Speaker-aware Deep Denoising Autoencoder with Embedded Speaker Identity for Speech Enhancement

Fu-Kai Chuang[1], *Syu-Siang Wang*[1,2], *Jeih-weih Hung*[3], *Yu Tsao*[4], *and Shih-Hau Fang*[1,2]

[1]Department of Electrical Engineering, Yuan Ze University, Taoyuan, Taiwan
[2]MOST Joint Research Center for AI Technology and All Vista Healthcare, Taipei, Taiwan
[3]Dept of Electrical Engineering, National Chi Nan University, Taiwan
[4]Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan

Abstract

Previous studies indicate that noise and speaker variations can degrade the performance of deep-learning–based speech–enhancement systems. To increase the system performance over environmental variations, we propose a novel speaker-aware system that integrates a deep denoising autoencoder (DDAE) with an embedded speaker identity. The overall system first extracts embedded speaker identity features using a neural network model; then the DDAE takes the augmented features as input to generate enhanced spectra. With the additional embedded features, the speech-enhancement system can be guided to generate the optimal output corresponding to the speaker identity. We tested the proposed speech-enhancement system on the TIMIT dataset. Experimental results showed that the proposed speech-enhancement system could improve the sound quality and intelligibility of speech signals from additive noise-corrupted utterances. In addition, the results suggested system robustness for unseen speakers when combined with speaker features.

Index Terms: additive noise, speech enhancement, deep denoise autoencoder, noise reduction, speaker identity

1. Introduction

In realistic environments, noise signals can deteriorate speech quality and intelligibility, and thereby limit for human-human and human-machine communication efficiency [1–4]. To address this issue, an important front-end speech process, namely speech enhancement, which extracts clean components from noisy input, can improve the voice quality and intelligibility of noise-deteriorated clean speech. These speech-enhancement approaches can be split into two categories: unsupervised and supervised. For an unsupervised speech-enhancement system, the noise-tracking and signal-gain estimation stages are included explicitly or implicitly [5], without employing information about the speech and noise components [6–9]. On the other hand, supervised speech-enhancement systems utilize a set of training data to prepare prior information about the speech and noise signals, which facilitates an effective denoising process at runtime. In recent years, most supervised speech-enhancement techniques have been based on deep-learning–based neural network architectures, which show strong regression capabilities from the source input to the target output [10–12, 12–14]. For example, the deep denoising autoencoder (DDAE) [15,16] technique was proposed to model the relationship between a noise-corrupted speech signal and its original clean counterpart, and to effectively reduce additive noises with a deep neural network (DNN) architecture. In addition, it was found that a DNN-based speech-enhancement system had good generalization capabilities in the unseen noise environments for models trained with data from various noisy conditions [17, 18].

To further improve the sound quality and intelligibility, several studies have incorporated information on speaker- and speaking-environment models into a supervised speech-enhancement model [19]. The speaking-environment information, e.g., signal-to-noise ratio (SNR) and noise types, has been used to improve the speech-enhancement model's denoising performance [20, 21]. In addition, visual cues, which provide complementary information to the speech signals, can be incorporated into the speech-enhancement system to more effectively suppress noise interference [22]. Several algorithms have also been derived to incorporate speaker information into a deep-learning–based speech-enhancement system. For example, works in [23, 24] characterize the speech signals of a target speaker using a statistical model, which is used to minimize the residual components from a preceding speech-enhancement system. Other works use the speaker identity as a prior knowledge for performing speech-enhancement [25–27]. For these approaches, the original training set is divided into several subsets, each of which corresponds to a single speaker. Then an individual speech enhancement model is created with each subset, and the ensemble of these speaker-specific models is used to perform speech enhancement. Although these approaches perform well, they usually require multiple speech-enhancement models, which may not be suitable for mobile or embedded devices. In this study, we investigate a novel speech-enhancement system that combines embedded speaker identities (code) to achieve robust enhanced performance for speaker variations.

Incorporating explicit/embedded speaker information into the main task is a common approach in speech-related frameworks. In [28], the speaker information is characterized by a speaker code, which guides a voice conversion system to generate target speech signals. In [29], the speaker-related i-code is extracted to perform speaker variation. Meanwhile, the speaker code is employed for supervised multi-speaker separation and effectively reduces the word error rate in a speech recognition system [30]. In this study, we proposed a novel architecture, termed a speaker-aware denoising autoencoder (with a shorthand notation "SaDAE"), to implement speaker-dependent speech-enhancement task. In SaDAE, two DNN-based models are created; the first DNN extracts the speaker representation from the input noisy spectra, while the second DAE enhances the speech from the output of the first DAE. Therefore, we expect that the presented SaDAE can further enhance noisy utterances since speaker cues are adopted. The objective evaluations conducted on the TIMIT corpus [31] showed that the presented

SaDAE can effectively improve the quality and intelligibility of the distorted utterances in the test set. In addition, SaDAE was shown to possess decent generalization capability since it also worked well for those utterances from unseen speakers.

The rest of this paper is organized as follows. Section 2 reviews the conventional DDAE-based speech-enhancement system. Then, section 3 introduces the proposed SaDAE architecture. Experiments and the respective analysis are given in Section 4. Finally, section 5 provides concluding remarks and a future avenue.

2. DDAE-based speech enhancement system

This section briefly reviews the process of a DDAE-based speech-enhancement system. Eq. (1) expresses how an additive-noise corrupted signal $\mathbf{y}$ is associated with the embedded clean signal $\mathbf{x}$ and noise $\mathbf{n}$ in the time domain:

$$\mathbf{y} = \mathbf{x} + \mathbf{n}. \tag{1}$$

A DDAE-based speech-enhancement system is applied to enhance $\mathbf{y}$ so as to reconstruct $\mathbf{x}$; the overall flowchart is depicted in Fig. 1. From this figure, the noisy spectrogram $\mathbf{Y}$ is first created from $\mathbf{y}$ using a short-time Fourier transform (STFT), and $\hat{\mathbf{Y}}_i$ denotes the magnitude spectrum of the i–th frame of $\mathbf{y}$. Then, the feature–extraction stage extracts the frame-wise logarithmic power spectra and concatenates adjacent frames to create a context feature $\tilde{\mathbf{Y}}_i$ for each frame, represented by $\tilde{\mathbf{Y}}_i = [\mathbf{Y}_{i-I}; \cdots; \mathbf{Y}_i; \cdots, \mathbf{Y}_{i+I}]$, where $\mathbf{Y}_i$ is the logarithmic power spectrum of the i–th frame, ";" denotes the vertical-concatenation operation, and $2I+1$ is the length of the context window. Next, each context feature $\tilde{\mathbf{Y}}_i$ is processed by the DDAE-based speech-enhancement algorithm, thereby producing its enhanced version, $\tilde{\mathbf{X}}_i$. The new context feature $\tilde{\mathbf{X}}_i$ is used to build the enhanced frame-wise logarithmic power spectrum $\mathbf{X}_i$, which is converted to the magnitude spectral domain and then combined with the preserved original noisy phase $\angle\mathbf{Y}_i$ to create the new spectrogram $\{\hat{\mathbf{X}}_i\}$. Finally, an inverse STFT (ISTFT) is applied to $\{\hat{\mathbf{X}}_i\}$ to produce the enhanced time-domain signal $\hat{\mathbf{x}}$.

For the DDAE block in Fig. 1, a deep neural network (DNN) is used to enhance the noisy input feature $\tilde{\mathbf{Y}}_i$. Consider a DNN that has L layers. For an arbitrary layer l of this network, the input-output relationship ($\mathbf{z}^{(l-1)}$, $\mathbf{z}^{(l)}$) is formulated by

$$\mathbf{z}^{(l)} = \sigma^{(l)}\left(h^{(l)}(\mathbf{z}^{(l-1)})\right), \quad l = 1, \cdots, L, \tag{2}$$

where $\sigma^{(l)}(\cdot)$ and $h^{(l)}(\cdot)$ are the activation function and linear regression function, respectively, for the l–th layer. Notably, the input and output layers correspond to the first and L-th layers, respectively. Therefore, for the DNN in the DDAE block, we have $\mathbf{z}^{(0)} = \tilde{\mathbf{Y}}_i$ and $\mathbf{z}^{(L)} = \tilde{\mathbf{X}}_i$.

To train the DDAE network, a training set consisting of noisy–clean ($\tilde{\mathbf{Y}}_i$–$\mathbf{X}_i$) pairs of speech features is first prepared. Then, the network parameters undergo supervised training by using the noisy feature $\tilde{\mathbf{Y}}_i$ as the input, and minimizing a loss function that measures the difference between the network output $\tilde{\mathbf{X}}_i$ and the noise-free counterpart $\mathbf{X}_i$. In this study, the mean squared error (MSE) is selected as the loss function.

3. The Proposed Algorithm

To increase the capability of a speech-enhancement system for utterances of different speakers, we propose a novel speaker-aware speech-enhancement architecture, namely SaDAE, which integrates DDAE with embedded speaker identity information. The SaDAE flowchart is depicted in Fig. 2. Similar to the DDAE-based speech-enhancement system, which was described in the previous section, the context feature $\tilde{\mathbf{Y}}_i$, composed of the neighboring frame-wise logarithmic power spectra for the input utterance, is selected as the main unit for enhancement in SaDAE. Specifically, the SaDAE scheme consists of two deep neural networks (DNNs), a speaker-embedded DDAE (SpE-DDAE) and a speaker-feature extraction (SFE) DNN, which will be described in the following two sub-sections.

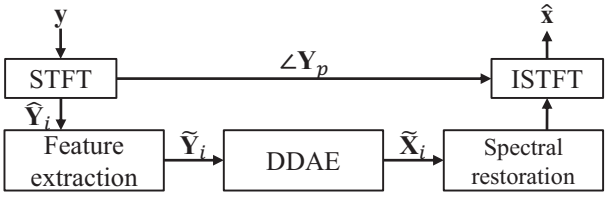

Figure 1: *The block diagram of a conventional DDAE-based speech-enhancement system.*

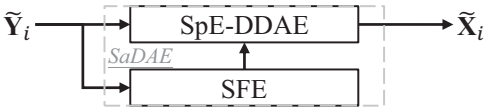

Figure 2: *The block diagram of the proposed SaDAE, which includes the SpE-DDAE and SFE components. The system input is the frame-wise noisy feature vector $\tilde{\mathbf{Y}}_i$, while the output is the enhanced feature vector $\tilde{\mathbf{X}}_i$.*

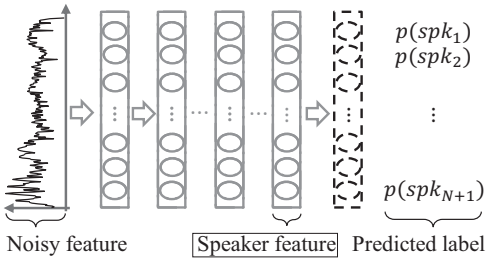

Figure 3: *The DNN model that extract frame-wise speaker features.*

3.1. The SFE module

In this sub-section, we present the method for creating a DNN that performs speaker–feature extraction (SFE), which is illustrated in Fig. 3. The objective of the SFE-based DNN is to classify each frame-wise speech feature $\tilde{\mathbf{Y}}_i$ into a certain speaker identity. Therefore, the dimension for the DNN output is set to the number of speakers, N, in the training set plus one that corresponds to the non-speech frames. In addition, the desired output for the DNN training is a one-hot $(N+1)$-dimensional vector, in which the single non-zero element corresponds to the speaker identity.

The input-output relationship for each layer of the SFE-based DNN is described in Eq. (2). Particularly, the activation function is set to softmax for the output layer, while the rectified linear units (ReLU) function is used for the input layer and all hidden layers. In addition, the categorical cross-entropy loss

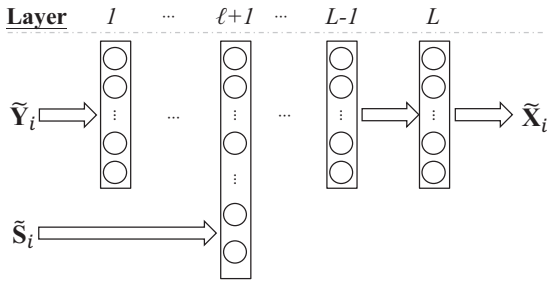

Figure 4: *The architecture of the SpE-DDAE model, where the noisy speech feature $\tilde{Y}_i$ is at the input of first input layer, the speaker feature $\tilde{S}_i$ is fed to the $(\ell+1)$-th layer, and the output is the enhanced speech feature $\tilde{Y}_i$.*

function is used for training this DNN.

Once the training of the SFE-based DNN is complete, we select the output of the last hidden layer (viz., the penultimate layer), denoted by $\tilde{\mathbf{S}}_i$, to be the speaker-feature representation for each frame-wise noisy input vector $\tilde{\mathbf{Y}}_i$; this speaker feature $\tilde{\mathbf{S}}_i$ will be fed into the subsequent SpE-DDAE network. $\tilde{\mathbf{S}}_i$ was selected because it possessed higher generalization ability for unseen speakers than the ultimate layer output, and it provided the proposed SaDAE system with a better speech-enhancement performance in our preliminary evaluations. Notably, the idea of employing a DNN to identify speakers is motivated by the speaker-verification task in [32], in which the input of the DNN is filterbank energy features. The respective *d*-vector speaker-verification system [32] behaves better than a conventional *i*-vector-based system [33].

3.2. The SpE-DDAE module

Compared with a conventional DDAE-based speech-enhancement system that uses noisy-speech features as the input, the presented SpE-DDAE additionally employs speaker features produced by the SFE-based DNN; its architecture is depicted in Fig. 4. From the figure, the SpE-DDAE network input contains the noisy-speech feature $\tilde{\mathbf{Y}}_i$ and the speaker feature $\tilde{\mathbf{S}}_i$. Specifically, $\tilde{\mathbf{Y}}_i$ is placed in front of the input layer, while $\tilde{\mathbf{S}}_i$ is concatenated with the output of a certain hidden layer, say, the ℓ–th layer. Hence, the input feature to the next hidden layer (the $(\ell+1)$-th layer) is denoted by $\mathbf{z}'^{(\ell)}_i = [\mathbf{z}^{(\ell)}_i; \tilde{\mathbf{S}}_i]$. As a result, the SpE-DDAE network is almost the same as a conventional DDAE network, except that SpE-DDAE incorporates the speaker feature at a certain hidden layer.

To train the SpE-DDAE, we first prepare the noisy-speech features $\{\tilde{\mathbf{Y}}_i\}$, the associated clean speech features $\{\tilde{\mathbf{X}}_i\}$, and the SFE-derived speaker features $\{\tilde{\mathbf{S}}_i\}$ to form the training set. Then, the training proceeds with $\{\tilde{\mathbf{Y}}_i\}$ and $\{\tilde{\mathbf{S}}_i\}$ on the input side to produce the enhanced output that approximates $\{\tilde{\mathbf{X}}_i\}$. As mentioned in Sec. 2, we choose the MSE as the loss function to be minimized during the training of the SpE-DDAE network.

3.3. The overall flow of the proposed SaDAE

The proposed SaDAE has offline and online stages. In the offline stage, we train the SFE-based DNN first, and the SpE-DDAE DNN separately. Both DNNs are then used in the online stage to perform the speaker-aware speech-enhancement task. According to Fig. 2, the frame-wise noisy input $\tilde{\mathbf{Y}}_i$ is fed into the SFE-based DNN to produce the speaker feature $\tilde{\mathbf{S}}_i$. Then, the SpE-DDAE DNN takes the augmented features that use $\tilde{\mathbf{Y}}_i$ and $\tilde{\mathbf{S}}_i$ as the input to ultimately generate the enhanced speech feature $\tilde{\mathbf{X}}_i$.

4. Experiment and Analysis

4.1. Experimental setup

We conducted evaluation experiments on the TIMIT database [31] of read speech, where utterances were recorded at a 16 kHz sampling rate. From this database, we randomly selected 486 native English speakers with each speaker pronouncing eight utterances; thus, 3,888 utterances were involved in the evaluations. Among these utterances, 3,696 utterances produced by 462 speakers (i.e., $N = 462$ in Sec. 3.1) are used as the training set, while the 192 utterances provided by the other 24 speakers serve as the test set. Next, 60 of 104 types of noise [34] were artificially added to the utterances in the training set at 21 SNRs ranging from 10 to -10 dB with 1 dB intervals, to generate the noisy training set. By contrast, three additive noises, "Car noise idle noise 60mph", "babble", and "street", were individually used to deteriorate the utterances in the test set at four SNR levels (-5 dB, 0 dB, 5 dB, and 10 dB); thus, the noisy test set consists of 2,304 utterances ($192 \times 3 \times 4$).

For the speech-feature preparation, each utterance in the training and test sets were first split into overlapped frames with a 32-ms-frame duration and 16-ms-frame shift. Then, a 512-point discrete Fourier transform (DFT) was conducted on each frame signal to produce the respective 257-dim acoustic spectrum. Following the procedures stated in Section 2, the context feature for each frame was created by concatenating the neighboring 11 frames of the logarithmic power spectra ($2I + 1 = 11$); thus, the corresponding dimension was 2,827 ($257 \times 11 = 2827$). Accordingly, the input-layer sizes of the three DDAE-related models (DDAE, SpE-DDAE, and SFE) were 2,827, while the output-layer sizes of DDAE, SpE-DDAE, and SFE were 257, 257 and 463 (i.e., $N + 1 = 462 + 1$), respectively.

The network configuration is arranged as follows:

- The SFE-based DNN consists of five layers with 1,024 nodes for each hidden layer.
- The SpE-DDAE DNN has seven layers, and the 1,024-dim speaker feature is fed into the third layer. Therefore, the number of nodes for the third layer is 3,072, while the number of nodes for the other six layers is 2,048.
- For the purpose of comparison, a DDAE DNN without speaker features is prepared; it is arranged to have seven layers and 2,048 nodes for each layer.

Notably, a dropout algorithm with a 67% drop rate is facilitated on all hidden layers in the DNNs for DDAE and SpE-DDAE during the training process to improve the generalization capability.

In this study, the performance of all systems was evaluated by three metrics: the quality test in terms of the perceptual evaluation of speech quality (PESQ) [35], the perceptual test in terms of short-time objective intelligibility (STOI) [36], and the speech distortion index (SDI) test [37]. The score ranges of PESQ and STOI are $[-0.5, 4.5]$ and $[0, 1]$, respectively. Higher scores for PESQ and STOI denote better sound quality and intelligibility. In contrast, the SDI measures the degree of speech distortion. Thus, a lower SDI indicates less speech distortion and a more enhanced performance.

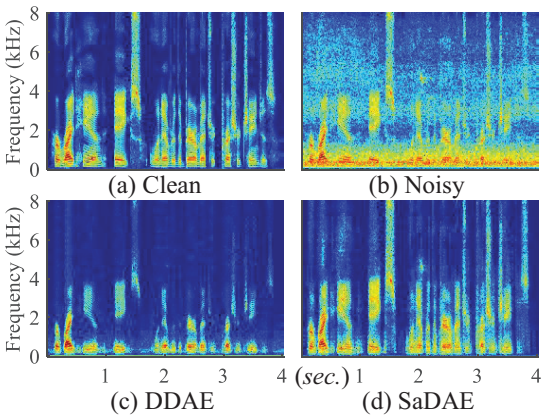

Figure 5: *The spectrograms of (a) a clean utterance x, (b) y, the noisy counterpart of x, (c) the DDAE enhanced version of y, and (d) the SaDAE enhanced version of y*

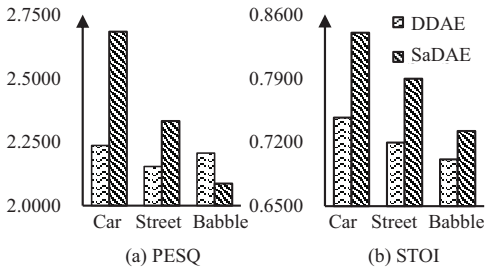

Figure 6: *The averaged PESQ and STOI results over noisy utterances with respect to three noisy environments, achieved by DDAE and SaDAE.*

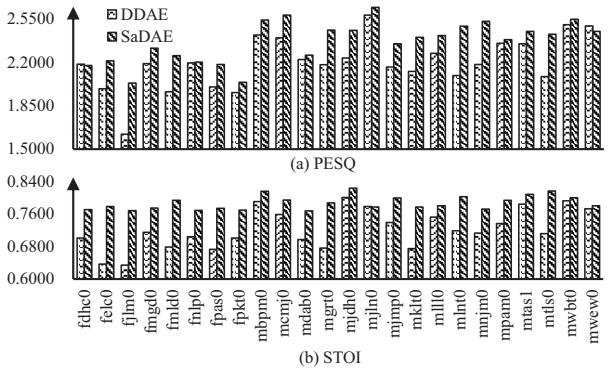

Figure 7: *The detailed results of (a) PESQ and (b) STOI with respect to different speakers achieved by DDAE and SaDAE.*

4.2. Experimental results

Figs. 5(a)(b)(c)(d) show the spectrograms of a clean utterance x, the corresponding noisy counterpart y, and y enhanced by either of DDAE and the presented SaDAE. From these figures, we find that the spectrogram of the SaDAE-processed utterance in Fig. 5(d) is quite close to that of the clean utterance in Fig. 5(a). In addition, comparing Fig. 5 (d) with Fig. 5(c) the harmonic structures of the spectrogram are revealed more clearly by SaDAE than DDAE.

Table 1 lists the averaged PESQ, STOI and SDI scores with respect to all tested utterances for noisy baseline and those processed by DDAE and SaDAE. From the table, we observe that both DDAE and SaDAE provide better results than the noisy baseline for all evaluation indices. In addition, SaDAE reveals superior scores when compared with DDAE. These observations clearly indicate that SaDAE can diminish the additive noise while simultaneously improving the speech quality and intelligibility.

In Fig. 6, we show the averaged PESQ and STOI scores for DDAE and SaDAE with respect to three noise environments. From this figure, SaDAE provides better metric scores than DDAE in almost all cases, except for the PESQ score in the babble noise environment. One possible explanation is that the babble noise contains multiple background speakers, which prevents the SFE module in SaDAE from producing reliable speaker features.

The detailed PESQ and STOI scores for DDAE and SaDAE with respect to the 24 testing speaker, are illustrated in Fig. 7. From the figure, SaDAE shows superior PESQ and STOI scores for most of the speakers when compared with DDAE. In addition, it is worth noting that all test speakers are not included in the training set; thus, they are unseen by the SaDAE model. Therefore, these results suggest the effectiveness of the SFE module in SaDAE since it provides a complete speech-enhancement process with robustness against speaker variation.

Table 1: *The averaged PESQ, STOI and SDI results over all noisy utterances in the test set, achieved by the noisy baseline, DDAE and SaDAE.*

Testing	PESQ	STOI	SDI.
Noisy	2.0280	0.7493	1.1450
DDAE	2.1987	0.7225	0.7501
SaDAE	**2.3715**	**0.7815**	**0.3228**

5. Conclusions and Future work

In this study, we proposed a novel speaker-aware speech enhancement system, termed SaDAE, to alleviate the distortion in noise-corrupted utterances from various speakers. SaDAE is composed of two DNNs: the first DNN extracts speaker-identity features, while the second DNN uses both speaker identity features and noisy speech features to restore the embedded clean utterance. The experimental results clearly indicated that the newly proposed SaDAE significantly reduced the noise in distorted utterances, and improved both the speech quality and intelligibility. It outperformed the conventional DDAE-based speech-enhancement system. Particularly, SaDAE was shown to work quite well when enhancing the utterances produced by unseen speakers. In the future, we plan to improve SaDAE under multiple-speaker situations, e.g., the babble noise environment. Furthermore, the presented SaDAE architecture will be tested on speaker-diarization and speech-source separation tasks.

6. Acknowledgment

The authors would like to thank the Ministry of Science and Technology for providing financial supports (MOST 107-2221-E-001-012-MY2, MOST 106-2221-E-001-017-MY2, MOST 108-2634-F-155-001)

7. References

[1] B. Jacob, M. Shoji, and C. Jingdong, "Speech enhancement (signals and communication technology): Chapter 1," 2005.

[2] S. Doclo, M. Moonen, T. Van den Bogaert, and J. Wouters, "Reduced-bandwidth and distributed mwf-based noise reduction algorithms for binaural hearing aids," *IEEE/ACM TASLP*, vol. 17, no. 1, pp. 38–51, 2009.

[3] Y.-H. Lai, F. Chen, S.-S. Wang, X. Lu, Y. Tsao, and C.-H. Lee, "A deep denoising autoencoder approach to improving the intelligibility of vocoded speech in cochlear implant simulation," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 7, pp. 1568–1578, 2017.

[4] Z.-Q. Wang and D. Wang, "A joint training framework for robust automatic speech recognition," *IEEE/ACM TASLP*, vol. 24, no. 4, pp. 796–806, 2016.

[5] J. Benesty, S. Makino, and J. Chen, *Speech enhancement*. Springer Science & Business Media, 2005.

[6] P. C. Loizou, "Speech enhancement based on perceptually motivated bayesian estimators of the magnitude spectrum," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 857–869, 2005.

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

[8] D. Malah, R. V. Cox, and A. J. Accardi, "Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments," in *Proc. ICASSP*, pp. 789–792, 1999.

[9] T. Lotter and P. Vary, "Speech enhancement by map spectral amplitude estimation using a super-gaussian speech model," *EURASIP journal on applied signal processing*, vol. 2005, pp. 1110–1126, 2005.

[10] D. Baby, J. F. Gemmeke, T. Virtanen, *et al.*, "Exemplar-based speech enhancement for deep neural network based automatic speech recognition," in *Proc. ICASSP*, pp. 4485–4489, 2015.

[11] A. J. R. Simpson, "Probabilistic binary-mask cocktail-party source separation in a convolutional deep neural network," *CoRR*, vol. abs/1503.06962, 2015.

[12] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM TASLP*, vol. 26, no. 10, pp. 1702–1726, 2018.

[13] K. Han, Y. Wang, D. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM TASLP*, vol. 23, no. 6, pp. 982–992, 2015.

[14] L. Sun, J. Du, L.-R. Dai, and C.-H. Lee, "Multiple-target deep learning for LSTM-RNN based speech enhancement," in *Proc. HSCMA*, pp. 136–140, 2017.

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

[16] 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.

[17] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, 2014.

[18] T. Gao, J. Du, L. Xu, C. Liu, L.-R. Dai, and C.-H. Lee, "A unified speaker-dependent speech separation and enhancement system based on deep neural networks," in *Proc. ChinaSIP*, pp. 687–691, 2015.

[19] P. Mowlaee and R. Saeidi, "Target speaker separation in a multi-source environment using speaker-dependent postfilter and noise estimation," in *Proc. ICASSP*, pp. 7254–7258, 2013.

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

[21] S.-W. Fu, Y. Tsao, and X. Lu, "SNR-aware convolutional neural network modeling for speech enhancement.," in *Proc. INTERSPEECH*, pp. 3768–3772, 2016.

[22] J.-C. Hou, S.-S. Wang, Y.-H. Lai, Y. Tsao, H.-W. Chang, and H.-M. Wang, "Audio-visual speech enhancement using multimodal deep convolutional neural networks," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 117–128, 2018.

[23] P. Mowlaee and C. Nachbar, "Speaker dependent speech enhancement using sinusoidal model," in *Proc. IWAENC*, pp. 80–84, 2014.

[24] R. Giri, K. Helwani, and T. Zhang, "A novel target speaker dependent postfiltering approach for multichannel speech enhancement," in *Proc. WASPAA*, pp. 46–50, 2017.

[25] T. Gao, J. Du, L.-R. Dai, and C.-H. Lee, "A unified dnn approach to speaker-dependent simultaneous speech enhancement and speech separation in low snr environments," *Speech Communication*, vol. 95, pp. 28–39, 2017.

[26] Y.-H. Tu, J. Du, and C.-H. Lee, "A speaker-dependent approach to single-channel joint speech separation and acoustic modeling based on deep neural networks for robust recognition of multi-talker speech," *Journal of Signal Processing Systems*, vol. 90, no. 7, pp. 963–973, 2017.

[27] Y. Wang, J. Du, L.-R. Dai, and C.-H. Lee, "A gender mixture detection approach to unsupervised single-channel speech separation based on deep neural networks," *IEEE/ACM TASLP*, vol. 25, no. 7, pp. 1535–1546, 2017.

[28] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks," in *Proc. INTERSPEECH*, pp. 3364–3368, 2017.

[29] H.-S. Lee, Y.-D. Lu, C.-C. Hsu, Y. Tsao, H.-M. Wang, and S.-K. Jeng, "Discriminative autoencoders for speaker verification," in *Proc. ICASSP*, pp. 5375–5379, 2017.

[30] Q. Wang, H. Muckenhirn, K. Wilson, P. Sridhar, Z. Wu, J. Hershey, R. A. Saurous, R. J. Weiss, Y. Jia, and I. L. Moreno, "Voicefilter: Targeted voice separation by speaker-conditioned spectrogram masking," *arXiv preprint arXiv:1810.04826*, 2018.

[31] J. S. Garofalo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "The darpa timit acoustic-phonetic continuous speech corpus cdrom," *Linguistic Data Consortium*, 1993.

[32] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *Proc. ICASSP*, pp. 4052–4056, 2014.

[33] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE/ACM TASLP*, vol. 19, no. 4, pp. 788–798, 2011.

[34] G. Hu and D. Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," *IEEE/ACM TASLP*, vol. 18, no. 8, pp. 2067–2079, 2010.

[35] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *ICASSP*, vol. 2, pp. 749–752, 2001.

[36] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE/ACM TASLP*, vol. 19, no. 7, pp. 2125–2136, 2011.

[37] J. Chen, J. Benesty, Y. Huang, and E. Diethorn, "Fundamentals of noise reduction in spring handbook of speech processing-chapter 43," 2008.