Text-Dependent Speech Enhancement for Small-Footprint Robust Keyword Detection

Meng Yu[1], *Xuan Ji*[1], *Yi Gao*[1], *Lianwu Chen*[1], *Jie Chen*[1], *Jimeng Zheng*[1], *Dan Su*[1], *Dong Yu*[2]

[1]Tencent AI Lab, Shenzhen, China
[2]Tencent AI Lab, Bellevue, WA, USA

{raymondmyu, helenji, jackyyigao, lianwuchen, leojiechen, jimzzheng, dansu, dyu}@tencent.com

Abstract

Keyword detection (KWD), also known as keyword spotting, is in great demand in small devices in the era of Internet of Things. Albeit recent progresses, the performance of KWD, measured in terms of precision and recall rate, may still degrade significantly when either the non-speech ambient noises or the human voice and speech-like interferences (e. g., TV, background competing talkers) exists. In this paper, we propose a general solution to address all kinds of environmental interferences. A novel text-dependent speech enhancement (TDSE) technique using a recurrent neural network (RNN) with long short-term memory (LSTM) is presented for improving the robustness of the small-footprint KWD task in the presence of environmental noises and interfering talkers. On our large simulated and recorded noisy and far-field evaluation sets, we show that TDSE significantly improves the quality of the target keyword speech and performs particularly well under speech interference conditions. We demonstrate that KWD with TDSE frontend significantly outperforms the baseline KWD system with or without a generic speech enhancement in terms of equal error rate (EER) in the keyword detection evaluation.

Index Terms: text-dependent, speech enhancement, keyword detection, keyword spotting

1. Introduction

With the proliferation of smart homes and mobile and automotive devices, speech-based human-machine interaction becomes prevailing, e.g., in Google Now, Microsoft Cortana, Amazon Alexa, and Apple Siri. To achieve hands-free speech recognition experience, the system continuously listens for specific wake-up keywords, a process often called keyword detection (KWD) or keyword spotting (KWS)[1], to initiate speech recognition. For the privacy concern, the wake-up KWD typically happens completely on the device with low footprint and power consumption requirement.

The KWD systems usually perform very well in relatively clean conditions. However, their performance degrades significantly under noisy conditions. In order to improve the robustness to the background noises, two major techniques, namely multi-condition training and frontend enhancement, have been proposed in recent years. Multi-condition training [2, 3, 4] pools data under different environments to train neural networks and often leads to more robust systems since unseen environments are more likely represented as interpolation, instead of extrapolation, of seen environments. Nevertheless, the feature representation learned in this way, and thus the performance of KWD, is still worse than desired because the size of KWD networks is limited by the platform memory and processing power. The frontend enhancement technique, on the other hand, filters out the interference signals from the target speech stream before passing it for KWD. Recent approaches treat speech enhancement as a supervised learning problem, especially in the context of monaural speech enhancement in non-stationary noise conditions. The problem has especially benefited from the rapid rise in deep learning [5, 6]. The frontend speech enhancement can be optimized independently [7, 8, 9] or jointly with the acoustic model [10, 11] to improve the robustness of speech recognition systems. However, these generic enhancement techniques are not optimal for KWD since in KWD the only target of interest is the keyword(s) speech signal, while in the conventional speech enhancement setup this is not the case.

In this work, we propose and develop the text-dependent speech enhancement (TDSE) technique that is specifically designed for small-footprint robust KWD. Unlike the generic deep learning based speech enhancement techniques used in speech recognition [9, 10, 11], the TDSE technique aims at recovering the target keyword speech signal from its corrupted observation by suppressing interference signals, no matter whether the interference is from speech-like or non speech-like noises. Since TDSE only cares about the target keywords, its footprint is significantly smaller than that of a conventional text-independent speech enhancement component. Equipped with TDSE based frontend, the resulting KWD system is much more robust to environmental interferences than prior arts, measured by the false reject (FR) rate and false alarm (FA) rate. To the best of our knowledge, this is the first work on text-dependent deep learning speech enhancement, especially under the keyword detection setup.

The rest of the paper is organized as follows. In Section 2, the baseline KWD system is reviewed. In Section 3, we present details of the proposed text-dependent speech enhancement technique and discuss its advantages. We describe our experimental setup in Section 4, and evaluate the effectiveness of the proposed system in Section 5. We conclude this work in Section 6.

2. Baseline Keyword Detection

A block diagram of the baseline KWD system employed in this work is shown in Fig. 1. It consists of convolutional and fully connected layers. From bottom to top, they are one convolutional layer, pooling layer, three fully connected layers with 384 hidden units/layer, one fully connected layer with 128 hidden units/layer and a softmax layer. 40 dimensional log-mel filterbank features with its delta and delta-delta appended are computed every 25ms with a 10ms frame shift. At each frame, we stack 10 frames to the left and 5 frames to the right as the input feature to the convolutional layer.

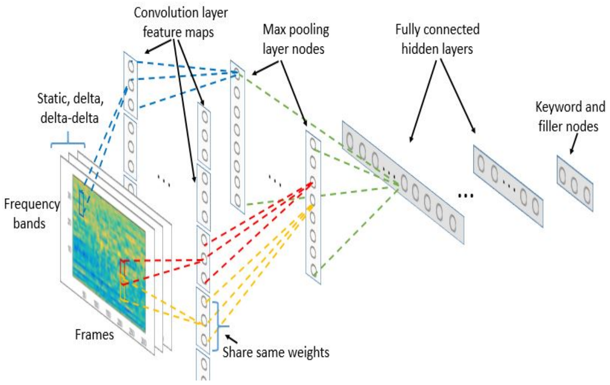

Figure 1: *Baseline KWD architecture.*

Limited weight sharing (LWS) scheme for CNNs [12] are used. We use separate sets of weights (filter size equals to 4) for totally 8 different regions of frequency bands in the convolutional layer since it allows for detection of distinct feature patterns in different frequency sub-bands. A total number of 384 filters are applied to 3 feature channels (static, delta, delta-delta), 16 context frames and 8 frequency regions, respectively. The stride within each frequency region equals to 1, while the max-pooling size is equal to 3.

Each hidden layer uses a sigmoid nonlinearity. The softmax output layer contains one output target for each of the words/characters in the keyword phrase to be detected, plus a single additional output target which represents all frames that do not belong to any of the words/characters (denoted as "filler"). The network weights are trained to optimize a cross-entropy criterion using stochastic gradient descent with momentum. Finally, in the posterior handling module, individual frame-level posterior scores from the neural network are combined into a single score corresponding to the keyword(s). We refer the readers to [13] for more details about posterior handling.

3. TDSE FRONTEND

The generic neural network based monaural speech enhancement problem is formulated as a regression task: determine a time-frequency mask for enhancing the speech source. Let us denote the speech signal in the time domain as $s(t)$ and the microphone received noisy signal as $y(t) = s(t) + n(t)$, where $n(t)$ is the interfering signal. The corresponding spectrum representation by short-time Fourier transform (STFT) is $Y(t, f) = S(t, f) + N(t, f)$ for each time frame t and frequency subband f. The goal of monaural speech enhancement is to recover speech signal $S(t, f)$ from $Y(t, f)$. More specifically, we train a deep learning model $g(\cdot)$ such that $g(log|Y|^2; \theta) = \hat{M}$, where θ is the model parameter. We use log power spectrum for representing input noisy signal and the model infers complex ideal ratio mask (cIRM) $\hat{M} = \hat{M}_r + i\hat{M}_i$ for all time-frequency bins (t, f). It has been shown that cIRM estimation produces higher quality speech than related methods [14]. We then estimate source spectrum $S(t, f)$ as $\hat{S} = \hat{M} \otimes Y$, where $\otimes$ is the element-wise product of two operands. Due to the issue of zero-division in silence segments for label preparation, the cost function for regular deep learning based monaural speech enhancement is

$$\mathcal{J} = \frac{1}{T \times F} \parallel \hat{M} \otimes Y - S \parallel_F^2, \tag{1}$$

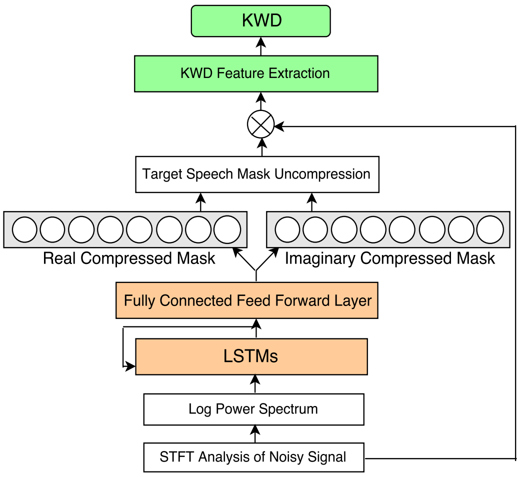
KWD
KWD Feature Extraction
Target Speech Mask Uncompression
Real Compressed Mask
Imaginary Compressed Mask
Fully Connected Feed Forward Layer
LSTMs
Log Power Spectrum
STFT Analysis of Noisy Signal

Figure 2: *Text-dependent speech enhancement architecture.*

where $\| \cdot \|_F$ is Frobenius norm.

Given an incoming stream of noisy speech, our model analyzes and separates a target keyword speech from its mixture with noise and other non-keyword speech interferences. The key training strategy that differentiates TDSE from generic speech enhancement models is that the training sets consist of mixtures of the target keyword speech and various of interfering signals, and as a result the training label is defined as below,

$$M = M_r + iM_i = \frac{Y_r S_r + Y_i S_i}{Y_r^2 + Y_i^2} + i\frac{Y_r S_i - Y_i S_r}{Y_r^2 + Y_i^2}, \tag{2}$$

where S is the keyword speech signal. The bin-wise label aims to infer the keyword signal as the only target. The complex mask may have large real and imaginary components with values in the range $(-\infty, \infty)$. Therefore, same as that in [14] we compress the cIRM with a hyperbolic tangent to limit mask values in a certain range. Unlike the training of generic speech enhancement, for TDSE the diversity of training speakers leads to better model generalization and the target speech content is fixed. Nevertheless, a sufficiently large amount of non-keyword speech materials, named as "negative samples", as well as environmental noises are required as the interfering signals for generating the noisy training data. Particularly, the diverse training sets of non-keyword speech samples improve the generalization capability of the networks. The way of TDSE training enables the model to extract the target speech signal through distinguishing a certain speech content from others, i.e. "text-dependent".

We train an LSTM-RNN model to jointly estimate the real and imaginary mask components. with deep LSTM models, as proposed in [9], both low-level and high-level features of the previous time step are carried forward to facilitate learning of long-term dependencies. The model maintains the text-sensitive information extracted from many previous frames to improve mask estimation for the current frame. The proposed system is illustrated in Fig. 2. We use a feature window of only one frame to estimate one frame of the cIRMs, which is defined on a 512-channel FFT with a 16-ms frame length and a 16-ms frame shift. The log power spectrum is computed and fed into the recurrent layer. Two stacked hidden LSTM layers with 256 hidden units are employed for temporal modeling, followed by one hidden fully connected layer with 512 rectified linear units (ReLUs) and one output layer of 512 units for mask estimation. Adam optimizer is incorporated with initial

Table 1: *Evaluation on simulated mixtures with non-keyword speech interference in terms of PESQ and SDR. The model of TDSE^1 is trained by adding extra interfering speech data, while TDSE^2 is trained by removing all the noisy data of speech interference.*

	PESQ/PESQ improvement					SDR/SDR improvement(dB)				
In. SNR	In. PESQ	SE	TDSE	TDSE^1	TDSE^2	In. SDR	SE	TDSE	TDSE^1	TDSE^2
-5 dB	1.15	1.15/0.0	1.52/0.4	1.57/0.4	1.20/0.1	-6.10	-6.55/-0.5	0.82 /6.9	1.05/7.2	-5.61/0.5
0 dB	1.42	1.40/0.0	1.87/0.5	1.92/0.5	1.52/0.1	-3.33	-3.54/-0.2	5.31/8.6	5.38/8.7	-1.72/1.6
5 dB	1.79	1.78/0.0	2.16/0.4	2.22/0.4	1.90/0.1	0.32	0.42/0.1	9.68/10.0	9.64/10.0	3.01/2.7
10 dB	2.14	2.11/0.0	2.40/0.3	2.45/0.3	2.23/0.1	4.71	5.05/0.3	13.95/9.2	13.87/9.2	8.20/3.5
15 dB	2.40	2.35/-0.1	2.57/0.2	2.63/0.2	2.47/0.1	9.41	10.00/0.6	17.92/8.5	17.89/8.5	13.30/3.9
20 dB	2.59	2.53/-0.1	2.71/0.1	2.77/0.2	2.66/0.1	14.36	15.17/0.8	21.58/7.2	21.72/7.4	17.85/3.5
25 dB	2.74	2.66/-0.1	2.81/0.1	2.88/0.1	2.80/0.1	19.33	19.90/0.6	24.57/5.2	24.91/5.6	21.08/1.8

Table 2: *Evaluation on simulated mixtures with environmental noise interference in terms of PESQ and SDR.*

	PESQ/PESQ improvement					SDR/SDR improvement(dB)				
In. SNR	In. PESQ	SE	TDSE	TDSE^1	TDSE^2	In. SDR	SE	TDSE	TDSE^1	TDSE^2
-5 dB	1.29	1.45/0.2	1.57/0.3	1.57/0.3	1.51/0.2	-4.07	1.50/5.6	2.55/6.6	2.02/6.1	1.14/5.2
0 dB	1.55	1.89/0.3	1.92/0.4	1.93/0.4	1.87/0.3	-1.84	6.79/8.6	7.25/9.1	6.70/8.5	6.09/7.9
5 dB	1.85	2.23/0.4	2.20/0.4	2.23/0.4	2.19/0.3	1.40	11.48/10.1	11.21/9.8	10.85/9.5	10.42/9.0
10 dB	2.14	2.47/0.3	2.41/0.3	2.45/0.3	2.43/0.3	5.62	15.61/10.0	14.68/9.1	14.54/8.9	14.22/8.6
15 dB	2.37	2.63/0.3	2.56/0.2	2.62/0.3	2.62/0.3	10.31	19.23/8.9	17.95/7.6	18.03/7.7	17.67/7.4
20 dB	2.54	2.73/0.2	2.68/0.1	2.75/0.2	2.76/0.2	15.26	22.48/7.2	21.01/5.8	21.24/6.0	20.56/5.3
25 dB	2.69	2.81/0.1	2.78/0.1	2.86/0.2	2.87/0.2	20.43	25.35/4.9	23.68/3.3	23.97/3.5	22.75/2.3

learning rate of 0.001. The dropout is applied at the first fully connected layer with rate 0.2. During training, the label mask is compressed. Thus the predicted mask in Fig. 2 is converted to linear one before element-wise multiplied with STFT of the input noisy signal.

4. Experimental Setup

A keyword of four Chinese characters is employed in this work, with their Chinese pinyin representation as “ni3”, “hao3”, “wei1” and “ling2”. We collected the keyword data in a few regular living rooms by a headset microphone, and omni-directional microphones placed at the distance $0.5m$, $1m$, $3m$, $5m$ and $7m$ to the speaker, respectively. The recordings are proceeded in quiet and noisy conditions, respectively. In noisy conditions, the noise source is played back through a loudspeaker at $1m$ to the distant microphone at each position. Noise types include kitchen, TV, home appliance, voice, music and other ambient noises sampled from recordings of “daily life” environments. Such recordings are used for preparing the following training and evaluation data sets.

4.1. Training Corpora of KWD

The KWD model is pretrained on a 100k-hour Chinese ASR multi-condition training set that contains both clean and far-field noisy data. The units in the output layer stand for Chinese syllables, plus silence and blank, summing up to 1434 output units. After one epoch, KWD training starts by reducing the number of output units to 5, representing the four Chinese characters of the keyword and one non-keyword filler. A keyword specific data set around 200 hours from 337 human speakers, which includes 45K utterances from headset recordings (near-field clean data) and 179K utterances from distant omni-directional microphones (far-field noisy data), is used as positive examples. A set of 139 hours’ 100K negative examples from a Mandarin speech database serves as negative examples.

4.2. Training Corpora of TDSE

From the living recordings, 45K headset recordings are used to create reverberant and noisy training data. The room simulator based on the image method [15] generates 15K room impulse responses (RIRs) from the source to the microphone. The reverberation time RT_{60}s range from 0 to 600 ms, with an average RT_{60} of 300 ms. Noise signals, including environmental noises and human speech signals (i.e. non-keyword negative samples used in KWD training), are mixed with the clean keyword utterances at uniformly distributed SNRs ranging from -5 to 25 dB. In total, we have 45K keyword utterances interfered by environmental noises and the same amount of keyword utterances interfered by the non-keyword speech.

4.3. Evaluation Corpora

We evaluate our models using simulated and real far-field noisy data. For the simulated sets, 4789 headset microphone recordings of clean keyword utterances by 36 human speakers are used. Noise signals, including environmental noises and human speech signals (non-keyword), which are different from those in training, are mixed with the clean utterances at SNRs ranging from -5 to 25 dB, using the room simulator with a room configuration distribution that approximately matches the training configurations. Therefore, for each SNR, we have 4789 simulated far-field noisy utterances. The real far-field set consists of 4570 utterances from 36 human speakers recorded by the distant microphones with a higher SNR range from 5dB to 20dB. No overlap of speakers exists between training and testing.

5. Results and Discussion

The TDSE model is first evaluated on its potential to improve the Signal-to-Distortion Ratio (SDR) [16] and the Perceptual

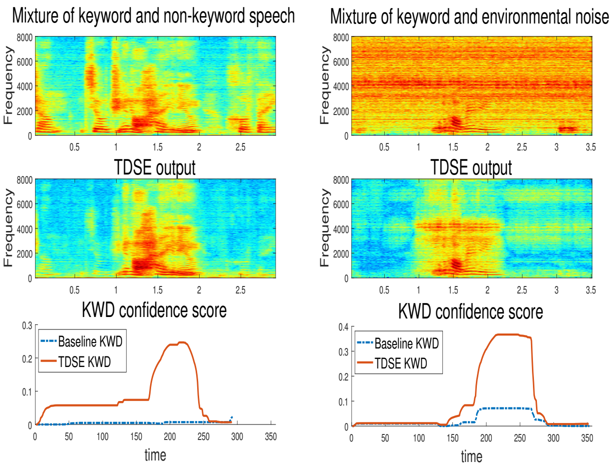

Figure 3: *Illustration of TDSE processing and its positive impact on KWD scores.*

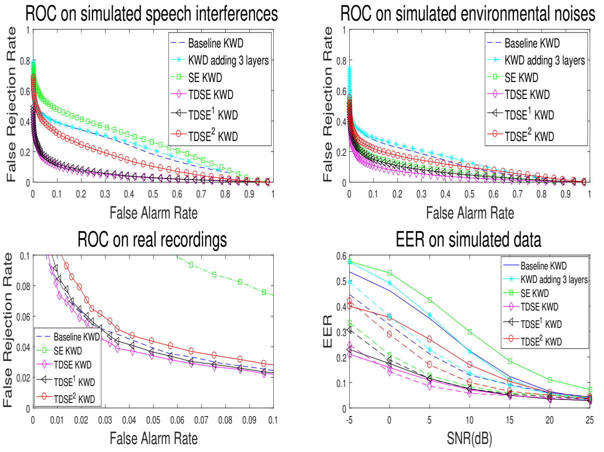

Figure 4: *ROC curves of KWD with different frontend enhancement approaches on simulated and real data: simulated data of speech interference (upper left), simulated data of non-speech environmental noise (upper right), real data of ambient noise (bottom left), and EERs on simulated data (solid line for speech interference and dash line for environmental noise) in each individual SNR condition (bottom right).*

Evaluation of Speech Quality (PESQ) score [17], both of which are metrics widely used to evaluate speech enhancement performance for speech denoising and multi-talker speech separation tasks. The number of parameters for TDSE is 1.4M while the number of parameters for the baseline KWD is 0.7M, which leads to small memory footprint and low computational cost and thus is favored for the embedded application.

The performance on simulated far-field keyword utterances interrupted by non-keyword background speech and environmental noises are shown in Table 1 and 2, respectively. Frontend TDSE processing shows significant improvement in both PESQ and SDR compared with the raw input. A generic speech enhancement (SE) model is trained using about 100 hours noisy speech signals with clean references based on the same LSTM model architecture. Such speech enhancement model is capable to infer original clean speech signal when it is mixed with environmental non-speech noises as summarized in Table 2. However, since it doesn't have the ability to distinguish speech contents, we observe from Table 1 that almost no improvement is achieved on the multi-talker speech mixtures.

Compared with target speech enhancement in non-speech noise conditions, more challenges have been observed in the tasks of separating speech signal from the interfering background competing speech noises. TDSE performs equally well on both evaluation sets due to its training on keyword utterances corrupted by non-keyword interfering speech signals. In order to examine the impact of speech interferences in the training set to the model, we trained the TDSE model by adding extra 5K (about 3.6 hours) negative samples of non-keyword speech and by removing all the 45K keyword utterances mixed with non-keyword speech signals in the training set, resulting two new models TDSE^1 and TDSE^2, respectively. Confirmed in Table 1 TDSE^1 performs slightly better in the speech interference condition, while TDSE^2 regresses on it.

KWD is applied on the TDSE processed data stream. The capability of recovering the clean keyword(s) signal enable KWD to achieve significantly better confidence scores. As illustrated in Fig. 3, by cleaning up the target keyword spectrum, TDSE based KWD scored much higher than the baseline KWD, which indicates greater detection accuracy in noisy conditions. KWD results are presented in the form of a modified receiver operating characteristic (ROC) curves, where we replace true positive rate with the false reject rate on Y-axis. Lower curves are better. Fig. 4 illustrates comparisons among the baseline KWD system, KWD system with the generic speech enhancement frontend (SE-KWD) and the proposed text-dependent speech enhancement based KWD system (TDSE-KWD). The ROC curves in the upper two panels represent average performance of all SNR conditions on simulated mixtures with speech interferences and environmental noises, corresponding to conditions in Table 1 and 2 respectively. Together with a detailed summarization of equal error rate (EER) for each SNR condition (bottom right panel), it shows that the KWD performance with the frontend processing coincides with the objective evaluation of the frontend speech quality. Consistent performance has been observed on real and simulated data, showing that KWD with TDSE frontend achieves the best EER. Even with a larger KWD model by adding 3 extra fully connected layers, each with 384 hidden units ("KWD adding 3 layers" in Fig. 4), we observed similar performance in speech interference and environmental noise conditions, respectively. The advantage of TDSE based KWD is greater when the keyword speech is corrupted by speech interferences, compared with that under non-speech environmental noise conditions. As a remark, SE-KWD performs the worst in interfering speech conditions possibly due to the conventional SE's training scheme where the keyword speech utterance is not required and thus the model inference may not aim to enhance the keyword speech when it is mixed with other speech signals.

6. Conclusions

In this paper, we proposed a text-dependent speech enhancement technique for recovering the original clean speech signal of a specific content. This technique is further explored and applied to keyword detection as a frontend processing component. Experimental results show that the proposed framework significantly outperforms the baseline KWD system and the KWD system that employs the generic speech enhancement under far-field noisy conditions. Particularly, KWD with TDSE frontend performs very robustly when the competing talker exists. We believe that the joint training of TDSE and the KWD model can further improve the overall performance.

7. References

[1] J. Rohlicek, W. Russell, S. Roukos, and H. Gish, "Continuous hidden markov modeling for speaker-independent word spotting," in *the Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE*, 1989, pp. 627–630.

[2] Y. Wang, P. Getreuer, T. Hughes, R. Lyon, and R. Saurous, "Trainable frontend for robust and far-field keyword spotting," in *the Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE*, 2017, pp. 5670–5674.

[3] R. Prabhavalkar, R. Alvarez, C. Parada, P. Nakkiran, and T. N. Sainath, "Automatic gain control and multi-style training for robust small-footprint keyword spotting with deep neural networks," in *the Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE*, 2015, pp. 4704–4708.

[4] T. Sainath and C. Parada, "Convolutional neural networks for small-footprint keyword spotting," in *Proc. Interspeech*, 2015.

[5] Y. Wang and D. Wang, "Towards scaling up classification based speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, pp. 1381–1390, 2013.

[6] J. Chen and D. Wang, "Dnn-based mask estimation for supervised speech separation," *in Audio source separation, S. Makino, Ed., Berlin: Springer, in press*, 2017.

[7] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[8] Y. Xu, J. Du, L. Dai, and C. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 1849–1858, 2015.

[9] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. L. Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr," *in Latent Variable Analysis and Signal Separation, Springer*, pp. 91–99, 2015.

[10] B. Li and K. Sim, "Improving robustness of deep neural networks via spectral masking for automatic speech recognition," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2013, pp. 279–284.

[11] A. Narayanan and D. Wang, "Improving robustness of deep neural network acoustic models via speech separation and joint adaptive training," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 92–101, 2015.

[12] O. Abdel-Hamid, A.-R. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 10, pp. 1533–1545, 2014.

[13] G. Chen, C. Parada, and G. Heigold, "Small-footprint keyword spotting using deep neural networks," in *the Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE*, 2014, pp. 4087–4091.

[14] D. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 3, pp. 483–492, 2015.

[15] J. B. Allen and D. A. Berkley, "Image method for efficiently simulation room-small acoustic," *Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[16] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, 2006.

[17] 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 *the Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE*, 2001, pp. 749–752.