INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Learning Speech Structure to Improve Time-Frequency Masks

Suliang Bu[1], *Yunxin Zhao*[1], *Shaojun Wang*[2], *Mei Han*[2]

[1]Dept. of Electrical Engineering and Computer Science, University of Missouri-Columbia, USA
[2]PAII Inc. Palo Alto, CA, USA

Abstract

Time-frequency (TF) masks are widely used in speech enhancement (SE). However, accurately estimating TF masks from noisy speech remains a challenge to both statistical or neural network approaches. Statistical model-based mask estimation usually depends on a good parameter initialization, while NN-based mask estimation relies on setting proper and stable learning targets. To address these issues, we propose a novel approach to extracting TF speech structures from clean speech data, and partition a noisy speech spectrogram into mutually exclusive regions of core speech, core noise, and transition. Using such region targets derived from clean speech, we train bidirectional LSTM to learn region prediction from noisy speech, which is easier to do than mask prediction. The predicted regions can further be used in place of masks in beamforming, or integrated with statistical and NN based mask estimation to constrain mask values and model parameter updates. Our experimental results on ASR (CHiME-3) and SE (CHiME-3 and LibriSpeech) have demonstrated the effectiveness of our approach of learning speech region structure to improve TF masks.
Index Terms: TF masks, speech enhancement, ASR

1. Introduction

Noise deteriorates speech quality and intelligibility, causing discomforts and difficulties to listeners in speech understanding, as well as significant performance degradation in automatic speech recognition (ASR). To mitigate this issue, SE is usually used to reduce noise in noisy speech. In both multiple-channel or single-channel SE (MCSE, SCSE), TF masks are widely used. For example, in SCSE, ideal binary masks (IBM) [1, 2] are used to label those TF points that are speech-dominant, and from which to derive an enhanced speech signal; in multiple-channel SE, TF masks are used to estimate speech and noise spatial covariances [3, 4, 5, 6, 7], from which minimum variance distortionless response (MVDR) beamformer is derived.

TF masks can be estimated by statistical or neural network (NN) methods. In the former category, [8, 6, 9, 10] used complex Gaussian mixture model (CGMM), and [11] proposed Watson mixture model. In the latter category, NN such as bidirectional Long Short-Term Memory network (BLSTM) was used successfully [12, 13, 14, 15]. As for the NN training targets, IBM, ideal ratio masks (IRM) [16], complex ideal ratio masks [17], spectral magnitude mask (SMM) [18] and phase sensitive mask [19] are commonly used.

The quality of statistical-model based masks often heavily relies on parameter initialization, and the mask values are usually estimated independently on individual TF elements instead of jointly, rendering the resulting mask values weak in spatial continuity. On the other hand, the NN-based approach appears to face two main issues. First, the mask target values for training may not be accurately determined. For example, in target calculation, IBM makes a comparison between a certain form of speech and noise energy at each TF point, while IRM/SMM uses a certain quotient of speech and noise/noisy speech energy, both are susceptible to artifacts and outliers. Because the non-speech TF points in clean speech often have non-zero values, given a noise recording, a non-speech point may be identified as speech for IBM, or be assigned a relatively high value for IRM or SMM. Second, the target values are sensitive to input scaling: a gentle scaling on either clean speech or noise may alter many TF target values. Due to the unstable target nature, a large amount of data is often needed to train a stable model.

It is well known that speech signals exhibit certain salient structures in the TF space. If such structural information is captured from clean speech, it can be used as an informative prior to facilitate estimating the speech and noise TF masks. Although voice activity detection (VAD) [20, 21, 22] can discriminate between speech and non-speech frames, it cannot do so for TF points, and thus its effect in mask refinement is limited. Along the direction of utilizing speech structure priors for SE, in [23], a speech model is learned from data produced by a generative NN, and the model is used as a speech prior in Bayesian multi-channel nonnegative matrix factorization based speech enhancement; in [24], a Markov Random Field based speech prior is used to model local TF contexts to improve spatial continuity of the estimated IBM and IRM. In [25], the speech temporal structure of pauses is used to estimate time-varying noise for end-to-end SCSE.

In this work, we focus on a basic issue in speech processing, i.e., how to improve TF masks. We investigate extracting speech structures in TF space from clean speech spectrogram, and use the structural prior to improve statistical and NN based mask estimation. Specifically, we divide a noisy speech spectrogram into mutually exclusive regions of core speech, core noise, and transition, which are defined by and estimated from its underlying clean speech spectrogram. These regions are used as the targets in training BLSTM to learn region prediction from noisy speech. We evaluate our method on both single-channel and multi-channel noisy speech data of CHiME-3 [26, 27] and LibriSpeech [28]. It is worth noting that for CHiME-3 or CHiME-4 tasks, various approaches have been proposed with good results achieved in ASR [29, 30, 31, 8, 9, 13, 12, 6], where the popular TF masks are IBM, IRM and CGMM-based ones. In this work, perceptual evaluation of speech quality (PESQ) [32] and short-time objective intelligibility (STOI) [33] are used to evaluate SE performance, where with regions, both PESQ and STOI scores are increased; ASR is performed on multi-channel CHiME-3 to evaluate the MVDR performance based on CGMM masks or IBM, where using regions also reduced word error rate (WER) over using CGMM/NN based mask estimation alone.

2. TF masks and speech enhancement

For clarity, we use bold font for vectors and regular font for scalars, with matrices specified explicitly.

http://dx.doi.org/10.21437/Interspeech.2021-1859

2.1. NN-based IBM estimation

Here we adopt the BLSTM method in [13, 12], where clean speech and noise are used to simulate noisy speech in training BLSTM for mask prediction. Let $x_{f,t}$ and $n_{f,t}$ denote speech and noise at a TF point (f,t), respectively. The targets for speech and noise masks, IBM_X and IBM_N, are defined by

$$IBM_X(f,t) = 1 \quad if \quad |x_{f,t}|/|n_{f,t}| > th_X(f) \quad else \quad 0$$
$$IBM_N(f,t) = 1 \quad if \quad |x_{f,t}|/|n_{f,t}| < th_N(f) \quad else \quad 0$$

where $th_X(f)$ and $th_N(f)$ are two thresholds and $|\cdot|$ denotes the magnitude of a complex number. When beamforming is performed in test, the multi-channel IBM_X and IBM_N estimates are respectively condensed to one using a median operation.

2.2. CGMM-based mask estimation

For multi-channel speech, CGMM [8, 9] can be used to estimate TF masks. Let $\boldsymbol{y}_{f,t}$, $\boldsymbol{x}_{f,t}$ and $\boldsymbol{n}_{f,t}$ denote M-channel observed signal, speech, and noise at (f,t), respectively, with $\boldsymbol{x}_{f,t} = s^x_{f,t}\boldsymbol{r}^x_f$ and $\boldsymbol{n}_{f,t} = s^n_{f,t}\,\boldsymbol{r}^n_f$, where $s^x_{f,t}$ is the speech component, and $\boldsymbol{r}^x_f$ is the acoustic transfer function vector, and $s^n_{f,t}$ and $\boldsymbol{r}^n_f$ are defined similarly. $s^x_{f,t}$ and $s^n_{f,t}$ are assumed to have complex Gaussian distributions: $s^x_{f,t} \sim \mathcal{CN}(0, \sigma^2_{x,f,t})$ and $s^n_{f,t} \sim \mathcal{CN}(0, \sigma^2_{n,f,t})$. Thus, $\boldsymbol{x_{f,t}} \sim \mathcal{CN}(\boldsymbol{0}, \sigma^2_{x,f,t}\boldsymbol{R}^x_f)$ and $\boldsymbol{n_{f,t}} \sim \mathcal{CN}(\boldsymbol{0}, \sigma^2_{n,f,t}\boldsymbol{R}^n_f)$ where $\boldsymbol{R}^x_f = \boldsymbol{r}^x_f\,(\boldsymbol{r}^x_f)^H$ and $\boldsymbol{R}^n_f = \boldsymbol{r}^n_f\,(\boldsymbol{r}^n_f)^H$, with $(.)^H$ the conjugate transpose. $\boldsymbol{y}_{f,t}$ has a mixture distribution of speech and noise. Let $z \in \{x, n\}$, the model parameters are updated by the EM algorithm [34]:

$$\sigma^2_{z,f,t} = \left(\boldsymbol{y}^H_{t,f}\left(\boldsymbol{R}^z_f\right)^{-1}\boldsymbol{y}_{t,f}\right)/M \tag{1}$$

$$\boldsymbol{R}^z_f = \frac{1}{\sum_t \lambda^z_{t,f}}\sum_t \frac{\lambda^z_{t,f}}{\sigma^2_{z,f,t}}\boldsymbol{y}_{t,f}\,\boldsymbol{y}^H_{t,f} \tag{2}$$

$$\hat{\lambda}^z_{t,f} = \frac{\lambda^z_f \cdot \mathcal{CN}(\boldsymbol{y}_{t,f}|\boldsymbol{0}, \sigma^2_{z,f,t}\boldsymbol{R}^z_f)}{\lambda^x_f \cdot \mathcal{CN}(\boldsymbol{y}_{t,f}|\boldsymbol{0}, \sigma^2_{x,f,t}\boldsymbol{R}^x_f) + \lambda^n_f \cdot \mathcal{CN}(\boldsymbol{y}_{t,f}|\boldsymbol{0}, \sigma^2_{n,f,t}\boldsymbol{R}^z_f)}$$

The estimated probability $\lambda^x_{t,f}$ becomes local speech mask.

2.3. Single-channel SE and multiple-channel SE

For SCSE, the estimated IBM_X is multiplied with the noisy speech spectrogram to estimate the enhanced spectrogram $\hat{x}_{f,t}$.

For multiple-channel SE, the noise and speech spatial covariances in a frequency bin are computed as $\boldsymbol{\Phi}^n_f = (\sum_t \lambda^n_{f,t}\boldsymbol{y}_{f,t}\,\boldsymbol{y}^H_{f,t})/(\sum_t \lambda^n_{f,t})$ and $\boldsymbol{\Phi}^x_f = \boldsymbol{\Phi}^{n+x}_f - \boldsymbol{\Phi}^n_f$, where $\boldsymbol{\Phi}^{n+x}_f = (\sum_t \boldsymbol{y}_{f,t}\,\boldsymbol{y}^H_{f,t})/T$. The eigenvector corresponding to the largest eigenvalue of $\boldsymbol{\Phi}^x_f$ is used as the steering vector (SV) $\boldsymbol{h}_f$ [8] for MVDR, with the spatial filter formed as

$$\boldsymbol{w}_f = \frac{(\boldsymbol{\Phi}^n_f)^{-1}\cdot\boldsymbol{h}_f}{\boldsymbol{h}^H_f\cdot(\boldsymbol{\Phi}^n_f)^{-1}\cdot\boldsymbol{h}_f} \tag{3}$$

The beamformed TF points are obtained by $\hat{x}_{f,t} = \boldsymbol{w}^H_f\,\boldsymbol{y}_{f,t}$, from which the enhanced speech waveform is obtained.

3. Proposed region learning

In this section, we cover region definition and rational, region target extraction, and learning region prediction by NN.

3.1. Region definition and merits

We divide a noisy speech spectrogram into three mutually exclusive regions: core speech region (CSR), transition region (TR), and core non-speech region (CNR). Together, the three regions fully cover the noisy spectrogram, but the regions are defined by the clean speech spectrogram underneath the noisy one. Qualitatively speaking, in a clean spectrogram, CSR consists of the TF points with large magnitudes, CNR the points with small magnitudes, and TR the remaining points.

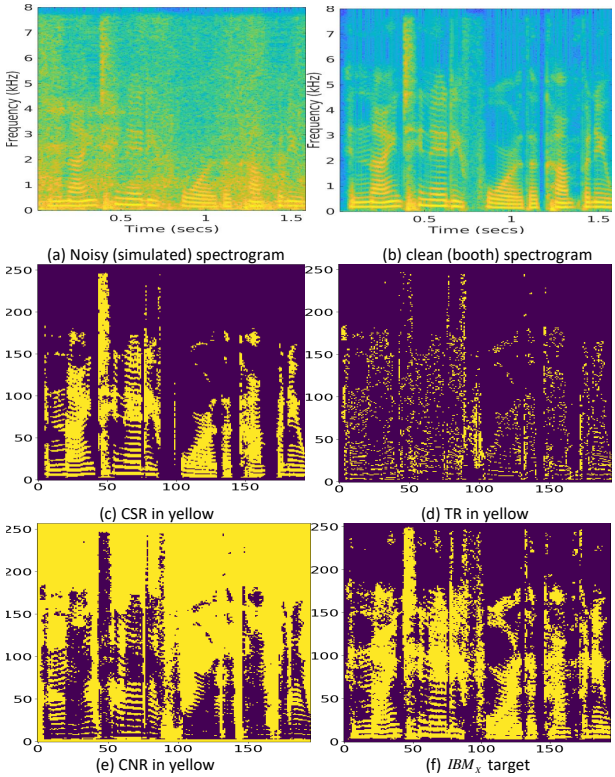

Figure 1: *Speech spectrogram, three regions, and IBM_X target for a speech utterance in CHiME-3*

Fig.1 illustrates the concept for three regions. Comparing (b) and (c), CSR clearly captures the main structure of speech. Comparing (d) and (c), the points in TR appear mainly on the edges of CSR. Comparing (b) and (e), speech magnitudes of the points in CNR are apparently negligible. Finally, (f) has many yellow points extraneous to (b), suggesting that some non-speech points are misclassified as speech in IBM_X target.

We find that learning the regions is in general easier than learning the masks, and the region targets provide informative speech structural prior knowledge in the TF space, which may improve the reliability of mask estimation. For example, in a CNR, the noise mask value may be simply set to 1. For statistical model based mask estimation, parameter updates can be constrained by the region types. Moreover, regions alone are already informative for computing the spatial statistics of speech and noise in beamforming, where using learned region information in place of NN masks can reduce WER (c.f. Sec. 4). In other words, regions can serve as a strong structural prior to replace masks to some extent, but masks do not conform to the structural constraints. For example, a TF point with a high noise mask estimate does not necessarily mean it belongs to a CNR.

3.2. Region target computation and model training

We extract region targets from clean speech, and train a BLSTM to learn region prediction from simulated noisy speech (adding noise to clean speech). The NN's inputs are noisy speech frames, and its outputs are predicted regions. Here, we sim-

ply view CNR as Core Noise Region, the union of CSR and TR as non core-noise region (NCNR), and the union of CNR and TR as non core-speech region (NCSR).

3.2.1. Region target computation

For preprocessing, each clean speech recording is scaled to have the maximum value of 16-bit signed integer of 32767, from which a spectrogram is computed. Our region target extraction procedure looks for speech structures in the spectrogram in three ways: in TF points, in time frames, and in each frequency bin across time. Extracting the structure of regions amounts to setting insignificant TF points or frames to zeros. The procedure determines CSR and NCNR targets by following the same 4 steps but with different threshold values:

Step-1. Zero out TF points: Set the TF points with magnitudes less than a threshold th_{gl1} to 0.

Step-2. Zero out a frame: Sum the magnitudes of the TF points in each frame, and if the sum of a frame is less than a threshold th_{sum}, then set all the TF points in the frame to 0.

Step-3. Zero out points in a frequency bin. Keep the TF points with the largest 20 percentile magnitude values for the target region by default. From the rest 80% TF points, compute the sum of magnitudes (SoM) S_f, and additionally keep the TF points with the largest q percentile magnitudes to make their SoM approximately equal to $S_f \times th_{gl2}$, where q depends on S_f and th_{gl2}. Set the value of the not chosen TF points to 0.

Step-4: Set the target values of the non-zero TF points to 1 for CSR or NCNR.

Denoting the full set of TF points by U, then we obtain CNR = U-NCNR, and TR = NCNR-CSR . Note that the values of TF points are cached, so that the computations for CSR and NCNR do not affect each other. Also note that our region method determines speech structures in 3 ways from clean speech alone, while IBM only compares speech and noise on (f, t) points.

3.2.2. Region prediction model training

For region prediction, we largely follow the BLSTM architecture and settings in [13, 12] of speech and noise mask prediction. However, our use of BLSTM differs from [13, 12] in prediction targets, data augmentation, and computation:

1) Prediction: two regions are predicted: CSR and NCNR;

2) Data augmentation: the block structure of frames marked by the zeroed-out frames (Step-2 of Sec.3.2.1) is used for data augmentation. For each utterance, 4 frame blocks are identified, the 4 clean speech blocks are shuffled 4 times with different positions for each block each time, while the noise ones are unchanged, and the targets of the blocks are adjusted accordingly;

3) Computation: to improve parallelism, two consecutive input frames are stacked to from a long frame, so are their targets.

4. Experiments and Results

We used CHiME-3 as the main dataset in our experiments. It covered four noisy environments: cafe (CAF), street (STR), bus (BUS) and pedestrian area (PED) and had 6 microphone channel recordings. The test datasets had real and simulated noisy speech, each consisting of 330 sentences per noise type. The details are as described in [26]. When training BLSTM for region prediction, the clean and noise data were selected from CHiME-3 to generate noisy speech. To assess the generalization ability of our region prediction method in SCSE, we also took about 14 hours of clean speech from LibriSpeech to generate simulated noisy test speech data and performed SCSE on these data by using the region predictor trained on CHiME-3.

4.1. Experiment Setup

When using CGMM or NN based masks, we followed the settings in [8, 13, 12] unless mentioned otherwise. DFT size was 512, frame shift was 25% of frame size. To compute the region targets, we empirically set the thresholds of Sec.3.2.1, with $th_{gl1}, th_{sum}, th_{gl2}$ respectively set to 0.07, 6 and 0.95 for NCNR, and to $0.1, 8$ and 0.8 for CSR. From a set of pilot speech samples, we verified that these thresholds allowed extracting clear speech structures, and the speech signals recovered from the speech regions had imperceptible distortion. We then applied the procedure of Sec.3.2.1 to the clean training speech to extract the region targets. For ASR, we used the popular CHiME-3 baseline backend in Kaldi [35], although we noticed recently appeared baseline systems for CHiME-3, such as [29]. We focused on extracting speech regions to improve TF masks, instead of obtaining the lowest WER, which would depend on the baseline and a host of other factors.

For SCSE, we considered the TF mask of IBM_X. The minimum value of IBM_X was set to 0.1 irrespective to the availability of region estimates. If region estimates were available, then the IBM_X values of all the TF points within NCSR were set to 0.1, and the IBM_X values in CSR were unchanged.

For M-channel noisy speech, the M spectrograms were first averaged to a mean spectrogram, and from which the regions were estimated. Upon available the region estimates, the noise mask values of the TF points within a CNR were set to 1, while the way the TF points within CSR were used depended on the settings of computing $\mathbf{\Phi}_f^n$, and the effects of these settings on WER were compared. If only region estimates were available (without masks), we used the TF points in CNR and NCNR to calculate $\mathbf{\Phi}_f^n$ and $\mathbf{\Phi}_f^{n+x}$ respectively, shown as MVDR_R in Table 2. When only CGMM masks were used, the utterance beginning and ending 20 TF points were used to initialize $\boldsymbol{R}_f^n$, and the remaining points were used to initialize $\boldsymbol{R}_f^x$. If region estimates were available, the TF points in CNR and CSR were used to initialize $\boldsymbol{R}_f^n$ and $\boldsymbol{R}_f^x$, respectively. In case that the number of TF points were insufficient, those in TR were additionally used for CNR and CSR. During the parameter updates, only the TF points in NCNR were used to update speech parameters. In estimating the CGMM parameters, we selectively used the TF points to update the parameters of the speech or noise component models based on the region belongings of these points.

4.2. Experiment Results

4.2.1. Speech Recognition

We compared MVDR performance on CHiME-3 in WER under eight settings: using our proposed regions, CGMM masks, NN masks, and combining regions with CGMM or NN masks, indicated in that order by $(\cdot)_R$, $(\cdot)_G$, $(\cdot)_N$, $(\cdot)_{GR}$, and $(\cdot)_{NR}$, respectively. In addition, $(\cdot)_{R0}$, $(\cdot)_{R1}$, and $(\cdot)_{R2}$ signifies the way the estimated regions were combined with mask estimation, as defined and summarized in Table 1:

a). The noise TF mask values in CNR were fixed to 1;

b). The TF points in CNR and CSR were used to initialize $\boldsymbol{R}_f^n$ and $\boldsymbol{R}_f^x$ in CGMM, respectively;

c). All TF points were used to update $\mathbf{\Phi}_f^n$, as well as to update noise parameters during EM iteration of CGMM;

d). Only TF points in NCSR were used to update $\mathbf{\Phi}_f^n$ and the CGMM noise parameters during EM iteration.

The WER results are given in Table 2. When using region

Table 1: *Methods of integrating regions with TF masks*

methods	NR1	NR2	GR0	GR1	GR2
constraints	a,c	a,d	b	a,b,c	a,b,d

Table 2: *WERs (%) of baseline, MVDR, w/wo regions, based on NN/CGMM TF masks on CHiME-3 test data*

2*	eval simu					eval real				
	BUS	CAF	PED	STR	AVG	BUS	CAF	PED	STR	AVG
baseline	8.7	13.1	12.9	14.9	12.4	18.8	10.5	10.3	9.8	12.4
MVDR_N	4.0	5.7	5.7	5.8	5.3	13.6	6.8	6.0	6.4	8.2
MVDR_R	3.7	5.2	4.6	5.5	4.8	11.2	6.6	6.0	5.6	7.4
MVDR_{NR1}	**3.6**	**4.4**	4.6	**4.9**	**4.4**	12.3	6.1	5.6	5.9	7.5
MVDR_{NR2}	3.8	4.7	**4.2**	5.3	4.5	**11.0**	**5.8**	**5.1**	**5.6**	**6.9**
MVDR_G	4.6	5.2	5.8	8.3	6.0	16.6	6.9	6.3	8.6	9.6
MVDR_{GR0}	3.9	**4.2**	4.9	6.2	4.8	12.6	5.6	6.2	6.7	7.8
MVDR_{GR1}	**3.8**	4.5	4.6	5.5	4.6	11.5	**5.1**	5.5	**5.8**	7.0
MVDR_{GR2}	3.8	4.4	**4.5**	**5.2**	**4.5**	**10.2**	5.5	**5.3**	5.9	**6.7**

estimates alone in estimating the spatial statistics for MVDR, i.e., MVDR_R, performance better than using NN masks in MVDR_N and CGMM masks in MVDR_G was already achieved.

The performance of CGMM was dependent on its parameter initialization. In Table 2, MVDR_G and MVDR_{GR0} differed only in parameter initialization, with the latter based on the region estimates. We observe that MVDR_{GR0} greatly reduced WER over MVDR_G, suggesting that using region constraints effectively improved CGMM parameter initialization.

Comparing MVDR_{NR1} with MVDR_N, or MVDR_{GR1} with MVDR_{GR0}, we can see that by setting the noise mask values of the TF points in CNR to 1, WER were largely reduced. This implies that the mask values in CNR were not estimated accurately by NN or CGMM mask methods, and imposing the CNR constraint reduced the uncertainty. Similarly, the mask values of TF points in NCNR may not be well estimated. But how to improve these masks remains an open question.

A common issue in beamforming is that $\mathbf{\Phi}_f^n$ may contain speech component, rendering the spatial filter inaccurate and often with numerical issues. When the TF masks were inaccurate, and all the TF points were used to update $\mathbf{\Phi}_f^n$, this issue was exacerbated. With the estimated regions, however, the TF points within CSR could be excluded from updating $\mathbf{\Phi}_f^n$. In Table 2, MVDR_{NR1} and MVDR_{GR1} included the TF points in CSR in computing $\mathbf{\Phi}_f^n$, but MVDR_{NR2} and MVDR_{GR2} did not. The WER reductions by $(\cdot)_{R2}$ over $(\cdot)_{R1}$ on the real noisy speech test set showed that excluding the TF points in CSR from computing $\mathbf{\Phi}_f^n$ improved MVDR. On the other hand, on the simulated noisy speech test set, the WER gap between using or excluding the TF points in CSR for $\mathbf{\Phi}_f^n$ was negligible. Possibly, the masks of CGMM or NN were more accurate on the simulated data than on the real data, since the simulated acoustic environments were simpler than the real ones.

When NN masks were used, MVDR_{NR2} achieved the lowest WER, where the estimated noise mask in CNR were set to 1, and the CSR points were excluded from updating $\mathbf{\Phi}_f^n$. Similarly, CGMM_{GR2} got the lowest WER with CGMM masks, where noise and speech model parameter updates only included the TF points in NCSR and NCNR, respectively. We can increase noise mask values in CNR, and constrain parameter updates in the matched regions, both being beneficial to MVDR, which could not be done by using masks alone.

4.2.2. Speech Enhancement

Here the simulated noisy test speech data were used to access the ground truth clean speech for computing PESQ and STOI. For SCSE, channel-4 of CHiME-3 was used, and LibriSpeech was also used to assess our region method's generalization ability. We observe that the average PESQ and STOI scores of NN-mask with region (NR) were higher than NN-mask (N) in all comparison cases, and this behavior was also held true on LibriSpeech. The relative gain of PESQ and STOI from using regions in beamformed signals was larger than in SCSE. One possibility was that in beamforming, the three types of regions could all be effectively used, but only CSR was used for SCSE. How to effectively use the different types of regions in SCSE deserves a further investigation. The proposed regions were also found complementary to other types of NN masks, such as IRM and SMM. Due to space limitation, only IBM results are provided here.

Table 3: *PESQ and STOI for MVDR, SCSE, w/wo regions, based on IBM_X on CHiME-3 test data & LibriSpeech*

	SE type	AVG PESQ	AVG STOI
noisy CH4 CHiME-3		2.1	0.88
CH4_N	SCSE	2.6	0.90
CH4_{NR}	SCSE	**2.7**	**0.91**
MVDR_N	beamforming	2.6	0.91
MVDR_{NR2}	beamforming	**2.8**	**0.96**
noisy LibriSpeech		1.9	0.79
LibriSpeech_N	SCSE	2.3	0.83
LibriSpeech_{NR}	SCSE	**2.4**	**0.84**

5. Discussion

One problem faced by setting the NN mask targets is the complexity of the combination patterns of clean speech and noise. For example, as shown in Fig.1.(f), some non-speech TF points were classified as speech due to the point-wise comparison on clean speech and noise power in IBM target calculation. These erroneous target values may cause difficulty for NN to reliably learn speech structure during training, and make NN predict wrong mask values during test. In contrast, the proposed region targets depend only on the structure of clean speech, and they are also immune to input scaling. Therefore, the region targets can be more reliably and accurately computed than NN-based mask targets. Accordingly, the probability space for regions is more likely to be correctly learnt than that for masks given the same model and amount of data, producing more accurate region estimates than masks during test. On the other hand, between region and mask, the former is more qualitative, while the latter is more quantitative, and the complementary natures of the two provides a promising potential in their integration, which has been verified by our results.

6. Conclusion

We have proposed a novel approach to extracting TF structure of clean speech for partitioning noisy speech spectrogram into mutually exclusive regions, which can be used in synergy with the TF masks for SCSE, or for estimating speech and noise statistics in beamforming. We have successfully trained a BLSTM to predict the regions of CSR and NCNR from noisy speech spectrograms, with the region targets obtained from clean speech. We have investigated using the predicted regions in place of TF masks in MVDR beamforming and integrating the regions with statistical and NN based mask estimations to constrain mask values and model parameter updates, as well as using the predicted regions with IBM in SCSE. Our results have demonstrated the effectiveness of our method in reducing WER of ASR and in improving PESQ and STOI in SE.

7. References

[1] G. Hu and D. Wang, "Speech segregation based on pitch tracking and amplitude modulation," in *IEEE WASPAA*, 2001, pp. 79–82.

[2] ——, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE TNN*, vol. 15, no. 5, pp. 1135–1150, 2004.

[3] K. Kumatani, T. Arakawa *et al.*, "Microphone array processing for distant speech recognition: Towards real-world deployment," in *APSIPA ASC*, 2012, pp. 1–10.

[4] L. Pfeifenberger, T. Schrank *et al.*, "Multi-channel speech processing architestures for noise robust speech recognition: 3-rd CHiME chanllenge results," in *INTERSPEECH*, 2015, pp. 452–459.

[5] T. Menne, J. Heymann *et al.*, "The RWTH/UPB/FORTH system combination for the 4th CHiME challenge evaluation," in *The 4th IWSPEE*, 2016.

[6] T. Yoshioka *et al.*, "The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices," in *IEEE ASRU*, 2015, pp. 436–443.

[7] H. Erdogan, T. Hayashi *et al.*, "Multi-channel speech recognition: LSTMs all the way through," in *Proc. CHiME-4 workshop*, 2016, pp. 1–4.

[8] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, "Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise," in *ICASSP*, 2016, pp. 5210–5214.

[9] T. Higuchi *et al.*, "Online MVDR beamformer based on complex Gaussian mixture model with spatial prior for noise robust ASR," *IEEE/ACM TASLP*, vol. 25, no. 4, pp. 780–793, 2017.

[10] S. Bu, Y. Zhao, M. Hwang, and S. Sun, "A probability weighted beamformer for noise robust ASR," *INTERSPEECH*, pp. 3048–3052, 2018.

[11] D. Vu and R. Haeb-Umbach, "Blind speech separation employing directional statistics in an expectation maximization framework," in *ICASSP*, 2010, pp. 241–244.

[12] J. Heymann *et al.*, "BLSTM supported GEV beamformer front-end for the 3rd CHiME challenge," in *IEEE ASRU*, 2015, pp. 444–451.

[13] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *ICASSP*, 2016, pp. 196–200.

[14] A. Subramanian *et al.*, "Student-teacher learning for BLSTM mask-based speech enhancement," *arXiv preprint arXiv:1803.10013*, 2018.

[15] X. Xiao, S. Zhao *et al.*, "On time-frequency mask estimation for MVDR beamforming with application in robust speech recognition," in *ICASSP*, 2017, pp. 3246–3250.

[16] C. Hummersone, T. Stokes, and T. Brookes, "On the ideal ratio mask as the goal of computational auditory scene analysis," in *Blind source separation*. Springer, 2014, pp. 349–368.

[17] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM TASLP*, vol. 24, no. 3, pp. 483–492, 2015.

[18] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM TASLP*, vol. 22, no. 12, pp. 1849–1858, 2014.

[19] H. Erdogan, J. R. Hershey *et al.*, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *ICASSP*, 2015, pp. 708–712.

[20] D. Ying, Y. Yan, J. Dang, and F. Soong, "Voice activity detection based on an unsupervised learning framework," *IEEE TASLP*, vol. 19, no. 8, pp. 2624–2633, 2011.

[21] Y. Suh and H. Kim, "Multiple acoustic model-based discriminative likelihood ratio weighting for voice activity detection," *IEEE SPL*, vol. 19, no. 8, pp. 507–510, 2012.

[22] H. Wang, Y. Xu, and M. Li, "Study on the MFCC similarity-based voice activity detection algorithm," in *IEEE AIMSEC*, 2011, pp. 4391–4394.

[23] K. Sekiguchi, Y. Bando *et al.*, "Bayesian multichannel speech enhancement with a deep speech prior," in *IEEE APSIPA ASC*, 2018, pp. 1233–1239.

[24] S. Liang, W. Liu *et al.*, "Integrating binary mask estimation with MRF priors of cochleagram for speech separation," *IEEE SPL*, vol. 19, no. 10, pp. 627–630, 2012.

[25] R. Xu, R. Wu, Y. Ishiwaka, C. Vondrick, and C. Zheng, "Listening to sounds of silence for speech denoising," *arXiv preprint arXiv:2010.12013*, 2020.

[26] J. Barker *et al.*, "The third CHiME speech separation and recognition challenge: Dataset, task and baselines," in *IEEE ASRU*, 2015, pp. 504–511.

[27] ——, "The third CHiME speech separation and recognition challenge: Analysis and outcomes," *Computer Speech & Language*, vol. 46, pp. 605–626, 2017.

[28] V. Panayotov, G. Chen *et al.*, "Librispeech: an ASR corpus based on public domain audio books," in *ICASSP*, 2015, pp. 5206–5210.

[29] S.-J. Chen *et al.*, "Building state-of-the-art distant speech recognition using the CHiME-4 challenge with a setup of speech enhancement baseline," *arXiv preprint arXiv:1803.10109*, 2018.

[30] J. Du *et al.*, "The USTC-iFlytek system for CHiME-4 challenge," *Proc. CHiME-4 workshop*, vol. 4, pp. 36–38, 2016.

[31] Y. Tu *et al.*, "On design of robust deep models for CHiME-4 multi-channel speech recognition with multiple configurations of array microphones." in *INTERSPEECH*, 2017, pp. 394–398.

[32] A. W. Rix *et al.*, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *ICASSP*, vol. 2, 2001, pp. 749–752.

[33] C. H. Taal *et al.*, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE/ACM, TASLP*, vol. 19, no. 7, pp. 2125–2136, 2011.

[34] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[35] P. Daniel, G. Arnab *et al.*, "The Kaldi speech recognition toolkit," in *IEEE ASRU*, 2011.