DEEP RECURRENT MIXTURE OF EXPERTS FOR SPEECH ENHANCEMENT

Shlomo E. Chazan, Jacob Goldberger and Sharon Gannot
Faculty of Engineering, Bar-Ilan University, Israel
{Shlomi.Chazan, Jacob.Goldberger, Sharon.Gannot}@biu.ac.il

ABSTRACT
deep neural networks (DNNs) have recently became a viable methodology for single microphone speech enhancement. The most common approach, is to feed the noisy speech features into a fully-connected DNN to either directly enhance the speech signal or to infer a mask which can be used for the speech enhancement. In this case, one network has to deal with the large variability of the speech signal. Most approaches also discard the speech continuity. In this paper, we propose a deep recurrent mixture of experts (DRMoE) architecture that addresses these two issues. In order to reduce the large speech variability, we split the network into a mixture of networks (denoted experts), each of which specializes in a specific and simpler task and a gating network. The time-continuity of the speech signal is taken into account by implementing the experts and the gating network as a recurrent neural network (RNN). Experimental study shows that the proposed algorithm produces higher objective measurements scores compared to both a single RNN and a deep mixture of experts (DMoE) architectures.

Index Terms— recurrent neural network; long short-term memory; speech presence probability

1. INTRODUCTION
Single microphone speech enhancement has been an active research field in the last four decades. A plethora of algorithms utilizing speech and noise characteristics can be found in the literature [1]. In recent years the field has witnessed a major paradigm shift from model-based to data-driven approaches, most notably DNN-based methods. DNN was used to directly filter the noisy speech features in [2, 3]. Although demonstrating good noise reduction capabilities, these methods still suffer from relatively high speech distortion. DNN was also trained to infer a mask which is used to enhance the speech [4, 5]. The main drawback with these approaches is the mismatch between train and test conditions. To circumvent this mismatch, it was proposed in e.g. [6], to use a very large noise dataset to train the DNN. Nevertheless, applying the algorithm in unknown noisy environment is still an open challenge.

More recently, a combination between model-based and data-driven approaches was proposed in [7] to mitigate the mismatch problem. Speech enhancement methods that are entirely based on DNNs are presented in [8, 9]. These methods are based on the mixture of experts (MoE) approach that was introduced more than twenty years ago [10, 11]. MoE combines the decisions of several experts, each of which specializes in a different component of the input space. In [8, 9] each expert was responsible for a different speech component (e.g. phoneme). High noise reduction and low distortion are demonstrated even if trained with a small amount of single noise type training data. The time smoothness of the speech signal is simply addressed by using past and future context frames.

Recently, a related approach was presented in [12], where a series of DNNs were separately trained on different noise types to estimate the power spectral density (PSD) of the clean speech, and a deep auto-encoder (DAE) was used as a selector to choose the best expert. This approach showed good results. However, since only the selected expert is considered, the method does not generalize well to an unfamiliar noisy environment.

Speech signal exhibits time-continuity. This important speech property was exploited in e.g. [13–15] by applying the recursive Kalman filter. RNNs are also suitable structures for dealing with time-series. In the context of speech processing they were mainly applied in automatic speech recognition (ASR) systems, where speech continuity is a major issue. RNNs were also recently applied in speech enhancement task in order to preserve smoothness of the enhanced speech signal. In [16, 17], a time-frequency mask was estimated using an RNN-based architecture rather than the commonly used DNN-based architectures. It was shown, in the context of single-channel speech separation, that the recursive network is better adjusted to unfamiliar speakers.

In this paper, we present the deep recurrent mixture of experts (DRMoE) algorithm for speech enhancement. Similar to [9], the proposed algorithm comprises a set of expert networks and a gating network. Each of the experts is responsible for a specific speech component (in [8] each of the experts was responsible to one phoneme, but other partitions, e.g. speech and non-speech components, can be used as well). The gating network is responsible for softly selecting the active experts. In our previously proposed multi-network architectures, both the gating and the experts networks were implemented using DNNs. Here, we propose to incorporate the time smoothness by substituting the DNNs with RNNs. This will enable a smoother switching between experts (in the gating network) and a better utilization of the smoothness of each speech component by the experts, and hence better preservation of the long-term speech characteristics, e.g. related to its harmonic structure. Since the proposed algorithms combines the benefits of DMOE architectures and RNNs, it is dubbed deep recurrent mixture of experts (DRMoE).

2. PROBLEM FORMULATION AND PROBABILISTIC MODELING
Let $z(t)$ be the observed noisy speech at time $t$, where $x(t)$ and $y(t)$ denote the speech and noise signals, respectively.

$$z(t) = x(t) + y(t).$$  \hfill (1)

Let $Z(n, k)$ denote the short-time Fourier transform (STFT) of $z(t)$, with $n$ the frame index and $k = 0, \ldots, L - 1$ the frequency index. The frame length is set to $L$, and the overlap between successive frames is set to $3L/4$ samples. Denote the log-spectrum vector at
frame $n$ by $z(n)$ with its $k$-th frequency component defined by:

$$z_k(n) = \log |Z(n,k)|, \quad k = 0, \ldots, L/2.$$  
(2)

Similarly, define the respective log-spectrum vectors of the clean speech and the noise signal by $x(n)$ and $y(n)$. Following Nádas et al. [18], the noisy log-spectrum vector can be approximated by:

$$z(n) \approx \max(x(n), y(n))$$
(3)

such that the maximization is component-wise over the elements of $x(n)$ and $y(n)$. The maximization approximation was used for speech recognition [18] and speech enhancement [7, 8, 19].

Following this approximation, a binary-mask can be constructed for each frequency $k$ at each frame $n$,

$$b_k(n) = \begin{cases} 1 & x_k(n) > y_k(n) \\ 0 & x_k(n) \leq y_k(n) \end{cases}.$$  
(4)

We denote the binary-mask vector of all the frequencies at a given time frame as $b(n) = [b_0(n), \ldots, b_{L/2}(n)]$. Define the speech presence probability (SPP) $\rho_k(n) \in [0,1]$ as the conditional probability, given the noisy frames $z(1:n) = [z(1), \ldots, z(n)]$, that the $k$-th frequency component of the $n$-th noisy frame is dominated by speech.

$$\rho_k(n) = p(b_k(n) = 1 | z(1:n)).$$  
(5)

Given the SPP, the $k$-th bin of the log-spectrum of the clean speech $\hat{x}(k)$, is estimated using soft attenuation:

$$\hat{x}_k(n) = z_k(n) - (1 - \rho_k(n)) \cdot \beta$$  
(6)

where $\beta$ is the noise attenuation level (in the log domain). In our implementation we set $\beta$ to correspond to attenuation of 20 dB, which yielded high noise suppression while maintaining low speech distortion. Respectively, in vector form:

$$\hat{x}(n) = z(n) - (1 - \rho(n)) \cdot \beta$$  
(7)

where $1$ is a vector of ones with the same dimensions as $\rho(n)$, the vector concatenation of $\rho_k(n)$, $k = 0, 1, \ldots, L/2$.

The observed noisy phase is used for reconstructing the time-domain speech signal, similarly to most speech enhancement algorithms. In this work we aim at finding an accurate estimate of $\rho$, using a special purpose neural network architecture.

### 3. DRMoE for Speech Enhancement

In this section we present the proposed enhancement algorithm. We start by surveying the DMoE architecture [9]. We then proceed to the derivation of the new DRMoE algorithm.

#### 3.1. Deep Mixture of Experts

The DMoE architecture consists of DNN experts alongside with a gating DNN. The gist of this architecture is the multi-network structure. Since a speech signal exhibit large variability, splitting it to its basic building blocks may simplify the processing carried out by each expert, specializing in a specific speech component. Thus, each expert is responsible for a simpler task. The gating DNN behaves as the controller of the composite network, by softly selecting the proper expert(s). Under this framework, the $i$-th expert infers a task-specific SPP, denoted $\rho_i$, and the gating DNN infers the probability that the current speech frame belongs to each one of the experts, denoted $(p_1, ..., p_m)$, where $m$ is the number of experts. The composite SPP is finally given by:

$$\rho(n) = \sum_{i=1}^{m} p_i(n) \cdot \rho_i(n).$$  
(8)

Let $\theta_i$ be the parameter set of expert $i$, $i = 1, \ldots, m$, $\theta$, the parameter set of the gating network, and $\theta = \{\theta_1, \ldots, \theta_m, \theta_g\}$, the entire parameter set. The objective function $L(\theta)$ is set to be:

$$\sum_{n=1}^{N} \log p(b(n)|z(n)) = \sum_{n=1}^{N} \log \left( \sum_{i=1}^{m} p(I(n) = i | z(n), \theta_{gi}) \prod_{k=0}^{L/2} p(b_k(n)|z(n), I(n) = i, \theta_i) \right),$$  
(9)

where, $I(n)$ is the unobserved identity of the expert selected by the gating network at time $n$ and $N$ is the batch size. We have recently shown [9] that this approach performs better than the standard network architecture which can be viewed as a decision based on a single expert. In a DMoE architecture each time-frame is separately enhanced and speech time-continuity is not taken into account. We next present an extension of the DMoE framework where time continuity is explicitly modeled.

#### 3.2. Deep recurrent mixture of experts

In this work, we merge the architecture of the DMoE with the power of the long short-term memory (LSTM) architecture. Fig. 1a depicts the architecture of the proposed DRMoE. The architecture of the proposed model is constructed from a number of experts and a gating network that jointly infer the SPP $\rho$. All networks are implemented using RNNs rather than the DNNs as in Sec. 3.1. All RNNs were designed as described in [17]. Fig. 1b depicts the architecture of each RNN. A fully-connected layer was first used for feature extraction. LSTM layers are then utilized for temporal modeling, and another fully-connected layer is used at the output.

The motivation behind this architecture is threefold: 1) the use of the multiple experts splits the cumbersome enhancement task to smaller and simpler sub-tasks; 2) the RNN implementation of the gating network preserves the continuity of the decision-making; and 3) the RNN implementation of the experts preserves the smoothness of the speech spectral structure.

The target of the DRMoE is set to the binary mask vector $b$ described in (4). The objective function for the training is therefore given by the following log-likelihood, where $z(n)$ in (9) was substituted by the set of measurement $z(1:n)$:

$$L(\theta) = \sum_{n=1}^{N} \log p(b(n)|z(1:n)) = \sum_{n=1}^{N} \log \left( \sum_{i=1}^{m} p(I(n) = i | z(1:n), \theta_{gi}) \prod_{k=0}^{L/2} p(b_k(n)|z(1:n), I(n) = i, \theta_i) \right).$$  
(10)
The network is jointly trained to maximize the likelihood function (10). At the end of the training we obtain the parameters of the gating LSTM and the parameters of each expert LSTM.

3.3. Practical considerations

The TIMIT database [20] was used to train the DRMoE architecture. Each sentence was contaminated with a speech-like noise drawn from the NOISEX-92 database [21] with SNR = 10 dB. No other noise types and SNR levels were required. Additional noise types are not required, due to the proposed special architecture in which each expert specializes in a different structure of the speech.

The input to the experts is log-spectrum frame \( x(n) \) and the input to the gating network is the corresponding noisy mel-frequency cepstral coefficients (MFCC) vector of the current frame \( v(n) \), which is known for its better representation of sound classes [22]. Less speech distortion was noticed utilizing the MFCC features rather than the log-spectrum features. Note that the output dimension of the experts is identical to the dimensions of the input \( z(n) \), and the output dimension of the gating network is equal to the number of experts \( m \). Due to space limitations, the number of experts is set to \( m = 2 \). An analysis of the performance as a function of \( m \) can be found in [9].

To alleviate the mismatch between the training and the test conditions, each utterance was normalized both in the training and test phases, such that the sample-mean and sample-variance of the utterance were set to zero and one, respectively [23]. In order to circumvent over-fitting of the DRMoE to the training database, we first applied the cepstral mean and variance normalization (CMVN) procedure to the input, prior to the training and test phases [23]. Additionally, the dropout method [24] was utilized in each layer. Finally, the batch-normalization method was applied to accelerate the training phase in each layer [25]. The adaptive moment estimation (ADAM) optimizer [26] was used.

4. EXPERIMENTAL STUDY

4.1. Experimental setup

To test the proposed DRMoE algorithm we contaminated clean speech signals from the test set of the TIMIT database [20] with several types of noise from the NOISEX-92 database [21], namely Speech-like, Babble, and Factory. Note that only the Speech-like noise is known to the network from the training stage. The noise was added to the clean signal drawn from the test set of the TIMIT database (24-speaker core test set), with 4 levels of signal to noise ratio (SNR) at 0 dB, 5 dB, 10 dB and 15 dB chosen to represent various practical conditions. Sampling rate is equal to 16 Khz and the frame length was set to \( L = 512 \), with overlap of 75% between two successive frames. The size of the input to the experts, \( z \) thus equals to \( L/2 + 1 = 257 \), and the size of the input of the gating network, \( v \), is 39. To assess the performance of the proposed algorithm we have used the standard perceptual evaluation of speech quality (PESQ) measure, which is known to have a high correlation with subjective score [27].

4.2. Compared methods

The DRMoE was implemented with two experts and a gating network. All sub-networks comprised one fully-connected layer constructed from 512 rectified linear unit (ReLU) neurons and then three LSTM layers each with 512 neurons. The size of the output layer of all experts is equal to \( L/2 + 1 = 257 \), with a sigmoid activation function. The size of the output layer of the gating is equal to the number of experts \( m = 2 \), with a softmax activation function.

We compared the proposed method with three competing methods: 1) Single fully-connected DNN denoted FC-DNN; 2) a fully connected RNN denoted FC-LSTM; and 3) DMoE described in Sec. 3.1 with \( m = 2 \) experts. The FC-DNN architecture consists of 4 hidden layers each with 512 ReLU neurons and an output layer similar to the DRMoE. The FC-LSTM architecture is constructed as a single expert of the proposed DRMoE. The DMoE was identical to the method described in [9]. All methods were trained with the same database as described in Sec. 3.2.

4.3. Results

Fig. 2 depicts the PESQ results. First it is evident that the FC-LSTM outperforms the FC-DNN, confirming the importance of the speech continuity to the enhancement task. Additionally, the DMoE outperforms the FC-LSTM, probably since each expert specializes in a simpler task. Finally, the proposed DRMoE, which combines the advantages of the DMoE architecture and the LSTM capabilities of maintaining speech smoothness, obtains the best results in all test conditions.
Fig. 2: PESQ results for different noise types.

Fig. 3 depicts the clean, noisy, the SPP and the enhanced signal produced by the proposed DRMoE algorithm for one speech utterance drawn from the test set of the TIMIT database contaminated by Babble noise with SNR=5 dB. It is evident that the DRMoE manages to find an accurate SPP and is hence capable of significantly attenuating the noise level, while preserving the speech structure.

As explained in Sec. 3.2, the gating network is responsible for weighting the outputs of the experts. We further analyzed the gating decisions. Fig. 3c depicts the estimated SPP $\rho$ (upper panel), and the gating decisions related to the first expert, $p(I(n) = 1)$ (lower panel). The decisions of the gating network can be regarded as a soft version of a voice activity detector (VAD) system. When speech is active, the gating produces high probability for the first expert (and vice versa to the second expert). It can hence be deduced that the first expert is mainly responsible to non-speech frames while the second expert is mainly responsible for speech frames. Consequently, the task carried out by each expert is easier than the task carried out by a fully-connected system, which does not distinguish between the speech components.

5. CONCLUSIONS

In this paper, we proposed the deep recurrent mixture of experts (DRMoE) algorithm for speech enhancement. This architecture takes care of the large variability using expert networks, each of which specializing in a different speech component. Using RNNs for implementing the experts and the gating network preserves the continuity of the speech signal. Experiments have shown that this approach outperforms a fully-connected DNN or RNN as well as the DMoE architecture that does not take time axis into account.
6. REFERENCES


