A Robust Acoustic Feature Extraction Approach Based On Stacked Denoising Autoencoder

J. H. Liu, W. Q. Zheng, Y. X. Zou*
ADSPLAB/ELIP, School of Electronic and Computer Engineering
Peking University
Shenzhen, China
*{zouyx@pkusz.edu.cn}

***Abstract*—Acoustic feature extraction (AFE) is considered as one of the most challenging techniques for speech applications since the adverse environment noises always cause significant variation on the extracted acoustic features. In this paper, we propose a systematical AFE approach which based on stacked denoising autoencoder (SDAE) aiming at extracting acoustic features automatically. Denoising autoencoder (DAE), which is trained to reconstruct a clean "repaired" input from a corrupted version of it, works as the basic building block to form SDAE. Besides, the training set with clean and noisy speech ensures the SDAE has much powerful ability to extract the robust features under different noise conditions. Considering the speaker classification task using features extracted by the proposed approach for evaluation, intensive experiments have been conducted on TIMIT and NIST SRE 2004 to show SDAE with 3 hidden layers (3L-SDAE) gives better performance than shallow layers. The results also show that the features extracted by 3L-SDAE performs better than MFCC features when SNR is lower than 6dB and act more robustly when SNR decreases. What's more, for different types of noises at SNR of 0dB, the accuracy of speaker classification using 3L-SDAE features is higher than about 84% while MFCC features is lower than 77%.**

Keywords- robust acoustic feature extraction; stacked denoising autoencoder; noisy environment; speaker classification

I. INTRODUCTION

In traditional realization of speech applications, there are several famous and commonly used acoustic features, such as MFCC [1], LPCC [2] and PLP [3]. Among them, the most widely used acoustic features is Mel-frequency cepstral coefficients (MFCC for short), which takes advantage of source/filter deconvolution from the cepstral transform and perceptually realistic compression of spectra from the Mel pitch scale [4]. Obviously, MFCC is a kind of hand-crafted feature representations. Researches show that the hand-crafted features may not always achieve good performance in practical applications due to the complex application scenarios, consisting of different types of noises, non-stationary condition or different noise levels. Hence, it is desirable that the intrinsic acoustic features of the speech can be automatically extracted and robust to the noise, especially non-stationary noise.

Recent researches show that deep neural networks (DNNs) have strong modelling ability and have been successfully applied in various fields, such as face recognition [6], speech enhancement [7] and image classification [8]. Recent insightful progress in training deep architecture by Hinton [5] using a greedy layer-wise unsupervised learning procedure has resurrected the interest of the DNN. Specifically, there are three important stages in this strategy: firstly, pre-training one layer at a time in a greedy manner; then, using unsupervised learning at each layer in order to preserve information from the input; finally, fine-tuning the whole network with respect to ultimate criterion of interest. It has been reported that, in general, training DNN with greedy layer-wise algorithm can avoid local minima and gradient diffusion problems effectively.

Autoencoder is a kind of simple DNN models which aims to transform input into output with the least possible amount of distortion [9]. Recently, Y. Bengio proposed a new model called Denoising AutoEncoder (DAE) which is trained to reconstruct a clean "repaired" input from a corrupted version of it, and provide a new solution to extract robust features. In practical, several DAEs are stacked to form SDAE which is trained in an unsupervised bottom-to-up manner, and then a supervised learning is conducted to train the top layer and fine-tune the entire architecture, which has been verified significant effect of denoising and successfully applied in noisy image classification and recognition [10].

Motivated by the key points discussed above, in this paper, we focus on developing a robust acoustic feature extraction system based on stacked DAE (AFE-SDAE). The unlabeled spectrograms are used in SDAE pre-training stage in an unsupervised manner, and then the AFE-SDAE system is fine-tuned with a few labeled spectrograms in a supervised manner. The effectiveness of the extracted acoustic features as well as its robustness to noise has been evaluated under different SNR levels and different noise types. It is encouraged to see that, compared to MFCC feature, the acoustic features extracted by 3L-SDAE are more robust when SNR is lower than 6dB, and perform better under different noisy condition at SNR of 0dB.

The rest of this paper is organized as follows. Sect. II introduces the architecture of denoising autoencoder; the proposed acoustic feature extraction system is described in Sect. III; Sect. IV shows the experimental setup and results; Sect. V concludes the paper.

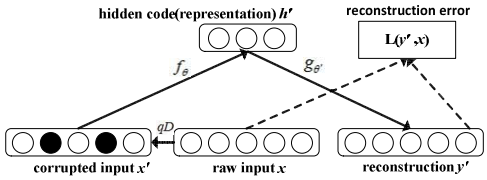


Figure 1. The architecture of denoising autoencoder

978-1-4799-8688-0/15 $31.00 © 2015 IEEE
DOI 10.1109/BigMM.2015.46

II. DENOISING AUTOENCODER

A denoising autoencoder is shown in Fig. 1, which is trained to reconstruct a clean "repaired" input from a corrupted version of it. This is done by first corrupting the initial input $\boldsymbol{x}$ into its corrupted version $\boldsymbol{x'}$ by means of a stochastic mapping $\boldsymbol{x'} \sim qD(\boldsymbol{x'}|\boldsymbol{x})$. Corrupted input $\boldsymbol{x'}$ is then taken as the input of an autoencoder [9] where the desired output is $\boldsymbol{x}$. As shown in Fig.1, a hidden representation will be $\boldsymbol{h'}=f_\theta(\boldsymbol{x'})$ and the reconstruction is $\boldsymbol{y'}=g_{\theta'}(\boldsymbol{h'})$. The parameters of the network (θ and θ') will be trained to minimize the average reconstruction error over a training set, that is, to make $\boldsymbol{y'}$ as close as possible to the uncorrupted input $\boldsymbol{x}$. Compared to the standard autoencoder, the key difference is that $\boldsymbol{y'}$ is now a deterministic function of $\boldsymbol{x'}$ rather than $\boldsymbol{x}$. Obviously, this approach achieves a far more flexible mapping than that of standard autoencoder and leads to suppress the adverse effect of the noise in feature extraction [9, 11]. As the result, the reconstruction error can be described as $L(\boldsymbol{y'},\boldsymbol{x})=\|\boldsymbol{y'}-\boldsymbol{x}\|^2$ with an affine decoder.

III. THE PROPOSED AFE-SDAE SYSTEM

The proposed AFE-SDAE system is illustrated in Fig. 2. In the feature learning stage, the SDAE is pre-trained and fine-tuned with a collection of spectrograms corresponding to clean and noisy speech without label or with label. Specifically, the clean and noisy speech is pre-processed into spectrograms, and then the unlabeled spectrograms are used in SDAE pre-training while the labeled spectrograms are used in SDAE fine-tuning. In the feature extraction stage, spectrogram of noisy speech is fed to the well-trained SDAE and the output of the SADE is the extracted acoustic features.

A. Pre-processing

Pre-processing of speech waveform is shown in Fig. 3. The step "processing" in Fig. 3 contains sampling of the speech waveform, quantization, pre-emphasis and windowing. The output of the pre-processing block is a spectrogram of clean or noisy speech. Namely, each utterance will be transformed into a spectrogram $\boldsymbol{x}$ ($\boldsymbol{x} \in \mathbb{R}^{m \times n}$), where m and n represents frequency and frame number respectively. Then the mean normalization of $\boldsymbol{x}$ is performed and a normalized spectrogram is denoted as:

$$\boldsymbol{x\_valid} = (\boldsymbol{x} - \boldsymbol{\mu}) / \boldsymbol{\delta} \quad (1)$$

where $\boldsymbol{\mu}$ is the mean vector and $\boldsymbol{\delta}$ is the standard deviation. $\boldsymbol{x\_valid}$ is used as the input of SDAE.

B. SDAE Pre-training

The block of SDAE pre-training is illustrated in Fig. 4. Specifically, Fig. 4 (a) shows the pre-training process of the

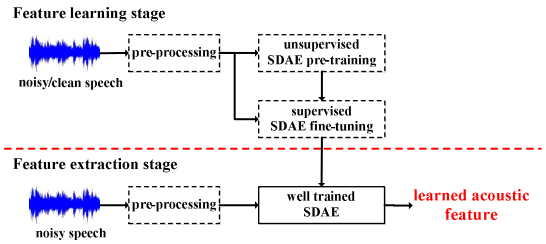


Figure 2. Block diagram of the proposed AFE-SDAE system

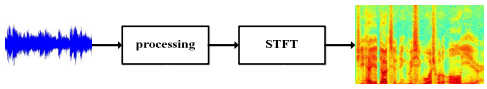


Figure 3. Pre-processing of each noisy utterance

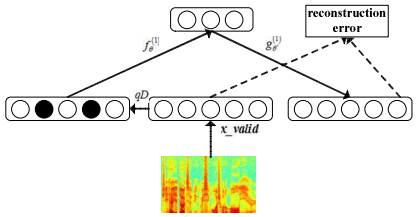
first layer of SDAE, of which the input is spectrogram corresponding to noisy speech. The noisy spectrogram as corruption is only used for the initial denoising-training of each individual layer. Hence, once the weight matrix W^1 representing the mapping $f_\theta^{(1)}$ of the first layer has been learnt, the resulting representation of the first layer is applied as the input of the second layer of SDAE. The process of pre-training the second layer is shown in Fig. 4 (b), in which the training algorithm is similar to that of the first layer. The SDAE is trained layer-by-layer in an unsupervised greedy fashion. Essentially, the pre-training step is to offer good initialization parameters for the fine-tuning stage.

C. SDAE Fine-tuning

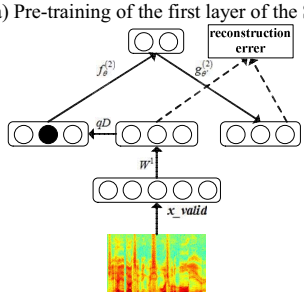
After the pre-training stage, the parameters of all the layers of SDAE which denote as W^1, W^2, ⋯, W^{hy_num} are well pre-trained. Then the output of the last layer on the top can be used as feature representation to a stand-alone classifier. As illustrated in Fig. 5, a softmax classifier [12] is added to the top of the SDAE in this paper, in which the parameters of all layers including classification layer namely W^1, W^2, ⋯, W^{hy_num}, W^{sup} can then be jointly fine-tuned to minimize the error in predicting the supervised target (e.g., speaker class) by performing BP algorithm. With the accomplishment of the fine-tuning process, the well-trained SDAE is formed and can be used to extract acoustic features from spectrogram.

IV. EXPERIMENTS AND RESULT ANALYSIS

In this section, we mainly conduct the experiments to demonstrate the performance of the proposed AFE-SDAE system. Only for the purpose of visualization, we take the speaker classification task as an example.



(a) Pre-training of the first layer of the SDAE



(b) Pre-training of the second layer of the SDAE

Figure 4. Description of SDAE pre-training

A. Experiment Setup

Without loss of generality, the following experiments are conducted on the commonly used databases TIMIT and NIST SRE 2004 corpus. For simplicity, the subset of TIMIT database [13] and the subset of the NIST SRE 2004 corpus have been considered. The subset we choose contains 168 speakers with 10 utterances per speaker and 168 speakers with training segment duration of about 5 minutes of a conversation, respectively. Besides, in order to verify the robustness of the proposed feature learning approach, the additive noises from NoiseX-92 database have been used to generate the noisy data with different type noises at different SNR levels.

The proposed AFE-SDAE system consists of more than one hidden layer with 150 units in each layer. The optimal number of hidden layer is discussed in our experiments. The number of epoch for each layer's pre-training is 50. Learning rate of pre-training is set to 0.5. We use the minimum mean square error as the loss function of fine-tuning. The activation function is sigmoid function. The input of SDAE is a 161-dimension vectors corresponding to the spectrogram of the utterance which is sampled at 8000Hz and segmented by a 20ms Hamming window with 10ms overlaps. Meanwhile, the input vectors are normalized to zero mean and unit variance. In the training phase, the training set which being a collection of 7 utterances per speaker randomly selected from the database is used to train the SDAE model. For speaker classification application, we utilize the softmax classifier consisting of 168 units to classify 168 speakers [13] and the learned acoustic feature is viewed as the input of classifier. In the speaker classification trails, the rest utterances of each speaker from database with combination of noisy types and SNR levels are selected to generate the test set. The classification accuracy is termed as the performance measure for the proposed learned features. The experimental results are averaged over 10 trials.

B. The Robustness to Noise

In order to show the denoising ability of the AFE-SDAE on speaker classification task, the results of AFE-SAE (SAE is short for Stacked Autoencoders) are compared. The difference between AFE-SDAE and AFE-SAE lies on whether the input is corrupted by noise. The experiment is conducted on the subset of TIMIT database under both noise-free and 0dB additive random noise conditions.

The speaker classification accuracy is shown in Table. I. The results illustrate that the performance of the acoustic features extracted by AFE-SDAE is comparable to that of AFE-SAE under noise-free condition. But the performance of the acoustic features extracted by AFE-SAE greatly degrades in 0dB noise environment, while the one by AFE-SDAE almost maintains the performance under noise-free condition. These experimental results indirectly show that the AFE-SDAE is able to extract the acoustic features robust to noise.

C. The Impact of the Depth of AFE-SDAE

As discussed above, AFE-SDAE is a deep network. Experiment in this section aims at evaluating the impact of the depth of AFE-SDAE on its performance, which means to find the optimal number of hidden layers for the specific speaker

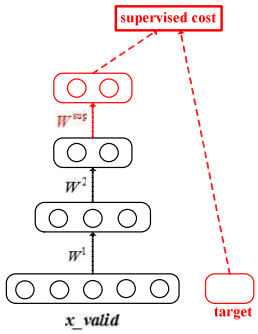


Figure 5. Illustration of two-layer SDAE fine-tuning for classification

classification task. The dataset used here is the subset of both noiseless TIMIT speech and noisy NIST2004 SRE corpus. Three network structures are considered: the single-layer, two-layer and three-layer AFE-SDAE, which are represented as SL-SDAE, 2L-SDAE and 3L-SDAE for short, respectively. The performance of the speaker classification experiment is given in Table. II, in which the result shows that the performance on different database is slightly different, but we observe the similar fact that the classification accuracy gets better with the increase of the number of hidden layers. Generally speaking, for the speaker classification task, 3L-SDAE gives the best result. The reason may be that more hidden layers increase the depth of the network which can learn more intrinsical information from original speech data.

D. Performance Comparision with MFCC

Following the results in Section C, we will further evaluate the performance of the SL-SDAE, 2L-SDAE and 3L-SDAE under different noisy conditions on speaker classification task, where the SNR level varies from 0 to 25dB incremented by 5dB. As comparison, we take the conventional MFCC feature [14] into consideration. The experimental results are presented in Fig. 6. It is quite clear to see that the accuracy of speaker classification with features extracted by 3L-SDAE is higher than that with features extracted by 2L-SDAE and SL-SDAE at all SNR level. It also shows that the accuracy of speaker classification using the 3L-SDAE features is much higher than that using MFCC features when SNR is less than about 6dB (refers to low SNR noise condition). However, the case goes opposite when SNR goes higher. This further validates that MFCC features hold intrinsic characteristics of speech, but it is easily corrupted by strong noise. It is also encouraged to see the feature extraction capability of the AFE-SDAE under low SNR

TABLE .I COMPARISONS OF ACCURACY OF AFE-SAE AND AFE-SDAE

Network Architecture	**Experiment Condition**	
	without noise	***SNR at 0dB noise***
AFE-SAE	87.16%	68.75%
AFE-SDAE	**87.41%**	**83.86%**

TABLE .II SPEAKER CLASSIFICATION ACCURACY VS DEPTH

Database Type	**number of hidden layer**		
	SL-SDAE	***2L-SDAE***	***3L-SDAE***
TIMIT	87.24%	89.72%	**91.01%**
NIST 2004	83.75%	85.19%	**86.74%**

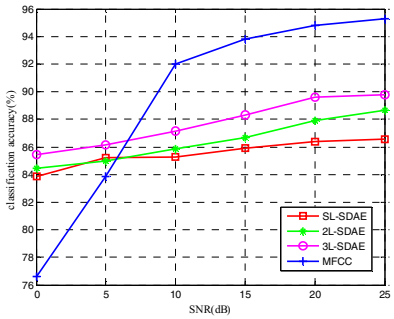


Figure 6. speaker classification accuracy using three different features

condition.

E. Performance Under Different Types of Noises

From Section D we know that the acoustic features extracted by 3L-SDAE have got the best performance in speaker classification tasks, so we further evaluate the performance of AFE-SDAE by comparing 3L-SDAE features and MFCC on speaker classification task under different types of noises at the SNR of 0dB. Other settings are the same as Section D. The noises consist of white noise, babble noise, buccaneer noise, destruction noise and factory noise taken from NoiseX-92. The average accuracy with 10 trials on each group was plotted in Fig. 7. From the experimental results, it is clear to see that for different types of noises, the average accuracy using 3L-SDAE features ranges from about 83% to 86%, while that using MFCC features is below about 77%. It tells that the acoustic features extracted by 3L-SDAE is quite robust to different noise types and outperform MFCC under low SNR level conditions, which is a desirable properties for real applications since the noise environment is non-stationary and unpredictable.

V. CONCLUSION

In this paper, a robust acoustic feature extraction approach based on stacked denoising autoencoder (AFE-SDAE) has been developed and investigated. The training strategy is derived following the recent proposed greedy layer-wise training algorithm. In training stage, the spectrograms generated from the subset of TIMIT and NIST2004 SRE corpus have been used to train the proposed AFE-SDAE in both unsupervised and supervised manners. In speech applications, the system takes spectrogram as input and extracts acoustic features automatically. Intensive experimental results have validated that the acoustic features extracted by the proposed AFE-SDAE system performs better and more robust than MFCC when SNR decreases less than 6dB. Furthermore, the accuracy of speaker classification using the proposed features is higher than about 84% (while lower than about 77% for MFCC features) under different noise conditions at the SNR of 0dB, such as babble noise, buccaneer noise, destroy noise and factory noise. The further research will focus on improving the performance of AFE-SDAE under different SNR conditions.

ACKNOWLEDGMENT

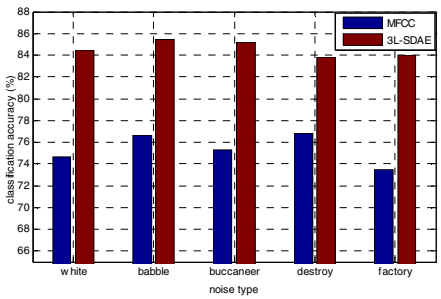


Figure 7. speaker classification accuracy under different type of noises

This work is partially supported by National Natural Science Foundation of China (No: 61271309) and Shenzhen Science Research Program (No. CXZZ20140509093608290).

REFERENCES

[1] W. Han, C.-F. Chan, C.-S. Choy, and K.-P. Pun, "An efficient MFCC extraction method in speech recognition," in Circuits and Systems, 2006. ISCAS 2006. Proceedings. 2006 IEEE International Symposium on, 2006, p. 4 pp.

[2] Y. Yujin, Z. Peihua, and Z. Qun, "Research of speaker recognition based on combination of LPCC and MFCC," in Intelligent Computing and Intelligent Systems (ICIS), 2010 IEEE International Conference on, 2010, pp. 765-767.

[3] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," the Journal of the Acoustical Society of America, vol. 87, pp. 1738-1752, 1990.

[4] P. Hamel and D. Eck, "Learning features from music audio with deep belief networks," in ISMIR, 2010, pp. 339-344.

[5] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," Science, vol. 313, pp. 504-507, 2006.

[6] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back, "Face recognition: a convolutional neural-network approach," Neural Networks, IEEE Transactions on, vol. 8, pp. 98-113, 1997.

[7] Y. Xu, J. Du, L. Dai, and C. Lee, "An experimental study on speech enhancement based on deep neural networks," 2014.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in neural information processing systems, 2012, pp. 1097-1105.

[9] P. Baldi, "Autoencoders, Unsupervised learning, and deep architectures," in ICML Unsupervised and Transfer Learning, 2012, pp. 37-50.

[10] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in Proceedings of the 25th international conference on Machine learning, 2008, pp. 1096-1103.

[11] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion," The Journal of Machine Learning Research, vol. 11, pp. 3371-3408, 2010.

[12] S. Gold and A. Rangarajan, "Softmax to softassign: neural network algorithms for combinatorial optimization," Journal of Artificial Neural Networks, vol. 2, pp. 381-399, 1996.

[13] H. Lee, P. Pham, Y. Largman, and A. Y. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," in Advances in neural information processing systems, 2009, pp. 1096-1104.

[14] L. Wang, K. Minami, K. Yamamoto, and S. Nakagawa, "Speaker identification by combining MFCC and phase information in noisy environments," in Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on, 2010, pp. 4502-4505.