INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Frame-level Signal-to-Noise Ratio Estimation using Deep Learning

Hao Li^1*, DeLiang Wang*^{2,3}*, Xueliang Zhang*^{1}*, and Guanglai Gao*^{1}

^{1} Department of Computer Science, Inner Mongolia University, China
^{2} Department of Computer Science and Engineering, The Ohio State University, USA
^{3} Center for Cognitive and Brain Sciences, The Ohio State University, USA

lihao@mail.imu.edu.cn, dwang@cse.ohio-state.edu, cszxl@imu.edu.cn, and csggl@imu.edu.cn

Abstract

This study investigates deep learning based signal-to-noise ratio (SNR) estimation at the frame level. We propose to employ recurrent neural networks (RNNs) with long short-term memory (LSTM) in order to leverage contextual information for this task. As acoustic features are important for deep learning algorithms, we also examine a variety of monaural features and investigate feature combinations using Group Lasso and sequential floating forward selection. By replacing LSTM with bidirectional LSTM, the proposed algorithm naturally leads to a long-term SNR estimator. Systematical evaluations demonstrate that the proposed SNR estimators significantly outperform other frame-level and long-term SNR estimators.

Index Terms: frame-level SNR estimation, long short-term memory, recurrent neural networks, feature combination.

1. Introduction

As speech is almost always interfered by various noises in real life, speech processing is a challenging task. SNR, which is defined as the ratio of signal power to noise power, provides information about the level of noise present in an original noisy signal. SNR knowledge is useful for many speech applications, such as hearing aids [1] and speech enhancement [2].

There are two categories for measuring SNR in a noisy signal. The first category is short-term SNR, where the short-term duration is usually in the range of 20 ms to several seconds. The $a\ priori$ SNR is a widely used short-term subband SNR measure, and can be used for speech enhancement directly. The decision-directed (DD) estimator of Ephraim and Malah [3] is a common method to estimate the $a\ priori$ SNR. It performs a weighted linear combination of two $a\ priori$ SNR-like components. This approach can greatly reduce the variations of $a\ priori$ SNR estimates, which helps to reduce musical tones significantly [4]. In [1], the authors extend three short-term subband SNR estimation algorithms to short-term broadband SNR estimation under different frame lengths.

The second category defines SNR at the utterance level, referred to as global or long-term SNR. Long-term SNR considers the entire signal and provides noise level information for the whole mixture. The widely used NIST SNR estimator [5] builds a histogram of short-term signal power of noisy speech to estimate noise and noisy speech distribution. The peak SNR is calculated from estimated distributions. Obviously, the peak SNR is an overestimation of the real SNR. In [6], Narayanan and Wang employ computational auditory scene analysis (CASA) for filtered global SNR estimation. An estimated ideal binary mask (IBM) is utilized to classifying time-frequency (T-F) units of noisy speech as noise-dominant or speech-dominant. Energy within each of these classes of T-F units is summated to derive the filtered global SNR within the bandwidth of a filterbank. They also design an SNR converter to transform the estimated filtered SNR into broadband SNR.

Recently, supervised learning algorithms are proposed to perform SNR estimation and have achieved substantial improvements over traditional methods. Suhadi $et\ al.$ [7] proposed a data-driven approach where two trained neural networks are used to estimate the $a\ priori$ SNR. In [8], Papadopoulos $et\ al.$ proposed a channel adapted deep neural network (DNN) that uses energy ratio features and i-vectors to train a DNN model for global SNR estimation in known and unknown channel conditions. In [9], the authors estimate the global SNR using a two-stage approach. The first stage produces noise residuals from a speech enhancement model. The second stage uses the noise residuals and a DNN to predict the global SNR.

In this paper, we investigate short-term broadband SNR estimation, where the duration is the frame length (20 ms) or frame-level SNR estimation. Unidirectional and bidirectional RNNs are proposed for causal and non-causal SNR estimation, respectively. Broadly speaking, a deep learning-based model consists of two main components: models and features [10]. While RNNs are powerful learning machines, input features need to be sufficiently discriminative [11, 12]. In this paper, we systematically examine a wide variety of monaural features for SNR estimation under the LSTM-based and BLSTM-based models. In addition, since each feature reveals certain characteristics of the speech signal, a set of features can be leveraged to boost SNR estimation performance. Hence, we further investigate feature combinations using Group Lasso [13, 11] and sequential floating forward selection (SFFS) [14, 12] methods. By substituting BLSTM for LSTM, the proposed algorithm naturally becomes a long-term SNR estimator.

The rest of this paper is organized as follows. We present our proposed algorithm in Section 2. Experimental setup and results are presented in Sections 3 and 4, respectively. We conclude this paper in Section 5.

2. Algorithm Description

2.1. Computational Objectives

In this study, we aim to predict the frame-level SNR, defined as

$$\mathrm{SNR}(m) = 10 \log 10 \frac{\sum_c |S(m,c)|^2}{\sum_c |N(m,c)|^2}, \tag{1}$$

where $S(m,c)$ and $N(m,c)$ refer to clean speech and noise, respectively, for the T-F unit at time frame m and frequency c. In this paper, the frame length is 20 ms with 10 ms frame shift and all mixtures are sampled at 16 kHz. The SNR value to be estimated is limited to the dB range of $[-30, 30]$, i.e., it will be

http://dx.doi.org/10.21437/Interspeech.2020-2475

set to -30 dB for any values lower than -30, and to 30 dB for any values higher than 30.

In order to convert a frame-level SNR estimator to a long-term SNR estimator, we assume speech and noise are uncorrelated, which is a common assumption. Based on this assumption, we have

$$|Y(m,c)|^2 = |S(m,c)|^2 + |N(m,c)|^2, \tag{2}$$

where Y denotes the mixture. According to Equation (1) and (2), the estimated noise energy at frame m can be estimated by Equation (3),

$$\hat{E}_N(m) = \frac{E_Y(m)}{10^{\frac{\widehat{SNR}(m)}{10}} + 1}, \tag{3}$$

where $E_Y(m) = \sum_c |Y(m,c)|^2$ and $\widehat{SNR}(m)$ denotes the estimate of $SNR(m)$. Then, the long-term SNR can be estimated as:

$$\widehat{SNR} = 10 \log_{10} \frac{\sum_m \left(E_Y(m) - \hat{E}_N(m) \right)}{\sum_m \hat{E}_N(m)}. \tag{4}$$

2.2. Acoustic Features

In this paper, we systematically examine 18 monaural features that have been introduced in different areas of speech processing:

- Waveform signal (WAV) [12].
- Mel-frequency cepstral coefficient (MFCC).
- Log-Mel filterbank feature (LOG-MEL).
- Multiresolution cochleagram (MRCG) [11].
- MRCG-causal.
- Perceptual linear prediction (PLP) [15].
- Relative spectral transform of PLP (RASTA-PLP) [16].
- Gammatone feature (GF).
- Gammatone frequency cepstral coefficient (GFCC) [17].
- Gammatone frequency modulation coefficient (GFMC) [18].
- Relative autocorrelation sequence MFCC (RAS-MFCC) [19].
- Autocorrelation sequence MFCC (AC-MFCC) [20].
- Power normalized cepstral coefficients (PNCC) [21].
- Gabor filterbank feature (GFB) [22].
- Amplitude modulation spectrogram (AMS) [23].
- Pitch-based feature (PITCH) [11].
- Magnitude spectral feature (MAG).
- Suppression of slowly-varying components and the falling edge of the power envelope (SSF) [24].

The MRCG is calculated at each frame by smoothing 11 past and future frames in a 64-channel cochleagram. For causal SNR estimation, we propose the MRCG-causal feature, which is the same as MRCG except for using 22 past frames and no future frame for smoothing.

All the features are normalized to zero mean and unit variance by using the statistics of the training data.

2.3. Network Architecture

An overview of the proposed RNN is shown in Fig. 1. The RNN has an input layer, four LSTM (or BLSTM) layers, and an output layer. The output layer is a linear layer that is used to map the output dimension to one. Each LSTM layer has 512 units. Each BLSTM layer has 300 units. The numbers of parameters in the LSTM-based and BLSTM-based model are comparable.

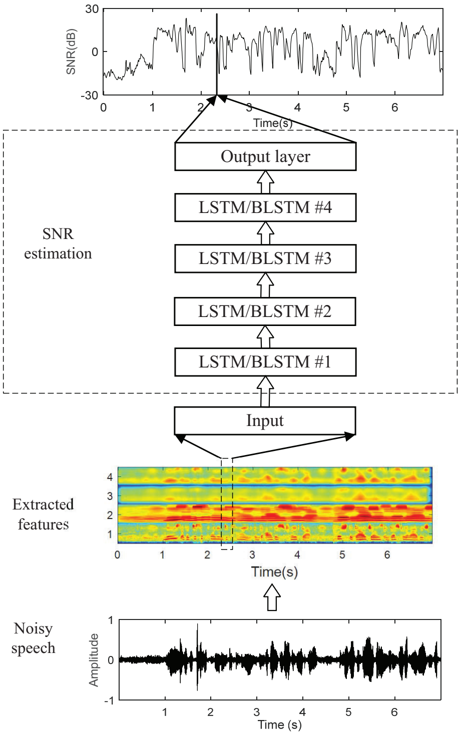

Figure 1: *Diagram of the proposed frame-level SNR estimation model. The input to the model is a noisy speech signal. The output is the frame-level SNR.*

The models are trained using the Adam optimizer [25] with a learning rate of 0.001 and the L_1-norm is used to define as the loss function. The minibatch size is set to 64 at the utterance level. The algorithms are run for 50 epochs, and the best model is selected by cross-validation.

3. Experimental Setup

3.1. Data Preparation

We evaluate the proposed models on the WSJ0 SI-84 dataset [26], which includes 7138 utterances from 83 speakers (42 males and 41 females). We randomly select and set aside six (three males and three females) of these speakers for testing. In other words, 77 remaining speakers are used to training the models. Of the utterances from the 77 training speakers, we hold out 150 randomly selected utterances to create a validation set with a babble noise from the NOISEX-92 dataset [27]. For training, we use the 10,000 noises from a sound effect library[1], which has a total duration of about 126 hours. For testing, we use six noises, i.e., babble and cafeteria noise from an Auditec CD[2], factory and speech shape noise (SSN) from NOISEX-92, and park and traffic noise from the DEMAND noise set [28]. Our training set contains 100,000 mixtures, and the total duration is about 160 hours. To create a training mixture, we mix a randomly selected training utterance and a random segment from the 10,000 training noises. The SNR is randomly sampled from -5 dB to 10 dB with a 1 dB increment. The validation set contains 800 utterances. The SNR of the validation utterances is randomly selected from -5 dB to 10 dB with a 1 dB increment,

[1] https://www.soundideas.com

[2] http://www.auditec.com

which is the same as in the training set. The test set includes 1200 mixtures that are created from 25 × 6 utterances of the 6 untrained speakers. The test set SNR is randomly selected from -10 dB to 15 dB with a 5 dB increment. Note that speech and noise signals are different between training and testing, and some of the test SNRs are not used during training.

3.2. Metrics

The accuracy of SNR estimation is measured by mean absolute error (MAE) between the true SNR and an estimated SNR:

$$\mathrm{MAE} = \frac{1}{M}\sum_{m=1}^{M}\left|\mathrm{SNR}(m) - \widehat{\mathrm{SNR}}(m)\right|. \tag{5}$$

For frame-level SNR estimation, M indicates the total number of frames of all the utterances in an evaluation corpus. For long-term SNR estimation, MAE measures the average of all utterances of an evaluation corpus.

3.3. Baseline Systems for Comparison

We compare the proposed frame-level SNR estimator with two baselines. The first algorithm uses a DNN to predict the ideal ratio mask (IRM) for speech enhancement (SE-based). The deep learning model used is the same as the BLSTM-based model, except that the output layer has 161 units with the sigmoidal activation function. The input feature is MRCG, which is concluded to be the best feature in [11]. After obtaining an estimated IRM, the speech and noise energy in a frame can be readily estimated to calculate the frame-level SNR. The second algorithm uses a minimum mean-square error (MMSE) estimator to predict the clean speech power spectral density (PSD) [1] . The ratio of speech PSD and noisy speech power in each frame is utilized to estimate the frame-level SNR, hence referred to as PSD-based.

We compare the proposed long-term SNR estimator with four baselines, i.e., WADA [29], CASA [6], SE-based, and Residual-based [9]. The SE-based baseline is the same as the corresponding baseline in frame-level SNR estimation. After obtaining an estimated IRM, the energy of speech and noise at the utterance level can be estimated to calculate long-term SNR.

4. Evaluation Results

In this section, we first evaluate individual features and perform feature combinations. We then evaluate the performance of the proposed frame-level and long-term SNR estimators and compare with the baseline models.

4.1. Feature Evaluations

4.1.1. Single Features

Table 1 shows the SNR estimation results of the individual features using the LSTM-based and BLSTM-based RNN, in the MAE order for the LSTM model. The BLSTM-based model performs better than the LSTM-based model, to be expected as BLSTM can captures both past and future information. The best single features for LSTM-based and BLSTM-based models are MRCG and GF, respectively. MRCG performs 0.1 dB better than MRCG-causal for the LSTM-based model. For the BLSTM-based model, the performances of these two features are close, as expected.

The noise-robust features of SSF, RAS-MFCC, PNCC, RASTA-PLP, GFMC, and PITCH are generally worse than other features. The reason may be that these features are designed for robust speech separation or automatic speech recognition (ASR), potentially making them relatively insensitive to the level of noise in a noisy speech signal. The sensitivity to noise level is important for SNR estimation, as SNR is determined by both speech and noise levels. Therefore, the noise-robust features are not robust to SNR estimation. It is interesting to note that the WAV feature, the raw waveform input, performs quite well for SNR estimation. This is consistent with the above argument regarding noise robustness.

Table 1: *SNR estimation results in terms of MAE for LSTM-based and BLSTM-based models evaluated on individual features. The 'Causal' column indicates whether the algorithm is causal.*

Feature	LSTM model		BLSTM model	
	MAE	Causal	MAE	Causal
MRCG	**3.197**	N	2.633	N
MAG	3.276	Y	2.641	N
MRCG-causal	3.288	Y	2.638	N
WAV	3.291	Y	2.748	N
LOG-MEL	3.353	Y	2.727	N
MFCC	3.592	Y	2.921	N
SSF	3.620	N	3.025	N
PLP	3.625	Y	2.796	N
AC-MFCC	3.729	Y	3.108	N
GF	3.777	Y	**2.556**	N
GFB	3.838	Y	3.516	N
GFCC	3.843	Y	2.694	N
RAS-MFCC	3.923	Y	3.303	N
RASTA-PLP	4.344	Y	3.452	N
AMS	4.825	Y	4.067	N
PNCC	4.970	N	4.660	N
GFMC	5.375	Y	4.049	N
PITCH	6.930	N	6.412	N

4.1.2. Feature Combinations

Individual features are designed to reveal certain characteristics of noisy speech. The combination of features may boost SNR estimation performance. In this paper, two feature selection algorithms are used to explore feature combinations. For the LSTM-based model, we only use causal features to ensure that the algorithm can work in real time. The feature set used is {WAV, MFCC, LOG-MEL, MRCG-causal, PLP, PASTA-PLP, GF, GFCC, GFMC, AC-MFCC, GFB, AMS, MAG}. All features except MRCG-causal are used in the BLSTM-based model.

The first feature combination algorithm is Group Lasso [13]. In [12, 11, 30], Group Lasso is used to find complementary relations between features. After performing Group Lasso on the features, we find that MRCG-causal, GFB, and MAG are the only features with significant responses and all other features have zero or negligible responses. Accordingly, we use MRCG-causal+GFB+MAG as the complementary feature set from Group Lasso in the LSTM-based model. For the BLSTM-based model, MRCG and PITCH are the features with significant responses. Hence, MRCG and PITCH are used as the complementary feature set from Group Lasso in the BLSTM-based model.

The SFFS algorithm [14] systematically adds and drops features until a desired number of features is selected. In this paper, the desired number of the features is unknown. We adopt the modified version proposed in [12], where the algorithm will stop when no improvement is achieved by adding the next feature. For the LSTM-based model, the feature set obtained by SFFS consists of MAG, GFB, MRCG-causal, GF, and GFCC. For the BLSTM-based model, the feature set selected is

Table 2: *SNR estimation results in MAE for feature combinations with LSTM-based model.*

Method	Noise						Avg.
	babble	cafeteria	park	traffic	factory	SSN	
MAG	3.98	3.80	2.73	2.46	3.77	2.96	3.28
Group Lasso	3.87	3.46	2.46	2.18	3.42	2.71	3.02
SFFS	**3.44**	**3.42**	**2.38**	**1.91**	**3.17**	**2.33**	**2.77**

Table 3: *SNR estimation results in MAE for feature combinations with BLSTM-based model.*

Method	Noise						Avg.
	babble	cafeteria	park	traffic	factory	SSN	
GF	3.04	2.85	2.61	1.92	2.79	2.15	2.56
Group Lasso	6.83	4.28	3.23	2.47	3.51	2.82	3.86
SFFS	**2.90**	**2.81**	**2.03**	**1.82**	**2.75**	**2.15**	**2.41**

GF+LOG-MEL+AMS+PLP.

We compare the performance of feature combinations with that of the best single feature. It should be noted that since MRCG is a non-causal feature, we select MAG as the best single feature associated with the LSTM-based model. The results of the LSTM-based and BLSTM-based model are shown in Tables 2 and 3, respectively. In both algorithms, the feature sets from the SFFS algorithm obtain the best performance under all noise conditions. For the LSTM-based model, the SFFS feature set has two more features than the Group Lasso feature set. By adding GF and GFCC, the average MAE is reduced by 0.25 dB. Babble noise achieves the biggest improvement among all noises, and it has decreased MAE by 0.43 dB. On average, the SFFS feature set results in 0.51 dB improvement over the MAG feature. In Table 3, the BLSTM algorithm using SFFS feature set is 0.15 dB better than using the single best feature of GF. The average MAE with the MRCG feature is 2.63 dB (see Table 1). After combining the PITCH feature, the average MAE is increased to 3.86 dB. The reason is pitch is hard to track, especially for babble noise which combines many speech utterances, and the inaccurate PITCH feature would decrease the performance of a feature set. As Group Lasso is a linear regression algorithm, it may not be strong enough to handle the nonlinear relationship between input features and the target SNR.

4.2. SNR Estimation

4.2.1. Frame-level Estimation

Table 4 shows the frame-level results in terms of MAE of the proposed LSTM-based algorithm and the baseline models for different noises. In the table, each result represents the average of the test SNRs. The features used are the best feature sets selected through SFFS feature combinations, which are MAG+GFB+MRCG-causal+GF+GFCC for the LSTM-based algorithm and GF+LOG-MEL+AMS+PLP for the BLSTM-based algorithm (see Sect. 4.1).

The proposed BLSTM-based algorithm shows the best performance under all noise conditions. On average, the MAE value of the proposed BLSTM-based algorithm is 2.41 dB, which is 4.28 dB better than the PSD-based algorithm and 1.9 dB better than the SE-based algorithm. The algorithm based on BLSTM is about 0.36 dB better than the algorithm based on LSTM. However, the LSTM-based algorithm is a causal system that can estimate SNR in real time. When a background noise is non-stationary or SNR is low, the PSD-based algorithm makes large estimation errors. Compared with the SE-based algorithm which predicts the IRM as an intermediate result, the proposed algorithms directly estimate the SNR and obtain better results.

Table 4: *Frame-level SNR estimation results in MAE for different methods under different noise conditions.*

Method	Noise						Avg.
	babble	cafeteria	park	traffic	factory	SSN	
PSD-based	8.70	6.39	7.96	5.11	6.56	5.41	6.69
SE-based	5.54	5.36	2.83	2.93	4.65	4.54	4.31
LSTM-based	3.44	3.42	2.38	1.91	3.17	2.33	2.77
BLSTM-based	**2.90**	**2.81**	**2.03**	**1.82**	**2.75**	**2.15**	**2.41**

Table 5: *Long-term SNR estimation results in MAE for different methods.*

Method	SNR (dB)						Avg.
	-10	-5	0	5	10	15	
WADA	5.78	2.41	0.97	0.95	1.11	1.76	2.16
CASA	2.44	1.29	0.78	0.91	1.39	2.12	1.49
Residual-based	2.47	1.37	2.42	2.49	2.08	2.63	2.24
SE-based	0.71	0.35	0.19	0.25	0.32	0.48	0.38
Proposed	**0.58**	**0.34**	**0.15**	**0.13**	**0.15**	**0.29**	**0.27**

4.2.2. Long-term Estimation

The estimation results of long-term SNR for different methods are shown in Table 5, where each MAE value represents the average of different noises. The results of the proposed method are calculated using the BLSTM-based model and the feature set of GF+LOG-MEG+AMS+PLP (see Sect. 4.1). The proposed algorithm achieves the best results across all test SNRs. The second best algorithm is SE-based. It is around 0.1 dB worse than the proposed algorithm. If we look further at each SNR condition, the MAEs of the SE-based and the proposed methods are closest at -5dB, which are 0.35 and 0.34 dB, respectively. As the SNR increases or decreases, the performance of SE-based method gradually becomes worse than the proposed method. The proposed method obtains an optimal MAE performance in the case of 5 dB, where the MAE is 0.13 dB. The CASA-based algorithm depends on whether noisy T-F units can be accurately classified, and it does not always work well, especially in low SNR conditions. WADA performs reasonably at relatively high SNRs. But in low SNR conditions, noisy speech does not follow the Gamma distribution assumed by WADA, resulting in poor results. The Residual-based method first uses a DNN to predict a complex ratio mask, and then uses noise residuals to predict global SNR. It is difficult to estimate global SNR by using noise residuals alone.

5. Conclusion

In this paper, we have proposed a deep learning algorithm for frame-level SNR estimation. Our algorithm shows clear improvements over previous methods. We have also examined a wide range of acoustic features for their effectiveness in SNR estimation and investigated feature combinations using Group Lasso and SFFS. We have found that feature combinations can boost SNR estimation performances. Based on the frame-level SNR estimator, we have additionally derived a long-term SNR estimator, which outperforms other long-term SNR estimation methods.

6. Acknowledgments

This work was performed while the first author was a visiting scholar at the Ohio State University. This research was supported in part by the Ohio Supercomputer Center.

7. References

[1] T. May, B. Kowalewski, M. Fereczkowski, and E. N. MacDonald, "Assessment of broadband SNR estimation for hearing aid applications," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 231–235.

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

[3] 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, pp. 1109–1121, 1984.

[4] O. Cappé, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 345–349, 1994.

[5] NIST, "NIST speech singal to noise ratio measurements," *https://www.nist.gov/itl/iad/mig/nist-speech-signal-noise-ratio-measurements*.

[6] A. Narayanan and D. Wang, "A CASA-based system for long-term SNR estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 20, pp. 2518–2527, 2012.

[7] S. Suhadi, C. Last, and T. Fingscheidt, "A data-driven approach to a priori SNR estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 186–195, 2010.

[8] P. Papadopoulos, R. Travadi, and S. S. Narayanan, "Global SNR Estimation of Speech Signals for Unknown Noise Conditions Using Noise Adapted Non-Linear Regression," in *Interspeech*, 2017, pp. 3842–3846.

[9] X. Dong and D. S. Williamson, "Long-term SNR estimation using noise residuals and a two-stage deep-learning framework," in *International Conference on Latent Variable Analysis and Signal Separation*, 2018, pp. 351–360.

[10] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, pp. 1702–1726, 2018.

[11] J. Chen, Y. Wang, and D. Wang, "A feature study for classification-based speech separation at low signal-to-noise ratios," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, pp. 1993–2002, 2014.

[12] M. Delfarah and D. Wang, "Features for masking-based monaural speech separation in reverberant conditions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, pp. 1085–1094, 2017.

[13] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, pp. 49–67, 2006.

[14] P. Pudil, F. J. Ferri, J. Novovicova, and J. Kittler, "Floating search methods for feature selection with nonmonotonic criterion functions," in *IAPR International Conference on Pattern Recognition*, 1994, pp. 279–283.

[15] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, pp. 1738–1752, 1990.

[16] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 578–589, 1994.

[17] Y. Shao and D. Wang, "Robust speaker identification using auditory features and computational auditory scene analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 1589–1592.

[18] H. K. Maganti and M. Matassoni, "An auditory based modulation spectral feature for reverberant speech recognition," in *Interspeech*, 2010, pp. 570–573.

[19] K. H. Yuo and H. C. Wang, "Robust features for noisy speech recognition based on temporal trajectory filtering of short-time autocorrelation sequences," *Speech Communication*, vol. 28, pp. 13–24, 1999.

[20] B. J. Shannon and K. K. Paliwal, "Feature extraction from higher-lag autocorrelation coefficients for robust speech recognition," *Speech Communication*, vol. 48, pp. 1458–1485, 2006.

[21] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, pp. 1315–1329, 2016.

[22] M. R. Schädler, B. T. Meyer, and B. Kollmeier, "Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition," *Journal of the Acoustical Society of America*, vol. 131, pp. 4134–4151, 2012.

[23] G. Kim, Y. Lu, Y. Hu, and P. C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *Journal of the Acoustical Society of America*, vol. 126, pp. 1486–1494, 2009.

[24] C. Kim and R. M. Stern, "Nonlinear enhancement of onset for robust speech recognition," in *Interspeech*, 2010, pp. 2058–2061.

[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[26] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," in *Workshop on Speech and Natural Language*, 1992, pp. 357–362.

[27] 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, pp. 247–251, 1993.

[28] J. Thiemann, N. Ito, and E. Vincent, "The Diverse Environments Multi-channel Acoustic Noise Database (DEMAND): A database of multichannel environmental noise recordings," in *International Congress on Acoustics*, 2013.

[29] C. Kim and R. M. Stern, "Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis," in *Interspeech*, 2008, pp. 2598–2601.

[30] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, pp. 1849–1858, 2014.