An Adaptation Method in Noise Mismatch Conditions for DNN-Based Speech Enhancement

Siying Xu [a)], Dan Qu [b)], Xingyan Long [c)]

China National Digital Switching System Engineering & Technological R&D Center, 450002, China

[a)] Corresponding author: xuisying_2015@163.com
[b)] qudanqudan@sina.com
[c)] lxy120999@qq.com

Abstract. The deep learning-based speech enhancement has shown considerable success. However, it still suffers performance degradation under mismatch conditions. In this paper, an adaptation method is proposed to improve the performance under noise mismatch conditions. Firstly, we advise a noise aware training by supplying identity vectors (i-vectors) as parallel input features to adapt DNN acoustic models with the target noise. Secondly, given a small amount adaptation data, the noise-dependent DNN is obtained by using Euclidean distance regularization from a noise-independent DNN, and forcing the estimated masks to be close to the unadapted condition. Finally, experiments were carried out on different noise and SNR conditions, and the proposed method has achieved significantly 29% benefits of STOI at most and provided consistent improvement in PESQ and segSNR against the baseline systems.

Key words: Enhancement; adaptation method; DNN; PESQ; segSNR; STOI.

INTRODUCTION

Speech enhancement is an essential part of speech signal processing in noisy environment, aiming at separating useful clean speech from noisy speech and improving the intelligibility and quality of speech contaminated. It is broadly applied in many domains, such as automatic speech recognition (ASR), speaker identification systems etc. Classical speech enhancement methods including several monaural methods and recently proposed deep neural network-based enhancement [1] usually degrade rapidly in mismatch conditions. So far, speech enhancement in mismatch condition is still a very challenging problem.

In the situation of environment mismatch, noise adaptation can help to improve the modeling accuracies under the unseen type of noise by using a small amount of adaptation data. There are mainly two kinds of noise adaptation algorithms, the one is using Gaussian mixture models (GMMs) and the other is based on DNN framework. Traditional GMM based noise adaptation methods include parameter adaptation method like feature normalization such as feature-space maximum likelihood linear regression (fMLLR) [2] and noise-aware training [3], etc. FMLLR applies an affine transform to the feature vector so that the transformed feature better matches the model. For GMM-HMMs, fMLLR transforms are estimated to maximize the likelihood of the adaptation data given to the model. [2] proposed feature-space discriminative linear regression (fDLR) in DNN framework, where cross-entropy (CE), a discriminative function, is used as optimization criterion. In [3], noise-aware training (NaT) is proposed, which the DNN is being given noise estimation in order to automatically learn the mapping from the noisy speech and noise to the ideal mask labels, implicitly through a clean speech estimation. Noise information can be jointly learned with the rest of the model parameters, or it can be estimated completely independent of the DNN training. In recent years, some speaker adaptation methods have been successfully utilized for noise adaptation. Among the speaker adaptation methods, in [4], i-vector [5] method is used. I-vector is a popular technique for speaker verification and recognition. It encapsulates the most important information about a speaker's, noise's or device's identity in a low-dimensional fixed-length representation and thus is an attractive tool for speaker adaptation techniques for ASR. We consider that when the data is composed of few speakers, the i-vectors represent for noise information mostly. I-vector has become the common speaker adaptation feature, rarely used in speech enhancement as a representation of environment information.

In this paper, a noise adaptation method based on DNN for speech enhancement is proposed to ameliorate the mismatching problem under multi-type noise condition. We combine noise-aware training (NaT) and Euclidean distance (ED) regularization. We use NaT to obtain noise information as auxiliary features, and the DNN can tune model parameters with it. And ED regularization is added to original adaptation criterion, so that we can use a small amount of adaptation data, improving the performance of mismatching system and ameliorating the noise adaptation.

The rest of the paper is organized as follows. Section 2 describes the DNN framework. We introduce noise-aware training (NaT) with identity-vector (i-vector) and Euclidean distance regularization for speech enhancement in Section 3 and Section 4 respectively. In Section 5, we show experimental settings and report some results. Section 6 concludes this paper.

SYSTEM OVERVIEW

Speech enhancement can be interpreted as the process that maps a noisy signal to a separated signal with improved intelligibility and/or perceptual quality. Without considering the impact of phase, this is often treated as the estimation of clean speech magnitude or ideal masks. Acoustic features extracted from a mixed signal, along with the corresponding desired outputs are fed into a learning machine for training. Enhanced speech is obtained by sending estimated outputs and mixture phase into a resynthesizer. FIGURE.1 shows the diagram of the evaluation system.

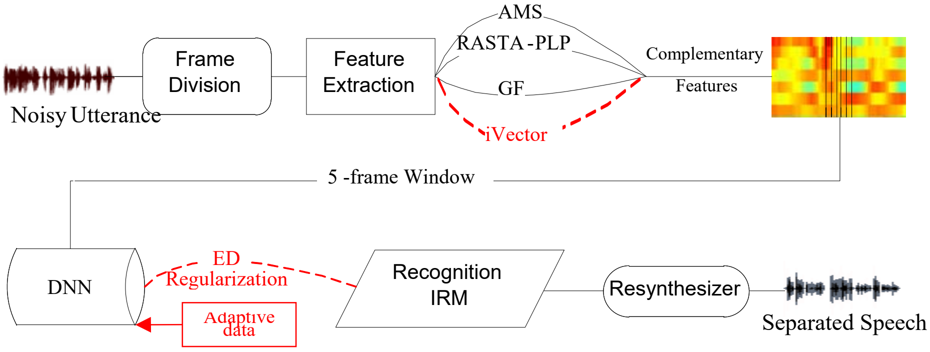

FIGURE 1. Proposed System Overview

Firstly, the noisy utterances are passed through the predesigned 64-channel gammatone filters and then divided into frames every 20ms, 10ms overlap. Secondly, several classical acoustic features are used in the baseline system, including Amplitude Modulation Spectrogram (AMS) [6], Relative Spectral Transformed Perceptual Linear Prediction Coefficients (RASTA-PLP) [7], Gamma-tone filter bank power spectra (GF) [8] and i-vectors. To further incorporate temporal context, a 5-frame window of features are input to the DNNs. Except i-vectors are utterance-level, other features are all frame-level. So, the i-vectors are concatenated with 5-frame window of features. Thirdly, DNNs are used as the discriminative learning machine. The output of the network is composed of the corresponding 5-frame window of IRM, the training target proved best in [9]. After training, a small amount of adaptive data is fed into DNNs and tune the parameters. The enhanced signals are resinized by the IRM prediction.

NaT can help bring noise and channel information into account, and Euclidean distance regularization can use data in adaptive set to adapt the DNN. The combinations of these two methods can effectively utilize the advantages of each other, leading to better speech enhancement results.

NOISE-AWARE TRAINING (NAT) WITH I-VECTORS

To reduce the influence of noise mismatch conditions, i-vector is extracted as a long-time feature to represent noise characteristics. The typical process of extracting i-vector feature and adding it to DNN input is showed in FIGURE.2.

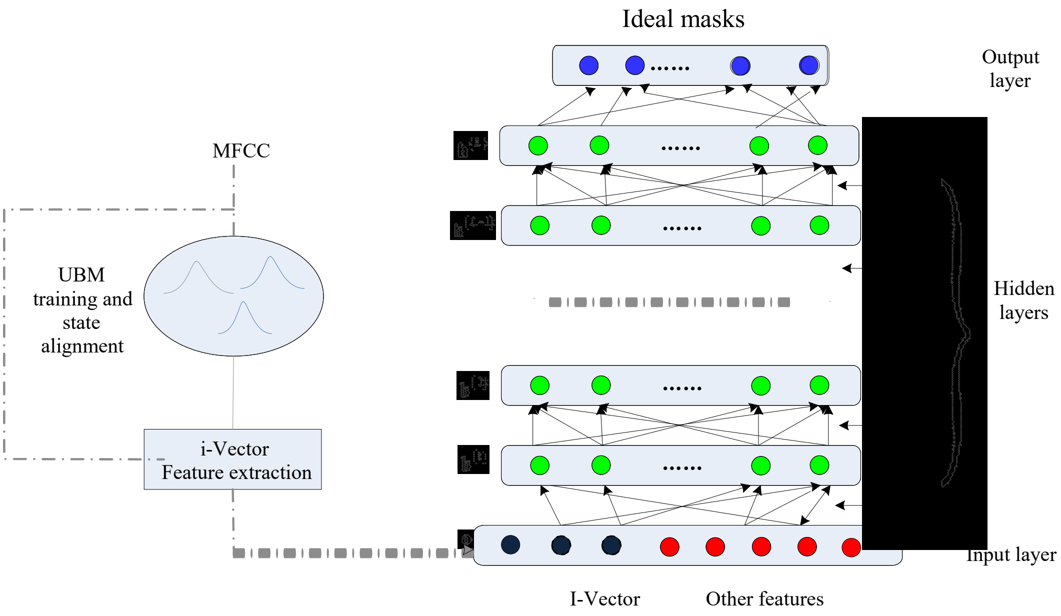

FIGURE 2. Extracting i-vector based on DNN

The right network in Fig.2 shows the DNN estimating ideal masks for speech enhancement. The universal background model (UBM) are constructed using all the training data with feature of MFCCs. Utterance-level C-dimensional i-vectors are extracted. And for both training and test, the i-vector for a given noise condition is concatenated to every frame. I-vector, as an additional utterance-level feature, is conducive to speaker, channel and background normalization. Recently, i-vector has achieved great success in speech recognition domain. In this paper, we apply i-vector to speech enhancement.

EUCLIDEAN DISTANCE REGULARIZATION

To do adaptation conservatively, the ED regularization is done with a small amount of adaptation data. Comparing to KLD normalization [10], we select a more suitable normalization function for mask estimating and apply it to speech enhancement domain firstly.

The intuitive explanation of Euclidean distance regularization is: the distance of the output vectors estimating from adaptive model should not differ too greatly from that estimating from unadopted model. The output of DNN is a vector, a measurement of whose distance is Euclidean distance. We add Euclidean distance to the adaptive criteria as a regular term to obtain the regular term as Equation (1):

$$J_{ED}(W,b;N)=(1-\lambda)J(W,b;N)+\lambda R_{ED}(W_{NI},b_{NI};W,b;N) \quad (1)$$

where λ is the regularization weight:

$$R_{ED}(W_{NI}, b_{NI}; W, b; N) = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{2}\| v_{NI}^{L} - v_{m}^{L} \|^2 \tag{2}$$

where v_{NI}^{L} and v_{m}^{L} are probability that the m^{th} output vector estimating from noise independent DNN and adaptation DNN respectively. The mean square error (MSE) criterion for regression tasks is usually used.

$$J_{MSE}(W, b; N) = \frac{1}{M}\sum_{m=1}^{M} J_{MSE}(W, b; v^m, y^m) \tag{3}$$

Where

$$J_{MSE}(W, b; v^L, y) = \frac{1}{2}\| v^L - y \|^2 = \frac{1}{2}(v^L - y)^T (v^L - y) \tag{4}$$

After adding Euclidean distance regularization, the regularized adaptation criterion can be converted to

$$\begin{aligned} J_{ED-MSE}(W, b; N) &= (1-\lambda) J_{MSE}(W, b; N) + \lambda R_{ED}(W_{NI}, b_{NI}; W, b; N) \\ &= \frac{1}{2M}\sum_{m=1}^{M} (1\text{-}\lambda) \| v_m^L - y \|^2 + \lambda \| v_{NI}^L - v_m^L \|^2 \end{aligned} \tag{5}$$

EXPERIMENTS AND RESULTS

Experimental Settings

Data

All the experiments are based on TIMIT corpus speech data set. We choose seven types of noises from NOISEX-92[10]: factory1, destroyer engine, Volvo, m109, and white for baseline system. And we choose 2520 clean speech utterances of 6 speakers from TIMIT corpus to mix with the noises aforementioned at different SNRs to obtain a parallel training dataset and choose 100 utterances to obtain adaptive dataset. Other 200 clean utterances are chosen to mix with these noises at different SNRs to constitute the core testing dataset.

Parameters Settings

The DNN includes 3 hidden layers, each of which has 1024 units, and an input layer and an output layer. It is trained to predict the desired outputs across all frequency bands, and the mean squared error (MSE) is used as the cost (loss) function for this regression task. The output of the DNN is 64-dimensional Gamma-tone filter bank, and sigmoid function is chosen to be the output function. Mean squared error (MSE) is used as the cost (loss) function for this regression task.

Evaluation Criteria

Three evaluation metrics are used in this paper, including the Short-Time Objective Intelligibility score (STOI) [13], Perceptual Evaluation of Speech Quality (PESQ) [14] and Segmental SNR (segSNR)[15].

STOI was proposed recently to evaluate the intelligibility of speech. The two aspects of speech are the quality and intelligibility. It scores from 0 to 1.

PESQ, an application guide for objective quality measurement, is calculated from the separated speech and the corresponding clean speech, scoring from -0.5 to 4.5.

Segmental SNR (segSNR) is used to evaluate SNRs of every segment.

$$SegSNR = \frac{1}{T}\sum_{t=0}^{T-1} \varsigma_1 \left\{ 10\log \frac{\sum_{l=0}^{L-1}\left[x^t(l+tR)\right]^2}{\sum_{l=0}^{L-1}\left[x^t(l+tR)-\hat{x}^t(l+tR)\right]^2} \right\} \tag{6}$$

where T represents for the number total frames, and ς is a meaningful range of SNRs for human auditory. Segmental SNR, the ratio of the signal to its delta, is a time-domain metric representing for the extent of denoising.

Results and Analysis

To compare the performance in different situations, we train two sets of DNNs. The first set uses mixtures at variable SNRs and test at different SNRs. The second set uses mixtures of variable types of noises and test with different types of noise. For every experiment, we test the performance after adjusting the weight of normalization λ and choose the best evaluation results to fill up the tables. Mixture means the scores before enhancement.

Comparison Between Various Systems in Mismatching SNRs

For the first set, aiming at the SNR mismatching problem, we make 2 experiments. In the 1st experiment, the DNN is trained with 0, 5 and 10 dB mixtures of the same type of noise and tested by -5 dB mixtures. 6 types of noises are used to separately do the experiment and the results are showed in Table 5.1. In the 2nd experiment, the training mixtures are 5, 10 and 15 dB mixtures, and the testing mixtures are 5 dB mixtures of the same type of noise. Other experimental setups are the same as the 1st experiment.

TABLE 1. Performance Comparisons between Various Systems in Mismatching SNRs
(0, 5, 10 dB training and -5 dB testing)

System	M109			White			Volvo			Destroyer engine		
	STOI	PESQ	segSNR	STOI	PESQ	segSNR	STOI	PESQ	segSNR	STOI	PESQ	segSNR
mixture	0.6790	1.0516	-6.0876	0.6048	1.0337	-6.3173	0.8881	1.2875	-5.3890	0.5711	1.0818	-6.3445
baseline	0.6795	1.1237	-3.2301	0.6403	1.0700	-4.2564	0.8964	1.9586	5.5892	0.5765	1.0876	-5.4134
baseline+iVector	0.6786	1.1226	-3.3632	0.6402	1.0682	-4.3443	0.8938	1.9468	5.3933	0.5779	1.0858	-5.4668
Baseline+ED	0.7449	1.3504	-0.4237	0.6838	1.1570	-1.3488	0.9135	2.3869	7.4149	0.6724	1.2666	-1.1864
Baseline+iVector+ED	0.7434	1.3160	-0.7653	0.6765	1.1223	-1.4014	0.9125	2.3729	7.3493	0.6665	1.2217	-2.0689

TABLE 2. Performance Comparisons between Various Systems in Mismatching SNRs
(10, 15, 20 dB training and 5 dB testing)

System	M109			White			Volvo			Destroyer engine		
	STOI	PESQ	segSNR	STOI	PESQ	segSNR	STOI	PESQ	segSNR	STOI	PESQ	segSNR
mixture	0.8532	1.2944	-1.0543	0.8324	1.0853	-1.4292	0.9522	2.6021	8.9953	0.8032	1.2496	-1.3883
baseline	0.8492	1.4834	2.1093	0.8847	1.7131	4.1640	0.9527	2.6823	10.0384	0.8110	1.2855	-0.5887
baseline+iVector	0.8508	1.4836	2.1526	0.8844	1.7129	4.1624	0.9508	2.6541	9.8302	0.8050	1.2755	-0.7296
Baseline+ED	0.8824	1.9730	4.9517	0.8867	1.8480	4.6512	0.9589	3.0688	12.2628	0.8624	1.6364	2.0512
Baseline+iVector+ED	0.8796	1.8994	4.7737	0.8866	1.8469	4.6234	0.9579	3.0263	12.1459	0.8543	1.6316	1.9150

Comparing the results in Section 5, we can reach the following conclusions.

For the noise mismatch condition of SNRs in the 1st set, on all of these noises, ED regularization improves the performance consistently while the i-vector method does not. The improvements show that i-vectors cannot reflect the variation of the SNRs, while ED can take adaptation information into account to make the original model adapt the test set well. And the decrease of i-vector system also shows that as the dimensional of features increases, it becomes harder to tune the network for adaptation.

Comparing Table 5.1~5.2, we can see the promotion of the percentage of STOI is getting smaller as the SNR increases. The promotion in the 2nd experiment is smaller than that in the 1st experiment. The promotion of the percentage of STOI in the 2nd experiment can reach 16.6% (Table 5.1, destroyer engine) at most, while that in the 3rd experiment can reach 6.5% (Table 5.2, Volvo) at most. The trend of the other two criteria, PESQ and segSNR, is consistent with the STOI.

In Table 5.1~5.2, the performance of Volvo noise is better than other noises in all the evaluation criteria. For example, in Table 5.1, the STOI of Volvo noise can reach more than 0.9 while others can reach only 0.7 at most. It is probably because the Volvo noise is a rather stationary noise which can be seen from its spectrum.

Comparison Between Various Systems on Mismatching Noise Types

For the second set, aiming at the multi-type noises problem, a DNN is trained with three types of noises, including Volvo, factory1 and f16 noises, and tested with m109 noise, destroyer engine noise and white noise, the results of which results are showed in Table 5.3~5.5 respectively.

In the second set, comparing Table 5.3~5.5, we use the same model trained by three types of noises to test other 3 types of noises separately. Among the test noises, factory1 noise, destroyer engine noise, Volvo noise, m109 noise and f16 noise have similarities in spectrum. White noise is a stationary noise acquired by sampling high-quality analog noise generator. So, in the second set, three of them are chosen to train the DNN, while the other two, along with white noise are chosen to test.

TABLE 3. Performance Comparisons between Various Systems in Mismatching SNRs (Volvo, factory1, f16 training and m109 testing)

System	-5dB			0dB			5dB			10dB		
	STOI	PESQ	segSNR	STOI	PESQ	segSNR	STOI	PESQ	segSNR	STOI	PESQ	segSNR
mixture	0.6790	1.0516	-6.0876	0.7694	1.1156	-3.9650	0.8532	1.2944	-1.0543	0.9153	1.6010	2.3919
baseline	0.6812	1.1200	-3.2957	0.7714	1.2492	-0.7065	0.8882	1.8982	4.6374	0.9345	2.3742	7.6857
baseline+iVector	0.6821	1.1219	-3.2725	0.7733	1.2456	-0.7079	0.8901	1.9740	5.1397	0.9349	2.4620	8.2688
Baseline+ED	0.7316	1.3515	-0.4835	0.8146	1.5958	1.8081	0.8910	2.0187	5.0738	0.9350	2.4681	8.0933
Baseline+iVector+ED	0.7319	1.3655	-0.2407	0.8150	1.6089	2.0690	0.8913	2.0467	5.5093	0.9354	2.5365	8.5326

TABLE 4. Performance Comparisons between Various Systems in Mismatching SNRs (Volvo, factory1, f16 training and destroyer engine testing)

System	-5dB			0dB			5dB			10dB		
	STOI	PESQ	segSNR	STOI	PESQ	segSNR	STOI	PESQ	segSNR	STOI	PESQ	segSNR
mixture	0.5711	1.0818	-6.3445	0.6906	1.1321	-4.2368	0.8532	1.2944	-1.0543	0.8897	1.4805	2.0355
baseline	0.5665	1.1265	-6.1808	0.6865	1.4425	-4.0871	0.7995	1.3332	-1.3552	0.8861	1.5686	1.8701
baseline+iVector	0.5681	1.1299	-6.1139	0.6908	1.1711	-4.1593	0.8001	1.3358	-1.2231	0.8863	1.5718	1.9493
Baseline+ED	0.6643	1.1760	-2.5357	0.8002	1.5327	0.4656	0.8082	1.3157	-0.9444	0.9173	2.1726	6.1442
Baseline+iVector+ED	0.6724	1.2799	-1.0798	0.8054	1.1731	0.7320	0.8485	1.5246	0.9305	0.9233	2.2403	6.4789

TABLE 5. Performance Comparisons between Various Systems in Mismatching SNRs (Volvo, factory1, f16 training and white testing)

System	-5dB			0dB			5dB			10dB		
	STOI	PESQ	segSNR	STOI	PESQ	segSNR	STOI	PESQ	segSNR	STOI	PESQ	segSNR
mixture	0.6048	1.0339	-6.3173	0.7222	1.0467	-4.2430	0.8324	1.0853	-1.4292	0.9123	1.1944	1.9483
baseline	0.5976	1.0473	-5.5414	0.7173	1.0577	-4.0562	0.8269	1.1181	-0.7996	0.8863	1.5686	1.8701
baseline+iVector	0.6008	1.0485	-5.3931	0.7217	1.0569	-4.2098	0.8252	1.1191	-0.6143	0.8861	1.5718	1.9493
Baseline+ED	0.6581	1.0576	-4.9445	0.7760	1.0860	-2.3037	0.8506	1.1548	0.0094	0.9206	1.3021	3.2077
Baseline+iVector+ED	0.6680	1.0642	-4.7134	0.7770	1.0969	-1.7046	0.8528	1.1636	0.1558	0.9211	1.3091	3.3096

For the noise mismatch condition of noise types in the 2nd set, on all these types of noises, i-vectors and ED separately improve all the three evaluation criteria. And it can get better performance when combing these two methods. The factor contributed to this phenomenon is that i-vectors can reflect noise type information, but not so sensitive to SNRs. And ED, which brings about testing information, is still effective when the dimensional of features increases.

The results show test sets of similar noises perform better than white noise. In the second set, the promotion of SOTI of destroyer engine noise (-5dB) can reach 18.7% at most, while that of white noise (-5dB) and destroyer ops noise (-5dB) is 11.8% and 12.2% respectively. Probably the similarities contribute to congenial DNN parameters, which is helpful for enhancement.

In the second set, the promotion of the percentage of STOI is getting smaller as the SNR increases on all the testing noise types, which is similar to that in the first set. For example, in Table 5.3, the largest promotion of the percentage of STOI is 7.4% at -5dB, while it is less than 0.1% when the SNR is higher than 5dB. Note that in low SNR conditions, STOI improvement is more meaningful. The trend of PESQ and segSNR are consistent with STOI.

From the all the results above, some conclusions can be summarized for i-vector and ED systems. On all the situations above, the improvements decrease as the SNRs increase. Among all experiments, we find that more robust performance and better results appear when λ ranges from 0.125 to 0.5.

CONCLUSION

To overcome the mismatching problem between training and testing sets in speech enhancement, we have presented an effective way to perform noise adaptation for neural network acoustic models. We proposed to use NaT in ideal mask estimation system based on DNN to bring environment information into account. We also test the performance of ED regularization, which using a small amount of adaptation data to adapt the network. The two methods can be combined to take advantages of both.

In this study, i-vector is extracted by MFCC and considered as an utterance-level feature. Future works we may use dynamic i-vectors to further improve the method. And we can try other CT methods for comparison.

REFERENCES

1. Bingyin Xia and Changchun Bao. Wiener filtering based speech enhancement with weighted denoising auto-encoder and noise classification. Speech Communication, 60:13-29 (2014).
2. Seide, F., Li, G., Chen, X., Yu, D.: Feature engineering in context-dependent deep neural networks for conversational speech transcription. In: Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ARSU), pp. 24-29 (2011).
3. Seltzer, M., Yu, D., Wang, Y.: An investigation of deep neural networks for noise robust speech recognition. In: Proc. International Conference on Acoustics, Speech and Signal Process ing (ICASSP) (2013).
4. George Saon, Hagen Soltau, David Nahamoo, and Michael Picheny. Speaker adaptation of neural network acoustic models using i-vectors. IEEE Workshop on Automatic Speech Recognition and Understanding (ARSU), pp.55-59 (2013).
5. Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., Ouellet, P.: Front-end factor analysis for speaker verification. IEEE Transactions on Audio, Speech and Language Processing 19(4), 788-798 (2011).

6. G. Kim, Y. Lu, Y. Hu, and P. C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," J. Acoust. Soc. Amer., 126: 1486–1494 (2009).
7. H. Hermansky and N. Morgan, "RASTA processing of speech," IEEE Trans. Speech, Audio Process., 2(4): 578–589 (1994).
8. Timo Gerkmann and Richard C Hendriks. Unbiased mmse-based noise power eatimation with low complexity and low tracking delay. IEEE Transactions on Audio, Speech and Language Processing, 20(4): 1383-1393 (2012).
9. Yuxuan Wang, Arun Narayanan, Deliang Wang. On Training Target for Supervised Speech Separation. IEEE/ACM Trans Audio Speech Lang Process. 22(12): 1849-1858 (2014).
10. Yu, D., Yao, K., Su, H., Li, G., Seide, F.: Kl-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition. I n: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7892-7897 (2013).
11. J. Garofolo, DARPA TIMIT acoustic-phonetic continuous speech corpus. Gaithersburg, MD, USA: Nat. Inst. of Standards Technol., 1993.
12. A.Varga and H. Steeneken, "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., 12:247–251 (1993).
13. C. Taal, R. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," IEEE Trans. Audio, Speech, Lang. Process., 19(7): 2125–2136 (2011).
14. 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 Proc. ICASSP, 749–752 (2001).
15. R. Talmon and S. Gannot, "Single-channel transient interference suppression with diffusion maps," IEEE Trans. Audio, Speech, and Lang. Process., 21(1), pp. 132–144 (2013).