Multi-objective Learning and Mask-based Post-processing for Deep Neural Network based Speech Enhancement

Yong Xu[1*], *Jun Du*[1], *Zhen Huang*[2], *Li-Rong Dai*[1], *Chin-Hui Lee*[2]

[1]National Engineering Laboratory for Speech and Language Information Processing,
University of Science and Technology of China, China
[2]School of Electrical and Computer Engineering, Georgia Institute of Technology, USA
xuyong62@mail.ustc.edu.cn, jundu@ustc.edu.cn, chl@ece.gatech.edu

Abstract

We propose a multi-objective framework to learn both secondary targets not directly related to the intended task of speech enhancement (SE) and the primary target of the clean log-power spectra (LPS) features to be used directly for constructing the enhanced speech signals. In deep neural network (DNN) based SE we introduce an auxiliary structure to learn secondary continuous features, such as mel-frequency cepstral coefficients (MFCCs), and categorical information, such as the ideal binary mask (IBM), and integrate it into the original DNN architecture for joint optimization of all the parameters. This joint estimation scheme imposes additional constraints not available in the direct prediction of LPS, and potentially improves the learning of the primary target. Furthermore, the learned secondary information as a byproduct can be used for other purposes, e.g., the IBM-based post-processing in this work. A series of experiments show that joint LPS and MFCC learning improves the SE performance, and IBM-based post-processing further enhances listening quality of the reconstructed speech.

Index Terms: speech enhancement, deep neural network, minimum mean square error, multi-objective learning, binary mask

1. Introduction

Classical speech enhancement (SE) approaches, such as spectral subtraction [1], MMSE-based spectral amplitude estimator [2, 3] and optimally modified log-MMSE estimator [4, 5], are considered as unsupervised techniques having been studied extensively for several decades. Based on key assumptions for the interactions between speech and noise, the tremendous progress has been made for those techniques in the past. However some issues, such as fast changing noise (e.g., *machine gun* [6]) and negative spectrum estimation, still need to be addressed.

On the other hand, supervised machine learning approaches have also been developed in recent years. They were shown to generate enhanced speech with good qualities [7]. Nonnegative matrix factorization (NMF) based speech enhancement [7, 8] was one notable example in which speech and noise basis models were learned separately from training speech and noise databases. Then the clean speech could be decomposed given the noisy speech. However, speech and noise are assumed uncorrelated and it limited the quality of the enhanced speech signals. Following recent successes in deep learning based speech processing [9, 10, 11] we have recently proposed a deep neural network (DNN) based speech enhancement framework [12, 13, 14] in which DNN was regarded as a regression model to predict the clean log-power spectra (LPS) features [15] from noisy LPS features. DNN also acts as a mapping function to learn the relationship between clean and noisy speech features without imposing any assumption. Similar DNN-based speech denoising methods were also proposed in [16, 17]. In [18, 19], DNN-based method was demonstrated to be better than the NMF-based methods in speech separation. In DNN-based speech enhancement, the minimum mean square error (MMSE) between the target features and the predicted features was always used as the objective function. It is difficult to design a better cost function to directly optimize the DNN model, especially for features that are correlated. In [19] it was shown that other cost functions, such as the Kullback Leibler divergence [20] or the Itakura-Saito divergence [21], all performed worse than the MMSE.

In this paper, a multi-objective learning framework is proposed to optimize a joint objective function, encompassing errors not only for the primary clean LPS features but also errors in secondary targets for continuous features, such as MFCC, and for categorical information, such as ideal binary mask (IBM) [22]. This joint optimization of different but related targets can potentially improve the DNN prediction performance of the primary target LPS which is then used to reconstruct the enhanced waveform. In the LPS domain, the target values of different frequency bins were predicted independently without any correlation constraint, and some knowledge in auditory perception [23] is not easily utilized. Nonetheless in the MFCC domain, mel-filtering is first applied and the correlation of each channel is represented in the MFCC coefficients. Furthermore, IBM is the most important concept in the computational auditory scene analysis (CASA) [23]. IBM which represents the noise-dominant or speech-dominant meta information can also improve DNN training and the estimated IBM could further be used for post-processing. Finally, MFCC and IBM can be combined together to help predict the target clean LPS features.

In our SE experiments, we find that learning MFCC and/or IBM as secondary tasks provides improvements to DNN-based speech enhancement. Furthermore, IBM-based post-processing also gives an additional 1.5 dB improvement of segmental signal-to-noise ratio (SSNR) [15].

2. Multi-objective Learning for DNN-based Speech Enhancement

In [12, 13], DNN is adopted as a mapping function to predict the clean LPS features from the noisy LPS features. The relationship between the clean and noisy speech features can be

[*] This work is done while Yong Xu was visiting Georgia Tech in 2014-2015.

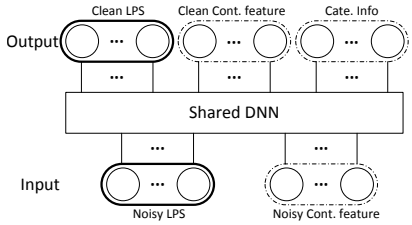

Figure 1: *The structure of the multi-objective learning.*

well learned because nearly no assumptions were imposed during the training process. However, other DNN-based methods, such as binary or soft mask [24, 25] based speech enhancement, assume that speech and noise are independent [12] at each time-frequency (T-F) unit.

Normalized MMSE is used to update the DNN weights,

$$Er = \frac{1}{N}\sum_{n=1}^{N}\frac{\|\hat{\mathbf{X}}_n(\mathbf{Y}_{n\pm\tau},\mathbf{W},\mathbf{b})-\mathbf{X}_n\|_2^2}{\|\mathbf{X}_n\|_2^2}. \quad (1)$$

where Er is the normalized mean squared error and it can also be treated as the reciprocal of signal-to-noise ratio (SNR). This normalized squared error always reduces the distribution diversity of the clean training data and makes DNN training more stable. It should be noted that all the input and output features are normalized with a global mean and variance of the noisy training data. Hence, $\hat{\mathbf{X}}_n$ and $\mathbf{X}_n$ denote the estimated and clean normalized LPS at sample index n, respectively, with N representing the mini-batch size, $\mathbf{Y}_{n\pm\tau}$ being the noisy LPS feature vector where the window size of the context is $2*\tau+1$, with $(\mathbf{W},\mathbf{b})$ denoting the weight and bias parameters to be learned.

In this study, multi-objective learning is proposed to jointly predict the primary LPS features together with other secondary continuous features, such as MFCC, or/and some discrete category information, such as IBM, to enhance DNN learning as follows,

$$\begin{aligned} Er = &\frac{1}{N}\sum_{n=1}^{N}\frac{\|\hat{\mathbf{X}}_n(\mathbf{Y}_{n\pm\tau},\mathbf{Y}^{\text{cont}}_{n\pm\tau},\mathbf{W},\mathbf{b})-\mathbf{X}_n\|_2^2}{\|\mathbf{X}_n\|_2^2}+\\ &\alpha*\frac{1}{N}\sum_{n=1}^{N}\frac{\|\hat{\mathbf{X}}^{\text{cont}}_n(\mathbf{Y}_{n\pm\tau},\mathbf{Y}^{\text{cont}}_{n\pm\tau},\mathbf{W},\mathbf{b})-\mathbf{X}^{\text{cont}}_n\|_2^2}{\|\mathbf{X}^{\text{cont}}_n\|_2^2}+\\ &\beta*\frac{1}{N}\sum_{n=1}^{N}\|\hat{\mathbf{X}}^{\text{cate}}_n(\mathbf{Y}_{n\pm\tau},\mathbf{Y}^{\text{cont}}_{n\pm\tau},\mathbf{W},\mathbf{b})-\mathbf{X}^{\text{cate}}_n\|_2^2. \quad (2)\end{aligned}$$

where $\hat{\mathbf{X}}^{\text{cont}}$ and $\mathbf{X}^{\text{cont}}$ denote the estimated and clean continuous features (also normalized), respectively. $\mathbf{Y}^{\text{cont}}$ represents the second noisy continuous feature. $\hat{\mathbf{X}}^{\text{cate}}$ and $\mathbf{X}^{\text{cate}}$ denote the estimated and target meta category information, respectively. α and β are the weighting coefficients of this two other error parts, respectively. Unlike linear continuous features, meta information just has binary values, which makes the normalization not necessary for squared error related with the category part. Fig. 1 presented the structure of the proposed multi-objective learning. In fact, it was similar to the multi-task learning [26], but different from the multi-task learning in recent DNN-based speech recognition [27, 28] with only one input feature type. The prediction for the secondary continuous feature should be complementary with the prediction for the primary LPS using the shared DNN. The learning for the category information with linear activation function should also promote the prediction of clean LPS. Overall, multi-objective learning can improve the generalization capability of DNN for the clean LPS estimation.

2.1. Joint Prediction of LPS with MFCC

MFCC is one of the most popular speech features used in speech recognition [29], speaker recognition [30] and music modeling [31]. Mel-filtering is applied to make it consistent with human auditory perception. However there is so far no prior auditory knowledge adopted in the LPS domain except for the log-compression. We believe the clean LPS features would be better predicted with a MFCC constraint imposed at the output layer. Furthermore, the discrete cosine transformation (DCT) [32] operation in MFCC can incorporate the correlation information of different channels into each MFCC coefficient. We therefore expect correlated and consistent distortion across different frequency bins can be learned when predicting the clean LPS. Noted that DCT here is not performing dimension reduction which means the same dimensional MFCC features as the Mel-filter bank features are extracted.

One similar work in [33] showed that the concatenation of different input features could improve the performance of DNN-based speech separation. However the motivation of our work is multi-objective learning with a novel architecture in both input and output layers, which is totally different from the motivation of feature fusion in [33]. It is expected that the enhancement of MFCC would be complimentary to the enhancement of LPS.

2.2. Joint Prediction of LPS with IBM

IBM [22] is one type of category information often used to represent the noise-dominant or speech-dominant nature at a certain T-F bin [23]. If the local SNR of a T-F bin is greater than a threshold, the IBM is set to one otherwise it is set to zero. Just like MFCC, IBM is also used as a constraint term in the joint objective function. IBM explicitly offers the additional speech presence information at T-F units. With this discriminative information, the speech components would be emphasized while reducing more noise components.

In addition, the joint prediction of clean LPS with clean MFCC and IBM can be combined together. The noisy MFCC augmented in the input with the noisy LPS can also improve the IBM-based post-processing performance with an accurate IBM estimation to be discussed in the next section.

2.3. IBM-based Post-processing

The direct prediction of the clean LPS using DNN may lead to an overestimate or underestimate problem at some T-F units. The estimated IBM can be used for post-processing to simultaneously control the noise reduction level and speech distortion as follows,

$$\hat{X}'_n(d)=\begin{cases} Y_n(d) & \widehat{\text{IBM}}_n(d)\geq\gamma \\ \frac{(Y_n(d)+\hat{X}_n(d))}{2} & \varepsilon<\widehat{\text{IBM}}_n(d)<\gamma \\ \hat{X}_n(d) & \text{otherwise}\end{cases} \quad (3)$$

where $\widehat{\text{IBM}}_n(d)$ denotes the estimated IBM at time frame n and frequency bin d. Noted that the estimated IBM is close to the range $[0,1]$. If the estimated IBM value is very large indicating that it has very high SNR at certain T-F unit, it is not necessary to perform noise reduction which can potentially result in the speech distortion. This is also the mask concept in [23]. If the estimated IBM has a medium value, the average value

between the noisy LPS and the estimated LPS was used. Otherwise, the DNN predicted LPS was adopted. The proposed IBM post-processing scheme in Eq. (3) is therefore different from [22] where the estimated soft mask was used as a Wiener gain to perform speech enhancement. In contrast to adopting DNN to learn the mask [22, 24] there is no independence assumption between speech and noise in our DNN based mapping strategy.

3. Experimental Results and Analysis

In [12, 13], all experiments were conducted on waveforms with 8kHz sample rate, in this work we extended it to 16kHz sample rate. 104 noise types were used in [12], however, in this study 115 noise types including some musical noises were adopted to further improve the generalization capacity of DNN. These 115 noise types include 100 noise types recorded by G. Hu [34] and 15 home-made noise types[1]. And the clean speech data is derived from the TIMIT corpus [35]. All 4620 utterances from the training set of the TIMIT database were corrupted with the abovementioned 115 noise types at six levels of SNR, i.e., 20dB, 15dB, 10dB, 5dB, 0dB, and -5dB, to build 80 hours multi-condition training set, consisting of pairs of clean and noisy speech utterances. The 192 utterances from the core test set of TIMIT database were used to construct the test set for each combination of noise types and SNR levels. As we only conduct the evaluation of unseen noise types in this paper, three other noise types, namely Buccaneer1, Destroyer engine and HF channel were adopted for testing. All of them are collected from the NOISEX-92 corpus [6]. An improved version of OM-LSA [5], denoted as **LogMMSE**, was used for performance comparison with our DNN approach.

A short-time Fourier analysis was used to compute the DFT of each overlapping windowed frame. Then 257 dimensions LPS features [15] were used to train DNNs. Segmental SNR (SSNR in dB) [15], perceptual evaluation of speech quality (PESQ) [36], and short-time objective intelligibility (STOI) [37] were used to assess the quality and intelligibility of the enhanced speech. Frequency-dependent log-spectral distortion, defined as subtracting estimated LPS from clean LPS at each frequency bin, was also proposed to analyze the consistency of distortion across frequencies. Rectified linear units (ReLU) [38] was used as the activation function of DNN, and the DNN was initialized with random weights. Dropout [39] and static noise aware training as in [12, 40] were used to improve its generalization capacity for unseen noise environments. Mean and variance normalization was applied to the input and target feature vectors of the DNN. All DNN configurations were fixed at $L = 3$ hidden layers, 2500 units at each hidden layer, and 7-frame input. The MFCC used in Section 2.1 had 40 dimensions of static feature and one energy dimension using 40 Mel-filters. The empirical value of α and β in Eq. (2) are set to 0.1 and 0.002, respectively. The empirical value of γ and ε in Eq. (3) are set to 0.9 and 0.6, respectively.

[1]The 115 noise types for training are N1-N17: Crowd noise; N18-N29: Machine noise; N30-N43: Alarm and siren; N44-N46: Traffic and car noise; N47-N55: Animal sound; N56-N69: Water sound; N70-N78: Wind; N79-N82: Bell; N83-N85: Cough; N86: Clap; N87: Snore; N88: Click; N88-N90: Laugh; N91-N92: Yawn; N93: Cry; N94: Shower; N95: Tooth brushing; N96-N97: Footsteps; N98: Door moving; N99-N100: Phone dialing; N101: AWGN; N102: Babble; N103-N105: Car; N106-N115: musical instruments. And all of them can be downloaded at http://home.ustc.edu.cn/˜xuyong62/demo/115noises.html

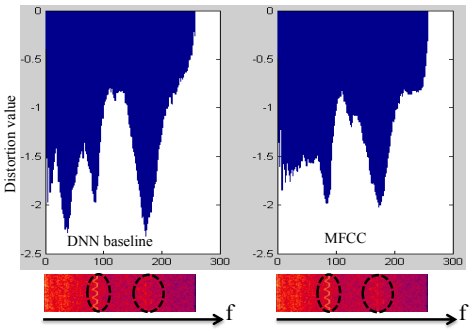

Figure 2: *Frequency-dependent log-spectral distortion between the DNN baseline and MFCC systems calculated from 192 testing utterances at SNR=0dB corrupted by the Buccaneer1 noise (shown in the spectrogram above). And the x-axis is frquency.*

3.1. Joint Prediction of LPS and MFCC

In Table 1, average PESQ and SSNR comparison on the test set at different SNRs of the three unseen noise environments among: DNN baseline, MFCC augmented in the output (denoted as MFCC-o) and MFCC augmented in both the input and output (denoted as MFCC), were given. MFCC-o system consistently outperformed the DNN baseline in PESQ and SSNR which indicated that the simultaneous prediction of MFCC was beneficial for the estimation of clean LPS. Furthermore, the noisy MFCC was complementary with the noisy LPS in the input to improve the prediction of clean LPS. And the MFCC system got the best performance, such as the average PESQ improved from 2.508 to 2.664. The multi-task of MFCC enhancement and LPS enhancement shared the DNN weights and promoted each other. The frequency-dependent log-spectral distortion between the DNN baseline and MFCC systems calculated from 192 testing utterances at SNR=0dB corrupted by the Buccaneer1 noise was also given in Fig. 2. The overall shape of this log-spectral distortion is determined by the noise type, such as here the Buccaneer1 noise has two continual and high energy parts at frequencies shown in the ellipses. But with the constraint of MFCC, the speech distortion at low frequencies where the most of speech info located was largely reduced and more consistent. This was because MFCC emphasized the info at low frequencies with the Mel-filtering.

3.2. Joint Prediction of LPS and IBM with Post-processing

Table 1 also presented the average PESQ and SSNR comparison for joint prediction of LPS and IBM on the test set at different SNRs of the three unseen noise environments. With the IBM constraint in the output, better average PESQ and SSNR performance could be obtained compared with the DNN baseline, especially in SSNR which improved from -0.084 dB to 0.251 dB at SNR=0dB. Moreover, the IBM-based post-processing can obtain large PESQ and SSNR improvements, especially at high SNRs, e.g., SSNR improved from 7.641 dB to 11.455 dB at SNR=20dB which implies that the baseline DNN might hurt the speech components due to under-estimation, especially at the T-F units with high SNRs. Hence, IBM-based post-processing is crucial in achieving less speech distortion. This also conformed the mask concept in [23] that it was not necessary to reduce noise when the speech energy is much larger than the noise energy at the certain T-F unit. In addition, IBM could be combined with MFCC. Compared with the performance of

Table 1: *Average PESQ and SSNR comparison on the test set at different SNRs of the three unseen noise environments, among: DNN baseline, MFCC-augmented output (denoted as MFCC-o), MFCC augmented in the input and output (denoted as MFCC), IBM augmented in the output of the DNN baseline without post-processing (denoted as IBM), IBM with post-processing (denoted as IBM+PP), MFCC and IBM without post-processing (denoted as MFCC+IBM) and MFCC and IBM with post-processing (denoted as MFCC+IBM+PP).*

	Baseline		MFCC-o		MFCC		IBM		IBM+PP		MFCC+IBM		MFCC+IBM+PP	
SNR	PESQ	SSNR	PESQ	SSNR	PESQ	SSNR	PESQ	SSNR	PESQ	SSNR	PESQ	SSNR	PESQ	SSNR
20	3.287	7.403	3.324	7.592	3.387	8.199	3.309	7.641	3.358	11.455	3.391	8.147	**3.424**	**11.862**
15	3.014	5.721	3.051	5.936	3.128	6.637	3.029	5.987	3.083	8.616	3.135	6.610	**3.167**	**9.164**
10	2.713	3.812	2.748	4.087	2.852	4.762	2.722	4.097	2.791	5.808	2.861	4.782	**2.895**	**6.418**
5	2.387	1.825	2.414	2.204	2.551	2.662	2.384	2.131	2.463	3.145	2.567	2.770	**2.597**	**3.673**
0	2.030	-0.084	2.045	0.413	2.217	0.534	2.013	0.251	2.084	0.811	2.238	0.759	**2.261**	**1.102**
-5	1.617	-1.693	1.624	-1.171	1.847	-1.433	1.603	-1.314	1.679	**-1.036**	1.868	-1.086	**1.887**	-1.054
Ave	2.508	2.831	2.534	3.177	2.664	3.560	2.510	3.132	2.576	4.800	2.677	3.664	**2.705**	**5.194**

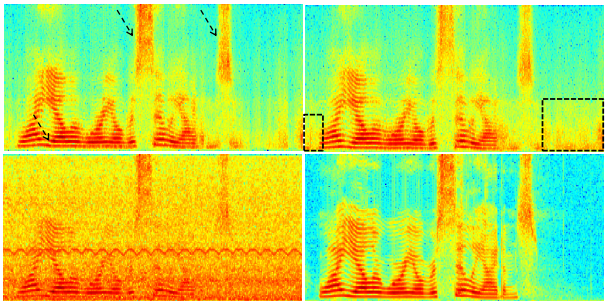

Figure 3: *Comparison of four spectrograms of a 16kHz TIMIT utterance corrupted by Buccaneer1 noise at SNR=5dB: proposed DNN (upper left, PESQ=2.815), DNN baseline (upper right, PESQ=2.585), Noisy (bottom left, PESQ=1.591) and clean speech (bottom right, PESQ=4.5).*

MFCC system, the combined system (MFCC+IBM in Table 1) gave slightly better results at all SNR levels. For example SSNR was improved from -1.433 dB to -1.086 dB at SNR=-5dB. Finally, the average SSNR of the best MFCC+IBM+PP system was improved from 3.664 dB to 5.194 dB.

3.3. Overall Performance Comparison

PESQ and STOI are often adopted to represent the objective quality and intelligibility of the enhanced speech, respectively. And STOI is often more meaningful at lower SNRs. An overall PESQ and STOI comparison of different SE techniques discussed in this study on the test set at different SNRs of the three unseen noise environments is displayed in Table 2. Compared with the noisy speech results, LogMMSE could yield 0.418 PESQ improvement while only 0.011 STOI improvement on average. The DNN baseline improved the LogMMSE with an average STOI from 0.801 to 0.845 across six SNRs. Our proposed MFCC+IBM+PP system overwhelms LogMMSE at all SNRs, especially at low SNRs, e.g., 0.163 STOI improvement and 0.626 PESQ improvement at SNR=-5dB. Fig. 3 presented spectrograms of an utterance. The non-stationary noise was successfully reduced in the DNN-enhanced spectrum, while LogMMSE could not well track the non-stationary Buccaneer1 noise (its spectrogram can be seen at the demo website[2]). Compared with the baseline DNN-enhanced spectrogram, the improved DNN can enhance the speech with less speech distortion shown in the three dashed arrow areas, especially at the consonant portions which are similar to noise. Furthermore the improved DNN can also reduce noise shown in the rectangle highlight segments. More enhanced waveforms of real-world noisy speech can also refer to the website.

[2]http://home.ustc.edu.cn/~xuyong62/demo/IS15.html

Table 2: *Average PESQ and STOI comparison on the test set at different SNRs of the three unseen noise environments, among: Noisy, LogMMSE [5], DNN baseline and the proposed MFCC+IBM+PP in Table 1 (denoted as Proposed).*

	Noisy		LogMMSE		DNN Baseline		Proposed DNN	
SNR	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI
20	2.834	0.971	3.349	0.975	3.287	0.963	**3.424**	**0.979**
15	2.481	0.934	3.049	0.945	3.014	0.944	**3.167**	**0.960**
10	2.133	0.868	2.711	0.890	2.713	0.908	**2.895**	**0.928**
5	1.793	0.772	2.299	0.800	2.387	0.849	**2.597**	**0.876**
0	1.482	0.656	1.798	0.669	2.030	0.762	**2.261**	**0.796**
-5	1.235	0.541	1.261	0.525	1.617	0.645	**1.887**	**0.688**
Ave	1.993	0.790	2.411	0.801	2.508	0.845	**2.705**	**0.871**

4. Conclusion

In this paper, multi-objective learning is proposed to improve DNN training for speech enhancement. Adding constraints from features like MFCC or IBM in the objective function is shown to obtain more accurate estimation of clean LPS. MFCC can make the log-spectral distortion more consistent across low frequencies; IBM can explicitly represent the speech presence information at T-F units, so higher SSNR could be obtained. Furthermore, the estimated IBM can be adopted to do post-processing to alleviate the over-estimate or under-estimate problems in regression-based DNN. And IBM-based post-processing was crucial to reduce speech distortion, especially at high SNR T-F units. Compared with DNN baseline, about 0.2 PESQ and 0.03 STOI improvements were obtained on average. In the future, other continuous features and meta information will be further explored.

5. Acknowledgement

This work was partially supported by the National Nature Science Foundation of China (Grant Nos. 61273264 & 61305002).

6. References

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[3] ——, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[4] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal processing*, vol. 81, no. 11, pp. 2403–2418, 2001.

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

[6] A. Varga and H. J. 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 communication*, vol. 12, no. 3, pp. 247–251, 1993.

[7] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 21, no. 10, pp. 2140–2151, 2013.

[8] K. W. Wilson, B. Raj, and P. Smaragdis, "Regularized non-negative matrix factorization with temporal dependencies for speech denoising." in *INTERSPEECH*, 2008, pp. 411–414.

[9] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[10] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.

[11] X.-L. Zhang and J. Wu, "Denoising deep neural networks based voice activity detection," in *ICASSP*, 2013, pp. 853–857.

[12] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 23, no. 1, pp. 7–19, 2015.

[13] ——, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, 2014.

[14] ——, "Dynamic noise aware training for speech enhancement based on deep neural networks." in *INTERSPEECH*, 2014, pp. 2670–2674.

[15] J. Du and Q. Huo, "A speech enhancement approach using piecewise linear approximation of an explicit model of environmental distortions." in *INTERSPEECH*, 2008, pp. 569–572.

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

[17] 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.

[18] P. S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *ICASSP*, 2014, pp. 1562–1566.

[19] D. Liu, P. Smaragdis, and M. Kim, "Experiments on deep learning for speech denoising," in *INTERSPEECH*, 2014, pp. 2685–2689.

[20] S. Kullback, *Information theory and statistics*. Courier Corporation, 1997.

[21] F. Itakura and S. Saito, "Analysis synthesis telephony based on the maximum likelihood method," in *Proceedings of the 6th International Congress on Acoustics*, 1968, pp. 17–20.

[22] Y. X. Wang, A. Narayanan, and D. L. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Acoustics, Speech and Signal Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[23] D. L. Wang and G. J. Brown, *Computational auditory scene analysis: Principles, algorithms, and applications*. Wiley-IEEE Press, 2006.

[24] Y. X. Wang and D. L. Wang, "Towards scaling up classification-based speech separation," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 21, no. 7, pp. 1381–1390, 2013.

[25] A. Narayanan and D. L. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *ICASSP*, 2013, pp. 7092–7096.

[26] R. Caruna, "Multitask learning: A knowledge-based source of inductive bias," in *ICML*, 1993, pp. 41–48.

[27] M. L. Seltzer and J. Droppo, "Multi-task learning in deep neural networks for improved phoneme recognition," in *ICASSP*, 2013, pp. 6965–6969.

[28] Z. Huang, J. Li, S. M. Siniscalchi, I.-F. Chen, J. Wu, and C.-H. Lee, "Rapid adaptation for deep neural networks through multi-task learning," 2015, submitted to INTERSPEECH.

[29] R. Vergin, D. O'shaughnessy, and A. Farhat, "Generalized mel frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 5, pp. 525–532, 1999.

[30] K. S. R. Murty and B. Yegnanarayana, "Combining evidence from residual phase and mfcc features for speaker recognition," *IEEE Signal Processing Letters*, vol. 13, no. 1, pp. 52–55, 2006.

[31] B. Logan *et al.*, "Mel frequency cepstral coefficients for music modeling." in *ISMIR*, 2000.

[32] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete cosine transform," *IEEE Transactions on Computers*, vol. 100, no. 1, pp. 90–93, 1974.

[33] Y. X. Wang, K. Han, and D. L. Wang, "Exploring monaural features for classification-based speech segregation," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 21, no. 2, pp. 270–279, 2013.

[34] G. Hu, "100 nonspeech environmental sounds, 2004 [online]," http://web.cse.ohio-state.edu/pnl/corpus/HuNonspeech/HuCorpus.html, 2004.

[35] J. S. Garofolo *et al.*, "Getting started with the darpa timit cd-rom: An acoustic phonetic continuous speech database," *National Institute of Standards and Technology (NIST), Gaithersburgh, MD*, vol. 107, 1988.

[36] 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*, 2001, pp. 749–752.

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

[38] G. E. Dahl, T. N. Sainath, and G. E. Hinton, "Improving deep neural networks for lvcsr using rectified linear units and dropout," in *ICASSP*, 2013, pp. 8609–8613.

[39] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

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