Improvements of Acoustic Features for Speech Separation

Lujun Li[1,a*], Chuxiong Qin[1,b], Dan Qu[3,c]

[1]China National Digital Switching System Engineering & Technological R&D Center, China

[a]elegant_llj@126.com, [b]qinchuxiong911@163.com, [c]qudanqudan@sina.com

Keywords: speech separation; DNN; complementary features; RBM pre-training; dropout

Abstract. Acoustic features chosen from the input utterances play crucial roles in speech separation. In this paper, we propose a novel complementary feature approach that performs speech separation by combining five promising features, including Gammatone filterbank power spectra (GF) and multi-resolution cochleagram (MRCG) proposed recently especially for speech separation, as a super-vector fed into deep neural network (DNN). Additionally, based on the complementary features, we do experiments with two DNN training strategies, which are restricted Boltzmann machine (RBM) pre-training and dropout combined with Rectified Linear Units (ReLU), to optimize the performance of DNN. The experiment results, obtained in IEEE and TIMIT corpora using four different noises at low SNR levels of 0dB and -5dB, indicate that complementary features and RBM model improve all evaluation metrics. By contrast, dropout combined with ReLU system specializes in noise suppression and objective intelligibility more.

1 Introduction

Speech separation has been applied to various domains, such as the hands-free vehicle equipment, mobile communication, teleconferencing and hearing-aids, aiming to separating clean utterances from noisy mixtures. As the frontend of automatic speech recognition (ASR), it plays a significant role in improving the performance in the noisy environment and overcoming the mismatching between training and test conditions, which contributes a lot in declining the word error rate (WER).

Speech separation has been investigated intensively for several decades. Classical separation methods include subspace methods, non-negative matrix factorization (NMF), hidden Markov model (HMM). Signal-subspace methods are based on the sparseness of the speech signals, which divide the noisy utterances into noise subspace and speech-combined-with-noise subspace, and then remove the noise in the speech-combined-with-noise subspace in order to obtain the estimated clean speeches. In non-negative matrix factorization (NMF) [1,2], we model noisy utterances as non-negative source bases. NMF learns the speech and noise models respectively, searches the optimal way to combine and eventually factorizes to obtain the clean speech. HMM has been the most common acoustic model for past two decades, in which noisy and noise signals can be established models adequately and target speech can be obtained by relieving the noise from noisy signals. Even under the circumstance, where there are only noisy signals, speech enhancement can be achieved by filtering parts of the signal which are probably noises.

However, the aforementioned methods always have the problem of musical noise, especially in low signal-to-noise ratio (SNR) environments. Furthermore, they cannot suppress non-stationary noises effectively, such as the factory noise and the destroyer engine noise. Eventually, there are many unreasonable hypotheses in the traditional methods, which establish the ceiling for its performance. Therefore, how to break the born limitation of the model and improve the generalization ability has always been the core challenge.

Based on this insight, we have focused speech separation on DNN [3], whose deep structures are good at learning the nonlinear relationship between noisy utterances and clean speeches and perform better at low SNRs and non-stationary noises markedly. Besides, there are few musical noises remained in the utterances.

The input feature set is significant for DNN model, bearing direct effects on the performance. Besides, the training strategy influences the system as well.

In order to obtain rich complementary information, we use a series of frame-level features for DNN training. The common features include amplitude modulation spectrogram (AMS), relative spectral transformed perceptual linear prediction coefficients (RASTA-PLP), mel-frequency cepstral coefficients (MFCC). In this paper, we combine Gammatone filterbank power spectra (GF) and multi-resolution cochleagram (MRCG) with the traditional separation features. Both GF and MRCG are derived from cochleagrams. GF was first proposed in [4] and it outperformed conventional MFCC feature in noisy conditions. MRCG was specially designed for speech separation, which was evaluated with a series of exiting acoustic features systematically and achieved the best performance eventually in [5].

Based on complementary features, we explore two methods for classifier structure optimization. On the one hand, we introduce RBM pre-training instead of randomly initialized network to overcome the poor local minima. On the other hand, dropout is employed to avoid over-fitting with the aid of its sub-model averaging. Additionally, we substitute Rectified Linear Units for sigmoid activation functions to maximize the training effect and minimize the training time.

Experiments are implemented to test the performance of the proposed methods in both IEEE and TIMIT corpora using four different noises at challenging SNR mixtures. Results show that the proposed methods make contribution to performance improvements in terms of widely-used evaluation metrics.

This paper is organized as follow. Section 2 describes DNN framework. We introduce the complementary features and DNN training strategies in section 3 and 4 respectively. Section 5 covers experiment settings and the simulation results. Section 6 concludes this paper.

2 System Overview

Speech separation is a process that separates the target speech from the noisy observations. Simultaneously, the speech intelligibility and quality are improved. Since the mapping is explicitly data driven, supervised speech separation is formulated as a supervised learning problem.

The diagram of DNN based speech separation is showed in Fig. 1.

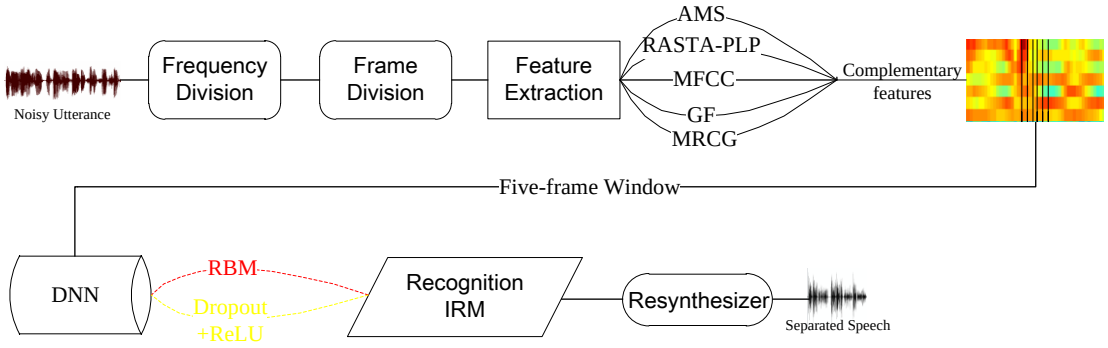

Figure 1 Diagram of DNN based speech separation

Firstly, the noisy utterances are passed through the predesigned 64-channel gammatone filters.

Secondly, frame-level features are extracted and concatenated with their corresponding delta portions, as the input of the DNN. In this study, to encode more useful information, a fixed set of complementary features [6] are used including MFCC, RASTA-PLP, AMS, GF and MRCG, which are illustrated respectively in the following section.

Thirdly, DNNs are used as the discriminative learning machine, which are amenable for speech separation [7,3]. The ideal ratio mask (IRM) is used as the training target, which is exemplified in encoding rich useful information, especially when the SNR is comparatively low [8]. In this paper, two DNN training optimization methods are used to cooperate with complementary features. One is RBM pre-training, which protects the system from falling into its local optima. The other is dropout combined with ReLU. The dropout is used to prevent the model from over-fitting while ReLU is introduced to optimize the training effect of DNN. All the algorithms will be described in detail below.

Eventually, the ERM is used on the noisy T-F units to synthesize the processed sound.

3 Feature Descriptions

3.1 Traditional Features. In the previous study, many features have been developed to train the DNN. Several classical features are chosen for the baseline.

Mel-Frequency Cepstral Coefficient (MFCC). MFCC reflects human ears' nonlinear correspondings towards high or low frequencies and speech short-time amplitude spectrums. MFCC is more in accord with characteristics of human hearing and approximately corresponds to log distribution of the real frequencies. To obtain MFCC features, discrete cosine transform (DCT) and log compression are employed.

Relative Spectral Transform PLP (RASTA-PLP). In order to obscure the differences between speakers and maintain the significant formant structure, perceptual linear prediction (PLP) is created [9]. RASTA-PLP introduces RASTA filtering to PLP [10]. By contrast, it lays more emphases on the modulation frequency which is quite relevant to human speech.

Amplitude Modulation Spectrogram (AMS). AMS is a common feature used in speech separation [11], which is neurophysiologically and psychoacoustically motivated and robust in noisy or reverberation environments. In order to compute AMS, the "short-term" Fourier transformation (STFT) is applied on each Bark frequency band of the non-logarithmic energy spectrogram to analyze the time trajectories. Additionally, principal component analysis (PCA) is employed to descend the feature dimensionality [12].

3.2 Auditory Features.

Gammatone Feature (GF). GF is extracted by passing the input signal through a bank of gammatone filters with the window shift of 10ms. Then a cubic root operation is applied to loudness compression over the magnitudes of the decimated outputs.

$$G_m[\tau] = \left| g|_{decimate}[\tau, m] \right|^{\frac{1}{3}}. \quad (1)$$

Where $G_m[\tau]$ indicates the T-F units of the input, and τ and m represent frequency channel index and time frame index respectively. We call a time slice of the above matrix GF, and $G[\tau]$ denotes its τ th channel. Frequency overlapping being applied among neighboring filter channels leads to correlations between GF components.

Multi-Resolution Cochleagram Feature (MRCG). Based on the cochleagram representation, MRCG represents the excitation pattern of the inner ear basilar membrane as a function of time [13].

There are five specific steps for computing MRCG. Firstly, we pass the input mixture through 64 gammatone filterbank and use the 20ms frame window with 10ms shift over the responding signals. Compute the power of each T-F unit [13] to derive 64-channel cochleagram, CG1. Each T-F unit is applied to a log operation. Secondly, CG2 is similar to CG1 except using 200ms frame window. Thirdly, a square window, which is 11 frequency channels and 11 time frames, is applied to average CG1 to derive CG3. Zero padding is used when the window slides out of the given cochleagram [5]. Fourthly, CG4 is computed in the same way except using a 23×23 square window. Finally, MRCG is obtained by concatenating CG1-4, each time frame of which has 64×4 dimensions.

It can be seen that CG1 captures the local information while the other three low-resolution ones encode spectrotemporal contexts to different degrees.

4 DNN Training Strategies

In order to cooperate with complementary features, two training strategies are used to optimize the model. One is RBM pre-training, and the other one is dropout combined with ReLU.

4.1 Restricted Boltzmann Machines Pre-Training. If the DNN training starts by a randomly initialized network, it is more likely to fall in local optima [14]. Therefore, the deep generative model

is trained by a stacking of multiple restricted Boltzmann machines (RBMs) [15]. As showed in Fig. 2, the bottom is a Gaussian-Bernoulli RBM [15] with a real-number visible layer and a two-valued hidden layer is connected through sigmoid activation function, above which are a stacking of Bernoulli-Bernoulli RBMs. Afterwards, unsupervised greedy fashion [14] is used for training layer by layer and contrastive divergence (CD) algorithm is used to update the parameters [15].

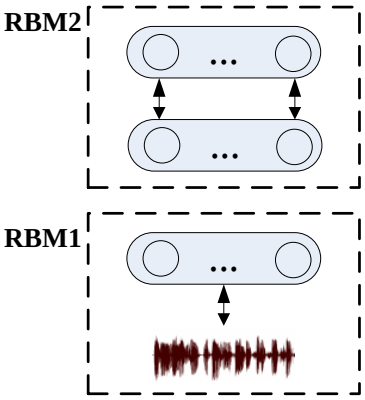

Figure 2 Illustrations of RBM pre-training

4.2 Dropout. Dropout is an effective strategy for overcoming the over-fitting in DNN training. In a feed-forward network, dropout discards a certain percentage (hidden drop factor) of the neural units randomly. In other words, dropout uses sub-model averaging to improve the generalization capability of the DNN. This advantage prevents complex co-adaptations between hidden units, forcing every hidden unit not to rely on each other.

The output of the hidden layer is

$$\mathbf{y}^{l} = \frac{1}{1-p}\mathbf{r}.* f\left(\mathbf{W}^{l}\mathbf{y}^{l-1} + \mathbf{b}^{l}\right). \tag{2}$$

Where $\mathbf{y}^{l}$ and $\mathbf{y}^{l-1}$ denote the outputs of l th and $l-1$ th layer respectively, $\mathbf{W}^{l}$ is the weight matrix of l th layer and $\mathbf{b}^{l}$ is its bias vector. f represents the nonlinear activation (e.g., sigmoid, ReLU) and $\mathbf{r}$ is the binary mask, which determines the real output of each layer, wherein r_j obeys Bernoulli distribution ($r_j \sim Bernoulli(p)$), so the factor during training is $\frac{1}{1-p}$. Fig. 3 shows DNN structure with the dropout of 4 inputs and 3 outputs.

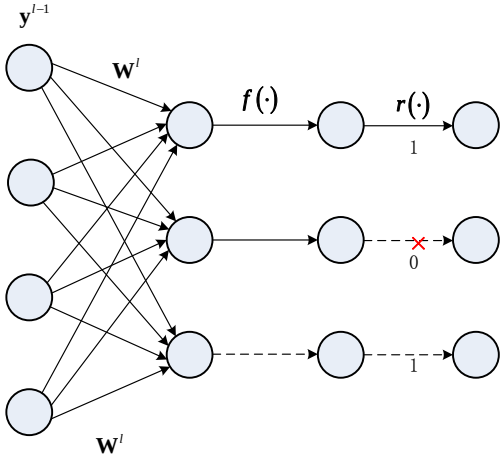

Figure 3 The diagram of dropout

[16] pointed out that dropout contributed to improving generalization error of DNN. [17] applied dropout to speech recognition, which led to the great improvements in DNN-HMM performance.

4.3 Rectified Linear Units. ReLU was first proposed by Nair and Hinton in 2010 [18] and applied to Restricted Boltzmann Machine (RBM). The function expression of ReLU is

$$u^{(i)} = \max\left(w^{(i)T}, 0\right) = \begin{cases} w^{(i)T}x & w^{(i)T}x > 0 \\ 0 & else \end{cases}. \tag{3}$$

Where $u^{(i)}$ indicates the activation function of a hidden unit and $w^{(i)}$ represents the weight. i is the number. Fig. 4 is the function graph of ReLU.

Due to its linearity and unsaturation, ReLU can realize parameter sparsification through simple thresholding activation. Hence, it can speed DNN training, improve generalization and avoid gradient disappearance problem. Bengio [19] claims that deep structure using ReLU can achieve similar or even advanced performance even if there is no pre-training.

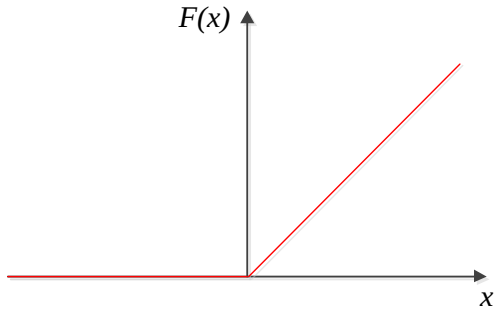

Figure 4 The diagram of ReLU function

5 Experiments and Results

5.1 Experiments Setup. The speech utterances are chosen from the IEEE [20] and TIMIT [21] corpora respectively. The noises are derived from the NOISEX-92 dataset [22].

The IEEE corpus contains 720 utterances wherein 600 utterances are for training and the other 120 utterances for test. We use them as our training and test set respectively. In TIMIT corpus, we use 1848 utterances, which are chosen from TIMIT training set randomly, as our training set. The TIMIT core test, which is composed of 192 utterances of both genders, is used as our test set. We use four noises from the NOISEX dataset for the experiment, including white Gaussian noise, babble noise, factory noise (called "factory 1") and destroyer engine room noise. Except white Gaussian noise, the other three noises are all non-stationary.

All noises have the length of 235 seconds. In order to ensure that the noises used for training and testing are not overlapped, we cut every noise into two halves. To generate the training set, we mix the training utterances with cuts randomly chosen from the first half at -5dB and 0dB. Similarly, to construct test mixtures, we mix randomly-chosen cuts from the other part with test utterances, at -5dB and 0dB as well. It is a challenging task to separate broadband noises at low SNRs [23].

5.2 Evaluation Criteria. Three evaluation metrics are used in this paper, including segmental Signal-to-Noise Ratio (segSNR), Perceptual Evaluation of Speech Quality (PESQ), Short-Time Objective Intelligibility (STOI). segSNR is used to measure the noise suppression performance [24,25],

$$segSNR = avg\left(10\log_{10}\frac{P(X_t)}{P(N_t)}\right). \tag{4}$$

Where $P(X_t)$ represents the power spectral density of the clean speech x at time t and $P(N_t)$ indicates the power spectral density of the noise n at time t.

STOI [26], as an objective intelligibility measurement, indicates the relevancy between clean and separated speech in the term of short-time temporal envelopes. STOI is highly correlated to human speech intelligibility. STOI scores from 0 to 1.

PESQ [27], belonging to objective Mean Opinion Score (MOS), is used to evaluate objective speech quality. PESQ is calculated from the separated speech and the corresponding clean speech, scoring from -0.5 to 4.5.

5.3 Baseline System. In the baseline system, DNN is used as the discriminative learning machine for predicting the desired estimated ratio mask (ERM) across all frequency bands, which has been exemplified to improve speech separation well [7,3]. The DNN has four hidden layers, which has 1024 sigmoid units each, an input layer and an output layer. The input feature set consists of 15 dimensional AMS, 13 dimensional RASTA-PLP and 31 dimensional MFCC [8]. To encode more temporal context, we splice a five-frame window as the input of the DNN. Therefore, with their delta components and five-frame window, the dimension of the input layer is 590 in total, while the output layer of DNN is 64, corresponding to the number of Gammatone filterbank. Since targets are in the range [0,1], we choose sigmoid activation functions for the output layer. According to [5], a second order ARMA filter is used to improve separation performance.

No pre-training is used. The standard backpropagation (BP) algorithm is used to train the networks. As for the optimization technique, we use the adaptive gradient descent [28] along with mini-batch size 1024 and a momentum term. The momentum rate for the first 5 epochs is 0.5 and the rate increases to and keeps at 0.9 after that. The learning rate is set to 0.8 at beginning and gradually reduces to 0.001. The maximum number of epochs is set to 20 and a DNN model is saved after every iteration. The minimum mean squared error (MMSE) is used as the cost function. We set aside 10 percent of the training set as a cross validation set. HIT-FA has shown nice correlation with human intelligence [11], where HIT denotes the accuracy rate and FA refers to false alarm rate. Hence in the validation data, we use HIT-FA to evaluate afore obtained 20 DNNs. Finally, the DNN with the highest HIT-FA rate is chosen as the final model.

5.4 Proposed Methods.

Complementary Features. In order to obtain richer feature information, we use a set of complementary frame-level features and its delta components [6] for mask estimation. We concatenate 64 dimensional GF and 256 dimensional MRCG with the aforementioned three features. With their delta components and five-frame window, the dimension of feature set is 3790.

DNN Training Strategy optimization. For the RBM pre-training, the maximum epoch is set to 25 with mini-batch size 1024. The learning rates of the input layer and hidden layer are set to be 0.004 and 0.01 respectively. The momentum rate of the first 5 epochs is 0.5, the rate increasing to and keeping at 0.9 after that.

Except for RBM, another strategy is coupling the complementary features with dropout and substituting 1024 ReLU hidden units for sigmoid activation. The dropout rate is 0.2.

5.5 Results and Analyses. The results of different systems in various noise conditions are shown in Tables 1 and 2, at different mixture SNRs.

Table 1 Performance Comparisons Between Various Systems on 0dB Mixtures

Corpus	System	White			Factory1			Engine			Babble		
		segSNR	STOI	PESQ	segSNR	STOI	PESQ	segSNR	STOI	PESQ	segSNR	STOI	PESQ
IEEE	baseline	4.0123	0.8545	1.4615	1.3605	0.7972	1.2556	2.7128	0.8648	1.4847	0.8051	0.7682	1.2703
	baseline+GF+MRCG	4.0928	0.8630	1.5520	1.7544	0.7999	1.2963	2.9881	0.8732	1.5219	1.1830	0.7742	1.2857
	baseline+GF+MRCG+RBM	4.0930	0.8643	1.5699	2.0284	0.8066	1.3097	3.0666	0.8753	1.5442	1.3366	0.7814	1.3060
	baseline+GF+MRCG+ReLU+Dropout	4.6494	0.8643	1.4716	2.5758	0.8156	1.2655	3.6234	0.8788	1.4895	1.3411	0.7961	1.2789
TIMIT	baseline	2.1377	0.8335	1.5499	0.5740	0.7679	1.3289	1.4472	0.8300	1.5049	0.3343	0.7514	1.3350
	baseline+GF+MRCG	2.3698	0.8369	1.5781	0.8932	0.7705	1.3625	1.9058	0.8431	1.6699	0.7399	0.7520	1.3544
	baseline+GF+MRCG+RBM	2.4128	0.8378	1.5813	1.0897	0.7762	1.3835	1.9313	0.8443	1.6888	0.8575	0.7549	1.3665
	baseline+GF+MRCG+ReLU+Dropout	3.1661	0.8404	1.5527	1.5346	0.7855	1.3337	3.0516	0.8523	1.5289	1.3770	0.7648	1.3367

Table 2 Performance Comparisons Between Various Systems on -5dB Mixtures

Corpus	System	White			Factory1			Engine			Babble		
		segSNR	STOI	PESQ	segSNR	STOI	PESQ	segSNR	STOI	PESQ	segSNR	STOI	PESQ
IEEE	baseline	1.5311	0.7849	1.2270	-1.3103	0.6745	1.1340	0.5583	0.7813	1.2670	-1.7193	0.6205	1.0904
	Baseline+GF+MRCG	1.7196	0.7893	1.2770	-0.7761	0.6863	1.1427	0.7656	0.8000	1.3043	-1.3054	0.6298	1.1133
	baseline+GF+MRCG+RBM	1.7804	0.7985	1.2794	-0.4673	0.6961	1.1536	0.8217	0.8004	1.3143	-1.0087	0.6349	1.1197
	baseline+GF+MRCG+ReLU+Dropout	2.7035	0.7993	1.2628	0.3691	0.7021	1.1372	1.8670	0.8062	1.2709	-0.0971	0.6399	1.0969
TIMIT	baseline	-0.1450	0.7350	1.2587	-1.9970	0.6311	1.1678	-0.5325	0.7550	1.3544	-2.3126	0.6025	1.1459
	baseline+GF+MRCG	-0.0656	0.7393	1.2726	-1.6730	0.6498	1.1688	-0.0599	0.7741	1.4036	-1.8765	0.6109	1.1581
	baseline+GF+MRCG+RBM	0.1387	0.7426	1.2845	-1.6210	0.6539	1.1749	0.1506	0.7772	1.4196	-1.5502	0.6192	1.1651
	baseline+GF+MRCG+ReLU+Dropout	0.8969	0.7468	1.2594	-0.8628	0.6673	1.1679	1.4349	0.7870	1.3635	-0.9842	0.6251	1.1464

From the results, some conclusions can be summarized for the complementary feature system. For the noise suppression measurement, segSNR, large improvements are obtained in the non-stationary noises, factory1, destroyer engine and babble noise conditions, except white Gauss noise, the improvement of which is comparatively limited. On white Gauss, factory1, destroyer engine noises, five complementary features yield STOI improvements universally. However, on babble noise, the STOI performance is slightly worse. There are two factors contributed to this phenomenon. For one thing, GF and MRCG are both based on human auditory perception properties, so they are more conducive to improve speech quality, not objective intelligibility. For another, babble noise itself is a speech noise. Hence even advanced complementary-feature system is difficult to distinguish target speech from speech noise. These analyses are exemplified by the results of PESQ, speech quality measurement. PESQ improvements on white Gauss, factory1 and destroyer engine noises outperform that of babble noise.

Advancing from five-complementary-feature system to RBM training strategy improves all objective evaluation criteria. Generally speaking, the gains of -5dB mixtures is better than that of 0dB mixtures; besides, the improvements gained in IEEE corpus is larger than that obtained in TIMIT corpus. One reason is that DNN mapping relation is more complex at low SNR environment. Another reason in point is that the quantity of training utterances in IEEE corpus is smaller than that of TIMIT corpus, so backpropagation algorithm is more likely to be trapped in local maxima. Therefore, RBM plays a more significant role in these two more challenging conditions.

Compared to the complementary feature system and the RBM system, substituting the dropout combined with ReLU system for the RBM system achieves noticeable improvements in segSNR, especially in the case of destroyer engine noises, where for example, in TIMIT corpus at 0dB, the segSNR of the dropout combined with ReLU system is 60% higher than that of complementary feature system while in IEEE corpus at -5dB, the gain between these two systems increases to 1.44 times. Similarly, there are universal STOI improvements in all the four noise conditions and the improvement of factory1 noise is the most appreciable. These are consistent with the common point of view that dropout can enhance the robustness of the model, avoid over-fitting and reduce this relatively stationary residue noise, especially for the non-stationary noises like the destroyer engine noise or the factory1 noise. Besides, the results prove that deep structure using ReLU can achieve advanced performance even if there is no pre-training. What is beyond our expectation is PESQ performance degradation. A comparatively reasonable explanation for this phenomenon is that GF and MRCG are conducive to improve PESQ performance based on human auditory perception properties. The sub-model averaging of dropout weakens this superiority although it can avoid complex co-adaptations wherein the activations of neurons are highly correlated. This is why PESQ performance meliorates compared to the baseline system, but deteriorates compared to the complementary model. Nevertheless, we could not deny the advantage of dropout. As [29] remarks, dropout might deteriorate the performance for matching noise cases, whereas it could optimize the robustness in mismatched types.

6 Conclusion

In this study, we have proposed a complementary feature method to perform speech separation by combining five features as a super-vector fed into DNN. We also test the performance of two DNN training strategies. The experiments suggest that the complementary feature and the RBM system achieve improvements in terms of all three evaluation indicators while dropout combined with ReLU system specializes in segSNR and STOI more.

In this study, the noise types and SNR environments are the same in training and testing set. Future work needs to choose unseen noises and SNR mixtures for test to experiment the potential of these systems to further improve performance.

7 Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (No. 61673395, No. 61302107 and No. 61403415).

References

[1] N. Mohammadiha, P. Smaragdis, and A. Leijon, “Supervised and unsupervised speech enhancement approaches using nonnegative matrix factorization,” IEEE Trans. Audio, Speech, Lang. Process., vol. 21, no. 10, pp. 2140–2151, Oct. 2013.

[2] T. Virtanen, J. Gemmeke, and B. Raj, “Active-set Newton algorithm for overcomplete non-negative representations of audio,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 11, pp. 2277–2289, Nov. 2013.

[3] Y. Wang and D. Wang, “Towards scaling up classification-based speech separation,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 7, pp. 1381–1390, Jul. 2013.

[4] X. Zhao, Y. Shao, and D. L.Wang, “CASA-based robust speaker identification,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 5, pp. 1608–1616, Jul. 2012.

[5] Chen, J., Wang, Y., and Wang, D. "A feature study for classification-based speech separation at very low signal-to-noise ratio."*Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014: 7039-7043.

[6] Y. Wang, “Exploring monaural features for classification-based speech segregation,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 2, pp. 270–279, Feb. 2013.

[7] Y. Wang and D. Wang, “Cocktail party processing via structured prediction,” in *Proc. NIPS*, 2012, pp. 224–232.

[8] Wang, Z. "Robust Speech Recognition From Ratio Masks." Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on. IEEE, 2016: 5720-5724.

[9] H. Hermansky, “Perceptual linear predictive (PLP) analysis of speech,” J. Acoust. Soc. Amer., vol. 87, pp. 1738–1752, 1990.

[10] H. Hermansky and N. Morgan, “RASTA processing of speech,” IEEE Trans. Speech, Audio Process., vol. 2, no. 4, pp. 578–589, Oct. 1994.

[11] G. Kim, Y. Lu, Y. Hu, and P. C. Loizou, “An algorithm that improves speech intelligibility in noise for normal-hearing listeners,” J. Acoust. Soc. Amer., vol. 126, pp. 1486–1494, 2009.

[12] B. Kollmeier, and R. Koch, “Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction,” J. Acoust. Soc. Am. 95(3), pp. 1593-1602, 1994.

[13] D. L.Wang and G. J. Brown, “Computational auditory scene analysis,” in Principles, algorithms and applications.. Hoboken, NJ, USA: Wiley-IEEE Press, 2006.

[14] G. E. Hinton and R. R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” Science, vol. 313, no. 5786, pp. 504–507, 2006.

[15] Y. Bengio, “Learning deep architectures for AI,” *Found. Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, 2009.

[16] Hinton G.E, Srivastava N, Krizhevsky A, Sutskever I and Salakhutdinov R, Improving neural networks by preventing co-adaptation of feature detectors[Z/OL]. Canada: Cornell University,[2013-07-3].http://arxiv.org/abs/1207.0580.

[17] Miao Yajie, Metze Florian, Improving Low-Resource CD-DNN-HMM using Dropout and Multilingual DNN Training, Proceedings of 14th Annual Conference of the International Speech Communication Association (INTERSPEECH)[C]. Lyon, France: ISCA, 2013.2237-2241.

[18] Nair V, Hinton G E. Rectified linear units improve restricted Boltzmann machines[C]//Proceedings of the 27th International Conference on Machine Learning (ICML-10). 2010: 807-814.

[19] Glorot X , Bordes A , Bengio Y Deep sparse rectifier neural networks[C]//International Conference on Artificial Intelligence and Statistics. 2011: 315-323.

[20] IEEE Subcommittee (1969). IEEE Recommended Practice for Speech Quality Measurements. *IEEE Trans. Audio and Electroacoustics*, AU-17(3), 225-246

[21] J. Garofolo, DARPA TIMIT acoustic-phonetic continuous speech corpus. Gaithersburg, MD, USA: Nat. Inst. of Standards Technol., 1993.

[22] A.Varga and H. 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., vol. 12, pp. 247–251, 1993.

[23] E. Healy, S. Yoho, Y. Wang, and D. Wang, “An algorithm to improve speech recognition in noise for hearing-impaired listeners,” *J. Acous. Soc. Amer.*, pp. 3029–3038, 2013.

[24] Cohen and B. Berdugo, “Speech enhancement for nonstationary noise environments,” Signal processing, vol. 81, no. 11, pp. 2403–2418, 2001.

[25] R. Talmon and S. Gannot, “Single-channel transient interference suppression with diffusion maps,” IEEE Trans. Audio, Speech, and Lang. Process., vol. 21, no. 1, pp. 132–144, 2013.

[26] C. Taal, R. Hendriks, R. Heusdens, and J. Jensen, “An algorithm for intelligibility prediction of time-frequency weighted noisy speech,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

[27] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, “Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs,” in *Proc. ICASSP*, 2001, pp. 749–752.

[28] J. Duchi, E. Hazan, and Y. Singer, “Adaptive subgradient methods for online learning and stochastic optimization,” *J. Mach. Learn. Res.*, pp. 2121–2159, 2011.

[29] Xu Y, Du J, Dai L R, et al. A regression approach to speech enhancement based on deep neural networks[J]. Audio, Speech, and Language Processing, IEEE/ACM Transactions on, 2015, 23(1): 7-19.