



# Global Signal-to-noise Ratio Estimation Based on Multi-subband Processing Using Convolutional Neural Network

Nan Li<sup>1</sup>, Meng Ge<sup>1,\*</sup>, Longbiao Wang<sup>1,\*</sup>, Masashi Unoki<sup>2</sup>, Sheng Li<sup>3</sup>, Jianwu Dang<sup>1,2</sup>

<sup>1</sup>Tianjin Key Laboratory of Cognitive Computing and Application,  
College of Intelligence and Computing, Tianjin University, Tianjin, China

<sup>2</sup>Japan Advanced Institute of Science and Technology, Ishikawa, Japan

<sup>3</sup>National Institute of Information and Communications Technology (NICT), Kyoto, Japan

{linan.tju, gemeng, longbiao.wang}@tju.edu.cn, sheng.li@nict.go.jp,  
{unoki, jdang}@jaist.ac.jp

## Abstract

The global signal-to-noise ratio (gSNR) is defined as the ratio of speech energy to noise energy in whole noisy audio. However, due to the increase in noise interference, the generalization ability declines when the traditional features (e.g., raw waveforms and MFCCs) are fed directly to the statistical model to estimate a single fullband gSNR. In this paper, we propose a multi-subband-based gSNR estimation network (MSGNet). Specifically, we split the noisy speech waveforms into Bark-scale subbands to obtain higher resolution signals to the middle and low frequencies. Then, convolutional neural networks (CNNs) are used to learn a non-linear function to estimate the speech and noise energy ratio of each subband from the input multi-subband features. Finally, by integrating subbands with different speech and noise energies, gSNR in the fullband is calculated. Extensive experimental results on the AURORA-2J dataset demonstrate that the proposed MSGNet significantly reduces the mean absolute error compared to other baseline gSNR estimation methods.

**Index Terms:** Global signal-to-noise ratio, multi-subband, Bark-scale, non-linear mapping

## 1. Introduction

The signal-to-noise ratio (SNR), as a measure of the noise level in a speech signal, helps guide the design and improvement of many speech applications, such as speech enhancement [1, 2], speech recognition [3], and the speech transmission index (STI) [4]. However, in real-world speech communication, SNR estimation is a challenging task, since speech signals are usually corrupted by various unknown noises [5].

Existing SNR estimation generally falls into two categories, namely local SNR and global SNR (gSNR). Local SNR estimation [1, 6], a frame-level decision, has attracted increasing attention in recent decades because of its direct applications in noise estimation and speech enhancement [2]. gSNR estimation [7, 8] is usually defined at the utterance level. It considers the entire signal and provides information about the effect of noise on the entire observed recording. In this study, we address the problem of gSNR estimation, as many SNR-specific speech applications (e.g., robust speech recognition [3], and speech intelligibility [4]) are strongly affected by the noise level over the entire signal.

Many statistics-based approaches have been proposed for gSNR estimation [9, 10, 11]. Kim et al. [10] proposed a wave-

form amplitude distribution analysis (WADA) approach that uses maximum likelihood estimation to determine the gSNR based on the assumption that the amplitudes of the speech and noise follow Gamma and Gaussian distributions, respectively. Morita et al. [11] designed a single global threshold to detect speech and noise regions under all test conditions using multi-subband voice activity detection (Multi-Sub). These methods are effective under high-SNR conditions but are not robust enough to compute the noise energy under low-SNR cases because of the statistical assumptions about the signal distribution and single global threshold.

Recently, data-driven gSNR estimation based on deep learning (DL), including deep neural networks (DNNs) [12, 13] and long short-term memory networks (LSTMs) [14] have attracted attention as a means of eliminating the errors caused by statistical assumptions. The key idea is to view gSNR estimation as a regression problem that maps noisy speech features to the fullband local SNR values and then calculates the gSNR. Benefiting from the non-linear mapping ability of neural networks, the energy content of speech and noise regions can be accurately identified in the raw waveforms [14, 15]. However, the original fullband waveform is subjected to strong fluctuations caused by noise, this will cause the statistical distributions of speech and noise to overlap because they can have similar statistical properties, which further results in poor gSNR estimation generalization [16, 17] in a real-world environment.

We propose a multi-subband-based gSNR estimation network (MSGNet) in the time domain to address the above problem. Motivated by psychoacoustic studies [18, 19], the noisy signal is first divided into subbands of non-uniform frequency division based on the Bark-scale, giving higher resolution to the middle and low frequencies, which is similar to the human auditory perception that pays more attention to the energy-intensive parts of speech. We then determine the noise level in each subband using an auditory encoder based convolution perception network. Due to the noise imbalance over frequency, existing fullband methods only use one static ratio of noise and speech in the whole audio, while bark-scale subband decomposition can dynamically estimate the ratio of noise and speech in each subband. Finally, by summarizing subbands with different speech and noise energy, the gSNR in fullband is calculated.

## 2. Multi-subband-based global signal-to-noise ratio estimation network

The fullband waveform is widely used to estimate fullband gSNR and other speech applications [14, 15]. The fullband raw

\* Corresponding author.

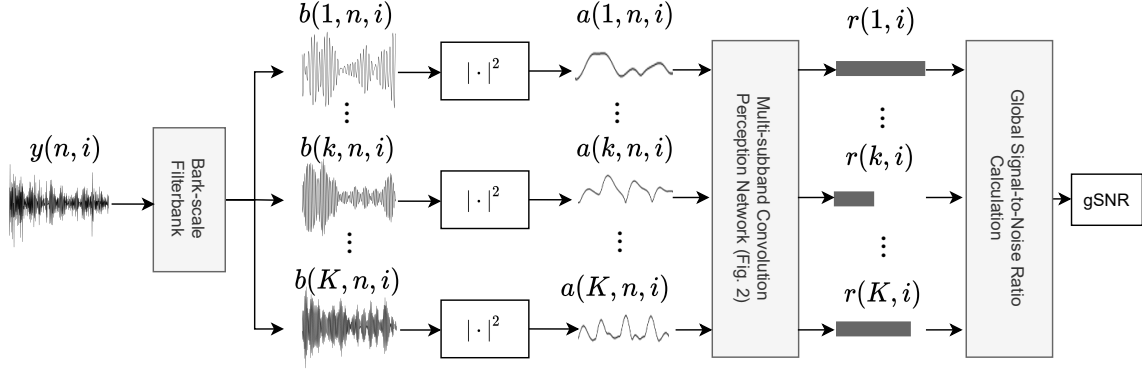


Figure 1: The diagram of the proposed multi-subband-based gSNR estimation network using CNN.

speech waveform, however, is subject to strong fluctuations because of noise in the signal. This causes the statistical distributions of speech and noise to overlap, because they have similar statistical properties. Inspired by the signal processing of the human cochlea and auditory mechanism, we propose the MS-GNet approach for gSNR estimation. Fig. 1 shows a diagram of the proposed method, containing a Bark-scale filterbank-based front-end processing, multi-subband convolution perception network, and multi-subband gSNR calculation.

### 2.1. Front-end processing

In this paper, we first use the Bark-scale filterbank to split noisy speech into different subbands. This Bark-scale filterbank approximates the process of sound perception in the human cochlea by designing several high-pass filters (HPF) and low-pass filters (LPF). The output of the  $k$ -th subband signal can be expressed as

$$b(k, n, i) = \text{HPF}_{\mathbf{F}(k-1)}(\text{LPF}_{\mathbf{F}(k)}(y(n, i))) \quad (2 \leq k \leq 16), \quad (1)$$

where  $y(n, i)$  is the  $i$ -th frame of noisy speech and  $n$  is the sample index. Here, we set the boundary of the filterbank to  $\mathbf{F}=[200 \ 300 \ 400 \ 510 \ 630 \ 770 \ 920 \ 1080 \ 1270 \ 1480 \ 1720 \ 2000 \ 2320 \ 2700 \ 3150 \ 3700]$  Hz. In the last filter, there is a 3700-Hz HPF; the first filter has a 200-Hz LPF. Furthermore, to effectively simulate the mechanical-to-neural signal transduction from the signal to the inner hair cells, the modulated signal is calculated [18, 20]. We simulate the modulated signal by computing the signal energy of different subbands. The  $k$ -th subband modulation power envelope is

$$a(k, n, i) = |b(k, n, i)|^2. \quad (2)$$

Finally, the auditory front can be denoted as

$$\mathbf{a} = [a(1, n, i), \dots, a(k, n, i), \dots, a(K, n, i)] \in \mathbb{R}^{B \times F \times D \times K}, \quad (3)$$

where  $B$  is the batch size for parallel calculations,  $F$  is the number of context frames,  $D$  is the frame length, and  $K$  is the subband number which is 17 in this paper.

### 2.2. Global signal-to-noise ratio calculation

The gSNR is defined as

$$\text{gSNR} = 10 \cdot \log_{10} \left( \frac{P_S}{P_N} \right), \quad (4)$$

where  $P_S$  and  $P_N$  represent all the speech and noise energy in a noisy speech signal, respectively. By dividing the signal

into multi-subbands, we further define the multi-subband-based gSNR as

$$\text{gSNR} = 10 \cdot \log_{10} \left( \frac{\sum_{k=1}^K P_S(k)}{\sum_{k=1}^K P_N(k)} \right), \quad (5)$$

where  $P_N(k)$  and  $P_S(k)$  are the total powers of additive noise and clean speech in the  $k$ -th sub-band, respectively. These quantities can be approximated through the following calculations

$$P_N(k) = \frac{\sum_{i=1}^L E_N(k, i)}{L_N} \cdot L(\hat{r}(k, i) > C), \quad (6)$$

$$P_S(k) = \sum_{i=1}^L E_T(k, i) - P_N(k), \quad (7)$$

where  $L_N$  is the frame number of  $E_N$  in which the estimated subband noise ratio  $\hat{r}(k, i)$  is greater than the hyperparameter  $C$ .  $L$  is the total frame number of utterance. We assume that  $E_T(k, i)$  can be approximately represented by

$$E_T(k, i) = E_S(k, i) + E_N(k, i), \quad (8)$$

where  $E_T(k, i)$  is the total energy of the  $i$ -th frame in the  $k$ -th subband of a long noisy speech utterance

$$E_T(k, i) = \sum_{n=1}^N |b(k, n, i)|^2, \quad (9)$$

where  $N$  is the frame length. The speech energy  $E_S(k, i)$  and the noise energy  $E_N(k, i)$  of the  $i$ -th frame in the  $k$ -th subband can be further represented by

$$\hat{E}_N(k, i) = \hat{r}(k, i) \cdot E_T(k, i), \quad (10)$$

$$\hat{E}_S(k, i) = (1 - \hat{r}(k, i)) \cdot E_T(k, i), \quad (11)$$

where  $\hat{r}(k, i)$  is the estimated noise ratio of the  $i$ -th frame in the  $k$ -th subband. In this paper, we propose convolution perception network to estimate these values.

### 2.3. Multi-subband convolution perception network

As shown in Fig. 2, our basic framework is based on a 2D CNN layer [21, 22], which provides an excellent approach for learning the temporal relationship of noisy speech contexts. However, simply applying the original noisy auditory front as the CNN encoder does not lead to a sufficiently accurate gSNR estimation. To adapt to the input of the neural network, we further propose a parallel CNN-based auditory encoder.

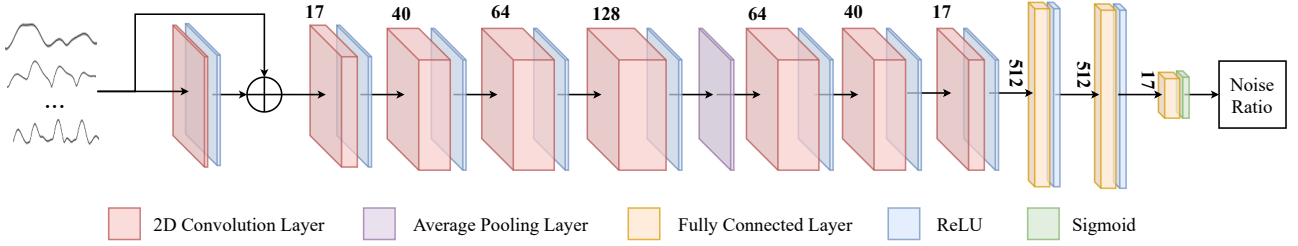


Figure 2: The diagram of the convolution perception network for subband noise ratio estimation.

### 2.3.1. Auditory encoder

Before  $\mathbf{a}$  is input to the auditory encoder, we first divide the auditory front into different channels according to the  $K$  subband dimensions, and concatenate them to the batch  $B$  dimension (batch data will be calculated in parallel), to give  $\mathbf{a} \in \mathbb{R}^{(K \cdot B) \times F \times D \times 1}$ . Next, we use a convolutional layer to map every  $a(k, n, i)$  to a more perceptual space

$$A(k, n, i) = \text{Conv}(a(k, n, i)). \quad (12)$$

After dividing the batch size dimension and recovering the multi-subband signal for  $A$ , we obtain the output of the auditory encoder  $\mathbf{A} \in \mathbb{R}^{B \times F \times D \times K}$  as a mediator between mechanical signals and nerve cells (CNN). Finally, a residual connection is used to avoid information loss

$$\mathbf{A} = \mathbf{A} + \mathbf{a}. \quad (13)$$

### 2.3.2. Convolution perception network

Using a CNN with seven 2D convolutional layers having a  $3 \times 3$  kernel size and 32 filters, we can learn  $\mathbf{A}$  using a high-dimensional encoder vector. The structure of the CNN is [17, 40, 64, 128]. There is an average pooling layer with a kernel size of [1, 1, 3, 1] behind the first four CNN layers for down-sampling to reduce the redundant information. Finally, a CNN with 64 and 40 channels in two layers and a CNN with 17 channels in one layer are used as decoders.

A post-mapping network, which contains two fully connected layers, is used in the multi-subband convolution perception network to estimate the noise more correctly. The network can predict more precise subband noise ratios through deep non-linear operations. Finally, through an output layer with a sigmoid activation function, we obtain the estimated subband noise ratios  $\mathbf{r} = [\hat{r}(1, i), \dots, \hat{r}(k, i), \dots, \hat{r}(K, i)]$ , which range from 0 to 1.

### 2.3.3. Loss function

To estimate the subband noise ratios, the mean squared error (MSE) loss is used

$$\text{MSE}(\mathbf{r}) = \frac{1}{K} \sum_{k=1}^K (r(k, i) - z(k, i))^2, \quad (14)$$

where  $r(k, i)$  is the learned subband noise ratio.  $z(k, i)$  is the true noise ratio of the  $k$ -th subband and  $i$ -th frame

$$z(k, i) = \frac{E_T(k, i)}{E_N(k, i)}, \quad (15)$$

where  $E_T(k, i)$  and  $E_N(k, i)$  represent the total energy and true noise energy in the  $k$ -th subband and  $i$ -th frame, respectively.

In the decoding stage, the auditory front-end of the test signal is input into the trained convolution perception network, and we obtain the final estimated proportion of  $k$ -th and  $i$ -th subband noise ratio  $\hat{r}(k, i)$ .

## 3. Experiments

### 3.1. Dataset descriptions and evaluation metric

In evaluation, we used the AURORA-2J [23] and NOISEX-92 [24] speech and noise data. 8,440 clean speech utterances from AURORA-2J were selected as clean data in the training stage. The noisy speech data was then artificially generated using the White, Pink, Factory, and Babble noises from NOISEX-92. The audio sample rate was resampled to 8 kHz. Noisy speech data with gSNRs of 15, 10, 5, 0, -5, and -10 was then generated. We selected 7,740 utterances as the training set and 700 utterances as the cross-validation set. The tests used the AURORA-2J test set, which contains 1,001 utterances, with noise added at the corresponding gSNRs for each noise type. The noise types for each utterance were the same as those of the training data and were randomly selected.

The evaluation criterion for gSNR is the mean absolute error (MAE) [9, 14] between the estimated gSNR ( $\text{gSNR}_i$ ) and the real gSNR ( $\text{gSNR}_i$ )

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N \left| \text{gSNR}_i - \text{gSNR}_i \right|, \quad (16)$$

where  $N$  indicates the total number of all the utterances in the test set.

### 3.2. Experimental setup

To verify the effectiveness of the proposed method compared with the existing methods, we compared LSTM [14] and basic CNN-based [19] fullband gSNR estimation methods using different features. The structure of the LSTM used in this research comprises three hidden layers, and each layer contains 512 cells. The CNN consists of seven 2D convolutional layers, and these convolutional layers have 17, 40, 64, 128, 64, 40 and 17 channels respectively. The convolution kernels of the 2D CNN are  $3 \times 3$  shaped. There are two hidden fully connected layers following the CNN with 512 units each. We set the same parameters in the LSTM as in the proposed method for fair comparisons. The LSTM and the CNN structures in the hidden layers all used the ReLU activation functions, the loss functions were the most basic MSE loss functions, and their outputs were fullband noise ratios. In practice, the learning rate for all methods in this research was 0.01, and the batch size was 64. The hyperparameter  $C$  was set to 0.88. Finally, the Adam optimizer was used to ensure steady learning.

Table 1: Comparison of the areas under the curve (MAE) obtained using LSTM and CNN structures.

Noise Type	Feature	LSTM	CNN
White	MFCC	0.32	0.33
	RW	0.61	0.30
Pink	MFCC	0.52	0.57
	RW	0.84	0.54
Factory	MFCC	1.00	1.17
	RW	1.20	1.03
Babble	MFCC	1.02	1.28
	RW	1.41	1.10

For the Mel-frequency cepstral coefficients (MFCCs) or Mel-based gSNR method [9], 128-dimensional Mel-spectrogram features were extracted from each frame of the noisy speech, and the frame length was 400. The raw waveform (RW) [14] features were divided into 400 samples in the experiment as neural network input. The ‘‘Multi-Sub’’ here refers to gSNR estimation proposed in Morita et al. [11].

### 3.3. Results and analysis

#### 3.3.1. Evaluation of fullband based methods

Table 1 presents the fullband-based gSNR estimation results given by the LSTM and CNN using the MFCC and RW features, and it is the average MAE from all the gSNR conditions (-10 dB to 15 dB). The results show the effectiveness of the fullband method [9, 14].

We note that gSNR can be estimated correctly in the presence of White and Pink noise. For Factory and Babble’s unsteady noise conditions, however, no matter which method is used, there is significant performance degradation. MAE using RW and CNN was lower than MFCC and CNN; MAE using RW and LSTM is higher than MFCC and LSTM results. Frequency domain features such as MFCC are more suitable for processing by LSTM, while RW features are more suitable for processing by CNN. In this research, RW-CNN was used as the baseline system. In addition, the fullband method usually has the problem of over-fitting, and the performance of gSNR estimation can not always be guaranteed at low SNR conditions. Hence, we need a method to improve the generalization ability of the algorithm. Similar to RW, our input feature is waveform signals, it is appropriate to select a framework like CNN for subband noise ratio estimation.

#### 3.3.2. Evaluation of proposed method

Fig. 3 shows the MAE of the estimated gSNRs under different noise and SNR environments. The analysis results indicate that the existing Multi-Sub [11] is closer to the ideal gSNR than the WADA-based [10] gSNR estimation method in the high gSNR conditions. However, as the noise levels increase, the performance decreases dramatically. The reason is that the multi-subband-based method utilizes a single threshold to calculate the noise energy in the entire length signal. The speech and noise waveforms are similar in a low-SNR environment, which makes it difficult to establish a single threshold. Our proposed methods learn a dynamic noise ratio in a short speech frame. It makes the learned threshold dynamic rather than static in an utterance. Therefore, the proposed Multi-CNN and MSGNet significantly improve performance over the Multi-Sub method,

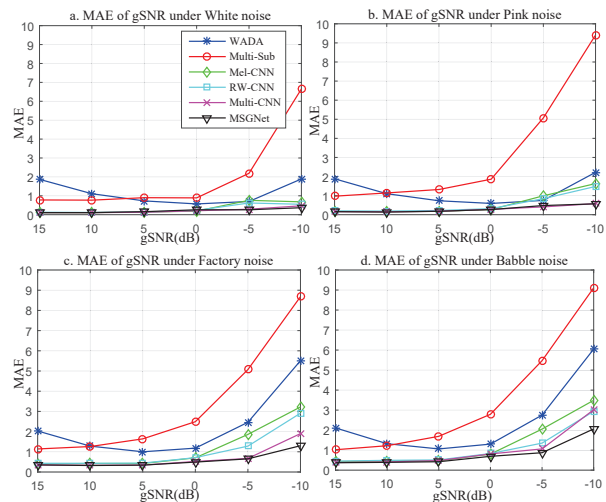


Figure 3: Comparison of the mean absolute error (MAE) obtained using different methods.

especially with White and Pink noise conditions.

Multi-CNN is the result of MSGNet in the absence of an auditory encoder, and the results obtained are better than the fullband methods based on Mel-CNN and RW-CNN. The above results are more obvious under Factory and Babble noise conditions, indicating the effectiveness of bark-scale filter bank strategy for gSNR estimation. Because noise and speech have different frequency distributions, the generalization of the fullband method can be improved by multi-subband analysis. Using the multi-subband strategy, the frequency information of speech signals can be fully used. However, the proposed multi-CNN method has limitations under low SNR and unsteady noise. Through analysis, it is found that CNN cannot fully adapt to the use of multi-subband signals as input. Therefore, Multi-CNN with an auditory encoder (MSGNet) is further proposed. The MSGNet-based method is more stable than Multi-CNN; this indicates that the auditory encoder between the convolution perception network and the auditory front end is helpful for subband noise estimation. The auditory encoder will map the auditory features to a more perceptual space as a medium between mechanical signals and neural signals. From Figure 3. c and d, we can find the decline of MAE is more obvious in the -5 dB and -10 dB low-SNR noisy environments.

## 4. Conclusion and future work

The gSNR, based on deep learning and fullband analysis, has achieved remarkable success. In this paper, we proposed a high-performance multi-subband-based gSNR estimation network using a Bark-scale filterbank. The performance of our proposed gSNR estimation method is obviously better than existing fullband methods. In future work, we will apply the idea of the proposed gSNR estimation method to speech enhancement and robust speech recognition tasks.

## 5. Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grant 62176182 and Alibaba Group through Alibaba Innovative Research Program. I would like to thank my wife LI Xue for her support to my research.

## 6. References

- [1] J. Lim, Oppenheim, and A., "All-pole modeling of degraded speech," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 1978.
- [2] K. Paliwal, B. Schwerin, and K. Wójcicki, "Speech enhancement using a minimum mean-square error short-time spectral modulation magnitude estimator," *Speech Communication*, vol. 54, no. 2, pp. 282–305, 2012.
- [3] X. Cui and A. Alwan, "Noise robust speech recognition using feature compensation based on polynomial regression of utterance snr," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 6, pp. 1161–1172, 2005.
- [4] Nakatsui and Mamoru, "Subjective speech-to-noise ratio as a measure of speech quality for digital waveform coders," *The Journal of the Acoustical Society of America*, vol. 72, no. 4, pp. 1136–44, 1982.
- [5] A. Narayanan and D. Wang, "A casa-based system for long-term snr estimation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 9, pp. 2518–2527, 2012.
- [6] O. Cappe, "Elimination of the musical noise phenomenon with the ephraim and malah noise suppressor," *IEEE Transactions on Speech, Audio Processing*, vol. 2, no. 2, pp. 345–349, 1994.
- [7] S. Standards, "Sound system equipment-part 16 : Objective rating of speech intelligibility by speech transmission index," *Iec*, 2003.
- [8] X. Dong and D. S. Williamson, "Long-term snr estimation using noise residuals and a two-stage deep-learning framework," in *Latent Variable Analysis and Signal Separation*, 2018, pp. 351–360.
- [9] A. H. Poorjam, M. A. Little, J. R. Jensen, and M. G. Christensen, "A supervised approach to global signal-to-noise ratio estimation for whispered and pathological voices," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 296–300.
- [10] C. Kim and R. M. Stern, "Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis," in *Interspeech*, 2008, pp. 2598–2601.
- [11] S. Morita, X. Lu, M. Unoki, and M. Akagi, "Method of estimating signal-to-noise ratio based on optimal design for sub-band voice activity detection," in *Journal of Information Hiding and Multimedia Signal Processing*, vol. 8, no. 6, 2009, pp. 2073–4212.
- [12] R. Aralikatti, D. K. Margam, T. Sharma, A. Thanda, and S. Venkatesan, "Global snr estimation of speech signals using entropy and uncertainty estimates from dropout networks," in *Proc. Interspeech*, 2018, pp. 1878–1882.
- [13] S. Suhadi, C. Last, and T. Fingscheidt, "A data-driven approach to a priori snr estimation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 186–195, 2010.
- [14] H. Li, D. Wang, X. Zhang, and G. Gao, "Frame-level signal-to-noise ratio estimation using deep learning," in *Interspeech*, 2020, pp. 4626–4630.
- [15] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, "SpEx+: A Complete Time Domain Speaker Extraction Network," in *Proc. Interspeech 2020*, 2020, pp. 1406–1410.
- [16] A. Défossez, "Hybrid spectrogram and waveform source separation," *arXiv e-prints*, 2021.
- [17] D. Ditter and T. Gerkmann, "A multi-phase gammatone filterbank for speech separation via tasnet," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 36–40.
- [18] M. L. Jepsen, S. D. Ewert, and T. Dau, "A computational model of human auditory signal processing and perception," *The Journal of the Acoustical Society of America*, vol. 124, pp. 422–38, 2008.
- [19] N. Li, L. Wang, M. Unoki, S. Li, W. Rui, M. Ge, and J. Dang, "Robust voice activity detection using a masked auditory encoder based convolutional neural network," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 01 2021.
- [20] Z. Peng, X. Li, Z. Zhu, M. Unoki, J. Dang, and M. Akagi, "Speech emotion recognition using 3d convolutions and attention-based sliding recurrent networks with auditory front-ends," *IEEE Access*, vol. 8, pp. 16 560–16 572, 2020.
- [21] N. Li, M. Ge, L. Wang, and J. Dang, "A fast convolutional self-attention based speech dereverberation method for robust speech recognition," in *Neural Information Processing*, 2019, pp. 295–305.
- [22] M. Ge, L. Wang, N. Li, H. Shi, J. Dang, and X. Li, "Environment-Dependent Attention-Driven Recurrent Convolutional Neural Network for Robust Speech Enhancement," in *Proc. Interspeech 2019*, 2019, pp. 3153–3157.
- [23] S. Nakamura, K. Takeda, K. Yamamoto, T. Yamada, S. Kuroiwa, N. Kitaoka, and et al., "Aurora-2j: An evaluation framework for japanese noisy speech recognition," *IEICE Transactions on Information and Systems*, vol. E88-D, no. 3, pp. p.535–544, 2005.
- [24] Andrew, Varga, Herman, M. J., and 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.