mask (IRM) [3], spectral magnitude mask (SMM) [4], complex ideal ratio mask (cIRM) [5], phase-sensitive mask (PSM) [6], etc. In the mapping based category, instead of predicting a mask, the DNNs are used to map the noisy speech features to the enhanced ones directly. The most commonly used features in this category of methods is the log-magnitude spectrogram [7][8].

Despite of all the differences among these methods, they all use one single DNN as the predicting model. Although the DNNs are becoming more and more powerful, it is believed that it would be easier for two or more DNNs working together to learn the varied patterns in speech enhancement than a single DNN. Several algorithms based on two or more DNNs have been explored during the last several years. Wang [9] proposed a fully discriminative approach. In the training phase, forty DNNS were trained, one for each phoneme, to predict the IRM. In the test phase, a robust automatic speech recognition (ASR) system was used to detect the phoneme label, and the mask generated by the corresponding DNN was applied to enhance the speech. Chazan [10] proposed a different method. He used thirty-nine phoneme-specific DNNs and one phoneme-classification DNN. A speech presence probability (SPP) based soft spectral subtraction algorithm was used to suppress the noise. The phoneme-specific DNNs were pre-trained according to phoneme labels, one for each label. A joint training step was then carried out to optimize the entire set of DNNs. These methods use phoneme information to train or pre-train the DNNs. However, phoneme-labeled database is not always tractable and a large amount of DNNs consume a lot of calculation and storage resources.

Chazan [11] proposed that instead of phoneme information based pre-training, a modified expectation maximization (EM) algorithm can also be used to train the DMoE system to avoid the convergence problem. A weighted back propagation equation was proposed, such that one sample would have different effect on each expert DNN. This method can be interpreted as a soft EM training strategy, while the weights have an inverse proportional relationship with the prediction error of each expert DNN. Two expert DNNs were used in the method, and simulation results showed that more experts were not helpful to the enhancement performance. Following Chazan's work, Karjol [12] proposed a hard expectation maximization pre-training strategy, and got better performance. All these DMoE

systems mentioned above use homogeneous expert, which means that the expert DNNs have same structure, and the inputs and outputs of them are similar with respect to their physical meaning and distribution. In this work, we propose a DMoE system with distinguishing experts. We name it as deep mixture of distinguishing experts (DMoDE). DMoDE system has two main advantages. Firstly, it is easier to train. DMoDE can be trained with normal back-propagation algorithm, while traditional DMoE systems need carefully designed training strategies to avoid the convergence problem. Secondly, it has better performance. Each expert has its own strength, and they can be complementary with each other.

It should be mentioned here that although our original intention of proposing the DMoDE system is to improve the DMoE system on both the training complexity and the speech enhancement performance, we find out later that the DMoDE system has some similarity with the ensemble methods. The ensemble methods, or the fusion methods, try to propose a framework of combining different speech enhancement methods to get better performance. Roux [13] proposed an ensemble learning method. Different speech enhancement methods based on binary masking can be combined by either a simple averaging strategy, or a learned classifier such as support vector machine or decision trees. Zhang [14] proposed a DNN based ensemble method. Several DNNs with different context window length are put together to form a module, and several modules are stacked together to get the final enhancement result. The work focusses on the information contained in the context windows with different length. The DNNs in the upper module take the outputs of the DNNs in the lower module as inputs, and there is only one DNN in the last module so that no extra strategy is needed to combine the outputs of different DNNs. Jaureguiberry [15] proposed a fusion method. Instead of the stacking framework used in the above mentioned methods, the final result is calculated as a weighted sum of the sub-methods, such as a NMF-based method and a DNN-base method. The weight factors are calculated by a single-layer network, which is claimed to have better performance than deeper networks. The network takes the original signal together with the outputs of the sub-methods with a certain context window as input. In DMoDE, we focus on taking the advantage of different DNN based methods. The weighted sum strategy is applied while the weight calculating network, which is named as gating DNN, takes only the original signal as input, and the context window length is the same with the expert DNNs, so that the whole system can be jointly optimized.

The rest of the paper is organized as follows. In section 2, we briefly introduce the HEM strategy, and describe our DMoDE system in detail. Then in section 3, we present some experiment results and discussion about DMoDE. Finally, we draw the conclusion and summarize the paper in section 4.

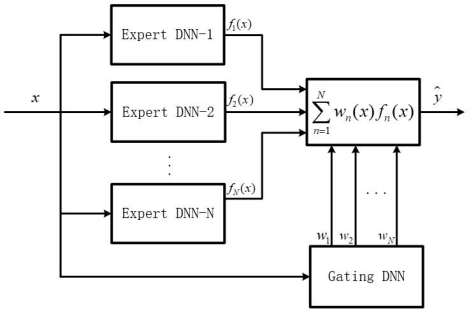


Fig. 1. Schematic diagram of DMoE.

## II. Deep mixture of distinguishing experts

Deep mixture of experts (DMoE) is a special case of mixture of experts where the experts employed are DNNs. The structure of DMoE is shown in Figure 1. $N$ expert DNNs are trained to solve the problem separately, and a gating DNN, which is actually a classifier, combines the outputs of the expert DNNs by a weighted-sum strategy. The output of such a system $\hat{y}$ is given by,

$$\hat{y} = \sum_{n=1}^{N} w_n(x) f_n(x) \tag{1}$$

where $N$ is the total number of the expert DNNs, $x$ is the input of the system, $f_n(x)$ is the output of the $n^{th}$ expert DNN, and $w_n(x)$ is the corresponding weight generated by the gating DNN. In the scope of speech enhancement, the input $x$ is a set of features of noisy speech and the target $y$ can be the corresponding features of clean speech, or the ideal masks. To train the system, we expect to reduce the error ($\epsilon$) between $\hat{y}$ and $y$. The objective function to be minimized can be written as,

$$\begin{aligned} \epsilon &= \frac{1}{M} \sum_{m=1}^{M} d(y_m, \hat{y}_m) \\ &= \frac{1}{M} \sum_{m=1}^{M} d(y_m, \sum_{n=1}^{N} w_n(x_m) f_n(x_m)) \end{aligned} \tag{2}$$

where $M$ is the total number of training samples, $x_m$ and $y_m$ are input features and output corresponding to the $m^{th}$ sample, and $d()$ is an error metric like mean square error or mean absolute error. However, it has been noted that training such a model with random initialization may lead to convergence problem [10]. Several algorithms have been proposed to solve the convergence problem, among which the hard EM pre-training strategy [12] has been proved the most successful. In the following section we will explain the hard EM pre-training strategy and show its defect which restricts it performance.

### A. Hard EM pre-training

Karjol proposed hard EM pre-training strategy to solve the convergence problem of DMoE system. To pre-train

the expert DNNs, a weighted loss function is used,

$$\epsilon_n = \frac{1}{M}\sum_{m=1}^{M} p_{n,m} d(y_m, f_n(x_m)), n = 1, 2, ..., N \quad (3)$$

where $M$ is the total number of the training samples, $N$ is the number of expert DNNs, $\epsilon_n$ is the loss of the $n^{th}$ expert DNN, $y_m$ is the target value of the $m^{th}$ sample, $f_n(x_m)$ is the output of the $n^{th}$ expert DNN corresponding to the $m^{th}$ sample, and $d()$ is the error metric. $p_{n,m}$ is the weight of the $m^{th}$ sample for the $n^{th}$ expert DNN.

$$p_{n,m} = \begin{cases} 1, if\ \ n = arg\min_n d(y_m, f_n(x_m)) \\ 0, otherwise \end{cases} \quad (4)$$

With $p_{n,m}$ determined, the expert DNNs can be optimized by finding the minimum of the objective function via back-propagation algorithm,

$$\theta_n = arg\min_{\theta} \epsilon_n, n = 1, 2, \cdots, N \quad (5)$$

It can be seen that finding the optimal $p$ and $\theta$ is a chicken-and-egg problem, because they depend on each other. Karjol proposed to use generalized EM method to solve this problem. The procedure is as follows:

1) Initialize the parameters $\theta^0$
2) Calculate $p^t$ based on $\theta^t$ according to Equation (4)
3) Train each expert DNN with $p$ fixed to $p^t$ and get $\theta^{t+1}$
4) Repeat step 2 and step 3 until Equation (5) converges

The DMoE system trained with hard EM pre-training strategy outperforms the single model based system and Chazan's soft EM based DMoE system. However, it still has two main disadvantages. The first one is that the EM procedure is very time-consuming. According to the algorithm, every time $p$ changes, the model needs to be trained thoroughly. Although the number of training epochs can be limited to a certain number, the whole iteration still costs much more time than training a single DNN-based system. What's more, as far as we know, no proof guarantees the convergence of this iteration because optimization of the DNN itself is a complex procedure. The second disadvantage is that it cannot detect or avoid the over-fit problem, which will be shown in section 3.

### B. Proposed method

As discussed in previous sections, all of the pre-training strategies based on hard EM algorithm, soft EM algorithm and phoneme information can train a DMoE system successfully. We think that the reason why these pre-training strategies can help to avoid the convergence problem is that the pre-training stage sets the expert DNNs to different status. The pre-trained expert DNNs have significantly different distributions of outputs corresponding to the same inputs. Based on this difference, the gating DNN can be trained easily and the whole system can converge to a certain point. According to this analysis, we propose a novel way to solve the convergence problem. Instead of using the same expert DNNs, we propose to using distinguishing expert DNNs and we name it as deep mixture of distinguishing experts (DMoDE).

Log-magnitude (or log-power) spectra has been widely used in mapping-based speech enhancement system. However, magnitude spectra can also be used to replace the log-magnitude spectra and has a different error distribution. Assuming that two DNNs performing the log-magnitude spectra mapping and magnitude spectra mapping respectively are trained equally well, which means that the outputs of the two DNNs have the same additive error $e$, the predicting error of the magnitude spectra mapping system is:

$$E_{mag} = e \quad (6)$$

and the predicting error of the log-magnitude spectra mapping system is:

$$\begin{aligned} E_{log} &= exp(log(s) + e) - s \\ &= s * (exp(e) - 1) \end{aligned} \quad (7)$$

where $s$ is the magnitude of the corresponding time-frequency unit. Noting that $s > 0$, $E_{log}$ and $E_{mag}$ have the same sign. We have:

$$\begin{aligned} |E_{log}| < |E_{mag}| \Leftrightarrow\ & \frac{e}{s * (exp(e) - 1)} > 1 \\ \Leftrightarrow\ & s < \frac{e}{exp(e) - 1} \end{aligned} \quad (8)$$

To make the condition more clearly, we draw the last inequation in figure 2. The solid line stands for the inequation in Equation (8). The dashed line stands for $s = -e$. This is because that in the magnitude spectra mapping system, the output $\hat{y}$ of the DNN cannot be negative.

$$\hat{y} = s + e \geq 0 \quad (9)$$

The shadow area in figure 2 shows the condition in equation (8). In [4], Wang proposes a similar deduction on the error relationship between FFT-mask and log-magnitude spectra mapping. Wang draws the conclusion that because in practice, $s \gg 1$ when speech is present, FFT-mask is more accurate than log-magnitude spectra mapping. However, high prediction accuracy of the large magnitude time-frequency points only is not enough for a good speech enhancement system. The time-frequency points of which the magnitudes are small should be considered as well. Besides, we normalize the magnitudes before calculating the log values. This results in more time-frequency points whose magnitudes are smaller than 1 and are more likely locating in the shadow area in Figure 2.

Figure 3 shows the simulation results of the relative relationship between $E_{log}$ and $E_{mag}$. The two DNNs are trained with the same setup which is described in section 3. The error results are calculated under the condition of 0dB babble noise with 100 sentences from TIMIT dataset. The bin-wise average absolute errors are divided by the errors of the magnitude spectra mapping model to get the

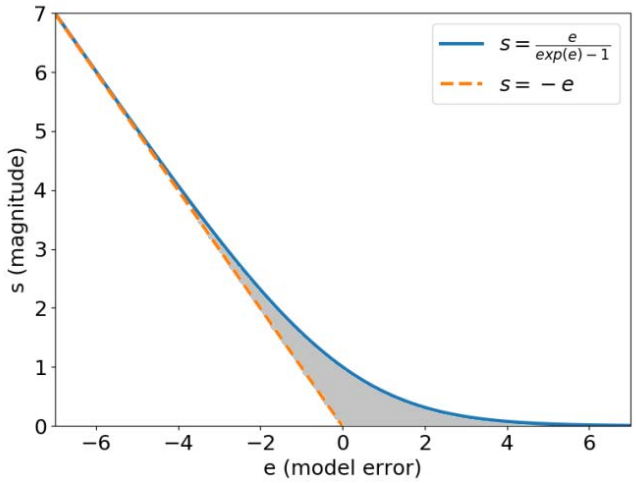


Fig. 2. Error relationship curve between $E_{log}$ and $E_{mag}$

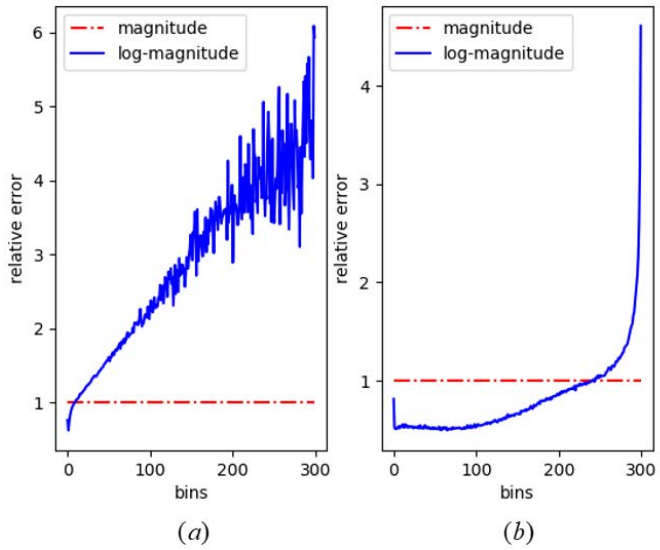


Fig. 3. simulation results of the relative relationship between $E_{log}$ and $E_{mag}$

relative errors. The dash-dot line which has a constant value of one stands for the magnitude spectra mapping model and the solid line stands for the log-magnitude spectra mapping model. The bins in Figure 3.a have the equally divided range of magnitude. It proves that only when the magnitudes are small, $E_{log}$ is supposed to be smaller than $E_{mag}$. The bins in Figure 3.b have the same number of time-frequency points. The points are sorted according to the magnitudes. It can be seen that more than seventy percent of the time-frequency points have smaller error with the log-magnitude spectra mapping system.

The proposed DMoDE system is as follows. It has the same global structure with DMoE system, which is depicted in Figure 1. It includes $N = 2$ expert DNNs. The expert DNN 1 is trained for magnitude spectra mapping and the expert DNN 2 is trained for log-magnitude spectra mapping. The input $x$ and the output $\hat{y}$ are the magnitude spectra of noisy and enhanced speech respectively. Note that $log()$ and $exp()$ are implemented as the input and output layer in expert DNN 2. In this way, the inputs and outputs of the two expert DNNs are consistent and the whole system can be optimized jointly. To train the DMoDE system, we first pre-train the two expert DNNs as two independent single DNN-based speech enhancement model. After that we fix the parameters of the two expert DNNs and train the gating DNN. At last we jointly train the whole system for a few epochs. The experiment results and discussions are presented in the next section.

## III. Experiment results

### A. experiment setup

We use TIMIT [16] and NoiseX-92 [17] databases for clean speech and noise recordings respectively. TIMIT database is divided into train and test categories containing 4620 and 1680 speech utterances respectively and the sampling rate is 16kHz. The noise recordings includes 15 types of noise, and the sampling rate is 19.92kHz. To generate the noisy speech signals, we down sample the noise recordings to the sampling rate of 16kHz. Five types of noise recordings are used for training, which are babble, white, factory1, buccaneer1 and destroyerops. Utterances chosen from the train category of TIMIT database are mixed with these five noise recordings at six different SNR levels, -5dB, 0dB, 5dB, 10dB, 15dB and 20dB. Thus, we have 30 configurations (5 noise types and 6 SNR levels). We randomly choose 200 utterances for each configuration. These 6000 utterances are divided into train and validation set at a ratio of 8:2. To generate the test set, we randomly choose 60 utterances from the test category of TIMIT database for each configuration. Extra four types of noise, namely f16, leopard, machinegun and pink, are used to test the generalization performance in unmatched situation.

The DMoDE system includes three DNNs, and each DNN has three hidden layers with 2048 units for each layer. ReLU is used as the activation function. The frame length is set to 512 points and the overlap length is 256 points. The input of the system consists of seven frames, with the target frame in the middle of the segment. The output of the system includes the target frame only. Except for the proposed DMoDE system, four other methods are implemented for comparison, which are the log-magnitude spectra mapping system, the magnitude spectra mapping system, the original DMoE system [12], and the magnitude mapping based DMoE system, donated as S-log, S-mag, D-log and D-mag respectively. The DNNs in these systems have the same setup with DMoDE system.

We compare the performance of the different systems in terms of PESQ, STOI and segSNR. PESQ measures the perceptual quality of speech, while STOI measures the intelligibility. SegSNR provides information about the

average reconstruction error across frames with respect to the clean speech.

B. results and discussion

TABLE I
Experiment results in matched situation

| | PESQ | | | | | |
|---|---|---|---|---|---|---|
| | -5dB | 0dB | 5dB | 10dB | 15dB | 20dB |
| noisy | 1.275 | 1.609 | 1.977 | 2.355 | 2.715 | 3.071 |
| S-log | 1.690 | 2.172 | 2.574 | 2.902 | 3.137 | 3.291 |
| S-mag | 1.917 | 2.340 | 2.682 | 2.936 | 3.109 | 3.228 |
| D-log | 1.765 | 2.244 | 2.640 | 2.966 | 3.207 | 3.360 |
| D-mag | 1.942 | 2.356 | 2.683 | 2.951 | 3.147 | 3.298 |
| ours | 1.932 | 2.359 | 2.711 | 3.010 | 3.235 | 3.389 |
| | STOI | | | | | |
| | -5dB | 0dB | 5dB | 10dB | 15dB | 20dB |
| noisy | 0.551 | 0.668 | 0.779 | 0.869 | 0.932 | 0.969 |
| S-log | 0.627 | 0.745 | 0.823 | 0.875 | 0.908 | 0.933 |
| S-mag | 0.692 | 0.802 | 0.874 | 0.920 | 0.948 | 0.965 |
| D-log | 0.641 | 0.756 | 0.835 | 0.885 | 0.917 | 0.941 |
| D-mag | 0.696 | 0.807 | 0.878 | 0.924 | 0.953 | 0.969 |
| ours | 0.695 | 0.807 | 0.879 | 0.926 | 0.954 | 0.971 |
| | SegSNR | | | | | |
| | -5dB | 0dB | 5dB | 10dB | 15dB | 20dB |
| noisy | -7.04 | -4.33 | -1.01 | 2.75 | 6.76 | 10.99 |
| S-log | 0.24 | 1.89 | 3.63 | 5.30 | 6.75 | 8.14 |
| S-mag | 0.32 | 2.53 | 4.96 | 7.42 | 9.50 | 11.11 |
| D-log | 0.42 | 2.08 | 3.87 | 5.52 | 7.04 | 8.55 |
| D-mag | 0.17 | 2.40 | 4.86 | 7.45 | 9.76 | 11.59 |
| ours | 0.50 | 2.71 | 5.21 | 7.80 | 10.13 | 12.07 |

TABLE II
Experiment results in unmatched situation

| | PESQ | | | | | |
|---|---|---|---|---|---|---|
| | -5dB | 0dB | 5dB | 10dB | 15dB | 20dB |
| noisy | 1.571 | 1.938 | 2.292 | 2.634 | 2.967 | 3.277 |
| S-log | 1.855 | 2.291 | 2.660 | 2.952 | 3.188 | 3.377 |
| S-mag | 1.836 | 2.254 | 2.606 | 2.889 | 3.095 | 3.250 |
| D-log | 1.874 | 2.340 | 2.709 | 3.005 | 3.248 | 3.440 |
| D-mag | 1.694 | 1.921 | 2.021 | 2.061 | 2.100 | 2.132 |
| ours | 1.990 | 2.411 | 2.767 | 3.051 | 3.276 | 3.467 |
| | STOI | | | | | |
| | -5dB | 0dB | 5dB | 10dB | 15dB | 20dB |
| noisy | 0.656 | 0.739 | 0.817 | 0.884 | 0.931 | 0.962 |
| S-log | 0.662 | 0.756 | 0.825 | 0.872 | 0.901 | 0.923 |
| S-mag | 0.658 | 0.779 | 0.860 | 0.910 | 0.939 | 0.957 |
| D-log | 0.673 | 0.768 | 0.835 | 0.881 | 0.911 | 0.932 |
| D-mag | 0.574 | 0.670 | 0.730 | 0.764 | 0.787 | 0.802 |
| ours | 0.698 | 0.802 | 0.873 | 0.919 | 0.947 | 0.965 |
| | SegSNR | | | | | |
| | -5dB | 0dB | 5dB | 10dB | 15dB | 20dB |
| noisy | -4.88 | -2.04 | 1.27 | 4.94 | 8.78 | 12.57 |
| S-log | 0.41 | 1.76 | 3.20 | 4.61 | 6.04 | 7.21 |
| S-mag | 0.23 | 2.49 | 4.75 | 6.93 | 9.01 | 10.61 |
| D-log | 0.81 | 1.99 | 3.42 | 4.90 | 6.38 | 7.71 |
| D-mag | -1.38 | 0.24 | 1.46 | 2.22 | 2.74 | 3.06 |
| ours | 0.83 | 3.11 | 5.37 | 7.55 | 9.72 | 11.58 |

Table 1 shows the experiment results in matched situation and table 2 shows the results in unmatched situation. It can be seen that our DMoDE system outperforms the other systems in most of the configurations. Roughly speaking, the DMoDE system is better than the DMoE systems, and the DMoE systems are better than the corresponding single DNN based systems. The performance of the magnitude spectra mapping based systems is better than that of the log-magnitude spectra mapping based ones in terms of STOI and SegSNR. However, in unmatched situation, log-based systems have better performance in terms of PESQ. Note that when SNR is high, for example, 20dB, the other systems tend to degrade the STOI or SegSNR of the speech, while the proposed DMoDE system are more likely to have a positive effect on the speech.

Another interesting thing is that the magnitude based DMoE system performs badly in unmatch situation but performs quite well in matched situation. We think this is because of the overfitting problem. If the expert DNNs are overfitted after the pre-training stage, the joint-training stage will make little modification to them, which will result in a overfitted system. As far as we know, there is no strategy in the DMoE system to avoid overfitting in the pre-training stage, and the complex pre-training strategy makes it difficult to test whether the expert DNNs are overfitted. In the proposed DMoDE system, this is avoided by early stopping strategy with a validation set in the pre-training stage.

Compared with the HEM pre-trained DMoE system, the DMoDE system has an average improvement of 0.076 in pesq, 0.043 in stoi and 1.82dB in SegSNR in matched situation, and 0.058 in pesq, 0.034 in stoi and 2.16dB in SegSNR.

## IV. Conclusion

In this paper, we proposed a DMoDE system which uses a log-magnitude spectra mapping DNN and a magnitude spectra mapping DNN as the experts. Such system solves the convergence problem in a novel way, and takes the advantages of the two different expert DNNs to improve the performance. The proposed system outperforms single DNN systems and DMoE systems both in matched and unmatched situation. The expert DNNs can be replaced by more competitive DNNs such as LSTM, CRNN, etc., which is a part of our future works.

## References

[1] P.C. Loizou. Speech Enhancement: Theory and Practice. CRC Press, USA, 2013.

[2] DeLiang Wang. On ideal binary mask as the computational goal of auditory scene analysis. In Speech separation by humans and machines, pages 181–197. Springer, 2005.

[3] Christopher Hummersone, Toby Stokes, and Tim Brookes. On the ideal ratio mask as the goal of computational auditory scene analysis. In Blind source separation, pages 349–368. Springer, 2014.

[4] Yuxuan Wang, Arun Narayanan, and DeLiang Wang. On training targets for supervised speech separation. IEEE/ACM transactions on audio, speech, and language processing, 22(12):1849–1858, 2014.

[5] Donald S Williamson, Yuxuan Wang, and DeLiang Wang. Complex ratio masking for monaural speech separation. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), 24(3):483–492, 2016.

[6] Hakan Erdogan, John R Hershey, Shinji Watanabe, and Jonathan Le Roux. Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 708–712. IEEE, 2015.
[7] Li Deng, Michael L Seltzer, Dong Yu, Alex Acero, Abdelrahman Mohamed, and Geoff Hinton. Binary coding of speech spectrograms using a deep auto-encoder. In Eleventh Annual Conference of the International Speech Communication Association, 2010.
[8] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee. An experimental study on speech enhancement based on deep neural networks. IEEE Signal processing letters, 21(1):65–68, 2014.
[9] Zhong-Qiu Wang, Yan Zhao, and DeLiang Wang. Phoneme-specific speech separation. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 146–150. IEEE, 2016.
[10] Shlomo E Chazan, Sharon Gannot, and Jacob Goldberger. A phoneme-based pre-training approach for deep neural network with application to speech enhancement. In 2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC), pages 1–5. IEEE, 2016.
[11] Shlomo E Chazan, Jacob Goldberger, and Sharon Gannot. Speech enhancement using a deep mixture of experts. arXiv preprint arXiv:1703.09302, 2017.
[12] Pavan Karjol and Prasanta Kumar Ghosh. Speech enhancement using deep mixture of experts based on hard expectation maximization. Proc. Interspeech 2018, pages 3254–3258, 2018.
[13] Jonathan Le Roux, Shinji Watanabe, and John R. Hershey. Ensemble learning for speech enhancement. In Applications of Signal Processing to Audio Acoustics, 2013.
[14] X. L. Zhang and D. Wang. A deep ensemble learning method for monaural speech separation. IEEE/ACM Transactions on Audio Speech Language Processing, 24(5):967–977, 2016.
[15] Xabier Jaureguiberry, Emmanuel Vincent, and Gaël Richard. Fusion methods for speech enhancement and audio source separation. IEEE/ACM Transactions on Audio Speech Language Processing, 24(7):1266–1279, 2017.
[16] Garofolo, John S., and et al. TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1, Web Download. Linguistic Data Consortium, Philadelphia, 1993.
[17] A. Varga and H. J. M. Steeneken. Assessment for automatic speech recognition: Ii.noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems. Speech Commun., 12(3):247–251, 1993.