Modern Physics Letters B
Vol. 31, Nos. 19–21 (2017) 1740096 (5 pages)

DOI: 10.1142/S0217984917400966

Deep neural network and noise classification-based speech enhancement

Wenhua Shi*,†,‡, Xiongwei Zhang*, Xia Zou* and Wei Han*

*Lab of Intelligent Information Processing,
PLA University of Science and Technology, Nanjing 210017, China
†Flight Instructor Training Base, Air Force Aviation University,
Bengbu 233000, China
‡whshi0919@163.com

Received 10 September 2016
Published 19 May 2017

In this paper, a speech enhancement method using noise classification and Deep Neural Network (DNN) was proposed. Gaussian mixture model (GMM) was employed to determine the noise type in speech-absent frames. DNN was used to model the relationship between noisy observation and clean speech. Once the noise type was determined, the corresponding DNN model was applied to enhance the noisy speech. GMM was trained with mel-frequency cepstrum coefficients (MFCC) and the parameters were estimated with an iterative expectation-maximization (EM) algorithm. Noise type was updated by spectrum entropy-based voice activity detection (VAD). Experimental results demonstrate that the proposed method could achieve better objective speech quality and smaller distortion under stationary and non-stationary conditions.

Keywords: Speech enhancement; Gaussian mixture model; deep neural network.

1. Introduction

Speech enhancement is to recover the original signal from contaminated speech with improved quality, clarity and intelligibility. It has been widely used in speech communication, speech coding and speech recognition. Monaural speech enhancement is more challenging since only one signal channel information is available and without any prior knowledge. Most noise reduction algorithms are based on the prior distribution assumption.[1,2] They are generally less effective in non-stationary noise condition or hard to use in real-world speech applications.[3] Recently, deep neural network[4] produces state-of-art results in complex signal processing depending on its high ability of extracting complex features and modeling structured information of data.

‡Corresponding author.

In this paper, a speech enhancement method using noise classification and deep neural network (DNN) is proposed. The proposed method consists of training and enhancement stage. In the training stage, MFCC is used as the feature to generate GMM for each type of noise. The corresponding DNN regression model is trained to obtain the complicated nonlinear mapping from noisy observation to clean signal. In the enhancement stage, the speech-absent frames are determined by spectrum entropy-based VAD algorithm. Noise type decision is made according to the minimum probability of classification error criterion during speech-absent frames. Then the corresponding DNN model is used to recover the clean speech from noisy speech.

2. The Proposed Method for Speech Enhancement

2.1. *GMM mixture model*

GMM is based on the assumption that arbitrary probability density function (PDF) can be modeled as a linear combination of several single Gaussian PDFs. The PDF of a GMM with M components can be described as $p(\mathbf{x}|\Theta) = \sum_{i=1}^{M} w_i G(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, where $\mathbf{x}$ is frame feature, w_i is the weight for the ith single Gaussian PDF which satisfied the conditions of $0 \leq w_i \leq 1$, $\forall\ i = 1, 2, \ldots, M$ and $\sum_{i=1}^{M} w_i = 1$. The PDF of ith single Gaussian function can be expressed as follows:

$$G(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_i|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^T \sum_{j}^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)\right]. \quad (1)$$

A GMM could be denoted by parameters, $\Theta = \{w_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}, i = 1, 2, \ldots, M$, where $\boldsymbol{\mu}_i$ is the mean vector and $\boldsymbol{\Sigma}_i$ is the covariance matrix, d is the dimension of frame feature vector. Mean and covariance parameters of the GMM are estimated in the *maximum likelihood* (ML) method. The parameters Θ are initialized by K-means cluster algorithm first. Then an iterative expectation-maximization (EM) algorithm is applied to maximize the likelihood function $J(\theta) = \ln[\prod_{i=1}^{n} p(\mathbf{x}_i; \Theta)]$ to update the model parameters Θ. n is the total number of samples in training stage.

2.2. *Noise type classification*

The noise type is determined during noise segments. Entropy-based VAD algorithm is expressed as[5]

$$H(t) = -\sum_{k=1}^{N}\left[|Y(t,k)|^2 \Big/ \sum_{k=1}^{N}|Y(t,k)|^2\right] \log\left[|Y(t,k)|^2 \Big/ \sum_{k=1}^{N}|Y(t,k)|^2\right], \quad (2)$$

where $Y(t,k)$ is the short-time Fourier transform (STFT) of noisy speech in the tth frame at the kth frequency bin. The entropy $H(t)$ has higher value in speech-absent frames. The beginning frame is usually considered as non-speech frame. The average entropy of the beginning frames is used as the threshold to discriminate speech-absent frames and speech frames.

For a c type of noise classification problem, denoted by $\{T_1, T_2, \ldots, T_c\}$, the GMM is established for each type of noise in the training stage. In the enhancement stage, given the input frame feature $\mathbf{x}$, assuming that each kind of noise has the same *a priori* probability $p(\Phi_i)$, the noise classification process is determined by finding which GMM has the largest *a posteriori* probability under minimum error rate criterion[6]:

$$T_i = \arg \max_{1 \le i \le c} p(\mathbf{x}|\Phi_i)\,. \tag{3}$$

2.3. *DNN model architecture*

The architecture adopted in this paper is the basic regression DNN model[7] with one input layer, L hidden layers and one output layer using log-power spectral as the feature for offering perceptually relevant parameters. All the parameters in the network are trained under MMSE criterion and updated using gradient descent algorithm:

$$J_{\mathrm{MSE}}(\mathbf{W}, \mathbf{b}) = \|\hat{\mathbf{y}}_t(\mathbf{x}_t, \mathbf{W}, \mathbf{b}) - \mathbf{y}_t\|_2^2\,, \tag{4}$$

$$(\mathbf{W}^l, \mathbf{b}^l) = (\mathbf{W}^l, \mathbf{b}^l) - \varepsilon \frac{\partial J(\mathbf{W}, \mathbf{b})}{\partial (\mathbf{W}^l, \mathbf{b}^l)} \quad 1 \le l \le L + 1\,, \tag{5}$$

where $\mathbf{x}_t$ is the input to the network at tth frame, $\hat{\mathbf{y}}_t$ and $\mathbf{y}_t$ denote the log-power spectral of the enhanced and target speech in the tth frame respectively. $\mathbf{W}^l$ is the weight matrix, $\mathbf{b}^l$ is the basis matrix, ε is a parameter determining the convergence rate.

The time domain enhanced speech could be reconstructed by applying inverse STFT using log-power spectral of enhanced speech and the phase information of the noisy speech.

3. Experiments and Results Analysis

3.1. *Dataset and setup*

Six types of noise including (a) babble, (b) white, (c) f16, (d) hfchannel, (e) pink, (f) Factory1 were used in the training of GMM. Training material of 200 s long was used for each kind of noise. Each GMM consisted of 10 single components. Compared with linear distribution of common cepstrum, mel-frequency cepstrum is more close to the response of human auditory system. Therefore, in the GMM training stage, the 34-dimensional mel representation and the first-order derivative were chosen as the input frame features. In the training of DNN, 1200 utterances were randomly selected from TIMIT database with 240 different speakers and concatenated into one utterance as the clean speech. Noisy speech was generated by adding six types of noise to clean speech. Each utterance was mixed with each noise at eight SNR levels from -6 dB to 15 dB spaced by 3 dB. All the sentences were down-sampled to 8 kHz. Log-power spectral feature was obtained using a 1024-point STFT with 50% overlap. Rectified linear function[8] $f(x) = \max(0, x)$ was adopted as the nonlinear activation function.

In the testing stage, another 30 s of each kind of noise was used to evaluate the performance of noise classification algorithm. 200 utterances with 120 different speakers over four input SNR conditions, i.e. from −5 dB to 10 dB spaced by 5 dB were selected as testing corpus to evaluate the proposed speech enhancement method.

3.2. *Results and analysis*

The classification result is presented in Table 1. It illustrates that only factory noise has a slightly lower classification accurate rate. The classification accuracy rate is f16 and the hfchannel noise is close to 100%. It demonstrates that the proposed algorithm is suitable for noise classification.

The proposed speech enhancement algorithm was compared with NMF approach and DNN method without noise classification. Perceptual evaluation of speech quality (PESQ) and log-spectral distortion (LSD) were employed as two objective measurements to evaluate the quality of enhanced speech. The results are presented in Fig. 1. The results suggest that the proposed method could bring improvement for PESQ and decrease for LSD for all types of noise, especially for white noise. The improvement of PESQ for Factory1 noise is smaller than other kind of noise, but it gets better LSD result compared with other five types of noise except white noise.

Table 1. The result of noise classification.

		Predicted class					
		Babble	White	f16	hfchannel	Pink	Factory1
Actual class	Babble	93.6	0.00	0.33	0.06	2.72	3.34
	White	0.21	92.3	2.12	0.54	0.30	4.56
	f16	0.00	1.72	98.3	0.00	0.00	0.00
	hfchannel	0.00	0.00	0.57	99.4	0.00	0.00
	Pink	0.56	1.21	1.48	0.00	94.4	2.34
	Factory1	6.38	0.36	0.00	0.00	7.86	85.4

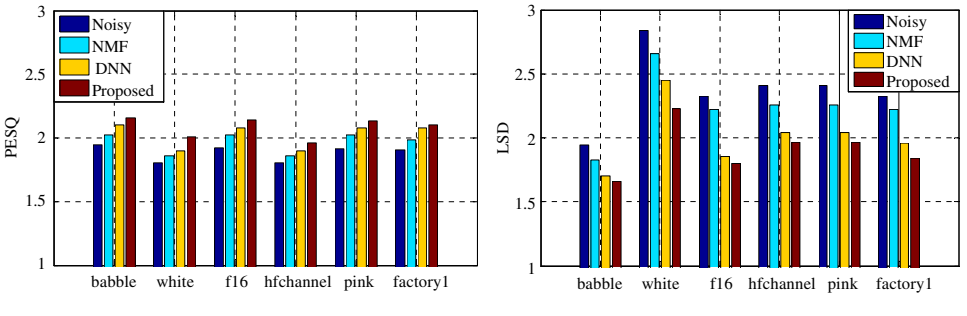

Fig. 1. (Color online) The PESQ and LSD scores of NMF, DNN and proposed method for six types of noise.

4. Conclusion

A speech enhancement method based on noise classification and DNN was proposed in this paper. Experimental results demonstrate that the proposed method could achieve better objective speech quality under stationary and non-stationary noise conditions. Our future work will focus on the generalization capability of the proposed system and study robust feature for noise classification and speech enhancement.

Acknowledgments

This work was supported partly by the National Nature Science Foundation of China (Nos. 61471394, 61402519) and the Natural Science Foundation of Jiangsu Province of China (Nos. BK2012510, BK20140071, BK20140074). The authors would like to thank the anonymous reviewer for helpful advice to improve the quality of this paper.

References

1. P. C. Loizou, *Speech Enhancement: Theory and Practice* (CRC Press, Boca Raton, 2007).
2. T. T. Vu, B. Bigot and E. S. Chng, in *Proc. Shanghai IEEE Int. Conf. Acoust. Speech, Signal Processing* (IEEE, USA, 2016), pp. 499–503.
3. Y. X. Wang, A. Narayanan and D. L. Wang, *IEEE Trans. Audio Speech* **22** (2014) 1849.
4. Y. Tu *et al.*, in *Proc. Hangzhou IEEE Int. Conf. Signal Processing* (IEEE, USA, 2014), pp. 532–536.
5. B. F. Wu and K. C. Wang, *IEICE Trans. Fund. Electr.* **89** (2006) 479.
6. B. Xia and C. Bao, *Speech Commun.* **60** (2014) 13.
7. Y. Xu *et al.*, *IEEE Trans. Audio Speech* **23** (2014) 7.
8. X. Glorot, A. Bordes and Y. Bengio, in *Proc. Int. Conf. AISTATS* (JMLR, USA, 2011), pp. 315–323.