INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

SNR-Aware Convolutional Neural Network Modeling for Speech Enhancement

Szu-Wei Fu [12], *Yu Tsao*[1], *Xugang Lu*[3]

[1] Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan
[2] Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan
[3]National Institute of Information and Communications Technology, Kyoto, Japan

{jasonfu,yu.tsao}@citi.sinica.edu.tw; xugang.lu@nict.go.jp

Abstract

This paper proposes a signal-to-noise-ratio (SNR) aware convolutional neural network (CNN) model for speech enhancement (SE). Because the CNN model can deal with local temporal-spectral structures of speech signals, it can effectively disentangle the speech and noise signals given the noisy speech signals. In order to enhance the generalization capability and accuracy, we propose two SNR-aware algorithms for CNN modeling. The first algorithm employs a multi-task learning (MTL) framework, in which restoring clean speech and estimating SNR level are formulated as the main and the secondary tasks, respectively, given the noisy speech input. The second algorithm is an SNR adaptive denoising, in which the SNR level is explicitly predicted in the first step, and then an SNR-dependent CNN model is selected for denoising. Experiments were carried out to test the two SNR-aware algorithms for CNN modeling. Results demonstrate that CNN with the two proposed SNR-aware algorithms outperform the deep neural network counterpart in terms of standardized objective evaluations when using the same number of layers and nodes. Moreover, the SNR-aware algorithms can improve the denoising performance with unseen SNR levels, suggesting their promising generalization capability for real-world applications.

Index Terms: speech enhancement, convolutional neural network, denoising autoencoder, multi-task learning

1. Introduction

The goal of speech enhancement (SE) is to improve the intelligibility and quality of a noisy speech signal. In the past, various SE approaches have been developed. Successful examples include spectral subtraction [1], minimum-mean square error (MMSE) based spectral amplitude estimator [2], Wiener filtering [3], Karhunen-Loéve transformation (KLT) [4], and non-negative matrix factorization (NMF) [5]. Although these approaches can effectively remove noise components from noisy speech signals, there is still room for further improvement of their performance, especially in very challenging acoustic conditions (e.g., low SNR and non-stationary noises [6]).

Recently, deep learning based SE approaches have been proposed and extensively investigated [7-10]. For this type of approach, a deep-structure model is used to predict the clean log-power spectra (LPS) features from noisy LPS features. In [7, 11], a deep neural network (DNN)-based SE approach was demonstrated to be better than traditional SE models. Moreover in terms of computational load, the DNN–based SE methods are more efficient when compared with the NMF-based SE algorithm, where an iterative optimization is required. Although DNN-based SE algorithms have achieved considerable successes, two significant issues remain unsolved and require further improvement: (1) characterization of the local temporal-spectral structures of speech signals, and (2) explicitly considering SNR information to achieve adaptive denoising.

For the first issue, because the DNN model processes speech signals in a fully-connected manner, the local temporal-spectral structures of speech signals may not be effectively characterized. Contrarily, the architecture of a convolutional neural network (CNN) is designed to take advantage of the 2D-structured input by using local connections to focus on local patterns. Compared to DNN, CNN may be more suitable for SE tasks since it can pay more attention to neighboring regions around each time-frequency (T-F) unit. Because CNN can model spatial and temporal correlations and reduce translational variance in signals [12], it has proved notably successful in the image recognition and computer vision fields [13]. Recently, CNN has also been applied to speech recognition [12, 14] where again it achieved better recognition accuracy than DNN. Meanwhile, Zhao *et al.*[15] proposed a music removal model based on CNN for speech recognition and obtained better recognition results compared with DNN. Hui et al. also employed CNN to separate speech and noise by estimating the ideal ratio mask of the time-frequency units [16]. Based on the successes of the abovementioned CNN for improving performance, this study investigates the capability of CNN for the SE task.

For the second issue, it has been shown that SE performance can be degraded by the mismatch of training and test conditions, particularly noise types and SNRs [11]. To deal with this issue, Xia *et al.* [17] employed GMM to classify the noise type before feeding the noisy speech signal into the DNN model. However, it may be difficult to generalize this approach to an unseen noise type since it is a classification problem. Meanwhile a noise-aware training criterion has been proposed to incorporate noise information in the input feature, thereby making DNN aware of the type of noise [18]. Aside from dealing with the noise type, this paper proposes two SNR-aware algorithms to enable the deep denoising model to effectively utilize the SNR information to achieve better SE performance. The first algorithm applies a multi-task learning (MTL) framework to estimate clean speech signals (the main task) together with the SNR level (the secondary task) for a noisy input speech. By the MTL, the trained CNN model can be implicitly aware of the SNR levels. The second algorithm is

http://dx.doi.org/10.21437/Interspeech.2016-211

an SNR adaptive denoising (SNR_AD), which consists of offline and online stages. In the offline stage, a set of SNR-specific denoising models is prepared, where each model is trained by the noisy/clean pair using noisy speech within a certain SNR range. Unlike predicting the noise types, which is a classification problem [17], estimating the SNR level is a regression problem. In other words, the estimated level has a numerical relationship, and thus, can be generalized for an unseen SNR level. In the online stage, SNR_AD first predicts the SNR level of the noisy input, and then an SNR-dependent CNN model is selected for denoising.

2. CNN model for denoising

2.1. Regression model for denoising

The SE framework used in this paper resembles the one used in [7]. In the training stage, a set of noisy/clean speech pairs is prepared. The noisy and clean speech signals are first converted into the frequency domain, and their LPS features are then placed as the input and output, respectively, to train a regression model. In the enhancement stage, the noisy LPS features are fed into the trained model to produce the enhanced LPS features. We can synthesize the enhanced speech signal in the time domain together with the phase information, which is borrowed from the original noisy speech. In [7], a DNN model is used as the regression model. CNN can deal with the local temporal-frequency structure of signal, which may be much more efficient in disentangling noise and speech features than the DNN model. Therefore, in this study, we chose to investigate a CNN model for SE.

2.2. The CNN SE framework

A CNN model may consist of one or more pairs of convolution and max-pooling layers. A convolution layer applies a set of filters to extract features, and a max-pooling layer generates a lower resolution version by taking the maximum filter activation. Max-pooling layers make the output of convolution networks translationally invariant. This is a desired property in speech (image) recognition since it can increase the robustness to speaker (object) variations and improve accuracy [13, 14]. However, pooling may also cause the network to lose information about the detailed structure and textures [19], and thus may not be suitable for SE applications. In Section 4, we will conduct a series of experiments to confirm this conjecture.

3. SNR-aware algorithms

3.1. Multi-task learning

One possible way to exploit the SNR information is to apply multi-task learning (MTL), with the intention that the learned model can be implicitly aware of the level of noise it faces. MTL learns a target problem together with other related problems at the same time, using a shared representation. This often leads to a better model for the main task, because it allows the model to use the commonality among the tasks [20]. Xu *et al.* [21] proposed a DNN structure by estimating clean LPS and MFCC features together to achieve better enhancement performance. To embed the ability of SNR estimation into the learned model, DNN/CNN should jointly estimate the primary LPS features together with a secondary task, namely the SNR level of the noisy input.

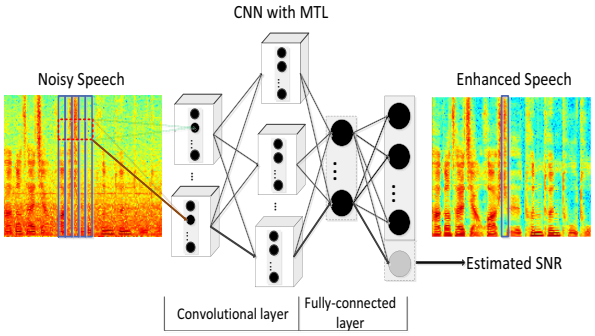


Figure 1: *Structure of the proposed CNN with MTL for SE.*

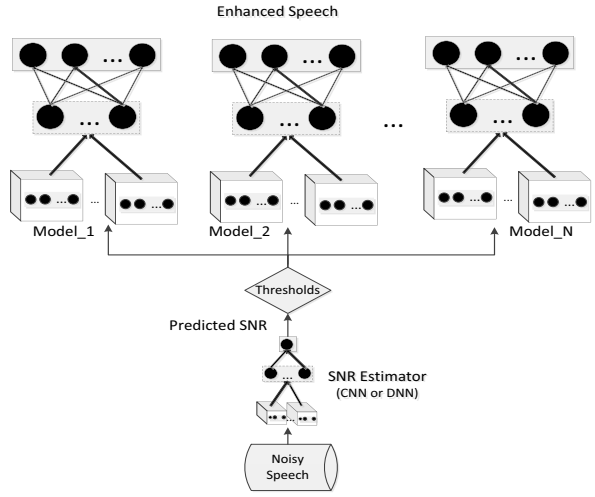


Figure 2: *Structure of CNN with SNR_AD for SE.*

The idea is illustrated in Fig.1. In this figure, we can see, in denosing, instead of applying conventional mean square error (MSE) between clean and estimated LPS as the cost function, the new objective to be minimized is augmented as follows:

$$J(\Theta) = \frac{1}{N}\left[\sum_{i=1}^{N}\|\mathbf{y}_i - \hat{\mathbf{y}}_i\|_2^2 + \lambda\sum_{i=1}^{N}(s_i - \hat{s}_i)^2\right] \qquad (1)$$

where $\mathbf{y}_i$ and $\hat{\mathbf{y}}_i$ denote the clean and estimated LPS feature vectors, respectively, at sample index i, with N representing the total training data size. Variables s_i and $\hat{s}_i$ denote the true and estimated SNR levels of the noisy input frame at index i, respectively, while λ is the weighting factor of the two targets..

3.2. SNR adaptive denoising

Another relatively intuitive method to exploit the SNR information is to enhance the noisy speech by using different denoising models according to the noise strength. The overall framework of the proposed SNR_AD scheme is illustrated in Fig. 2. To achieve this goal, an SNR estimator has to be applied beforehand. In this study, the estimator is another DNN/CNN model that is trained with noisy LPS features to predict its true SNR value. Then a decision is made by comparing the estimated level to some predefined thresholds to decide which denoising model is most suitable for the current input frame. The decision process is defined as follows:

$$m = \begin{cases} 1; & s_i > \rho_1 \\ k; & \rho_{k-1} > s_i > \rho_k \\ K; & \rho_{K-1} > s_i \end{cases}, \forall K > i > 1 \qquad (2)$$

where s_i is the estimated SNR level, K is the total number of denoising models, m is the decision output to apply the m-th model, and ρ_k is the k-th threshold which is dependent on K and the SNR level used in the training set. Thresholds are set to be in descending order, i.e., $\rho_1 > \rho_2 > \cdots > \rho_{K-1}$. Note that the SNR estimator itself is a regression model, and the final categorical result will be obtained only after the decision process in (2).

Intuitively, each denoising model should only be trained on the noisy speech within a certain range of SNR. Hence, a better performance at that specific noise strength can be expected. However, to achieve better generalization and to fully utilize the training data, the weights in each denoising model are initialized to those in the universal model that was trained using all the data. The weights are then fine-tuned based on the training data within the specific SNR range.

4. Experiments

4.1. Experimental setups

In our experiments, the Mandarin version of Hearing in Noise Test (MHINT) corpus [22] was used to prepare the training and test sets. The MHINT corpus includes 240 utterances, and we collected another 240 utterances from the same speaker to form the complete task in this study. Among these 480 utterances, 250 utterances were excerpted and corrupted with five noise types (Babble, Car, Jackhammer, Pink, and Street), at five SNR levels (-10dB, -5dB, 0dB, 5dB, and 10dB), to build a 3.5 hours training set. Another 50 utterances were mixed to form the test set. In this study, we consider a more realistic condition, where both noise types and SNR levels of the training and test sets were mismatched. Thus we intentionally adopted two other noise signals (White Gaussian noise (WGN), a stationary noise) and (Engine, a non-stationary noise), with another five SNR levels: -12dB, -6dB, 0dB, 6dB, 12dB (different from the training conditions) to form the test set.

In this work, 257 dimensional LPS were extracted from the speech waveforms as the acoustic features; in the meanwhile, to model the context information, multiple frame expansion was applied to extend the input to 5 frames [7]. Mean and variance normalization was applied to the input feature vectors to make training process more stable. To evaluate an SE algorithm, two aspects must be considered: noise reduction and speech distortion Therefore, segmental SNR (SSNR in dB) and mean square error (MSE) between the enhanced LPS and clean LPS were used as objective evaluations to measure noise reduction and speech distortion, respectively, of the enhanced speech. For a fair comparison, both CNN (3 convolutional layers plus 2 fully connected layers) and DNN were fixed to five hidden layers. Each corresponding layer has the same number of nodes. Because the CNN model has a weight sharing and local connectivity structure, the trainable weights in CNN are less than that of DNN. Therefore except the dropout rate in the first three layers, all the other hyperparameters were set to be the same in the DNN and CNN models.

For the experiments of MTL, the weighting factor, λ in Eq. (1), was set to 0.006 empirically. For the experiments of SNR_AD, the number of denoising models, K in Eq. (2), was set to 3, and the two thresholds were set as $\rho_1 = 5, \rho_2 = -5$, according to the SNR levels in the training set.

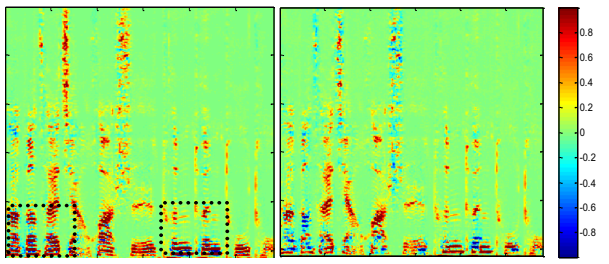
4.2. Experimental results

4.2.1. The effect of max-pooling on CNN-based SE

In this section, we investigate the SE performance of CNN denoising with and without max-pooling. The results of Engine and WGN noises have very similar trends, and thus only the Engine noise results are presented in this section. Table 1 demonstrates the average MSE and SSNR scores of CNN with and without max-pooling under the Engine noise over five SNR levels. From the table, we observe that CNN without max-pooling outperforms CNN with max-pooling in terms of both MSE and SSNR. Additionally, Fig. 3 presents the differences of clean spectrogram and CNN enhanced spectrogram (a) with max-pooling, and (b) without max-pooling. For a clearer comparison, both differences are processed by a tanh function to normalize the output within the range of -1 to +1. In the regions within the black rectangles, we note that blue (negative values) and red (positive values) parts appear alternately, implying that the denoising model perform SE in a relatively inaccurate T-F positions caused by the max-pooling process.

Table 1: *Average MSE and SSNR scores for the Engine noise over five SNR levels by CNN with and without max-pooling.*

Method	MSE	SSNR
With max-pooling	0.828	1.423
Without max-pooling	0.816	2.090



(a) (b)

Figure 3: *Normalized differences of the clean spectrogram and CNN enhanced spectrograms: (a) with max-pooling; (b) without max-pooling under Engine noise at 6 dB SNR.*

4.2.2. SNR estimation

For the SNR-aware algorithms, an accurate estimation of the SNR level given the test utterance is a critical requirement. Therefore in this section, we intend to investigate the SNR estimation accuracies of DNN and CNN. In the training phase, the target output of both models was a true SNR value given each noisy speech utterance input. It should be noted that the noise types of the training and test conditions were different. Figure 4 shows the mean and standard deviation (STD) of the predicted SNRs at target SNRs on the test set. From Fig 4, we observe two results: (1) the accuracies of DNN and CNN on SNR prediction are similar; (2) although the absolute SNR prediction accuracies are not perfect, we observe notable differences of predicted SNR results across five distinct SNR inputs. This set of results suggests that the DNN/CNN model can be suitably used to perform SNR estimation to implement the SNR-aware algorithms.

Table 2. *Average MSE and SSNR scores on the test set consists of five different SNRs with the two unseen noise environments, among: DNN baseline, proposed CNN, CNN with multi-task learning (CNN+MTL), CNN with SNR adaptive denoising (CNN+SNR_AD), and oracle SNR adaptive denoising with given true SNR level (CNN+SNR_AD(O)).*

	DNN (baseline)		CNN		CNN+MTL		CNN+SNR_AD		CNN+SNR_AD(O)	
SNR	MSE	SSNR	MSE	SSNR	MSE	SSNR	MSE	SSNR	MSE	SSNR
12	0.713	5.103	0.725	5.280	0.681	5.509	**0.677**	**5.639**	0.669	5.846
6	0.828	4.083	0.807	4.367	**0.785**	**4.550**	0.789	4.536	0.796	4.779
0	0.999	2.517	0.936	2.964	**0.930**	**3.053**	0.939	3.019	0.961	2.968
-6	1.371	0.331	**1.257**	1.047	1.268	**1.070**	1.282	0.980	1.271	1.054
-12	1.716	-2.166	**1.586**	-1.392	1.593	-1.344	1.591	**-1.310**	1.599	-1.330
Ave	1.125	1.973	1.062	2.453	**1.051**	2.567	1.055	**2.573**	1.059	2.663

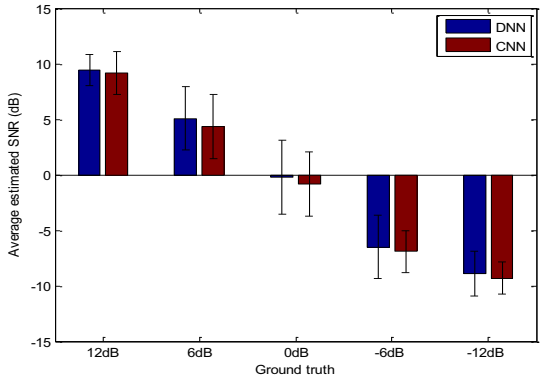


Figure 4: *Mean and standard deviation (STD) of the SNR estimations carried out by DNN and CNN.*

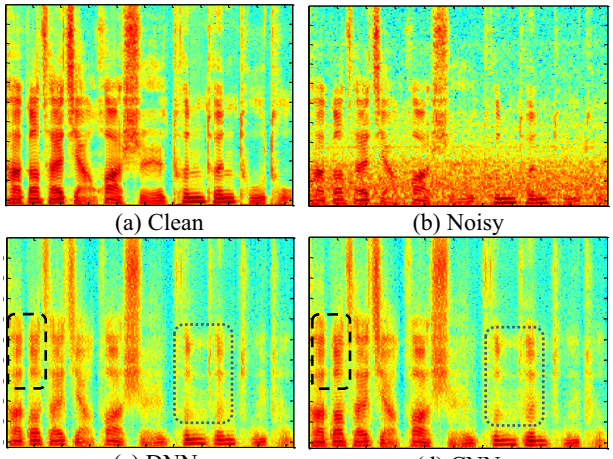


Figure 5: *Four spectrograms of a MHINT utterance corrupted by Engine noise at SNR=6dB: (a) clean speech (b) noisy speech (c) enhanced by DNN (d) enhanced by CNN.*

4.2.3. MTL and SNR_AD for CNN denoising

Figure 5 presents the spectrograms of (a) a clean utterance excerpted from MHINT, (b) the same utterance corrupted by the Engine noise at SNR=6dB, (c) DNN enhanced utterance, and (d) CNN enhanced utterance. Form the figure, we first note that both DNN and CNN can effectively remove noise components from the noisy utterance. Moreover, (d) CNN can keep more speech details around the formant parts (indicated by the black dash rectangles) when compared with (c) DNN. In addition to a qualitative (visual) comparison using Fig. 5, Table 2 shows the quantitative results of the average MSE and SSNR scores on the test set, among: DNN, CNN, CNN with MTL (denoted as CNN+MTL), CNN with SNR_AD (denoted as CNN+ SNR_AD). From the table, CNN provides lower MSE and higher SSNR scores compared to DNN in most SNR levels, with a 0.5 dB SSNR improvement on average. The result implies that CNN can more effectively model the local T-F characteristics. We also note that CNN+MTL outperforms CNN consistently among all of the SNR levels, confirming the effectiveness of incorporating the SNR information into the objective function for denoising. This may be owing to the second term (target) in (1) serve as a regularization term [23] to cause the CNN model to conservatively remove noise components while avoiding speech distortion.

From Table 2, we also note that CNN+SNR_AD outperforms CNN in most cases. This result is consistent with CNN+MTL: combining SNR information can effectively improve the performance of CNN denoising. The advantage of SNR_AD is especially noted at high SNR input (at 12 dB) , which implies that using a universal model for denoising would cause speech distortion. From Fig. 4, we note that the SNR estimation is not perfected in our current system. To further investigate the upper-bound performance of SNR_AD, we carried out an oracle experiment (CNN+SNR_AD(O) in Table 2, which assumes that the true SNR level was given (and thus the accurate SNR-dependent denoising model was selected) to perform SE. From Table 2, we note that CNN+SNR_AD(O) outperforms CNN+SNR_AD, showing that for the tasks where the accurate SNR level is accessible, the performance of CNN+SNR_AD can be further enhanced. From the results in Table 2, we conclude that both proposed SNR-aware algorithms can effectively improve the CNN denoising capability while SNR_AD requires more computational cost than MTL.

5. Conclusions

The contribution of this paper is two-fold. First, we confirm that CNN can effectively extract the local T-F features of speech signals and thus achieve better SE performance than DNN. Meanwhile, we find that max-pooling may not be necessary for SE tasks because of its reduced capability of characterizing detailed speech patterns. Second, we proposed two SNR-aware algorithms and proved that both algorithms can enable CNN model with improved denoising performance. In the future, we will explore other possible cost functions with advanced neural network architectures for the SE task.

6. References

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech and Signal Processing,* vol. 27, pp. 113-120, 1979.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing,* vol. 32, pp. 1109-1121, 1984.

[3] P. Scalart, "Speech enhancement based on a priori signal to noise estimation," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1996, pp. 629-632.

[4] Y. Hu and P. C. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Transactions on Speech and Audio Processing,* vol. 11, pp. 334-341, 2003.

[5] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2008, pp. 4029-4032.

[6] T. Gao, J. Du, Y. Xu, C. Liu, L.-R. Dai, and C.-H. Lee, "Improving Deep Neural Network Based Speech Enhancement in Low SNR Environments," in *Latent Variable Analysis and Signal Separation*, ed: Springer, 2015, pp. 75-82.

[7] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters,* vol. 21, pp. 65-68, 2014.

[8] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing* vol. 23, pp. 7-19, 2015.

[9] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *INTERSPEECH*, 2013, pp. 436-440.

[10] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for joint enhancement of magnitude and phase," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 5220-5224.

[11] D. Liu, P. Smaragdis, and M. Kim, "Experiments on deep learning for speech denoising," in *INTERSPEECH*, 2014, pp. 2685-2689.

[12] T. N. Sainath, A.-r. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for LVCSR," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 8614-8618.

[13] 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.

[14] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, and G. Penn, "Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 4277-4280.

[15] M. Zhao, D. Wang, Z. Zhang, and X. Zhang, "Music removal by denoising autoencoder in speech recognition," *APSIPA,* 2015.

[16] L. Hui, M. Cai, C. Guo, L. He, W.-Q. Zhang, and J. Liu, "Convolutional maxout neural networks for speech separation," in *2015 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, 2015, pp. 24-27.

[17] B. Xia and C. Bao, "Wiener filtering based speech enhancement with weighted denoising auto-encoder and noise classification," *Speech Communication,* vol. 60, pp. 13-29, 2014.

[18] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "Dynamic noise aware training for speech enhancement based on deep neural networks," in *INTERSPEECH*, 2014, pp. 2670-2674.

[19] Y. Taigman, M. Yang, M. A. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1701-1708.

[20] R. Caruana, "Multitask Learning: A Knowledge-Based Source of Inductive Bias1," in *Proceedings of the Tenth International Conference on Machine Learning*, 1993, pp. 41-48.

[21] Y. Xu, J. Du, Z. Huang, L.-R. Dai, and C.-H. Lee, "Multi-Objective Learning and Mask-Based Post-Processing for Deep Neural Network Based Speech Enhancement," in *INTERSPEECH*, 2015.

[22] L. L. Wong, S. D. Soli, S. Liu, N. Han, and M.-W. Huang, "Development of the Mandarin hearing in noise test (MHINT)," *Ear and hearing,* vol. 28, pp. 70S-74S, 2007.

[23] R. Caruana, "Multitask learning," *Machine learning,* vol. 28, pp. 41-75, 1997.