Role of Deep Neural Network in Speech Enhancement: A Review

Chapter · July 2019

DOI: 10.1007/978-981-13-9129-3_8

2 authors, including:

Judith Justin
Avinashilingam University
15 PUBLICATIONS **33** CITATIONS

SEE PROFILE

Role of Deep Neural Network in Speech Enhancement: A Review

D. Hepsiba[1(✉)] and Judith Justin[2]

[1] Department of Instrumentation Engineering,
Karunya Institute of Technology and Sciences, Coimbatore, Tamil Nadu, India
hepsiba@karunya.edu
[2] Department of Biomedical Instrumentation Engineering,
Avinashilingam Institute for Home Science and Higher Education for Women,
Coimbatore, Tamil Nadu, India
hodbmieaul@gmail.com

Abstract. This paper presents a review on different methodologies adopted in speech enhancement and the role of Deep Neural Networks (DNN) in enhancement of speech. Mostly, a speech signal is distorted by background noise, environmental noise and reverberations. To enhance speech, certain processing techniques like Short-Time Fourier Transform, Short-time Auto-correlation and Short-time energy can be adopted. Features such as Logarithmic Power Spectrum (LPS), Mel-Frequency Cepstral Coefficients (MFCC) and Gammatone Frequency Cepstral Coefficient (GFCC) can be extracted and given to DNN for noise classification, so that the noise in the speech can be eliminated. DNN plays a major role in speech enhancement by creating a model with a large amount of training data and the performance of the enhanced speech is evaluated using certain performance metrics.

Keywords: Speech enhancement · Deep Neural Network · Feature extraction · Background noise · Speech signal

1 Introduction

Speech enhancement plays an important role in processing of any speech signal because it tends to be easily affected by different problems such as interference due to environmental noise, background noise and reverberations. Speech enhancement techniques are implemented to eliminate the environmental noise that disturbs the target speech signal and to retrieve the clean speech for applications such as Automatic Speech Recognition (ASR) [21, 22], mobile speech communication, speaker recognition, hearing aids [25, 26] and speech coding [23, 24]. Speech enhancement [1, 2] helps in improving the intelligibility and perceptual quality and also helps in the reduction of noise distortion of a speech signal degraded by adverse conditions. The different types of speech enhancement techniques developed in the past years are spectral subtraction [4], iterative wiener filtering [5], minimum mean square error [6], Kalman Filtering [15] and optimally modified log spectral amplitude [19, 20]. The presence of musical noise in the enhanced speech is the major drawback of these traditional techniques.

J. Hemanth et al. (Eds.): SLAAI-ICAI 2018, CCIS 890, pp. 103–112, 2019.
https://doi.org/10.1007/978-981-13-9129-3_8

Compared to the other traditional techniques Minimum Mean Square Error (MMSE) gives better quality of enhanced speech with lower musical noise [38, 39].

Non-linear DNN based regression models [3] are developed with training data, depending on different conditions and considering factors such as types of noise, noisy speech, noise from speakers and Signal-to-Noise Ratios (SNRs). The performance of the DNN is limited in adverse conditions and in real time noisy situations. To overcome this limitation, and to improve the generalization capability for detecting varying inputs, the training set is formed with hundreds of different noise types. This attempt proved to be efficient in managing the non-stationary behavior of noise and the different categories of unseen noise. This is done by equalizing the global variance of enhanced speech features [6] and the reference clean speech features for reducing the over-smoothing problem. The drop out training [7] is applied on the datasets of neural network when overfitting problem arises. Noise Aware Training (NAT) is done [8] by adding noise information in the DNN inputs to improve the noise robustness and performance in DNN-based speech enhancement systems.

When the need is to separate the noise from the speech signal or to separate a target source from a mixture data, the Non-negative Matrix Factorization (NMF) plays a major role. NMF has a wide scope in acoustic signal detection, speech enhancement, speech recognition in adverse environment, acoustic source separation and many more [9–12]. To increase the performance of NMF target data extraction algorithm with source subspace overlap, the estimation of encoding vectors is done by DNN to reconstruct the desired source data vectors [13]. The mixture data given to the DNN for training includes the clean speech and the noise generated from the interfering sources. DNN modeling is done by mapping the data vectors to its corresponding encoding vectors. Instead of using NMF for separation of clean speech from the mixture data, DNN can be used in two stages: first for separation of clean speech from noisy speech and second for enhancing the clean speech [17]. Another approach for enhancing speech using NMF is the exemplar-based speech enhancement technique [14], where the training clean speech and noisy data are taken in time-frequency representations. Speech and noise have varied modulation frequency content, hence, the Modulation Spectrogram feature holds good in separating the speech and noise in an efficient manner.

Time-Frequency masking is another methodology implemented when background noise causes the major problem [27]. This method improves the magnitude and phase response of the noisy speech through estimating the complex ideal ratio mask in real and imaginary domains. Here the DNN is made to learn the mapping between the reverberant speech and the complex ideal ratio mask [28].

Improved Least Mean Square Adaptive Filtering (ILMSAF) [16] helps in overcoming the drawbacks such as reduced performance in low SNR environments and poor adaptability in different noisy environments. Adaptive filter coefficients estimated by Deep Belief Network (DBN) helps in efficient noise removal. DNN acts as a noise classifier and based on the noise classification the filter parameters are chosen for removing noise.

The most commonly occurring problem in the DNN based algorithm is the reduced performance in mismatched noise condition [3, 6]. To get rid of this problem it is mandatory to have more noise types in the training set. DNN based feature extraction

can also be done to achieve speech enhancement, by learning the mapping in linear-frequency spectral domain [18]. Applying pre-enhancement in the spectral features of the DNN input could help in recovering clean speech features.

The rest of the paper is organized as follows. Section 2 discusses on the different types of databases of clean speech and noise signals. Section 3 elaborates on the processing methodologies adopted for the speech signal. Role of DNN in speech enhancement is explained in Sect. 4.

2 Databases

Database refers to both the clean speech data and noisy speech data that can be utilized for the research findings. The clean speech data is taken from the TIMIT corpus [31]. NTT database [34] has clean speech utterances in eight different languages (English, American English, Japanese, German, Chinese, Spanish, French and Italian). The DARPA-RM [29] database is suitable for training the supervised learning system. Noisy data is taken from Noisex-92 [32], Aurora-2 [30] and Speechdat-Car US (SDC) database [33]. Common noise types taken for training and testing the DNN are Babble, Restaurant, Street, Cafeteria, Machine gun, White, Volvo, Factory1, Buccaneer, etc.

3 Processing

The properties of the speech signal vary with time, and hence, the short-time processing methods that periodically repeat for the waveform duration are utilized. The following are the different processing techniques adopted in the processing of speech signal.

3.1 Short-Time Energy

Short-time energy helps in differentiating the voiced and unvoiced sounds in a speech signal. Thus, the speech and the background noise can be easily detected. The variation in short-time energy [37] determines the difference between the voiced and unvoiced speech segments. The short-time energy is high for voiced segments and low for unvoiced segments and very low for silent speech.

The short-time energy is represented as given in Eq. (1)

$$E_{\hat{n}} = \sum_{m=-\infty}^{\infty} (x[m]w[\hat{n} - m])^2 \tag{1}$$

where

$E_{\hat{n}}$-Energy of the sample n in the signal x
w-Window
m-Number of frames in the signal

3.2 Short-Time Fourier Transform

The Short-Time Fourier Transform (STFT) is the most powerful tool in any audio signal processing, especially in speech signal processing [35]. When a signal with changing frequency such as music, audio signal and speech signal is taken for noise removal, instead of analyzing the whole signal, STFT helps in analyzing the smaller divisions of the signal. The STFT is a function of both time and frequency, therefore, it is represented as time-frequency distribution [36].

The Short-Time Fourier Transform is computed using Eq. (2)

$$X[n,\lambda] = \sum_{m=-\infty}^{\infty} x[n+m]w[m]e^{-j\lambda m} \tag{2}$$

where

$n \in Z$ is a time index and $\lambda \in R$ is a normalized frequency index

3.3 Short-Time Autocorrelation

Autocorrelation is a technique that compares the original signal with the time-delayed version of itself. The Short-Time Autocorrelation is the autocorrelation function of the windowed segment of the speech signal. The voiced and unvoiced speech can be decided based on the peaks of the autocorrelation function [16].

The Short-Time Autocorrelation is denoted as given in Eq. (3)

$$R_{\hat{n}}[k] = \sum_{m=-\infty}^{\infty} (x[m]w[\hat{n}-m])(x[m+k]w[\hat{n}-k-m]) \tag{3}$$

where

$R_{\hat{n}}$ – Short-time autocorrelation at sample n in the signal x
w – Window

4 Deep Neural Networks for Speech Enhancement

DNN has wide scope in audio recognition, speech recognition, speech enhancement and other domains. DNN is a feedforward network and has the capability to model non-linear relationships. The DNN is trained with a collection of data comprising the clean and noisy speech. Different features such as Log-Power Spectra (LPS), Mel Cepstral Coefficients (MFCC) and Gammatone Frequency Cepstral Coefficients (GFCC) are extracted from the speech signal to model the DNN. During training stage, the DNN is made to learn the mapping function and the relationship between the noisy and clean speech, where the noisy data with different levels of Signal-to-Noise Ratio (SNR) are considered. In some cases, DNN is used for noise classification, where adaptive filter coefficients are selected according to the determination of noise. DNN plays a major

role in separation of the source signal from the mixed signal by decreasing the interference and distortion.

4.1 Pre-training DNN with Noisy Data

A collection of clean speech and noisy speech data represented by the log spectra features are given to the regression based DNN model in the training phase. After training, the enhanced log power spectra features are given as input to the DNN model. The DNN concatenates the time axis information in the form of multiple frames and frequency axis information in the form of log spectral features as the input feature vector for DNN learning [3].

It is observed that the performance of DNN based method gives better results compared to the logarithmic minimum mean square error (L-MMSE) method [19, 20, 40] for estimating the noise corrupted target speech. The DNN enhanced spectrogram shows no musical noise and lies closer to the original clean speech spectrogram than the L-MMSE enhanced speech. From the study made on the subjective preference evaluation, it is observed that, on an average, 76.35% of subjects have preferred DNN-based enhanced speech instead of L-MMSE enhanced speech under one or two mismatched noisy environments [3].

4.2 Drop Out Training and Noise Aware Training in DNN

The main drawback in the estimated clean speech is over-smoothing. Equalizing the global variance of estimated clean speech and reference clean speech reduces this problem to an extent. In order to remove the mismatch between the training and testing conditions caused by the different types of noise and various SNR conditions, the drop out training methodology could be adopted. Drop out Training [6] is implemented in DNN by randomly removing certain percentage of neurons from the input, intermediate or hidden layer and treated as a model. Sometimes drop out training causes decrease in performance for matched conditions but gives robustness for mismatched conditions. To give a clean picture on the noise information, Noise Aware Training [6] is done by feeding the DNN with noisy speech samples and subsequent estimation of noise. Thus, the DNN gets trained to determine the clean speech signal.

The DNN enhanced speech suppresses the non-stationary noise and results in less residual noise compared to L-MMSE enhanced speech [41]. From the study made from the subjective preference evaluation, it is observed that, on an average, 78% of the subjects have preferred DNN enhanced speech over the L-MMSE enhanced speech. It is inferred that, the DNN-based speech enhancement system is more efficient in dealing with real world noisy speech in different languages and various recording conditions that is not included in the training [6].

4.3 DNN Based Encoding Vector Estimation

Non-negative Matrix Factorization (NMF) technique is a conventional method which is used to extract encoding data vectors [9, 12]. The performance of conventional NMF based method degrades as the strength of the noise sources increase. The concept of

regression is used for estimating the encoding vectors from a mixture of data given. The mixture data and encoding vectors are mapped and learned by DNN [13].

From performance metrics such as Signal to Distortion Ratio (SDR), Signal to Interference Ratio (SIR), Signal to Artifacts Ratio (SAR) [42] and Perceptual Evaluation of Speech Quality (PESQ) [43], it is observed that the performance of DNN based NMF is good compared to the conventional NMF based techniques and DNN based separation in both matched and mismatched conditions [13].

4.4 DNN Based Noise Classification

Filter parameters play a major role in removal of noise. The filter parameters vary depending on the type of noise. DNN helps in classification of noise and selection of the filter coefficients according to the noise type. In the training phase, the Improved Least Mean Square Adaptive Filtering (ILMSAF) model is trained for different noise types. The enhancement of speech is done by selecting the ILMSAF model according to noise type. The adaptive filter coefficients play a major role in improving the perceptual quality of enhanced speech [16].

The ILMSAF based speech enhancement algorithm with DNN gives better results in terms of speech objective quality measures than the Wiener filtering method used for speech enhancement [44]. The ILMSAF based speech enhancement algorithm with DNN gives a good response in high SNR conditions and extraordinary response in low SNR conditions [16].

4.5 Source Separation and Enhancement Using DNN

The Single Channel Source Separation (SCASS) helps to separate audio source from the mixed signal [45, 47]. The most popular method is the Non-negative Matrix Factorization (NMF) and nowadays DNN is implemented for source separation also [48, 50]. Source separation is adopted by two methods using DNN. The first method maps the features of the mixed signal onto features of the source signal [49, 50]. In the second method, the spectral mask of the mixed signal is mapped, and therefore it contributes to each source in the mixed signal [51]. These methods are used for separating the sources that is distorted due to interference by other sources and distortions. Distortion is eliminated in two stages: In the first stage, the signals are denoised from the background noise, and is termed as the separation stage. Quality of the signal is enhanced in the second stage, which is the enhancement stage [17].

The separation is either done by NMF or DNN and the enhancement is done by DNN using two methods. In the first method, the separated signal is enhanced individually for each source using its own trained DNN. In the second method, a single DNN is used to enhance all the separated sources together. In both the methods, discriminative training is adopted to train the DNN in the enhancement stage. The observations made from the SIR and SDR values show that the quality of the separated sources is improved by decreasing the interference and distortions [17].

4.6 DNN Based Speech Enhancement Systems

Generally, DNN is trained in different conditions such as noise type, gender of the speaker and Signal to Noise Ratio (SNR), to ensure the generalizing capability of the DNN based speech enhancement system [53, 56] in terms of Speech Quality (SQ) and Speech Intelligibility (SI) [2]. A comparison is made in terms of noise specific, speaker specific, and signal-to-noise (SNR) specific system performance with respect to noise general, speaker general and SNR general systems. A single DNN based Speech Enhancement (SE) system has been designed for a specific noise type, speaker & SNR, is compared with the general DNN based SE system designed for various noise types, speakers & SNR and the short-time spectral amplitude minimum mean square error (STSA-MMSE) based Speech Enhancement algorithm [58].

From the performance metrics speech quality and speech intelligibility, it is observed that the DNN based SE system has good generalizing capability when exposed to unseen noise types and speakers. The DNN trained with only one type of noise, one type of speaker and one type of SNR performs excellent when compared with the general DNN based SE system trained with a variety of noise types, speakers and SNR [52].

5 Conclusion

Deep Neural Network is an emerging technique in speech enhancement and has a wide scope for research. Various processing techniques applicable for the enhancement of speech are discussed. The DNN in speech enhancement can be trained in multiple conditions and tested in mismatched conditions to test the efficiency of the network. The performance of the enhanced speech signal is evaluated with different performance metrics such as Short-Time Objective Intelligibility (STOI) score, Perceptual Evaluation of Speech Quality (PESQ), Signal to Distortion Ratio (SDR), Signal to Interference Ratio (SIR) and Signal to Artifact Ratio (SAR).

Thus, it can be concluded that the Deep Neural Network plays a major role in speech recognition, speech enhancement, audio separation and noise classification.

References

1. Benesty, J., Makino, S., Chen, J.D.: Speech Enhancement. Springer, New York, NY (2005)
2. Loizou, P.C.: Speech Enhancement: Theory and Practice. CRC Press, Boca Raton, FL (2013)
3. Xu, Y., Du, J., Dai, L.-R., Lee, C.-H.: An experimental study on speech enhancement based on deep neural networks. IEEE Signal Process. Lett. **21**(1), 65–68 (2014)
4. Boll, S.: Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust. Speech Signal Process. **ASSP-27**(2), 113–120 (1979)
5. Lim, J.S., Oppenheim, A.V.: Enhancement and bandwidth compression of noisy speech. Proc. IEEE **67**(12), 1586–1604 (1979)

6. Xu, Y., Du, J., Dai, L.-R., Lee, C.-H.: A regression approach to speech enhancement based on deep neural networks. IEEE/ACM Trans. Audio Speech Lang. Process. **23**(1), 7–19 (2015)
7. Hinton, G.E., Srivastava, N., Krizheysky, A., Sutskever, I., Salakhutdinoy, R.R.: Improving neural networks by preventing co-adaptation of feature detectors (2012)
8. Seltzer, M., Yu, D., Wang, Y.: An investigation of deep neural networks for noise robust speech recognition. In: Proceedings of ICASSP, pp. 7398–7402 (2013)
9. Jin, Y.G., Kim, N.S.: On detecting target acoustic signal based on negative matrix factorization. IEICE Trans. Inf. Syst. **E93-D**(4), 922–925 (2010)
10. Wilson, K.W., Raj, B., Smaragdis, P., Divakaran, A.: Speech denoising using nonnegative matrix factorization with priors. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 4029–4032 (2008)
11. Weninger, F., Geiger, J., Wllmer, M., Schuller, B., Rigoll, G.: The Munich 2011CHiME challenge contribution: NMF-BLSTM speech enhancement and recognition for reverberated multisource environments. In: Proceedings of 1st International Workshop on Machine Listening in Multisource Environments (CHiME), pp. 24–29 (2011)
12. Grais, E.M., Erdogan, H.: Single channel speech music separation using non-negative matrix factorization and spectral masks. In: Proceedings of International Conference on Digital Signal Process, pp. 1–6 (2011)
13. Kang, T.G., Kwon, K., Shin, J.W., Kim, N.S.: NMF-based target source separation using deep neural network. IEEE Signal Process. Lett. **22**(2), 229–233 (2015)
14. Baby, D., Virtanen, T., Gemmeke, J.F., Van Hamme, H.: Coupled dictionaries for exemplar-based speech enhancement and automatic speech recognition. IEEE/ACM Trans. Audio Speech Lang. Process. **23**(11), 1788–1799 (2015)
15. Grancharov, V., Samuelsson, J., Kleijin, B.: On causal algorithms for speech enhancement. IEEE Trans. Audio Speech Lang. Process. **14**(3), 764–773 (2006)
16. Li, R., et al.: ILMSAF based speech enhancement with DNN and noise classification. Speech Commun. **85**, 53–70 (2016)
17. Grais, E.M., Roma, G., Simpson, A.J.R., Plumbley, M.D.: Two-stage single-channel audio source separation using deep neural networks. IEEE/ACM Trans. Audio Speech Lang. Process. **25**(9), 1773–1783 (2017)
18. Lee, H.-Y., Cho, J.-W., Kim, M., Park, H.-M.: DNN-based feature enhancement using DOA-constrained ICA for robust speech recognition. IEEE Signal Process. Lett. **23**(8), 1091–1095 (2016)

Cohen, I., Berdugo, B.: Speech enhancement for non-stationary noise environments. Signal Process. **81**(11), 2403–2418 (2001)

20. Cohen, I.: Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging. IEEE Trans. Speech Audio Process. **11**(5), 466–475 (2003)
21. Li, J., Deng, L., Haeb-Umbach, R., Gong, Y.: Robust Automatic Speech Recognition: A Bridge to Practical Applications, 1st edn. Academic, Orlando (2015)
22. Li, B., Tsao, Y., Sim, K.C.: An investigation of spectral restoration algorithms for deep neural networks-based noise robust speech recognition. In: Proceedings of Interspeech, pp. 3002–3006 (2013)
23. Li, J., et al.: Comparative intelligibility investigation of single-channel noise reduction algorithms for Chinese, Japanese and English. J. Acoust. Soc. Am. **129**(5), 3291–3301 (2011)
24. Li, J., Sakamoto, S., Hongo, S., Akagi, M., Suzuki, Y.: Two-stage binaural speech enhancement with Wiener filter for high-quality speech communication. Speech Commun. **53**(5), 677–689 (2011)

25. Levitt, H.: Noise reduction in hearing aids: an overview. J. Rehabil. Res. Dev. **38**(1), 111–121 (2001)
26. Chern, A., Lai, Y.H., Chang, Y.-P., Tsao, Y., Chang, R.Y., Chang, H.-W.: A smartphone-based multi-functional hearing assistive system to facilitate speech recognition in the classroom. IEEE Access **5**, 10339–10351 (2017)
27. Williamson, D.S., Wang, Y., Wang, D.: Complex ratio masking for monaural speech separation. IEEE/ACM Trans. Audio Speech Lang. Process. **24**(3), 483–492 (2016)
28. Williamson, D.S., Wang, D.L.: Time-frequency masking in the complex domain for speech dereverberation and denoising. IEEE/ACM Trans. Audio Speech Lang. Process. **25**(7), 1492–1501 (2017)
29. Price, P., Fisher, W.M., Bernstein, J., Pallet, D.: The DARPA 1000-word resource management database for continuous speech recognition. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, New York, NY, USA, pp. 651–654 (1988)
30. Hirschand, H.G.. Pearce, D.: The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions. In: Proceedings of ISCA ITRWASR, pp. 181–188 (2000)
31. Garofolo, J.S.: Getting started with the DARPA TIMIT CD-ROM: an acoustic phonetic continuous speech database. NIST Technical Report (1988)
32. Varga, A., Steeneken, H.J.M.: Assessment for automatic speech recognition: II. NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems. Speech Commun. **12**(3), 247–251 (1993)
33. Moreno et al.: Speech dat-car: a large database for automotive environments. In: Proceedings of International Conference on Language Resources and Evaluation, Athens, Greece, pp. 1–6 (2000)
34. Multi-Lingual Speech Database for Telephonometry, NTT Advanced Technology Corporation, San Jose, CA, USA (1994)
35. Allen, J.B.: Application of the short-time Fourier transform to speech processing and spectral analysis. In: Proceedings of IEEE ICASSP-82, pp. 1012–1015 (1982)
36. Cohen, L.: Time-Frequency Analysis. Englewood Cliffs, Prentice-Hall, Upper Saddle River (1995)
37. de-la-Calle-Silos, F., Stern, R.M.: Synchrony based feature extraction for robust automatic speech recognition. IEEE Signal Process. Lett. **24**(8), 1158–1162 (2017)
38. Cappe, O.: Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor. IEEE Trans. Speech Audio Process. **2**(2), 345–349 (1994)
39. Hussain, A., Chetouani, M., Squartini, S., Bastari, A., Piazza, F.: Nonlinear Speech Enhancement: An Overview. In: Stylianou, Y., Faundez-Zanuy, M., Esposito, A. (eds.) Progress in Nonlinear Speech Processing. LNCS, vol. 4391, pp. 217–248. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-71505-4_12
40. Cohen, I., Gannot, S.: Spectral Enhancement Methods. In: Benesty, J., Sondhi, M.Mohan, Huang, Y.A. (eds.) Springer Handbook of Speech Processing. SH, pp. 873–902. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-49127-9_44
41. Ephraim, Y., Malah, D.: Speech enhancement using minimum mean square log spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. **ASSP-33**(2), 443–445 (1985)
42. Vincent, E., Gribonval, R., Fevotte, C.: Performance measurement in blind audio source separation. IEEE Trans. Audio Speech Lang. Process. **14**(4), 1462–1469 (2006)
43. ITU, Perceptual Evaluation of Speech Quality (PESQ): an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs ITU-T Rec. p. 862 (2000)

44. Li, R., Bao, C., Xia, B., Jia, M.: Speech enhancement using the combination of adaptive wavelet threshold and spectral sub-traction based on wavelet packet decomposition. In: 2012 IEEE 11th International Conference on Signal Processing (ICSP), vol. 1, pp. 481–484 (2012)
45. Virtanen, T.: Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria. IEEE Trans. Audio Speech Lang. Process. **15**(3), 1066–1074 (2007)
46. Smaragdis, P.: Convolutive speech bases and their application to supervised speech separation. IEEE Trans. Audio Speech Lang. Process. **15**(1), 1–12 (2007)
47. Smaragdis, P., Shashanka, M., Raj, B.: A sparse non-parametric approach for single channel separation of known sounds. In: Neural Information Processing Systems, Vancouver, BC, Canada, Dec 2009, pp. 1705–1713
48. Grais, E.M., Roma, G., Simpson, A.J.R., Plumbley, M.D.: Single channel audio source separation using deep neural network ensembles. In: Proceedings of 140th Audio Engineering Society Convention, Paper no. 9494 (2016)
49. Huang, P.-S., Kim, M., Hasegawa-Johnson, M., Smaragdis, P.: Singing-voice separation from monaural recordings using deep recurrent neural networks. In: Proceedings of International Society for Music Information Retrieval Conference, pp. 477–482 (2014)
50. Huang, P.-S., Kim, M., Hasegawa-Johnson, M., Smaragdis, P.: Deep learning for monaural speech separation. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 1562–1566 (2014)
51. Weninger, F., Hershey, J.R., Roux, J.L., Schuller, B.: Discriminatively trained recurrent neural networks for single-channel speech separation. In: Proceedings of IEEE Global Conference on Signal and Information Processing, pp. 577–581 (2014)
52. Kolbæk, M., Tan, Z.-H.: Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems. IEEE/ACM Trans. Audio Speech Lang. Process. **25**(1), 153–167 (2017)
53. Lee, T., Theunissen, F.: A single microphone noise reduction algorithm based on the detection and reconstruction of spectro-temporal features. Proc. R. Soc. Lond. A Math. Phys. Sci. **471**, 2184 (2015)
54. Huang, P.-S., Kim, M., Hasegawa-Johnson, M., Smaragdis, P.: Joint optimization of masks and deep recurrent neural networks for monaural source separation. IEEE/ACM Trans. Audio Speech Lang. Process. **23**(12), 2136–2147 (2015)
55. Liu, D., Smaragdis, P., Kim, M.: Experiments on deep learning for speech denoising. In: Proceedings of INTERSPEECH, pp. 2685–2689 (2014)
56. Wang, Y., Chen, J., Wang, D.: Deep neural network based supervised speech segregation generalizes to novel noises through large-scale training. Dept. Comput. Sci. Eng. Ohio State Univ., Columbus, OH, USA, Technical Report OSU-CISRC-3/15-TR02 (2015)
57. Hendriks, R.C., Gerkmann, T., Jensen, J.: DFT-Domain Based Single-Microphone Noise Reduction for Speech Enhancement: A Survey of the State of the Art. Synthesis Lectures on Speech and Audio Processing, vol. 9, pp. 1–80. Morgan & Claypool, SanRafael, CA (2013)
58. Erkelens, J., Hendriks, R., Heusdens, R., Jensen, J.: Minimum mean square error estimation of discrete Fourier coefficients with generalized Gamma priors. IEEE/ACM Trans. Audio Speech Lang. Process. **15**(6), 1741–1752 (2007)