LETTER

A Perceptually Motivated Approach for Speech Enhancement Based on Deep Neural Network

Wei HAN[†a)], Xiongwei ZHANG[†], Gang MIN[†*], *Nonmembers*, *and* Meng SUN[†], *Member*

SUMMARY In this letter, a novel perceptually motivated single channel speech enhancement approach based on Deep Neural Network (DNN) is presented. Taking into account the good masking properties of the human auditory system, a new DNN architecture is proposed to reduce the perceptual effect of the residual noise. This new DNN architecture is directly trained to learn a gain function which is used to estimate the power spectrum of clean speech and shape the spectrum of the residual noise at the same time. Experimental results demonstrate that the proposed perceptually motivated speech enhancement approach could achieve better objective speech quality when tested with TIMIT sentences corrupted by various types of noise, no matter whether the noise conditions are included in the training set or not.

key words: *perceptually motivated, deep neural network, speech enhancement, masking residual noise*

1. Introduction

Speech enhancement is an important stage to suppress the noise and improve perceptual quality in several noisy environments such as the inside of a subway, in the street or inside an airport. Miscellaneous speech enhancement methods have been proposed, like spectral subtraction [1], minimum mean squared error (MMSE) estimation [2] and optimally-modified log-spectral amplitude (OM-LSA) speech estimator [3]. These methods are generally less effective in low SNR and usually assume that the noise is stationary, thus not good at dealing with non-stationary noise.

In contrast to above signal processing based methods, model-based methods build models of speech and/or noise using premixed signals and show promising results in challenging conditions. Non-negative matrix factorization (NMF) is a widely used model-based method for removing or separation of non-stationary noises in spectral domain [4]. In this approach, a non-negative data matrix is approximated by a product of a basis matrix and an encoding matrix with non-negative elements. However, NMF can be viewed as linear model, which performance of extract complex speech features not better than non-linear model.

Recently, a new approach named deep learning was applied in speech enhancement [5]. Deep learning methods are representation-learning methods with multiple levels representation, obtained by composing simple but nonlinear modules that each transform the representation at one level into a representation at a higher, slightly more abstract level. With the composition of enough such transformations, very complex functions and ultimately more useful representations can be learned. In [6], the authors propose a regression DNN, which is trained by large speech sentences and 104 noise types. The DNN can achieve better enhancement performance than the conventional MMSE based technique. In [7], the authors use DNN to train different targets for speech separation tasks, including binary mask, the ideal ratio mask, the short-time Fourier transform spectral magnitude, FFT-mask and the Gammatone frequency power spectrum. In [8], the authors explore joint optimization of masking functions and deep recurrent neural networks (RNN) for monaural source separation tasks. The joint optimization of the RNN with an extra masking layer enforces a reconstruction constraint and discriminative training criterion strategies for the neural networks to further enhance the separation performance. In general, this methods not consider the auditory perception.

This letter proposes a novel architecture of DNN for speech enhancement which using auditory perception. Auditory system is insensitive to the quantization noise which near the high-energy regions of the spectrum. Considering this good characteristics of auditory masking, we design a new DNN that can predict both the enhance speech and perceptual information, then the perceptual information could be used to further improve the speech enhancement quality.

The rest of the letter is organized as follows. In Sect. 2, we propose to use Perceptually motivated DNN to enhance noisy speech. The results of the experimental evaluation over the TIMIT database are outlined in Sect. 3. Finally, Sect. 4 conclude our work.

2. Proposed Method

2.1 Perceptual Weighting

In most low-rate speech codes (e.g.,CELP), the perceptually weighted error criterion was used to mask quantization noise. This is based on the fact that the auditory system has a limited ability to detect the quantization noise near the high-energy regions of the spectrum (e.g., near the formant peaks). Quantization noise near the formant peaks is masked by the formant peaks, and is therefore not audible [9]. In order to obtain good auditory masking performance, we can

Manuscript received October 1, 2015.

Manuscript revised December 22, 2015.

[†] The authors are with the Lab of Intelligence Information Processing, PLA University of Science and Technology, Nanjing, China.

[*] Presently, also with XI'AN Communications Institute, Xi'an, China.

a) E-mail: lan3533065@163.com

DOI: 10.1587/transfun.E99.A.835

shape the frequency spectrum of the error so that less emphasis is placed near the formant peaks and more emphasis is placed on the spectral valleys, where any amount of noise present will be audible. The following perceptual filter is used to shape the error.

$$P(z) = \frac{A(\frac{z}{\gamma_1})}{A(\frac{z}{\gamma_2})} = \frac{1-\sum_{k=1}^{p} a_k \gamma_1^k z^{-k}}{1-\sum_{k=1}^{p} a_k \gamma_2^k z^{-k}} \tag{1}$$

where $A(z)$ is the LPC polynomial, a_k are the short-term linear prediction coefficients, γ_1 and γ_2 ($0 \leq \gamma_2 \leq \gamma_1 \leq 1$) are parameters that control the energy of the error in the formant regions and p is the prediction order.

The frequency response of the perceptual filter as shown:

$$P(\omega) = P(z)|_{z=e^{j\omega}} \tag{2}$$

The perceptual weighting matrix W used in the frequency domain method was defined using the following diagonal matrix:

$$W_f = \begin{bmatrix} P(0) & 0 & \cdots & 0 \\ 0 & P(\omega_0) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & P((N-1)\omega_0) \end{bmatrix} \tag{3}$$

where $\omega_0 = 2\pi/N$ and N is the FFT length.

2.2 Perceptually Motivated Method for Speech Enhancement

The noisy signal $y = s + n$ is expressed in terms of additive components which are speech vector s and noise vector n. The Fourier transform of the noisy speech can be written as

$$\mathrm{Y}(\omega) = F^H \mathrm{y} = F^H \mathrm{s} + F^H \mathrm{n} = \mathrm{S}(\omega) + \mathrm{N}(\omega) \tag{4}$$

where F is the N-point discrete Fourier transform matrix, $\mathrm{S}(\omega)$ and $\mathrm{N}(\omega)$ are the $N \times 1$ vector containing the spectral components of clean speech s and noise n, respectively.

Let $\hat{\mathrm{S}}(\omega) = G \cdot \mathrm{Y}(\omega)$ be the linear estimator of $\mathrm{S}(\omega)$, where G is a $N \times N$ matrix. The error signal can be obtained as follows:

$$\begin{aligned} \varepsilon(\omega) &= \hat{\mathrm{S}}(\omega) - \mathrm{S}(\omega) = (G - I)\ \mathrm{S}(\omega) + G\mathrm{N}(\omega) \\ &= \varepsilon_{\mathrm{s}}(\omega) + \varepsilon_{\mathrm{n}}(\omega) \end{aligned} \tag{5}$$

where $\varepsilon_{\mathrm{s}}(\omega)$ represents the speech distortion and $\varepsilon_{\mathrm{n}}(\omega)$ represents residual noise. We define the energy of the spectral speech distortion and the energy of the spectral residual noise as $\overline{\varepsilon_{\mathrm{s}}^2(\omega)} = E(\varepsilon_{\mathrm{s}}^H(\omega) \cdot \varepsilon_{\mathrm{s}}(\omega)) = \mathrm{tr}((G - I) \cdot F^H \cdot R_{\mathrm{s}} \cdot F \cdot (G - I)^H)$ and $\overline{\varepsilon_{\mathrm{n}}^2(\omega)} = E(\varepsilon_{\mathrm{n}}^H(\omega) \cdot \varepsilon_{\mathrm{n}}(\omega)) = \mathrm{tr}(G \cdot F^H \cdot R_{\mathrm{n}} \cdot F \cdot G^H)$, respectively. The optimal linear estimator can be obtained by solving the following constrained optimization problem :

$$\begin{aligned} &\min \ \overline{\varepsilon_{\mathrm{s}}^2(\omega)} \\ &subject\ to\ :\ \tfrac{1}{N}\overline{\varepsilon_{\mathrm{n}}^2(\omega)} \leq \sigma^2 \end{aligned} \tag{6}$$

where σ^2 is a threshold, which is a positive number.

In order to shape the spectrum of the residual noise, the perceptually weighed residual noise $\varepsilon_{\mathrm{wn}}$ was used to replace the constraint in (6), $\varepsilon_{\mathrm{wn}}$ can be obtained by perceptual weighting matrix W_f in (3) as follows:

$$\varepsilon_{\mathrm{wn}} = W_f \cdot \varepsilon_{\mathrm{n}} \tag{7}$$

The energy of the perceptual weighed residual noise is defined as:

$$\overline{\varepsilon_{\mathrm{wn}}^2} = E(\varepsilon_{\mathrm{wn}}^H \varepsilon_{\mathrm{wn}}) = \mathrm{tr}(W_f E[\varepsilon_{\mathrm{n}} \varepsilon_{\mathrm{n}}^H] W_f^H) \tag{8}$$

the optimal linear estimator can be obtained by solving the following constrained optimization problem:

$$\begin{aligned} &\min \ \overline{\varepsilon_{\mathrm{s}}^2} \\ &subject\ to\ :\ \tfrac{1}{N}\overline{\varepsilon_{\mathrm{wn}}^2} \leq \sigma^2 \end{aligned} \tag{9}$$

The solution to (9) can be found in [9]. G is a stationary feasible point if it satisfies the gradient equation of the objective function

$$J(G, \mu) = \overline{\varepsilon_{\mathrm{s}}^2} + \mu(\overline{\varepsilon_{\mathrm{wn}}^2} - N\sigma^2)$$

and

$$\mu(\overline{\varepsilon_{\mathrm{wn}}^2} - N\sigma^2) = 0 \quad for\ \mu \geq 0 \tag{10}$$

where μ is the Lagrangian multiplier. From $\nabla_G J(G, \mu) = 0$ we have

$$\mu(W_f^H W_f) G\, F^H R_{\mathrm{n}} F + G\, F^H R_{\mathrm{s}} F = F^H R_{\mathrm{s}} F \tag{11}$$

To simplify matters, assuming that G is a diagonal matrix. The matrices $F^H R_{\mathrm{s}} F$ and $F^H R_{\mathrm{n}} F$ are asymptotically diagonal and the diagonal elements of R_{s} and R_{n} are the power spectrum components $S_{s}(\omega_i)$ and $S_{\mathrm{n}}(\omega_i)$ of the clean speech vector s and noise vector n, respectively. Denoting the diagonal element of G by $g(\omega_i)$, (11) can be simplified and the gain function $g(\omega_i)$ can be obtained by

$$g(\omega_i) = \frac{S_{\mathrm{s}}(\omega_i)}{S_{\mathrm{s}}(\omega_i) + \mu |P(\omega_i)|^2 S_{\mathrm{n}}(\omega_i)} \tag{12}$$

2.3 Perceptually Motivated Speech Enhancement Based on DNN

In this letter, we design a new DNN architecture for Perceptually motivated speech enhancement which is named PDNN and illustrated in Fig. 1.

As shown in Fig. 1, we direct train the PDNN and obtain the clean speech power spectrum $\hat{S}_{\mathrm{s}}(\omega)$, the noise power spectrum $\hat{S}_{\mathrm{n}}(\omega)$ and the frequency response of the perceptual filter $\hat{P}(\omega)$, while traditional methods always need accurate estimation of the clean speech spectrum, then the estimation of gain function $\hat{g}(\omega)$ in (12) together with the noisy speech spectrum $\mathrm{Y}(\omega)$ viewed as an extra layer and is placed on the top of the network as follows:

$$\begin{aligned} \hat{\mathrm{S}}(\omega) &= \hat{g}(\omega) \cdot \mathrm{Y}(\omega) \\ &= \frac{\hat{S}_{\mathrm{s}}(\omega)}{\hat{S}_{\mathrm{s}}(\omega) + \mu |\hat{P}(\omega)|^2 \hat{S}_{\mathrm{n}}(\omega)} \cdot \mathrm{Y}(\omega) \end{aligned} \tag{13}$$

the μ value used in above equation is calculated as follows:

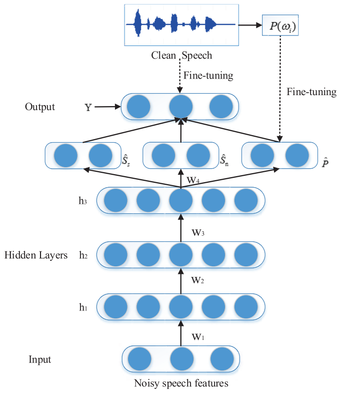

Fig. 1 Proposed PDNN architecture.

$$\mu = \begin{cases} \mu_0 - \frac{SNR_{dB}}{k} & , \quad -5 < SNR_{dB} < 20 \\ 1 & , \quad SNR_{dB} \geq 20 \\ \mu_{\max} & , \quad SNR_{dB} \leq 5 \end{cases} \tag{14}$$

where $\mu_{\max}$ is the maximum allowable value of μ, the value of μ controls the tradeoff between residual noise and speech distortion. $\mu_0 = (1 + 4\mu_{\max})/5$, $k = 25/(\mu_{\max} - 1)$, $SNR_{dB} = 10\log_{10}SNR$ and SNR is computed as:

$$SNR = \frac{\sum_{i=0}^{N-1} \hat{S}_s(\omega_i)}{\sum_{i=0}^{N-1} \hat{S}_n(\omega_i)} \tag{15}$$

The objective of training PDNN is minimizing mean squared error between the output and reference clean features both the speech and frequency response of perceptual filter as follows:

$$J_{MSE}(W, b) = \left\|\hat{S}_s(W, b) - S_s\right\|_2^2 + \left\|\hat{P}(W, b) - P\right\|_2^2 \tag{16}$$

where W and b are the network weight and bias parameters. This training stage is also called supervised fine-tuning. The reference frequency response of perceptual filter P is calculated by LPC polynomial of clean speech. The update rules for the W and b can be expressed as:

$$W^l = W^l - \varepsilon\frac{\partial J(W,b)}{\partial W^l} \qquad 1 \leq l \leq L + 1 \tag{17}$$

$$b^l = b^l - \varepsilon\frac{\partial J(W,b)}{\partial b^l} \qquad 1 \leq l \leq L + 1 \tag{18}$$

where ε is a learning rate, L indicate the total number of hidden layers. After many update iterations for parameters, we can get a good network mapping from noisy to clean speech.

The enhanced magnitude spectrum is transformed back into the time domain using the short-time inverse FFT and the original (noisy signal) phase information.

3. Experiments and Result Analysis

Experiments were conducted on TIMIT database. 1200 sentences were randomly selected from 240 different male and female speakers as the training set. All the sentences were down-sampled to 8 kHz. Fifteen types of noise, namely *babble, car, casino, cicadas, f16, factory1, frogs, hfchannel, jungle, restaurant, street, white, airport, pink* and *birds*, were considered. The 1200 sentences were added with the above-mentioned fifteen types of noise and eight levels of SNR, at −6 dB, −3 dB, 0 dB, 3 dB, 6 dB, 9 dB, 12 dB, 15 dB, to build a multi-condition stereo training set. The 257 dimensions spectrum as the input features were extracted using a 512 point short time Fourier transform (STFT) with 25% overlap.

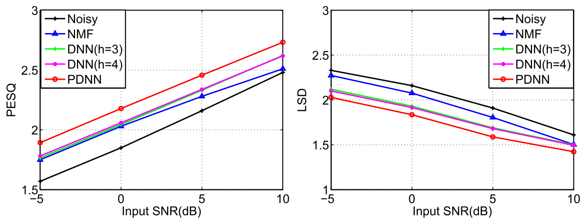

Fig. 2 The PESQ and LSD scores of NMF, DNN and PDNN at four different input SNR levels. For each condition, the numbers are the mean values over all the 20 noise types.

200 sentences were randomly selected from 120 different male and female speakers as the test set and five type of unseen noise namely *exhibition, subway, train, motorcycles* and *ocean* were used to evaluate the generalization ability of the proposed method.

The prediction order p was set to 10 and the $\mu_{\max}$ was set to 5 according to experimental results. The proposed PDNN was 3 hidden layers of 2048 hidden units and Rectified linear unit (ReLU) was chosen as the activation function [10]. We compared the method with NMF approach, traditional DNN with 3 hidden layers and DNN with 4 hidden layers. The dictionary of NMF was learned from the 1200 clean sentences and the speech base and noise base were chosen as 1000 and 100, respectively. Two objective measures, the perceptual evaluation of speech quality (PESQ) and log-spectral distortion (LSD) were used to evaluate the quality of the enhanced speech.

The general evaluation of the proposed method at different SNR levels is shown in Fig. 2. It is clear that PDNN has shown better performance than the reference methods, especially at low SNR levels. This due to speech has more noise at low SNR conditions, but our method have good masking ability.

To better understand the method performance on each noise type, Fig. 3 and Fig. 4 present the mean PESQ scores and LSD values over the four SNR levels, respectively. For the PESQ scores, from Fig. 3, we can see that, for the 15 noise conditions in the training set, PDNN outperforms the reference methods. Furthermore, for the noise conditions that are not included in the training data, the proposed methods still perform better. By contrasting DNN(h3) and DNN(h4), objective speech quality does not improve with the increase of the number of hidden layers. Compared DNN(h4) and PDNN, although we add more a hidden layer which equal to the extra layer of PDNN, the DNN(h4) performance is not good compared with PDNN. It prove that our method which using perceptual weighting matrix can mask more noise and achieve better performance than tradi-

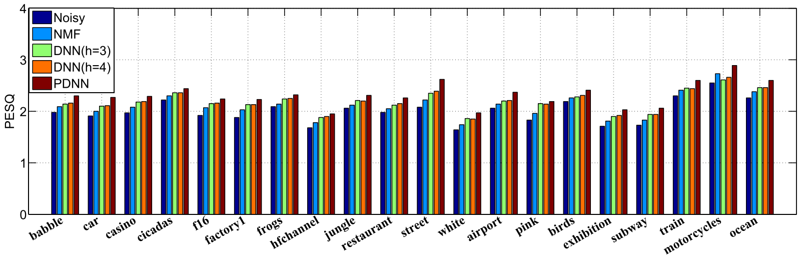

Fig. 3 The PESQ scores of NMF, DNN and PDNN for the 20 noise types. For each noise type, the numbers are the mean values over four input SNR conditions, i.e. from −5 dB to 10 dB spaced by 5 dB.

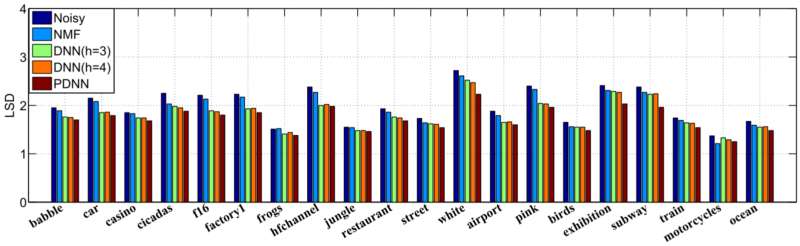

Fig. 4 The LSD values of NMF, DNN and PDNN for the 20 noise types. For each noise type, the numbers are the mean values over four input SNR conditions, i.e. from −5 dB to 10 dB spaced by 5 dB.

Table 1 The improved mean PESQ scores over four stationary noise (left side: *f16, hfchannel, white, train*) and four non-stationary noise (right side: *babble, factory1, restaurant, exhibition*).

Input SNR	-5dB	0dB	5dB	10dB
NMF	0.118/0.113	0.113/0.108	0.115/0.098	0.120/0.103
DNN(h4)	0.221/0.237	0.209/0.219	0.202/0.195	0.158/0.167
PDNN	0.313/0.350	0.318/0.333	0.305/0.310	0.235/0.273

tional DNN. A similar conclusion can be drawn for the LSD metric from Fig. 4.

Table 1 presents the improvements on PESQ of PDNN, DNN(h4) and NMF with respect to the noisy speech under stationary and non-stationary noise conditions, respectively. The proposed PDNN approach consistently outperformed the DNN and NMF for all the noise levels, especially at low SNR conditions. The results demonstrate that our proposed approach shows better ability on modeling non-stationary noise types than the stationary noise types.

4. Conclusion

We propose a novel speech enhancement approach that combines auditory perception and Deep neural network. Experimental results under 20 noise types at different SNR levels confirm the effectiveness of this method. We will focus on exploring more human auditory perception with DNN for speech enhancement in future work.

Acknowledgements

This work is partially supported by NSF of China (Grant No. 61471394, 61402519) and NSF of JIANG Su Province (Grant No. BK2012510, BK20140071, BK20140074). The authors would like to thank the anonymous reviewer for helpful advices to improve the quality of this letter.

References

[1] K. Paliwal, K. Wójcicki, and B. Schwerin, "Single-channel speech enhancement using spectral subtraction in the short-time modulation domain," Speech Commun., vol.52, no.5, pp.450–475, 2010.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," IEEE Trans. Acoust., Speech, Signal Process., vol.33, no.2, pp.443–445, 1985.

[3] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," IEEE Trans. Speech Audio Process., vol.11, no.5, pp.466–475, 2003.

[4] C. Sun, Q. Zhu, and M. Wan, "A novel speech enhancement method based on constrained low-rank and sparse matrix decomposition," Speech Commun., vol.60, pp.44–55, 2014.

[5] 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, 2013.

[6] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," IEEE/ACM Trans. Audio Speech Lang. Process., vol.23, no.1, pp.7–19, 2015.

[7] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," IEEE/ACM Trans. Audio Speech Lang. Process., vol.22, no.12, pp.1849–1858, 2014.

[8] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," IEEE/ACM Trans. Audio Speech Lang. Process., vol.23, no.12, pp.2136–2147, 2015.

[9] Y. Hu and P.C. Loizou, "A perceptually motivated approach for speech enhancement," IEEE Trans. Speech Audio Process., vol.11, no.5, pp.457–465, 2003.

[10] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," Proc. AISTATS, 2011.