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

Robust Example Search Using Bottleneck Features for Example-based Speech Enhancement

Atsunori Ogawa[1], *Shogo Seki*[1,2], *Keisuke Kinoshita*[1], *Marc Delcroix*[1], *Takuya Yoshioka*[1], *Tomohiro Nakatani*[1] *and Kazuya Takeda*[2]

[1]NTT Communication Science Laboratories, NTT Corporation
[2]Graduate School of Information Science, Nagoya University

Abstract

Example-based speech enhancement is a promising approach for coping with highly non-stationary noise. Given a noisy speech input, it first searches in noisy speech corpora for the noisy speech examples that best match the input. Then, it concatenates the clean speech examples that are paired with the matched noisy examples to obtain an estimate of the underlying clean speech component in the input. This framework works well if the noisy speech corpora contain the noise included in the input. However, it is impossible to prepare corpora that cover all types of noisy environments. Moreover, the example search is usually performed using noise sensitive mel-frequency cepstral coefficient features (MFCCs). Consequently, a mismatch between an input and the corpora is inevitable. This paper proposes using bottleneck features (BNFs) extracted from a deep neural network (DNN) acoustic model for the example search. Since BNFs have good noise robustness (invariance), the mismatch is mitigated and thus a more accurate example search can be performed. Experimental results on the Aurora4 corpus show that the example-based approach using BNFs greatly improves the enhanced speech quality compared with that using MFCCs. It also consistently outperforms a conventional DNN-based approach, i.e. a denoising autoencoder.
Index Terms: example-based speech enhancement, example search, bottleneck feature

1. Introduction

Speech enhancement is an essential technology for significantly improving the quality of speech-based applications in adverse environments. A lot of effort has been expended over the years on developing various types of effective speech enhancement approaches [1]. In particular, single-channel approaches have been extensively studied (e.g. [1–20]), since they impose very few hardware constraints compared with multi-channel approaches.

Conventional filtering-based single-channel approaches (e.g. [1–7]) estimate noise statistics and then use them to filter out the noise component from a noisy speech input. The advantage of these approaches is their low computational complexity. However, tracking the statistics of a highly non-stationary noise remains a difficult task.

Deep neural network (DNN) technology has led to a new trend in single-channel approaches, i.e. denoising autoencoders (DAEs) (e.g. [8–12]). A DAE is a DNN trained by using a noisy-clean parallel speech corpus to map noisy input features (e.g. log-power spectra) to clean feature estimates. Thanks to nonlinear feature transformations through stacked hidden layers in a DNN, a DAE has a superior mapping ability and shows a high denoising performance in various noisy environments.

In this paper, we focus on an *example-/corpus-based (or inventory-style)* approach (e.g. [13–20]). As with a DAE, an example-based approach directly estimates the underlying clean speech component in a given noisy input using a noisy-clean parallel speech corpus. However, it focuses strongly on exploiting *raw and precise data*, i.e. *examples*, included in the speech corpora. The example-based approach originally proposed in [13, 14] can be outlined as follows (see Section 2 for details). It prepares a clean speech corpus and the corresponding artificially contaminated noisy speech corpora. In the testing stage, given a noisy speech input, it first searches in the noisy speech corpora for noisy speech examples (segments) that best match the input. Then, it concatenates the corresponding clean speech examples included in the clean speech corpus to obtain an estimate of the underlying clean speech component in the input. Finally, it uses this clean speech estimate to denoise the input. The example search is performed based on a *longest matching* criterion. This criterion is important since longer speech examples can be identified more accurately in noisy environments than shorter examples because of their more distinct and richer spectral-temporal pattern information. As a result, the example-based approach exhibits higher enhancement performance than the conventional approaches especially in highly non-stationary noisy environments [13–17].

However, in previous studies of the example-based approach, the example search was not always performed accurately enough. Although it is desirable that the noisy speech corpora encompass all the noisy environments that we encounter at test time, in reality, this is infeasible. Moreover, since the example search is performed by evaluating similarity between an input and a noisy example both of which are represented by acoustic features, typically, *mel-frequency cepstral coefficients* (*MFCCs*), which are sensitive to noise, the search process can be greatly affected by noise. Therefore, a mismatch between an input and the noisy speech corpora is inevitable. This mismatch makes the example search less accurate and therefore degrades the quality of the enhanced speech. In addition, it is clear that the cost of the example search is very high, especially when large speech corpora are used [15, 16, 20].

In this paper, we propose the use of *bottleneck features* (*BNFs*) [21–23] extracted from a DNN acoustic model as a representation of a noisy speech input and a noisy speech example and perform a robust example search based on them (Section 3). Since BNFs have good *noise robustness* (*invariance*) [21, 24], the mismatch problem between an input and the noisy speech corpora can be mitigated, and thus a more accurate example search can be conducted. Experimental results on the Aurora4 corpus [25, 26] show that the example-based approach using BNFs greatly improves the enhanced speech quality compared with that using MFCCs (Section 4). It also consistently outperforms a DAE, i.e. a DNN-based strong competitor. In addition, because of the BNFs' *discriminative* property [21, 24], many unlikely example hypotheses can be pruned efficiently during the example search, and thus the example search can be greatly accelerated compared with when using MFCCs.

2. Example-based speech enhancement

This section briefly describes the basic framework of the example-based approach, which was originally proposed in [13, 14], using Fig. 1 (see [16] for further details), and elaborates the problems in the example search.

http://dx.doi.org/10.21437/Interspeech.2016-671

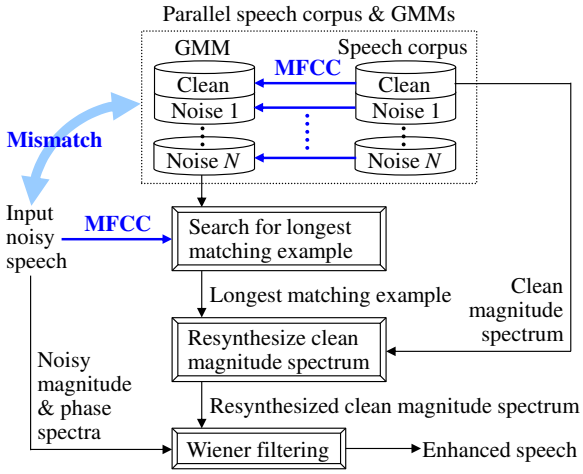

Figure 1: *Basic framework of example-based speech enhancment.*

2.1. Basic framework

In the training stage (dotted box in Fig. 1), a clean speech corpus is first prepared. It is then artificially contaminated with various types of noise to form a parallel speech corpus. Feature extraction is performed for all of these speech corpora. Here, the features are typically conventional *MFCCs*. As regards the clean corpus, the magnitude spectra are also extracted. Using the extracted MFCCs, Gaussian mixture models (GMMs) are trained that represent each of the corpora. To represent the precise spectral patterns of speech, the dimensionality of the MFCCs and the number of mixture components in the GMMs are set at large values (e.g. 80 and 4096).

In the example search stage, given a noisy speech input, we first extract its MFCC, magnitude spectrum and phase spectrum sequences. Using the GMMs and the noisy speech corpora, and based on an example evaluation function [16], we search for a sequence of *longest matching noisy examples* (*segments*) in the input noisy speech.

In the enhancement stage, using the found matching noisy example sequence and the clean speech corpus, we resynthesize a clean magnitude spectrum sequence by concatenating the corresponding clean speech magnitude spectra. Finally, we perform Wiener filtering to obtain the final enhanced speech using the resynthesized clean magnitude spectrum sequence and the magnitude/phase spectrum sequences extracted from the input noisy speech.

2.2. Problems in example search

As we have already mentioned in Section 1, it is impossible to prepare noisy speech corpora that cover all types of noisy environments. Moreover, the example search is usually performed using conventional noise sensitive MFCCs. Therefore, a mismatch between a noisy speech input and the noisy speech corpora is inevitable as shown in Fig. 1. This mismatch can degrade the accuracy of the example search and, as a result, the quality of the enhanced speech.

In addition, the cost of the example search is very high [15, 16, 20]. We greatly accelerated the example search in [16] by introducing a *tree and linear connected search space*. In this search space, we can perform a shared likelihood calculation for many example hypotheses with efficient pruning of unlikely hypotheses. However, the example search must be accelerated further, especially, when we use large speech corpora.

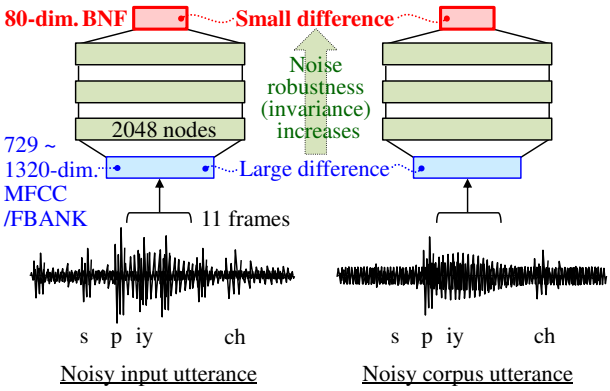

Figure 2: *Extraction of bottleneck features (BNFs) for the same two utterances "speech", which are contaminated with different noises.*

3. Robust example search using BNFs

The *noise robustness* of acoustic models used in automatic speech recognition (ASR) can be greatly increased by incorporating *DNN* technology, e.g. [21–25]. The noise robustness of a DNN acoustic model comes from multiple nonlinear feature transformations through its stacked hidden layers. In noisy environments, acoustic features (e.g. MFCCs or log-mel filterbank features (FBANKs)) are usually influenced by noise. However, if they are input into a DNN and layer-by-layer transformations are applied, the influence of the noise is significantly decreased. Consequently, the features extracted from the higher (deeper) hidden layers are less influenced by noise [21, 24]. It is reasonable to exploit this *noise robustness* (*noise invariance*) of DNN-based features in the example search to tackle the mismatch problem described in Section 2.2.

However, the direct use of DNN-based features for GMM training is difficult because of their high dimensionality (typically, 2048, i.e. the number of nodes in a hidden layer). Therefore, we use features extracted from a *bottleneck hidden layer* (e.g. [21–23]), which has a smaller number of nodes (e.g. 80) than the other layers. When using BNFs instead of MFCCs, we do not need to change the basic framework of the example-based approach described in Section 2.1.

Figure 2 shows why a more accurate example search can be performed with BNFs than with MFCCs. In this figure, we assume that an utterance is given as the input and the same utterance is included in the corpus, and that these two utterances are contaminated with different noises. Ideally, the input utterance should match the corpus utterance. However, if we use MFCCs, the input utterance would not match the corpus utterance, since MFCCs are influenced by noise, and the extracted MFCCs for the two utterances would be very different. In contrast, if we use BNFs transformed from the MFCCs (or FBANKs) by a DNN, the input utterance would match the corpus utterance, since the influences of the noises are mitigated by the layer-by-layer nonlinear feature transformations, and the extracted BNFs for the two utterances would be similar.

There are other advantages of using BNFs instead of MFCCs in the example search. Features input to a DNN (e.g. MFCCs or FBANKs) are spliced for several (typically, 11) frames, as a result, the typical dimensionality of the input features is around from 729 to 1320. Thus, we can consider a longer context of a given speech signal in the example search. A DNN used in ASR is trained to discriminatively predict triphone states. This indicates that BNFs are also *discriminative* features [21, 24]. Thus, we can perform the example search while predicting the content of a given utterance (in Fig. 2, "speech") at triphone state level granularity. In addition, be-

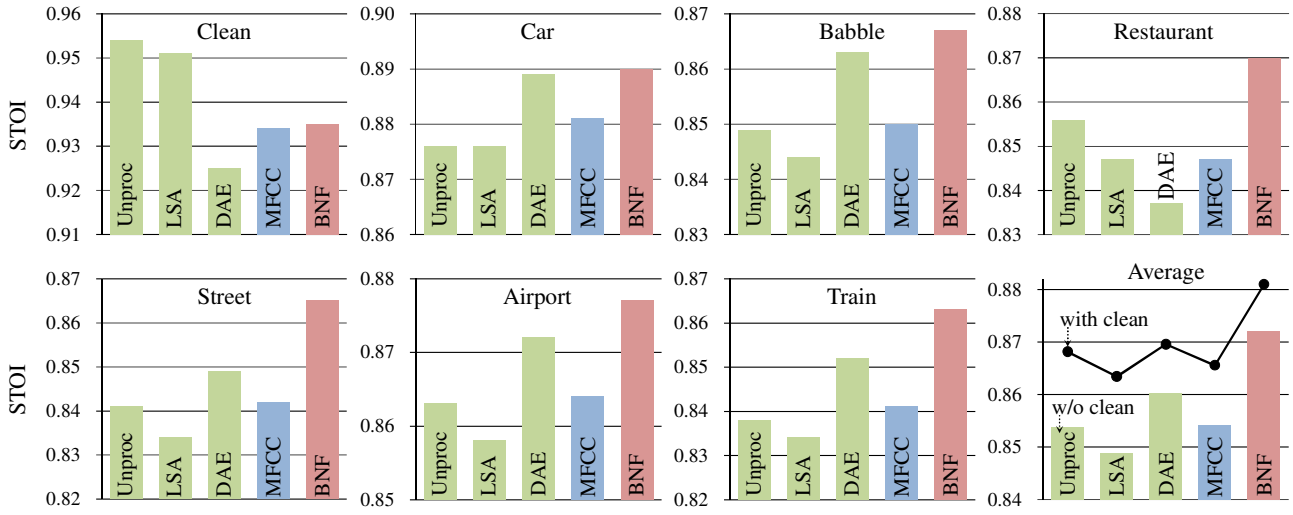

Figure 3: *Short-time objective intelligibility measure (STOI* $\in[0,1]$*, a correlation coefficient, larger is better) values of unprocessed input speech (referred to as Unproc in this figure), and those of the enhanced speech obtained by the optimally modified log-spectral amplitude estimator (LSA), the denoising autoencoder (DAE), and the example-based approaches with MFCCs (MFCC) and bottleneck features (BNF), for the clean and each of the six noise conditions (and their averages with and without the clean condition) in the Aurora4 corpus. The obtained values are averaged over the different microphones for each of the seven conditions.*

cause of the BNFs' discriminative property, during the example search, the likelihood of a top example hypothesis tends to become much higher than those of the other example hypotheses. Thus, by using BNFs, we can efficiently prune unlikely example hypotheses and greatly accelerate the example search.

On the other hand, there is a concern as regards using BNFs. BNFs are also *robust to speaker variability* [21, 24]. Because of this *speaker normalization* ability, it may be difficult to use speaker information included in given utterances in the example search. As a result, the enhanced speech may lose the original speaker characteristics. We will investigate this concern experimentally in the next section.

4. Experiments

We conducted experiments to evaluate the proposed example-based approach using BNFs in comparison with that using MFCCs, a conventional filtering-based approach, and a DAE.

4.1. Experimental settings

The Aurora4 multi-condition corpus [25, 26] was used in the experiments. The corpus is derived from the Wall Street Journal (WSJ0) 5k-word closed vocabulary ASR task. The training set consists of 7138 utterances spoken by 83 speakers (about 14 hours in total). Half of them were recorded with a close talking microphone while the other half were recorded with a desk mounted secondary microphone. Each part was further divided into seven subsets. One subset was left unprocessed while six different types of noise (car, babble, restaurant, street traffic, airport, train station) were added to each of the remaining subsets with 10-20 dB signal-to-noise ratios (SNRs). A clean training set corresponding to the above multi-condition training set was also used as a corpus providing clean magnitude spectrum examples (Fig. 1).

Aurora4 has 14 evaluation sets, each with different environmental conditions, to allow systems to be evaluated under different noise conditions. Each evaluation set contains 330 utterances from eight speakers different from those of the training set. As with the training set, seven of the 14 evaluation sets were recorded with a close talking microphone while the remaining sets were recorded with a secondary microphone. The same six different types of noise as used with the training sets were added to the six close talking and six secondary microphone evaluation sets with various SNRs ranging from 5 to 15 dB. There are 4620 evaluation utterances in total.

The sampling frequency, frame length, and frame shift were 16 kHz, 20 ms, and 10 ms, respectively. For the conventional example-based approach with MFCCs (hereafter, referred to as *ExB w/ MFCC*), we extracted 80-dimensional (including a one-dimensional log energy term) MFCCs for the multi-condition training set described above and trained a GMM with 4096 Gaussian mixture components used for the example search.

For the proposed example-based approach with BNFs (hereafter, referred to as *ExB w/ BNF*), we first trained a basic fully connected feedforward DNN used for ASR according to a standard recipe [27] using the multi-condition training set. 729-dimensional features input to the DNN were obtained by splicing 24-dimensional mean-and-variance-normalized FBANKs plus their delta and delta-delta features within an 11-frame context window. The output layer of the DNN corresponds to 3040 triphone states defined by a baseline GMM-hidden Markov model (HMM) system. The DNN has seven hidden layers. The sixth layer is a bottleneck layer that has 80 nodes and the other layers have 2048 nodes. Using this DNN, we extracted 80-dimensional BNFs for the multi-condition training set and trained a GMM with 4096 components that was used for the example search.

We also evaluated two other competitor approaches. One is the *optimally modified log-spectral amplitude estimator* (*OM-LSA*) [4, 5], which is a conventional filtering-based approach that is available as a MATLAB tool [28]. OM-LSA is a good competitor for our approach in Aurora4, since the noise of Aurora4 is *moderately* non-stationary and thus can be suppressed by filtering-based approaches. The other is a *DAE*, which is a DNN-based strong competitor. The DAE is based on a basic fully connected feedforward DNN. It has four 2048-node hidden layers and is trained using the Aurora4 corpus with a minimum mean square error criterion. The input/output features are log-power spectra. Details of this DAE are described in [12].

The objective evaluation measure is the *short-time objective intelligibility measure* (*STOI*) [29, 30]. A larger STOI ($\in[0,1]$, a correlation coefficient) indicates higher quality enhanced speech. STOI has a high correlation with speech intelligibility. The two example-based systems were implemented in C with our fast example search algorithm [16]. They were run on a Linux system with Intel Xeon CPU E5-2650 v2 2.60GHz

and their example search speeds were measured with the *real time factor* (*RTF*). A smaller RTF indicates a faster search.

4.2. Experimental results

Figure 3 shows the obtained STOI values. Looking at the "average" results, we can confirm that it is difficult to improve the STOI value by applying a speech enhancement technique since the values obtained by OM-LSA, DAE and ExB w/ MFCC are slightly better or worse than those of the unprocessed input speech. In contrast, ExB w/ BNF greatly improves the STOI values. Looking at each of the six noise conditions, with the exception of the "restaurant" noise, DAE shows the second best results. In contrast, ExB w/ BNF consistently shows the best results. Under stationary noisy conditions such as "car" and "babble", ExB w/ BNF performs only slightly better than DAE. However, under non-stationary noisy conditions such as "restaurant", "street" and "train", ExB w/ BNF performs much better than DAE. It should be noted that, although the STOI value for "clean" speech input obtained by ExB w/ BNF is not very good, its audible quality is perfectly fine.

Figure 4 shows spectrogram samples. We can confirm that appropriate denoising is realized by ExB w/ BNF. Comparing the spectrograms obtained with ExB w/ MFCC and ExB w/ BNF, we can confirm that ExB w/ MFCC fails to denoise in the silence region while ExB w/ BNF denoises successfully (surrounded by dotted boxes). We can also find similar results in the other spectrogram samples. These results are attributable to the ability of the DNN from which the BNFs are extracted. The DNN is trained to discriminatively predict the triphone states including the silence states (Section 3), and thus it can accurately discriminate silence regions from speech regions.

The audible quality of the enhanced speech obtained by ExB w/ BNF is steadily improved compared with that obtained by ExB w/ MFCC. We had a concern that the speaker characteristics may be lost in the enhanced speech obtained by ExB w/ BNF because of BNFs' *speaker normalization* ability [21, 24] (Section 3). However, it maintains the original speaker characteristics well. This can be attributed to the fact that, in the last stage of the enhancement, we perform Wiener filtering using an original noisy speech input [13, 14] (Section 2.1). To ensure that we exploit the speaker information in the example search, we can use i-vectors (e.g. [31–33]) along with BNFs.

Finally, Table 1 shows the RTF measurement results of the example search. We can confirm that ExB w/ MFCC is fast enough by employing our fast example search algorithm proposed in [16]. However, ExB w/ BNF is even faster. It is about 8.5 times faster than ExB w/ MFCC. Note that these RTF values do not include the time for feature extraction. Actually, the cost of the BNF extraction is higher than that of the MFCC extraction. However, the cost of the feature extraction is negligible compared with the potential cost of an example search (i.e. the cost of a frame-by-frame example search [16] in large speech corpora). In addition, the BNF extraction can be accelerated by using general purpose graphical processing units [21].

5. Relation to previous work

The mismatch problem between a noisy speech input and the noisy speech corpora is tackled in [15, 17] by exploiting filtering-based approaches that estimate noise statistics (Section 1). [15] proposes the interconnection of a filtering-based approach and the example-based approach. With this method, the filtering-based approach acts as a preprocessor that reduces the noise component in noisy inputs. As a result, the variety of noisy environments that the noisy speech corpora should cover is reduced. Our approach differs from this method since it does not change the basic framework of the example-based approach but simply changes the features used in the example search. Consequently, we can combine our approach with [15] to achieve further performance improvements.

[17] proposes an example-based approach that does not use the noisy speech corpora (but only uses the clean speech corpus). This approach estimates the noise statistics in a noisy speech input as with the filtering-based approaches and combines them with the examples in the clean speech corpus to reconstruct the noisy input. This approach can be understood as a hybrid of the filtering- and example-based approaches. This is essentially different from our approach and would be worth pursuing.

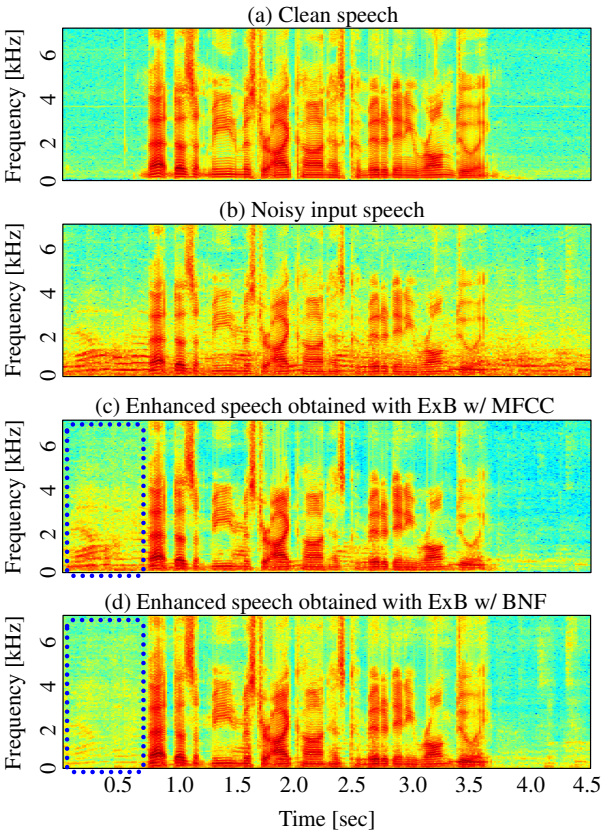

Figure 4: *Spectrogram samples of (a) clean speech, (b) noisy input speech, and enhanced speech obtained with the example-based approaches with (c) MFCCs and (d) bottleneck features (BNFs), for "airport" noise.*

Table 1: *Real time factors (RTFs) of the example search obtained with example-based approaches with MFCCs and bottleneck features (BNFs). A smaller RTF indicates a faster search.*

ExB w/ MFCC	0.825
ExB w/ BNF	0.096

6. Conclusion and future work

We have proposed using bottleneck features (BNFs) extracted from a DNN acoustic model for a robust example search in example-based speech enhancement. Because of the BNFs' noise robust and discriminative properties, the proposed approach provides high quality enhanced speech while achieving a very fast example search.

Future work will include an evaluation using the other corpora (e.g. [34–36]), a comparison with more sophisticated DAEs (e.g. [8–12]), an evaluation using i-vectors (e.g. [31–33]) along with BNFs, and an evaluation as an ASR frontend.

7. References

[1] P. Loizou, *Speech Enhancement: Theory and Practice, Second Edition*. CRC Press, 2013.

[2] R. Martin, "An efficient algorithm to estimate the instantaneous SNR of speech signals," in *Proc. Eurospeech*. ISCA, 1993, pp. 1093–1096.

[3] ——, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, Jul. 2001.

[4] I. Cohen, "Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator," *IEEE Signal Processing Letters*, vol. 9, no. 4, pp. 113–116, Apr. 2002.

[5] ——, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, Sep. 2003.

[6] R. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," in *Proc. ICASSP*. IEEE, 2010, pp. 4266–4269.

[7] M. Souden, M. Delcroix, K. Kinoshita, T. Yoshioka, and T. Nakatani, "Noise power spectral density tracking: A maximum likelihood perspective," *IEEE Signal Processing Letters*, vol. 19, no. 8, pp. 495–498, Aug. 2012.

[8] A. Maas, Q. Le, T. O'Neil, O. Vinyals, P. Nguyen, and A. Ng, "Recurrent neural networks for noise reduction in robust ASR," in *Proc. Interspeech*. ISCA, 2012.

[9] 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, no. 1, pp. 65–68, Jan. 2014.

[10] K. Han, Y. Wang, D. Wang, W. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 6, pp. 982–992, Jun. 2015.

[11] F. Weninger, S. Watanabe, Y. Tachioka, and B. Schuller, "Deep recurrent de-noising auto-encoder and blind de-reverberation for reverberated speech recognition," in *Proc. ICASSP*. IEEE, 2014, pp. 4623–4627.

[12] K. Kinoshita, M. Delcroix, A. Ogawa, and T. Nakatani, "Text-informed speech enhancement with deep neural networks," in *Proc. Interspeech*. ISCA, 2015, pp. 1760–1764.

[13] J. Ming, R. Srinivasan, and D. Crookes, "A corpus-based approach to speech enhancement from nonstationary noise," in *Proc. Interspeech*. ISCA, 2010, pp. 1097–1100.

[14] ——, "A corpus-based approach to speech enhancement from nonstationary noise," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 822–836, May 2011.

[15] K. Kinoshita, M. Souden, M. Delcroix, and T. Nakatani, "Single channel dereverberation using example-based speech enhencement with uncertainty decoding technique," in *Proc. Interspeech*. ISCA, 2011, pp. 197–200.

[16] A. Ogawa, K. Kinoshita, T. Hori, T. Nakatani, and A. Nakamura, "Fast segment search for corpus-based speech enhancement based on speech recognition technology," in *Proc. ICASSP*. IEEE, 2014, pp. 1576–1580.

[17] J. Ming and D. Crookes, "An iterative longest matching segment approach to speech enhancement with additive noise and channel distortion," *Computer Speech and Language*, vol. 28, no. 6, pp. 1269–1286, Nov. 2014.

[18] X. Xiao and R. Nickel, "Speech enhancement with inventory style speech resynthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1243–1257, Aug. 2010.

[19] R. Nickel, R. Astudillo, D. Kolossa, and R. Martin, "Corpus-based speech enhancement with uncertainty modeling and cepstral smoothing," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 983–997, May 2013.

[20] R. Nickel and R. Martin, "Memory and complexity reduction for inventory-style speech enhancement systems," in *Proc. EUSIPCO*. EURASIP, 2011, pp. 196–200.

[21] D. Yu and L. Deng, *Automatic speech recognition: a deep learning approach*. Springer-Verlag London, 2015.

[22] F. Grézl and P. Fousek, "Optimizing bottle-neck features for LVCSR," in *Proc. ICASSP*. IEEE, 2008, pp. 4279–4732.

[23] T. Yoshioka and M. Gales, "Environmentally robust ASR front-end for deep neural network acoustic models," *Computer Speech and Language*, vol. 31, no. 1, pp. 65–86, May 2015.

[24] D. Yu, M. Seltzer, J. Li, J. Huang, and F. Seide, "Feature learning in deep neural networks - studies on speech recognition tasks," *arXiv:1301.3605v3 [cs.LG]*.

[25] M. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Proc. ICASSP*. IEEE, 2013, pp. 7398–7402.

[26] G. Hirsch, "Experimental framework for the performance evaluation of speech recognition front-ends on a large vocabulary task, version 2.0, AU/417/02." ETSI STQ-Aurora DSR Working Group, 2002.

[27] G. Dahl, T. Sainath, and G. Hinton, "Improving deep neural networks for LVCSR using rectified linear units and dropout," in *Proc. ICASSP*. IEEE, 2013, pp. 8609–8613.

[28] "OM-LSA," http://webee.technion.ac.il/people/IsraelCohen/.

[29] C. Taal, R. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

[30] "STOI," http://www.ceestaal.nl/matlab-code/.

[31] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, May 2011.

[32] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *Proc. ASRU*. IEEE, 2013, pp. 55–59.

[33] C. Yu, A. Ogawa, M. Delcroix, T. Yoshioka, T. Nakatani, and J. Hansen, "Robust i-vector extraction for neural network adaptation in noisy environment," in *Proc. Interspeech*. ISCA, 2015, pp. 197–200.

[34] E. Vincent, J. Barker, S. Watanabe, J. L. Roux, F. Nesta, and M. Matassoni, "The second 'CHiME' speech separation and recognition challenge: datasets, tasks and baselines," in *Proc. ICASSP*. IEEE, 2013, pp. 126–130.

[35] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, R. Haeb-Umbach, V. Leutnant, A. Sehr, W. Kellermann, R. Maas, S. Gannot, and B. Raj, "The REVERB challenge: a common evaluation framework for dereverberation and recognition of reverberant speech," in *Proc. WASPAA*. IEEE, 2013.

[36] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: dataset, task and baselines," in *Proc. ASRU*. IEEE, 2015, pp. 504–511.