Recurrent Neural Networks and Acoustic Features for Frame-Level Signal-to-Noise Ratio Estimation

Hao Li, *Student Member, IEEE*, DeLiang Wang, *Fellow, IEEE*, Xueliang Zhang, *Member, IEEE*, and Guanglai Gao

***Abstract*—It is important to know the presence and the relative level of background noise for many speech processing tasks. Frame-level signal-to-noise ratio (SNR) provides a measure of instantaneous noise level of a noisy signal, and its estimation has been researched for decades. This problem can be approached from a supervised learning perspective by predicting SNR from features of noisy speech. In this study, we introduce a deep learning algorithm for frame-level SNR estimation. The proposed algorithm employs recurrent neural networks (RNNs) with long short-term memory (LSTM) to leverage contextual information. We also systematically examine a range of acoustic features and investigate feature combinations using Group Lasso and sequential floating forward selection (SFFS). The proposed algorithm naturally leads to an utterance-level SNR estimator. Systematical evaluations show that the proposed algorithm provides an accurate estimate of frame-level SNR, as well as utterance-level SNR, under different noise conditions, outperforming other estimators.**

***Index Terms*—Frame-level SNR estimation, feature combination, long short-term memory, recurrent neural networks.**

I. Introduction

Speech processing is a challenging task in real life since various types of noise interfere with a speech signal. Signal-to-noise ratio (SNR) indicates the amount of noise interference in an acoustic environment. Knowledge of the SNR has many applications, including speech enhancement [9], [10] and hearing aids [24].

There are typically two ways to measure SNR in a noisy signal. The first is short-term or frame-level SNR, also referred to as instantaneous SNR. The duration is usually in the range of tens to hundreds milliseconds. Short-term SNR can be narrowband or broadband, and the former is defined in dB as

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

where $S(m, c)$ and $N(m, c)$ denote the complex spectra of clean speech and noise, respectively, for the time-frequency (T-F) unit at time frame m and frequency c. Broadband SNR is defined as

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

These definitions correspond to *a priori* SNR in traditional speech enhancement literature. For example, the decision-directed method of Ephraim and Malah [8] is a common *a priori* SNR estimator, which helps to reduce musical tones [1]. Nemer *et al.* [26] make use of higher-order statistics of speech and noise, assuming a sinusoidal model for band restricted speech and a Gaussian model for noise. Other methods for short-term SNR estimation include a spectral histogram based method [15], energy clustering to distinguish speech and noise portions of the mixture [4], [5], and voice activity detection [21].

The second way measures SNR at the utterance level, referred to as global, long-term or utterance-level SNR. Utterance-level SNR considers the entire speech signal and provides noise level for the whole mixture. Similar to short-term SNR, utterance-level SNR can be narrowband,

$$\mathrm{SNR(c)} = 10 \log_{10} \frac{\sum_m |S(m, c)|^2}{\sum_m |N(m, c)|^2} \tag{3}$$

or broadband:

$$\mathrm{SNR} = 10 \log_{10} \frac{\sum_{m,c} |S(m, c)|^2}{\sum_{m,c} |N(m, c)|^2} \tag{4}$$

The widely used NIST SNR estimator [27] builds a histogram of short-term signal power using noisy speech to estimate noise and noisy speech distribution. The peak SNR is then calculated from the estimated distributions. Obviously, the peak SNR is an overestimate of global SNR. The method of Kim and Stern [17] is based on waveform amplitude distribution. It assumes that clean and noisy speech have Gamma distributions, and noise has a Gaussian distribution. The method uses a maximum likelihood estimation to find the parameters of the Gamma distribution to infer the utterance-level SNR. Narayanan and Wang employ computational auditory scene analysis (CASA) for global SNR estimation [25]. An estimate of the ideal binary mask (IBM) is utilized to classify T-F units of noisy speech as noise-dominant or speech-dominant. Energy within each of these classes is summated to derive the global SNR within the bandwidth of

Manuscript received March 24, 2021; revised August 16, 2021; accepted August 19, 2021. Date of publication August 30, 2021; date of current version September 11, 2021. This work was supported in part by the China National Nature Science Foundation under Grant 62066033 and Grant 61773224, and the Ohio Supercomputer Center. The work was conducted while Hao Li was visiting The Ohio State University. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Zheng-Hua *(Corresponding author: Hao Li.)*

Hao Li, Xueliang Zhang, and Guanglai Gao are with the Department of Computer Science, Inner Mongolia University, Hohhot 010021, China (e-mail: lihao@mail.imu.edu.cn; cszxl@imu.edu.cn; csggl@imu.edu.cn).

DeLiang Wang is with the Department of Computer Science and Engineering and the Center for Cognitive and Brain Sciences, Ohio State University, Columbus, OH 43202 USA (e-mail: dwang@cse.ohio-state.edu).

Digital Object Identifier 10.1109/TASLP.2021.3107617

a filterbank. They also design an SNR converter to transform the estimated band-limited SNR to the broadband SNR.

Recently, supervised learning algorithms are proposed to perform SNR estimation and have achieved substantial improvements over conventional methods. Suhadi *et al.* [35] propose a data-driven approach that employs two feedforward neural networks to estimate the *a priori* SNR, each with one hidden layer of 10 neurons and one output neuron. Papadopoulos *et al.* use energy ratio features to train regression models for different noises to estimate the utterance-level SNR [30]. In the test phase, if the noise type is known, the corresponding model is used to estimate the SNR; if the noise type is unknown, a deep neural network (DNN) model is used to find the "closest" regression model to estimate the SNR. In [29], they further use i-vectors to provide information about noise, as well as energy ratio features to train a feedforward DNN model for utterance-level SNR estimation in known and unknown channel conditions. The DNN model consists of 4 hidden layers where each has 1024 units. Dong and Williamson [7] propose a two-stage approach to estimate the utterance-level SNR. The first stage produces noise residuals from a speech separation model. The second stage uses the noise residuals and a feedforward DNN to predict utterance-level SNR. They use complementary features [39] extracted from residuals as inputs to the DNN. These features consist of amplitude modulation spectrogram, relative spectral transform perceptual linear prediction, and Mel-frequency cepstral coefficients. They also add delta features to capture temporal dynamics. Their DNN has three hidden layers with 512, 256, and 128 units, respectively.

This paper investigates short-term broadband SNR estimation. Unidirectional and bidirectional recurrent neural networks (RNNs) are proposed for causal and non-causal frame-level SNR estimation, respectively. Compared with the feedforward DNNs used in [35] [30] [29] [7], RNNs are better suited for modeling sequential data with long-term dependencies. In [42], [41], [46], RNNs with long short-term memory (LSTM) are used to perform speech enhancement. Chen and Wang [3] employ an RNN with four LSTM layers to address speaker generalization of noise-independent speech enhancement.

Broadly speaking, a deep learning based model consists of two components: models and features [38]. While RNNs are powerful learning machines, input features need to be sufficiently discriminative [2], [6]. To explore the influence of different acoustic features, we systematically examine acoustic features. As individual features reveal certain characteristics of noisy speech, it would be useful to leverage a combination of features. Hence, we further study different feature combinations using Group Lasso [44], [2] and sequential floating forward selection (SFFS) [32], [6]. As a frame-level estimator, the proposed algorithm naturally leads to an utterance-level estimator. We additionally evaluate the accuracy of the utterance-level SNR estimator under different SNR conditions.

A preliminary version of this paper is included in [22]. The current work provides a more detailed analysis, in addition to expanded evaluations and comparisons. In addition, we have documented SNR estimation results using a new metric and an untrained speech corpus.

The rest of the paper is organized as follows. Section II describes the acoustic features researched in this study. The proposed algorithm is detailed in Section III. The experimental setup is explained in Section IV. Performances of each individual feature and feature combinations are evaluated in Section V. SNR estimation compared with the baseline models are studied in Section VI. Section VII concludes the paper.

II. Acoustic Features

We systematically examine 18 acoustic features that have been used for many speech processing tasks:

- Waveform signal (WAV) [6].
- Mel-frequency cepstral coefficient (MFCC).
- Log-mel filterbank feature (LOG-MEL).
- Multi-resolution cochleagram (MRCG) [2].
- Causal MRCG (MRCG-causal).
- Perceptual linear prediction (PLP) [13].
- Relative spectral transform of PLP (RASTA-PLP) [14].
- Gammatone feature (GF).
- Gammatone frequency cepstral coefficient (GFCC) [43].
- Gammatone frequency modulation coefficient (GFMC) [23].
- Relative autocorrelation sequence MFCC (RAS-MFCC) [45].
- Autocorrelation sequence MFCC (AC-MFCC) [34].
- Power normalized cepstral coefficients (PNCC) [18].
- Gabor filterbank feature (GFB) [33].
- Amplitude modulation spectrogram (AMS) [19].
- Pitch-based feature (PITCH) [39].
- Magnitude spectral feature (MAG).
- Suppression of slowly-varying components and the falling edge of the power envelope (SSF) [16].

The input signal is divided into 20 ms frames with 10 ms frame shift to generate the WAV feature. A 320-point fast Fourier transform is applied to the WAV feature to obtain the spectrogram of the signal. The MAG feature is the magnitude response of the spectrogram.

The MRCG encodes multi-resolution power distributions in the T-F domain of a signal. Four cochleagrams at different resolutions are concatenated to construct the MRCG feature. A high resolution cochleagram (CG1) captures the local information while three lower resolution cochleagrams (CG2, CG3, CG4) capture spectrotemporal contexts at different scales. The cochleagrams are calculated by the following steps:

i) Compute the first 64-channel cochleagram with the frame length of 20 ms and frame shift of 10 ms. Then, a log operation is applied to the cochleagram to form CG1.
ii) Compute the first 64-channel cochleagram with the frame length of 200 ms and frame shift of 10 ms. Then, a log operation is applied to the cochleagram to form CG2.
iii) CG3 is calculated by smoothing CG1 using the square window of 11×11.
iv) CG4 is calculated by smoothing CG2 using the square window of 23×23.

The MRCG feature uses future frames in steps (iii) and (iv), making it non-causal. For causal SNR estimation, we construct

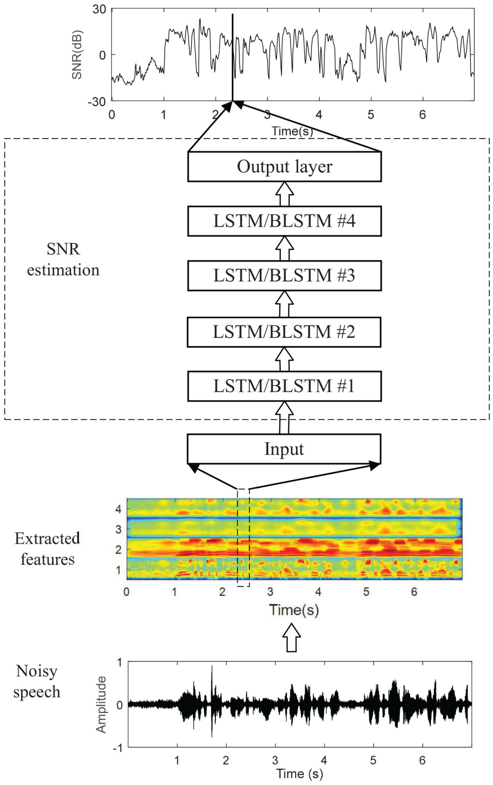

Fig. 1. Diagram of the proposed model. The input to the model is a noisy speech signal. The output is the frame-level SNR.

an MRCG-causal feature, which is the same as MRCG except for using the past 10 and 22 frames in steps (iii) and (iv), respectively, and no future frame for smoothing.

To calculate the PITCH feature, the cochleagram of a noisy signal is first calculated. We use the PEFAC algorithm [12] for pitch estimation at each time frame. Then six pitch-based features are extracted in each T-F unit [39]. Finally, the extracted features are concatenated across frequency to form the PITCH feature.

We use publicly available programs to extract GFB,[1] PNCC, SSF,[2] GF, GFCC and MRCG,[3] LOG-MEL, AMS, GFMC, and AC-MFCC,[4] and use the RASTAMAT toolbox[5] to obtain PLP, RASTA-PLP, MFCC and RAS-MFCC. All the features are normalized to zero mean and unit variance based on the statistics of the training data.

III. Proposed Algorithms

A. System Overview

An overview of the proposed framework is shown in Fig. 1. The framework has an input layer, four LSTM (or BLSTM) layers, and an output layer. The output layer is a linear layer to map the output dimension to one. Each LSTM layer has 512 units, and each BLSTM layer has 300 units. The LSTM- and BLSTM-based models are used to explore causal and non-causal SNR estimations, respectively. The numbers of trainable parameters for these two models under different acoustic features are listed in Table I, and they are comparable.

TABLE I
Numbers of Trainable Parameters for the LSTM-Based and BLSTM-Based Algorithms for Different Acoustic Features

Feature	Dimension	Trainable parameters (million)	
		LSTM-based	BLSTM-based
WAV	320	8.01	7.98
MFCC	31	7.42	7.29
LOG-MEL	40	7.44	7.32
MRCG	256	7.88	7.83
MRCG-causal	256	7.88	7.83
PLP	13	7.38	7.25
RASTA-PLP	13	7.38	7.25
GF	64	7.49	7.37
GFCC	31	7.42	7.29
GFMC	31	7.42	7.29
RAS-MFCC	31	7.42	7.29
AC-MFCC	31	7.42	7.29
PNCC	31	7.42	7.29
GFB	311	7.99	7.97
AMS	15	7.39	7.26
PITCH	384	8.14	8.14
MAG	161	7.69	7.61
SSF	34	7.43	7.30

The models are trained using the Adam optimizer [20] with a learning rate of 0.001. The minibatch size is set to 64 at the utterance level. Within a minibatch, all training samples are padded with zeros to have the same number of time steps as the longest sample. The algorithm is run for 50 epochs, and the best model is selected by cross validation. The L_1-norm is used to define the loss function:

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

where $\mathrm{SNR}(m)$ and $\widehat{\mathrm{SNR}}(m)$ denote the target and predicted SNR of frame m, respectively. M is the number of frames in an utterance.

B. Frame-Level SNR Estimation

We aim to predict the frame-level SNR defined in Eq. (2). The frame length is 20 ms with 10 ms frame shift, and all signals are sampled at 16 kHz. The range of the SNR value is $(-\infty, \infty)$. The infinite value range complicates SNR estimation. In this study, the SNR to be estimated is limited to the dB range of $[-30, 30]$, i.e., it will be set to -30 dB for any values lower than -30, and to 30 dB for any values higher than 30. This range should be sufficient in practice.

[1] [Online]. Available: https://github.com/m-r-s/reference-feature-extraction
[2] [Online]. Available: http://www.cs.cmu.edu/~chanwook
[3] [Online]. Available: http://web.cse.ohio-state.edu/pnl/software.html
[4] [Online]. Available: https://github.com/imu-HaoLi/Feature_tools
[5] [Online]. Available: http://labrosa.ee.columbia.edu/matlab/rastamat

C. Utterance-Level SNR Estimation

As a frame-level estimator, the proposed algorithm readily leads to an utterance-level SNR estimator. To calculate the utterance-level SNR, we assume speech and noise are uncorrelated, which is a common assumption. Based on this assumption, we have:

$$|Y(m,c)|^2 = |S(m,c)|^2 + |N(m,c)|^2, \tag{6}$$

where Y denotes the mixture. According to Eq. (2) and Eq. (6), the estimated noise energy at frame m is given by:

$$\widehat{E}_N(m) = \frac{E_Y(m)}{10^{\frac{\widehat{SNR}(m)}{10}} + 1} \tag{7}$$

where $E_Y(m) = \sum_c |Y(m,c)|^2$, and $\widehat{SNR}(m)$ denotes the frame-level SNR obtained by the proposed methods. Then, the estimated speech energy at frame m is obtained as follows:

$$\widehat{E}_S(m) = E_Y(m) - \widehat{E}_N(m) \tag{8}$$

Finally, the utterance-level SNR in dB is estimated by:

$$\widehat{SNR} = 10\log_{10}\frac{\sum_m \widehat{E}_S(m)}{\sum_m \widehat{E}_N(m)} \tag{9}$$

IV. Experimental Setup

A. Data Preparation

We evaluate the proposed algorithm on the WSJ0 SI-84 dataset [31], which includes 7138 utterances from 83 speakers (42 males and 41 females). Six (three males and three females) of these speakers are randomly selected and set aside for testing. In other words, 77 remaining speakers are used to train the model. We also hold out 150 randomly selected utterances from the 77 training speakers to create a validation set with a babble noise from the NOISEX-92 dataset [37]. We use the 10,000 noises from a sound effect library,[6] which has a total duration of about 126 hours, as the training noise set. For testing, we use six noises, i.e., babble and cafeteria noise from an Auditec CD,[7] factory and speech shape noise (SSN) from NOISEX-92, and park and traffic noise from the DEMAND noise set [36]. These test noises are selected to represent the kinds encountered in practical situations. The training set contains 100,000 mixtures, and the total duration is about 160 hours. To generate 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 step size of 1 dB. The validation set contains 800 utterances. The SNR of the validation utterances is randomly selected from -5 dB to 10 dB with a step size of 1 dB, which is the same as in the training set. The test set includes 1,200 mixtures generated from 25 × 6 utterances of the 6 untrained speakers. The test set SNR is randomly selected from -10 dB to 15 dB with a step size of 5 dB. Note that speech and noise signals are different between training and testing, and two test SNRs are not included in the training set.

B. Metrics

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

$$\mathrm{MAE} = \frac{1}{M}\sum_{m=1}^{M}\left|\mathrm{SNR}(m) - \widehat{\mathrm{SNR}}(m)\right| \tag{10}$$

For frame-level SNR estimation, M indicates the total number of frames of all the utterances in an evaluation corpus. For utterance-level SNR estimation, MAE measures the average error of all utterances of an evaluation corpus. The evaluation metric is aligned with the loss function (see Eq. (5)).

We also use Pearson's Correlation Coefficient (PCC) and Spearman's Rank Correlation (SRC) to complement MAE. PCC is a correlation coefficient between the true SNR and an estimated SNR given by Eq. (11), where an upper bar indicates the mean and $\widehat{SNR}$ indicates estimated SNR. The closer PCC is to 1, the stronger is the correlation between true SNRs and estimated SNRs. SRC is similar to the PCC, but measured in terms of ranked values between true SNRs and estimated SNRs. While PCC assesses the linear relationship, SRC evaluates the monotonic relationship, whether it is linear or not. In the following sections, PCC and SRC are displayed in percentage.

C. Baseline Systems for Comparison

We compare the proposed frame-level SNR estimators with three strong baselines.

1) Power spectral density (PSD) baseline. This recent baseline uses a minimum mean-square error estimator to predict the clean speech PSD [24]. The ratio of speech PSD and noisy speech power in each frame is utilized to estimate the frame-level SNR.
2) Speech enhancement (SE) baseline. To our knowledge, no frame-level SNR estimator exists that is based on deep learning. Since an SE algorithm outputs estimated clean speech at every frame, it can be readily converted to a frame-level SNR estimator. This SE baseline uses DNN to predict the the ideal ratio mask (IRM) of speech for enhancement. The deep learning structure 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, the best feature proposed in [2]. After obtaining an estimated IRM, the estimated energies of speech and noise in a frame are used to calculate frame-level SNR.
3) SE+ baseline. DNN based speech enhancement algorithms typically output a speech estimate, not a noise estimate [38]. As SNR is equally sensitive to speech and noise levels, we propose another SE baseline, called SE+, that predicts both the speech IRM and the noise IRM. We use LSTM and BLSTM to create causal and non-causal SE+ versions, referred to as the LSTM-SE+ and BLSTM-SE+, respectively. The corresponding input features are MRCG-causal and MRCG. Different from the SE baseline, the output layer has 161×2 units.

[6] [Online]. Available: https://www.soundideas.com

[7] [Online]. Available: http://www.auditec.com

For SE and SE+ baselines, the mean squared error loss is used, and it is defined in terms of the difference between the IRM and an estimated IRM. The models are trained using the Adam optimizer [20] with a learning rate 0.001, and run for 50 epochs. The best model is selected by cross validation.

Many methods have been proposed to perform utterance-level SNR estimation. We compare the proposed utterance-level SNR estimators with six representative baselines, including traditional and DNN based methods.

1) WADA baseline. This algorithm is a commonly used estimator based on Waveform Amplitude Distribution Analysis (WADA) [17].
2) CASA baseline. The CASA-based method [25] performs IBM estimation to identify speech-dominant and noise-dominant T-F units as described in Section I.
3) SE. This algorithm is the same as the corresponding baseline for frame-level SNR estimation. After obtaining an estimated speech IRM, the energies of speech and noise at the utterance level are estimated to calculate the utterance-level SNR.
4) BLSTM-SE+. As the utterance-level SNR is estimated over the whole mixture, the non-causal BLSTM-SE+ algorithm is proposed as another baseline.
5) Residual-based. As described in Section I, this baseline by Dong and Williamson [7] uses noise residuals and a two-stage deep-learning model to estimate utterance-level SNR.
6) Papadopoulos *et al*. baseline. As described in Section I, this channel adapted DNN method [29] employs energy ratio features and i-vectors for utterance-level SNR estimation.

V. Feature Evaluation Results

In this section, we systematically examine each feature individually and feature combinations. For feature evaluations, both the LSTM-based and BLSTM-based models are employed.

A. Single Features

TABLE II
SNR Estimation Results in Terms of MAE and PCC/SRC (In Parentheses) 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 (PCC/SRC)	Causal	MAE (PCC/SRC)	Causal
MRCG	**3.197 (93.3/94.2)**	N	2.633 (95.4/96.0)	N
MAG	3.276 (92.6/93.4)	Y	2.641 (95.3/95.9)	N
MRCG-causal	3.288 (93.1/93.9)	Y	2.638 (95.3/96.0)	N
WAV	3.291 (92.9/93.7)	Y	2.748 (95.1/95.9)	N
LOG-MEL	3.353 (92.4/93.3)	Y	2.727 (95.0/95.7)	N
MFCC	3.592 (91.9/92.8)	Y	2.921 (94.7/95.4)	N
SSF	3.620 (91.8/92.9)	N	3.025 (94.2/94.9)	N
PLP	3.625 (92.0/92.8)	Y	2.796 (94.5/95.3)	N
AC-MFCC	3.729 (91.5/92.4)	Y	3.108 (94.1/94.8)	N
GF	3.777 (91.7/92.4)	Y	**2.556 (95.5/96.1)**	N
GFB	3.838 (91.3/92.3)	Y	3.516 (93.0/93.8)	N
GFCC	3.843 (91.3/92.0)	Y	2.694 (95.2/95.9)	N
RAS-MFCC	3.923 (91.2/92.2)	Y	3.303 (93.6/94.5)	N
RASTA-PLP	4.344 (90.1/91.1)	Y	3.452 (93.5/94.3)	N
AMS	4.825 (87.0/87.7)	Y	4.067 (90.1/91.0)	N
PNCC	4.970 (89.1/90.2)	N	4.660 (90.5/91.5)	N
GFMC	5.375 (85.7/86.3)	Y	4.049 (91.4/92.2)	N
PITCH	6.930 (79.6/82.8)	N	6.412 (83.2/85.0)	N

Table II shows SNR estimation results in terms of MAE and PCC/SRC (shown in parentheses) evaluated on 18 individual features for the LSTM-based and BLSTM-based models. The features in the table are listed in the order of the MAE value from low to high for the LSTM-based model. For LSTM-based model, the three best features are MRCG, MAG, and MRCG-causal. MRCG is better than MRCG-causal on MAE, PCC and SRC by 0.1 dB, 0.2 and 0.3, respectively, which indicates that, for the LSTM-based model, future information is helpful for SNR estimation. Although MRCG achieves the best performance, it is a non-causal feature, unable to estimate frame-level SNR in real time. It is not surprising that the performances of MRCG and MRCG-causal are better than GF, which is used to build MRCG and MRCG-causal with additional contextual information. It is worth noting that the simple MAG feature is a pretty well causal feature. A log version of MAG is evaluated in [6] for speech separation and it is just slightly worse than LOG-MEL.

For the BLSTM-based algorithm, MAE and PCC/SRC results are better than those of the LSTM-based algorithm, to be expected as BLSTM captures both past and future information. The three best features are GF, MRCG, and MRCG-causal. The performances of MRCG and MRCG-causal are almost identical, different from the results in LSTM-based model. The likely reason is that BLSTM can make full use of future information, hence filling the gap between these two features. Interestingly, the GF feature performs better than MRCG and MRCG-causal, probably because contextual information extracted in MRCG and MRCG-causal is not as effective as learned by BLSTM.

Table II shows that the noise-robust features of SSF, RAS-MFCC, PNCC, RASTA-PLP, GFMC, and PITCH generally perform worse than other features. These features are designed for robust speech separation or automatic speech recognition (ASR), making them relatively insensitive to noise in a noisy speech signal. However, the sensitivity to noise is important for SNR estimation, as SNR is determined by both speech and noise levels. Thus the noise-robust features may not be suitable

$$PCC = \frac{\sum_{m=1}^{M}\left(SNR(m)-\overline{\mathrm{SNR}}(m)\right)\left(\widehat{SNR}(m)-\overline{\widehat{SNR}}(m)\right)}{\sqrt{\sum_{m=1}^{M}\left(SNR(m)-\overline{\mathrm{SNR}}(m)\right)^2}\sqrt{\sum_{m=1}^{M}\left(\widehat{SNR}(m)-\overline{\widehat{\mathrm{SNR}}}(m)\right)^2}} \tag{11}$$

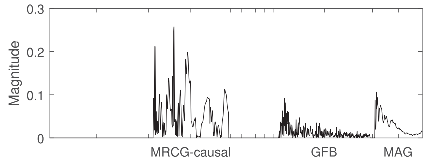

Fig. 2. Magnitude response of causal features under Group Lasso.

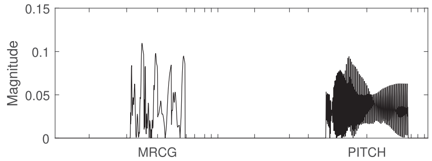

Fig. 3. Magnitude response of all features, except for MRCG-causal, under Group Lasso.

for SNR estimation. We point out that the WAV feature, the raw waveform input, performs quite well and achieves the best results in park and traffic noises for the LSTM-based model. But it is an ineffective feature for speech separation [6], consistent with our explanation regarding noise robustness.

B. Feature Combinations

One feature uncovers certain characteristics of noisy speech. For example, MRCG is designed for speech separation, and PNCC and SSF are designed for robust ASR. A proven way for further improvement is to combine features that complement each other. We evaluate feature combinations to boost SNR estimation performance. Here, two feature combination methods are employed. The first is Group Lasso [44], which has been used to find complementary features for speech separation [6], [2], [40]. The second is SFFS [32], which also has been used for feature selection in speech separation [6].

For the LSTM-based model, we only examine causal features to ensure that the algorithm can work in real time. So, the features are WAV, MFCC, LOG-MEL, MRCG-causal, PLP, RASTA-PLP, RAS-MFCC, GF, GFCC, GFMC, AC-MFCC, GFB, AMS, and MAG. For the BLSTM-based model, all features except MRCG-causal are investigated.

1) Group Lasso: The Group Lasso objective function is defined as:

$$\boldsymbol{\alpha}_{\lambda} = \arg\min_{\boldsymbol{\alpha}} \left\| \sum_{k=1}^{K} \mathbf{X}_k \boldsymbol{\alpha}_k - \mathbf{y} \right\|_2^2 + \lambda \sum_{k=1}^{K} \|\boldsymbol{\alpha}_k\|_2 \tag{12}$$

$$\boldsymbol{\alpha} = \left[\boldsymbol{\alpha}_1^T, \boldsymbol{\alpha}_2^T, \ldots, \boldsymbol{\alpha}_K^T\right]^T \tag{13}$$

where $\|\cdot\|_2$ denotes the Euclidean norm, and $\mathbf{y}$, a vector of N frames, is the desired response indicating target SNR in this paper. K is the number of the features. $\mathbf{X}_k$, a $N \times D_k$ matrix, indicates the k-th feature, where D_k indicates the feature length. $\boldsymbol{\alpha}_k$, a D_k-dimensional vector, indicates feature coefficients. λ is a parameter to control sparsity in groups of coefficients.

To do feature selection, all the features (14 types of features for causal SNR estimation and 17 types for non-causal SNR estimation) are concatenated together to form a long feature vector, and each feature type is defined as a group. Then for a fixed λ we can get $\boldsymbol{\alpha}_\lambda$ through Eq. (12) on the validation set. $\boldsymbol{\alpha}_k$ with small or zero values means that the feature $\mathbf{X}_k$ contributes little to the SNR estimation in the presence of the other groups. Features shall be selected if the magnitudes of the feature coefficients are greater than zero.

Fig. 2 and Fig. 3 show the magnitudes of average Group Lasso coefficients, where $\lambda = 0.2$ is used. In Fig. 2, 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 of Group Lasso for the LSTM-based model. In Fig. 3, MRCG and PITCH are the only features with significant responses. Hence, MRCG+PITCH are used as the complementary feature set of Group Lasso for the BLSTM-based model. It appears that Group Lasso favors higher-dimensional features (see Table I).

Algorithm 1: SFFS Algorithm.

```
Input: $\mathbf{Y} = \{\mathbf{X}_k | k = 1, 2, 3, ..., K\}$
    // Set of all features used in SFFS
Output: $\mathbf{O}_j = \{\mathbf{Z}_i | i = 1, 2, ..., j\}, \mathbf{Z}_i \in \mathbf{Y}$,
        $j = 1, 2, ...K$
    // Set of selected features
1  $j \leftarrow 0, \mathbf{O_j} \leftarrow \emptyset$
2  $\mathbf{X}^+ \leftarrow \arg\min_{\mathbf{X} \in (\mathbf{Y} - \mathbf{O}_j)} J(\mathbf{O}_j + \mathbf{X})$
3  if $J(\mathbf{O}_j + \mathbf{X}^+) < J(\mathbf{O}_j)$ then
4      $\mathbf{O}_{j+1} \leftarrow \mathbf{O}_j + \mathbf{X}^+$
5      $j \leftarrow j + 1$
6  else
7      close;
8  $\mathbf{X}^- \leftarrow \arg\min_{\mathbf{X} \in (\mathbf{O}_j)} J(\mathbf{O}_j - \mathbf{X})$
9  if $J(\mathbf{O}_j - \mathbf{X}^-) < J(\mathbf{O}_j)$ then
10     $\mathbf{O}_{j-1} \leftarrow \mathbf{O}_j - \mathbf{X}^-$
11     $j \leftarrow j - 1$
12     goto 8.
13 else
14     goto 2.
```

2) SFFS: The SFFS algorithm [32] starts with an empty set and systematically adds and drops features until a desired number of features is selected. Because the accuracy number of the features is unknown, we adopt the modified version proposed in [6], where the algorithm will stop when no improvement is achieved by adding more features. The SFFS algorithm is shown in Algorithm 1, where $J(\mathbf{O}_j)$ denotes the MAE performance on the entire validation set and $\mathbf{O}_j$ indicates the input feature set.

Fig. 4 and Fig. 5 show the state of the selected features in each step of the SFFS algorithm with the LSTM-based and BLSTM-based models, respectively. 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 features selected are GF, LOG-MEL, AMS, and PLP.

TABLE III
FRAME-LEVEL SNR ESTIMATION RESULTS IN MAE AND PCC/SRC FOR FEATURE COMBINATIONS WITH THE LSTM-BASED MODEL

Method	**Noise**						**Avg.**
	babble	**cafeteria**	**park**	**traffic**	**factory**	**SSN**	
MAG	3.98 (88.9/89,8)	3.80 (90.5/92.1)	2.73 (95.3/96.0)	2.46 (96.0/96.6)	3.77 (90.7/91.5)	2.96 (94.1/94.6)	3.28 (92.6/93.4)
Group Lasso	3.87 (89.4/90.5)	3.46 **(92.2/93.7)**	2.46 (96.2/**96.9**)	2.18 (97.0/97.4)	3.42 (91.6/92.6)	2.71 (94.7/95.2)	3.02 (93.5/94.4)
SFFS	**3.66 (89.9/90.9)**	**3.37** (92.0/93.6)	**2.44 (96.3/96.9)**	**2.15 (97.1/97.5)**	**3.41 (92.1/93.0)**	**2.41 (95.4/95.9)**	**2.90 (93.8/94.6)**

TABLE IV
FRAME-LEVEL SNR ESTIMATION RESULTS IN MAE AND PCC/SRC FOR FEATURE COMBINATIONS WITH THE BLSTM-BASED MODEL

Method	**Noise**						**Avg.**
	babble	**cafeteria**	**park**	**traffic**	**factory**	**SSN**	
GF	3.04 (93.1/93.6)	2.85 (94.1/95.4)	2.61 (96.8/97.6)	1.92 (**97.8**/98.0)	2.79 **(94.7/95.3)**	**2.15 (96.6/97.1)**	2.56 (95.5/96.2)
Group Lasso	6.83 (78.7/79.5)	4.28 (89.3/89.8)	3.23 (91.5/92.3)	2.47 (95.3/95.8)	3.51 (90.7/91.2)	2.82 (93.4/94.3)	3.86 (89.8/90.7)
SFFS	**2.90 (93.4/94.1)**	**2.81 (94.3/95.4)**	**2.03 (97.2/97.7)**	**1.82 (97.8/98.2)**	**2.75** (94.6/95.3)	**2.15** (96.5/97.0)	**2.41 (95.7/96.3)**

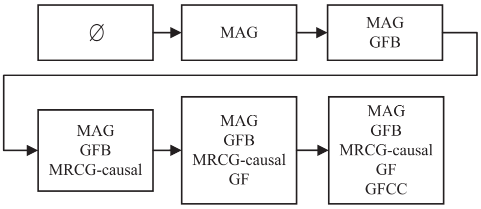

Fig. 4. Steps in the LSTM-based model with SFFS.

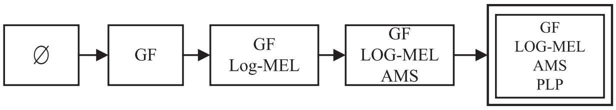

Fig. 5. Steps in the BLSTM-based model with SFFS.

3) Feature Combination Results: We compare the performances of complementary features with the best single feature. It should be noted that since MRCG is non-causal, we designate MAG as the best single feature associated with the LSTM-based model. The results of the LSTM-based and BLSTM-based models are shown in Tables III and IV, respectively. In both algorithms, the feature set obtained by the SFFS algorithm achieves the best performances in all noise conditions.

For the LSTM-based model, the feature set selected by Group Lasso is MAG+GFB+MRCG-causal, and the SFFS feature set is MAG+GFB+MRCG-causal+GF+GFCC. The SFFS selected set has two more features (i.e. GF and GFCC) than the Group Lasso selected set. By adding these two features, the average MAE is reduced by 0.12 dB. The biggest improvement occurs for SSN noise, where MAE is reduced by 0.3 dB. On average, the SFFS feature set reduces MAE by 0.38 dB over MAG, or about 12% relative improvement.

In Table IV, the MAE performance of the SFFS feature set is 0.15 dB better than the GF feature. The set selected by Group Lasso (MRCG+PITCH) is worse than the GF feature; the average MAE with MRCG is 2.63 dB (Table II), and after combining with PITCH, it increases to 3.86 dB. The reason is that pitch is difficult to track, especially for babble noise which combines many speech utterances, and the inaccurate PITCH feature drags the performance of the feature set. As Group Lasso is a linear regression algorithm, it does not seem to handle the nonlinear relationship between input features and the target SNR well.

VI. SNR ESTIMATION RESULTS

In this section, the evaluation and comparison results are first presented for frame-level SNR estimation, and then utterance-level SNR estimation. The generalization of the proposed methods in unseen corpus is also explored.

A. Frame-Level SNR Evaluation

Table V shows the frame-level MAE and PCC/SRC results under different noise conditions for the proposed models as well as four comparison baselines. For the LSTM- and BLSTM-based algorithms the input features are MAG+GFB+MRCG-causal+GF+GFCC and GF+LOG-MEL+AMS+PLP (see Section V-B), respectively.

The proposed algorithms obtain the best performances on most conditions in both causal and non-causal settings. For the causal system, the average MAE value of the LSTM-based algorithm is 2.9 dB, which is 3.8 dB and 0.25 dB better than the PSD and LSTM-SE+ baselines, respectively, representing 56.7% and 7.9% relative improvement. For the non-causal system, the average MAE value of the BLSTM-based algorithm is 2.41 dB, which is 0.37 better than the BLSTM-SE+ baseline. For PCC and SRC, the deep-learning based methods achieve comparable results, which are significantly better than PSD. When a background noise is non-stationary or SNR is low, the PSD baseline makes large estimation errors.

Fig. 6 plots the frame-level SNR estimation results in MAE for different deep learning based methods with respect to frame-level SNR. The proposed BLSTM-SE+ baseline obtains the best performance in the range of [-20 dB, 0 dB], and the proposed BLSTM-based algorithm achieves the best results in other ranges. The performances of the SE and the BLSTM-SE+ baselines are close in low SNRs (<5 dB). As SNR increases, the performance gap becomes larger, likely because the SNR is sensitive to the estimated noise, especially at high SNRs. Unlike the BLSTM-SE+ baseline, the SE baseline calculates a noise estimate from the predicted speech. On the other hand, the algorithms in Fig. 6 share a trend: when the SNR is less than 5 dB, MAE increases with the decrease of SNR; when SNR is greater than 10 dB, MAE increases with the increase of SNR. In low SNR conditions, speech is difficult to estimate, and in high SNR conditions, noise is difficult to estimate. Around 5-10 dB is where all SNR estimators achieve the best performance.

TABLE V
FRAME-LEVEL SNR ESTIMATION RESULTS IN MAE AND PCC/SRC UNDER DIFFERENT NOISE CONDITIONS

Method	Noise						Avg.	Causal
	babble	cafeteria	park	traffic	factory	SSN		
PSD	8.70 (70.9/68.4)	6.39 (81.4/81.7)	7.96 (78.5/80.0)	5.11 (87.3/88.5)	6.56 (79.9/78.6)	5.41 (87.7/86.7)	6.70 (81.0/80.7)	Y
LSTM-SE+	3.90 (**90.5/91.1**)	3.73 (**92.3**/93.4)	2.77 (95.9/96.8)	2.29 (96.8/97.4)	3.57 (**93.0/93.6**)	2.64 (**95.8/96.4**)	3.15 (**94.0/94.8**)	Y
LSTM-based	**3.66** (89.9/90.9)	**3.37** (92.0/**93.6**)	**2.44** (**96.3/96.9**)	**2.15** (**97.1/97.5**)	**3.41** (92.1/93.0)	**2.41** (95.4/95.9)	**2.90** (93.8/94.6)	Y
SE	3.50 (91.9/93.6)	3.39 (93.6/95.4)	2.37 (96.7/97.6)	2.31 (97.1/98.0)	3.22 (94.1/95.3)	2.73 (95.9/**97.1**)	2.92 (94.9/96.2)	N
BLSTM-SE+	3.46 (92.5/93.7)	3.27 (93.9/**95.5**)	2.25 (96.9/**97.7**)	2.10 (97.4/98.1)	3.09 (94.2/**95.4**)	2.52 (95.6/97.0)	2.78 (95.1/96.2)	N
BLSTM-based	**2.90** (**93.4/94.1**)	**2.81** (**94.3**/95.4)	**2.03** (**97.2/97.7**)	**1.82** (**97.8/98.2**)	**2.75** (**94.6**/95.3)	**2.15** (**96.5**/97.0)	**2.41** (**95.7/96.3**)	N

TABLE VI
FRAME-LEVEL SNR ESTIMATION RESULTS IN MAE AND PCC/SRC UNDER DIFFERENT NOISE CONDITIONS

Method	Noise						Avg.	Causal
	babble	cafeteria	park	traffic	factory	SSN		
LSTM-SFFS-SE+	3.89 (90.8/91.6)	3.46 (93.3/94.5)	2.27 (96.8/97.4)	2.13 (97.2/97.8)	3.40 (93.3/93.9)	2.70 (95.7/96.2)	2.98 (94.5/95.2)	Y
BLSTM-SFFS-SE	3.51 (92.8/94.4)	3.35 (94.1/95.8)	2.24 (96.9/97.8)	2.32 (97.1/98.1)	3.27 (94.4/95.6)	2.75 (95.9/97.3)	2.91 (95.2/96.5)	N
BLSTM-SFFS-SE+	3.48 (93.1/94.2)	3.18 (94.6/95.8)	2.10 (97.2/97.9)	2.13 (97.4/98.1)	3.08 (94.7/95.6)	2.61 (96.4/97.3)	2.76 (95.6/96.5)	N

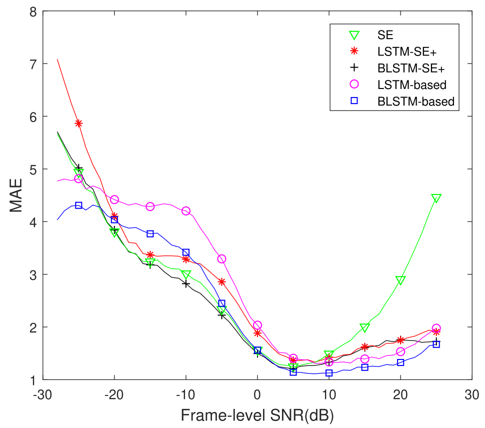

Fig. 6. Frame-level SNR estimation results in MAE for different methods with respect to frame-level SNR.

Furthermore, we use the feature set selected by SFFS as the input of the SE and SE+ baselines for additional comparisons. In the SE method, the BLSTM and the SFFS feature set are used to estimate the speech IRM (referred to as BLSTM-SFFS-SE). For SE+, the MRCG feature is replaced by the SFFS feature set to generate another two baselines: LSTM-SFFS-SE+ and BLSTM-SFFS-SE+. The results are shown in Table VI. Comparing with the results in Table V, we find that the SFFS feature set produces slightly better results than the MRCG feature. Directly estimating SNR achieves lower estimation errors than indirect SNR estimation by first estimating the IRM, although PCC/SRC results are similar. This is likely because MAE is used to measure training loss in direct estimation.

B. Utterance-Level SNR Estimation

In this part, we present utterance-level SNR estimation results for the proposed BLSTM-based algorithm and six comparison baselines. The results of the proposed method are calculated using the feature set of GF+LOG-MEG+AMS+PLP (see Section V-B). It should be noted that, for the comparison with Papadopoulos *et al.* [29], we use their test set in order to avoid SNR estimation inaccuracies caused by our implementation of their method.

TABLE VII
UTTERANCE-LEVEL SNR ESTIMATION RESULTS IN MAE AND PCC/SRC ACROSS NOISES

Method	Input SNR						Avg.	PCC/SRC
	-10	-5	0	5	10	15		
WADA	5.78	2.41	0.97	0.95	1.11	1.76	2.16	93.31/95.44
CASA	2.44	1.29	0.78	0.91	1.39	2.12	1.49	97.77/97.72
Residual-based	2.47	1.37	2.42	2.49	2.08	2.63	2.25	95.59/94.97
SE	0.59	0.36	0.28	0.45	0.80	1.56	0.67	99.68/98.57
BLSTM-SE+	**0.58**	**0.30**	0.24	0.29	0.37	0.48	0.38	99.78/98.57
BLSTM-SFFS-SE+	0.60	0.31	0.19	0.25	0.33	0.46	0.36	99.79/**98.59**
BLSTM-based	**0.58**	0.34	**0.15**	**0.13**	**0.15**	**0.29**	**0.27**	**99.82/98.59**

TABLE VIII
UTTERANCE-LEVEL SNR ESTIMATION RESULTS IN MAE FOR PAPADOPOULOS *et al.* AND PROPOSED METHODS UNDER EIGHT NOISES

Method Noise	Papadopoulos *et al.*	Proposed	Proposed-DNN
KITCHEN	2.38	**0.22**	0.75
LIV.ROOM	1.36	**0.27**	0.96
METRO	2.90	**0.27**	0.90
PARK	2.12	**0.17**	0.83
STATION	1.14	**0.17**	1.08
TRAFFIC	1.93	**0.11**	0.96
RESTAURANT	1.92	**0.53**	1.64
CAFE	1.11	**0.35**	0.87
Avg.	1.86	**0.26**	0.99

Utterance-level SNR estimation results are shown in Table VII and VIII. Table VII shows the average results across all test noises. To obtain PCC/SRC statistics at the utterance level, Eq. (11) is calculated over the entire evaluation set with 6 input SNRs. Overall the proposed algorithm achieves the best results across all SNR conditions. The second best algorithm is BLSTM-SFFS-SE+. The average MAE of the proposed algorithm is 0.27 dB, about 0.09 dB better than the BLSTM-SFFS-SE+ which represents 25% relative improvement. Looking closer at each SNR condition, the MAEs of the SE+ and the proposed algorithm are close at low SNRs. As the SNR increases, the MAE performance gap becomes larger, which is consistent

TABLE IX
UTTERANCE-LEVEL SNR ESTIMATION RESULTS IN MAE AND PCC/SNR FOR REVISED VERSIONS OF THE RESIDUAL-BASED METHOD

Method	**Input SNR**						**Avg.**	**PCC/SRC**
	-10	**-5**	**0**	**5**	**10**	**15**		
Residual-BLSTM	2.56	1.81	1.55	1.11	2.35	2.23	1.94	96.80/95.44
Residual-SFFS	3.07	2.33	2.55	1.53	2.32	3.68	2.58	93.32/91.19
Residual-BLSTM/SFFS	2.50	2.27	0.91	0.86	2.14	2.07	1.79	96.48/97.02

with the trends shown in Fig. 6. The proposed algorithm obtains the best performance in the range of 0−15 dB; for example, at 5 dB MAE is 0.13 dB. The CASA algorithm depends on whether noisy T-F units can be accurately classified, and it performs relatively poorly in low or high 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, yielding poor results. The Residual-based method shows difficulty to estimate utterance-level SNR using noise residuals alone. Table VIII shows that the proposed algorithm is substantially better than the Papadopoulos $et\ al.$ algorithm, which uses several energy ratio features and i-vectors extracted from a noisy utterance as the inputs of DNN for utterance-level SNR estimation. On average, the MAE of the proposed algorithm is 1.6 dB better than the Papadopoulos $et\ al.$ method. The results indicate that the energy ratio features extracted from noisy speech are insufficient for predicting utterance-level SNR accurately.

For further comparisons, we revise the Residual-based model in the following ways for a deeper examination: (1) Replace its DNN with the BLSTM used in our methods; (2) Replace its residual-based feature with the feature set selected by SFFS; (3) Perform both (1) and (2). These three revised methods are referred to as Residual-BLSTM, Residual-SFFS, and Residual-BLSTM/SFFS, respectively. The results are shown in Table IX. Residual-BLSTM/SFFS achieves the best results. Although they are better than the original Residual-based method, they are still substantially worse than the proposed method, demonstrating that our overall algorithm is responsible for its superior performance.

We also replace the BLSTM used in our model with the DNN used in Papadopoulos $et\ al.$ so that the two methods use the same neural network, but different acoustic features. This variant of our method is denoted by "Proposed-DNN," and its results are shown in Table VIII. The results in the last column of the table demonstrate that the BLSTM used in our model outperforms the DNN used by Papadopoulos $et\ al.$

C. Evaluation on Untrained Corpus

In the above experiments, the speakers and speech signals in the test set are not included in the training set, but the training and test utterances belong to the same corpus (WSJ0 SI-84) with similar recording conditions. In this section, we explore the robustness of supervised SNR estimation methods on an untrained corpus. The test set is generated in the same way except that 150 utterances are from the TIMIT corpus [11] rather than WSJ0 SI-84.

TABLE X
FRAME-LEVEL SNR ESTIMATION RESULTS IN MAE AND PCC/SRC ON TIMIT CORPUS

Method	**MAE (PCC/SRC)**	**Causal**
PSD	6.78 (81.0/80.9)	Y
SE	4.30 (89.3/91.8)	N
BLSTM-SFFS-SE	3.99 (91.0/**93.2)**	N
LSTM-SE+	5.61 (81.3/81.7)	Y
BLSTM-SE+	4.14 (89.5/91.5)	N
BLSTM-SFFS-SE+	3.78 (91.2/**93.2)**	N
LSTM-based	**5.24 (83.4/83.7)**	Y
BLSTM-based	**3.49 (91.3**/92.5)	N

TABLE XI
UTTERANCE-LEVEL SNR ESTIMATION RESULTS IN MAE AND PCC/SRC ON TIMIT CORPUS

Method	**Input SNR**						**Avg.**	**PCC/SRC**
	-10	**-5**	**0**	**5**	**10**	**15**		
WADA	5.05	2.51	1.22	0.95	0.94	1.51	2.03	94.59/96.76
CASA	2.70	1.25	0.63	0.66	0.76	0.93	1.15	**97.59**/97.71
Residual-based	3.30	1.43	2.82	2.81	2.67	2.02	2.51	93.10/93.18
SE	**2.34**	1.06	0.86	1.20	1.56	2.35	1.56	95.45/97.98
BLSTM-SE+	3.12	1.06	0.62	0.74	0.77	0.84	1.19	95.29/97.76
BLSTM-based	2.42	**0.75**	**0.36**	**0.27**	**0.37**	**0.46**	**0.77**	96.50/**98.38**

Frame-level and utterance-level SNR estimation results averaged across all noises are shown in Tables X and XI. Compared with Tables V and VII, the performance of deep learning based algorithms is decreased. On the other hand, the proposed methods still obtain the best results at most SNR conditions. These evaluation results are consistent with a recent observation that deep learning based methods have difficulty to generalize to new corpora, mainly due to channel mismatch [28].

VII. CONCLUDING REMARKS

In this paper, we have addressed frame-level SNR estimation, where recurrent neural networks with LSTM and BLSTM are trained to perform the estimation. Our algorithm shows substantial improvements over baseline models.

This study has examined a range of acoustic features for their effectiveness for SNR estimation. The best single features in the LSTM-based and BLSTM-based models are different, while MRCG is the best for the LSTM-based model, and GF performs the best for the BLSTM-based model. As SNR must be sensitive to noise, noise robust features tend not to work well for SNR estimation. On the other hand, relatively raw features, such as GF, WAV and MAG, perform well. We have also found that feature combinations significantly boost SNR estimation performance; the best feature combination for the LSTM-based model is MAG+GFB+MRCG-causal+GF+GFCC, and for the BLSTM-based model is GF+LOG-MEL+AMS+PLP.

In addition, we have extended the frame-level estimation to utterance-level SNR estimation, and the proposed method outperforms other utterance-level SNR estimation methods.

Future research will investigate cross-corpus generalization, and real-time, light-weight, robust SNR estimation that can be deployed in practical applications.

Acknowledgment

The authors would like to thank P. Papadopoulos for helpful discussions regarding [29] and X. Dong for discussions and sharing the software used in [7].

References

[1] O. Cappé, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Speech Audio Process.*, vol. 2, no. 2, pp. 345–349, Apr. 1994.

[2] J. Chen, Y. Wang, and D. L. Wang, "A feature study for classification-based speech separation at low signal-to-noise ratios," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1993–2002, Dec. 2014.

[3] J. Chen and D. Wang, "Long short-term memory for speaker generalization in supervised speech separation," *J. Acoustical Soc. Amer.*, vol. 141, no. 6, pp. 4705–4714, 2017.

[4] D. V. Compernolle, "Noise adaptation in a hidden Markov model speech recognition system," *Comput. Speech Lang.*, vol. 3, no. 2, pp. 151–167, 1989.

[5] T. H. Dat, K. Takeda, and F. Itakura, "On-line Gaussian mixture modeling in the log-power domain for signal-to-noise ratio estimation and speech enhancement," *Speech Commun.*, vol. 48, no. 11, pp. 1515–1527, 2006.

[6] M. Delfarah and D. L. Wang, "Features for masking-based monaural speech separation in reverberant conditions," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 5, pp. 1085–1094, May 2017.

[7] X. Dong and D. S. Williamson, "Long-term SNR estimation using noise residuals and a two-stage deep-learning framework," in *Proc. Int. Conf. Latent Variable Anal. Signal Separation*, 2018, pp. 351–360.

[8] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

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

[10] T. Gao, J. Du, L. Dai, and C. Lee, "SNR-based progressive learning of deep neural network for speech enhancement," in *Proc. Interspeech*, 2016, pp. 3713–3717.

[11] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM," NIST Internal Report 4930, 1993.

[12] S. Gonzalez and M. Brookes, "PEFAC-A pitch estimation algorithm robust to high levels of noise," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 2, pp. 518–530, Feb. 2014.

[13] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoustical Soc. Amer.*, vol. 87, no. 4, pp. 1738–1752, 1990.

[14] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Speech Audio Process.*, vol. 2, no. 4, pp. 578–589, Oct. 1994.

[15] H. G. Hirsch, "Estimation of noise spectrum and its application to SNR-estimation and speech enhancement," *Int. Comput. Sci. Inst.*, Berkeley, CA, USA, Tech. Rep. TR-93-012, 1993.

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

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

[18] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 7, pp. 1315–1329, Jul. 2016.

[19] G. Kim, Y. Lu, Y. Hu, and P. C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *J. Acoustical Soc. Amer.*, vol. 126, no. 3, pp. 1486–1494, 2009.

[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Int. Conf. Learning Representations*, 2015, *arXiv:1412.6980*.

[21] A. Korthauer, "Robust estimation of the SNR of noisy speech signals for the quality evaluation of speech databases," in *Proc. Workshop Robust Methods Speech Recognit. Adverse Condition*, Tampere, Finland, 1999, pp. 123–126.

[22] H. Li, D. L. Wang, X. Zhang, and G. Gao, "Frame-level signal-to-noise ratio estimation using deep learning," in *Proc. Interspeech*, 2020, pp. 4626–4630.

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

[24] T. May, B. Kowalewski, M. Fereczkowski, and E. N. MacDonald, "Assessment of broadband SNR estimation for hearing aid applications," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 231–235.

[25] A. Narayanan and D. L. Wang, "A CASA-based system for long-term SNR estimation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 9, pp. 2518–2527, Nov. 2012.

[26] E. Nemer, R. Goubran, and S. Mahmoud, "SNR estimation of speech signals using subbands and fourth-order statistics," *IEEE Signal Process. Lett.*, vol. 6, no. 7, pp. 171–174, Jul. 1999.

[27] NIST, "NIST speech signal to noise ratio measurements," Accessed: Oct. 25, 2020. 1994. [Online]. Available: https://www.nist.gov/itl/iad/mig/nist-speech-signal-noise-ratio-measurements

[28] A. Pandey and D. Wang, "On cross-corpus generalization of deep learning based speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2489–2499, 2020, doi: 10.1109/TASLP.2020.3016487.

[29] 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 *Proc. Interspeech*, 2017, pp. 3842–3846.

[30] P. Papadopoulos, A. Tsiartas, and S. Narayanan, "Long-term SNR estimation of speech signals in known and unknown channel conditions," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 12, pp. 2495–2506, 2016.

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

[32] P. Pudil, F. J. Ferri, J. Novovicova, and J. Kittler, "Floating search methods for feature selection with nonmonotonic criterion functions," in *Proc. IAPR Int. Conf. Pattern Recognit.*, 1994, pp. 279–283.

[33] M. R. Schädler, B. T. Meyer, and B. Kollmeier, "Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition," *J. Acoustical Soc. Amer.*, vol. 131, no. 5, pp. 4134–4151, 2012.

[34] B. J. Shannon and K. K. Paliwal, "Feature extraction from higher-lag autocorrelation coefficients for robust speech recognition," *Speech Commun.*, vol. 48, no. 11, pp. 1458–1485, 2006.

[35] S. Suhadi, C. Last, and T. Fingscheidt, "A data-driven approach to a priori SNR estimation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 1, pp. 186–195, Jan. 2010.

[36] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multichannel acoustic noise database (DEMAND): A database of multichannel environmental noise recordings," in *Proc. Int. Congr. Acoust.*, pp. 1–6, 2013.

[37] 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 Commun.*, vol. 12, no. 3, pp. 247–251, 1993.

[38] D. L. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, Oct. 2018.

[39] Y. Wang, K. Han, and D. L. Wang, "Exploring monaural features for classification-based speech segregation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 2, pp. 270–279, Feb. 2013.

[40] Y. Wang, A. Narayanan, and D. L. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1849–1858, Dec. 2014.

[41] F. Weninger *et al.*, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *Proc. Int. Conf. Latent Variable Anal. Signal Separation*, Springer, 2015, pp. 91–99.

[42] F. Weninger, F. Eyben, and B. Schuller, "Single-channel speech separation with memory-enhanced recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 3709–3713.

[43] Y. Shao and D. L. Wang, "Robust speaker identification using auditory features and computational auditory scene analysis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2008, pp. 1589–1592.

[44] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Roy. Stat. Soc.: Ser. B. (Stat. Methodol.)*, vol. 68, no. 1, pp. 49–67, 2006.

[45] K. H. Yuo and H. C. Wang, "Robust features for noisy speech recognition based on temporal trajectory filtering of short-time autocorrelation sequences," *Speech Commun.*, vol. 28, no. 1, pp. 13–24, 1999.

[46] Y. Zhao, D. L. Wang, B. Xu, and T. Zhang, "Late reverberation suppression using recurrent neural networks with long short-term memory," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5434–5438.