PAN: Phoneme-Aware Network for Monaural Speech Enhancement

Zhihao Du, Ming Lei, Jiqing Han, Shiliang Zhang

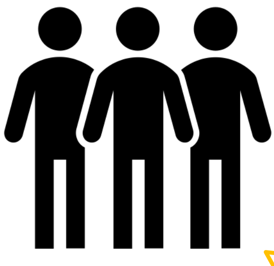
Harbin Institute of Technology & Machine Intelligence Technology

Accepted as a poster in ‘Speech Enhancement II: Network Architectures’, ICASSP 2020.

The work was performed while the first author was an intern at Machine Intelligence Technology, Alibaba Group. This research was supported by National Natural Science Foundation of China under Grant U1736210, National Key Research and Development Program of China under Grant 2017YFB1002102.

Contents

Introduction

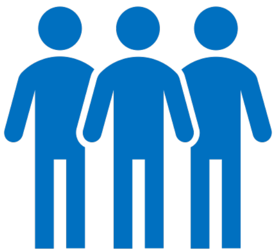


Speech Enhancement

Many acoustic features has been explored for monaural speech enhancement.

Significant process has been achieved by using the phonetic information via phonetic posteriorgram.

Cool!!! Inspired by the progress, we attempt to introduce the phonetic information into monaural speech enhancement by developing a phoneme-aware network.



Voice Conversation

Advantages

- Phonetic information provides more stationary cues than acoustic features.
- The solution space of an enhanced speech will be restricted given the phonetic posteriorgram as the condition.

Phonetic posteriorgram (PPG)

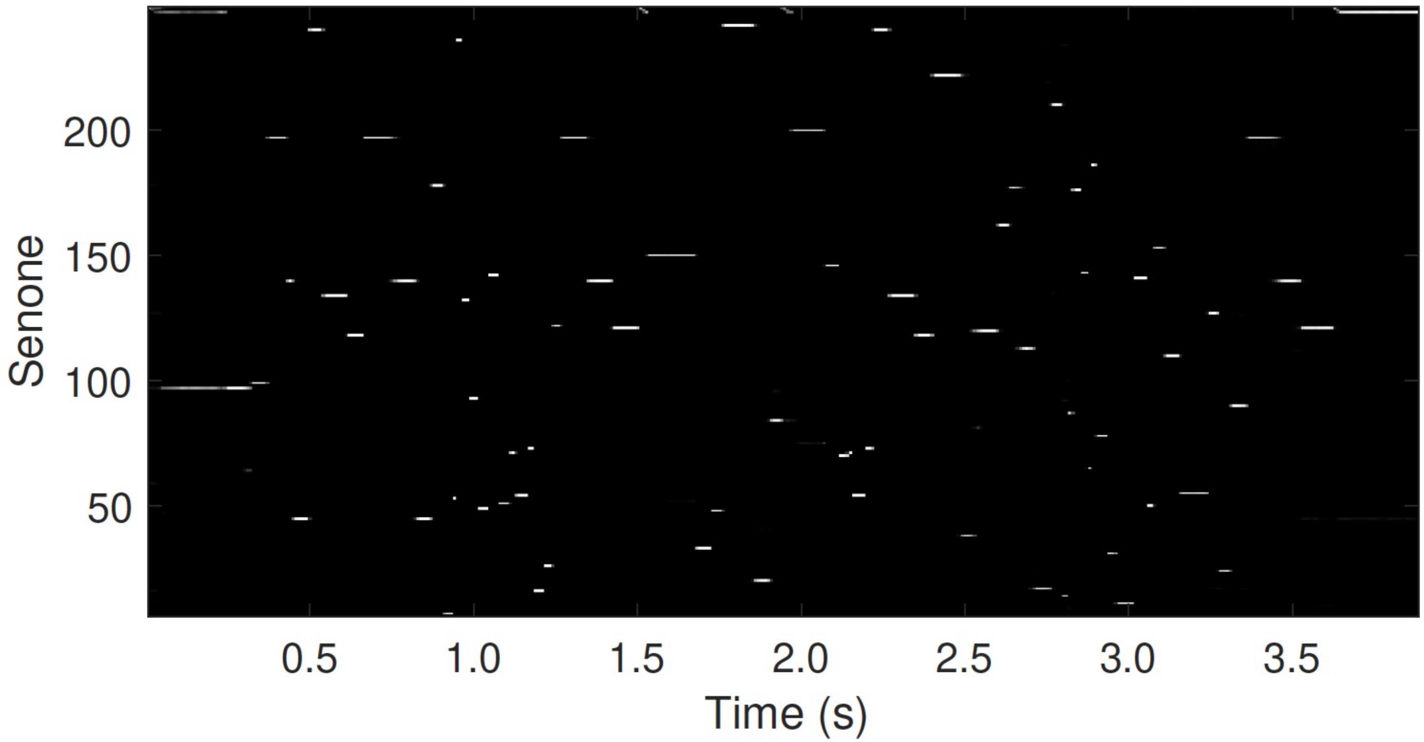
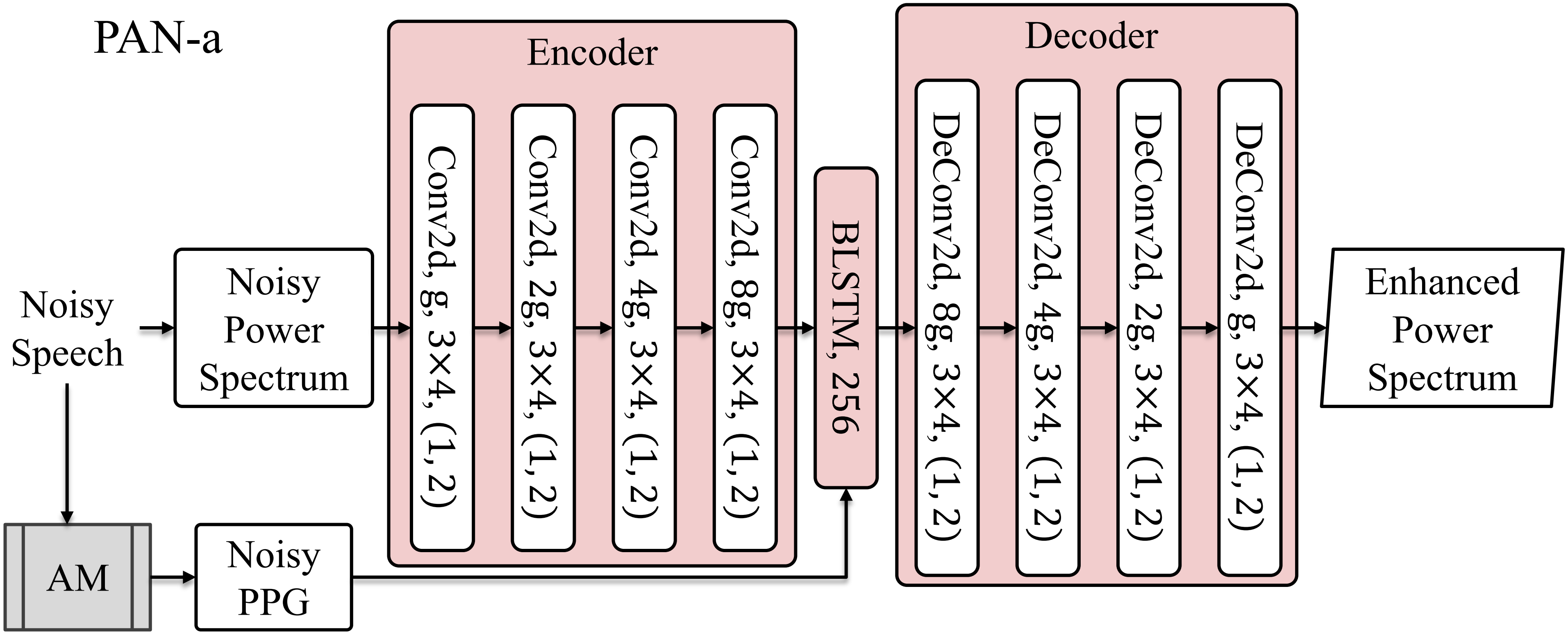


Fig. 1: PPG representation of a mandarin sentence. The horizontal axis represents time in seconds and the vertical axis denotes the phonetic class. Lighter shade implies a higher posterior probability.

Basic Phoneme-Aware Network

PAN-a



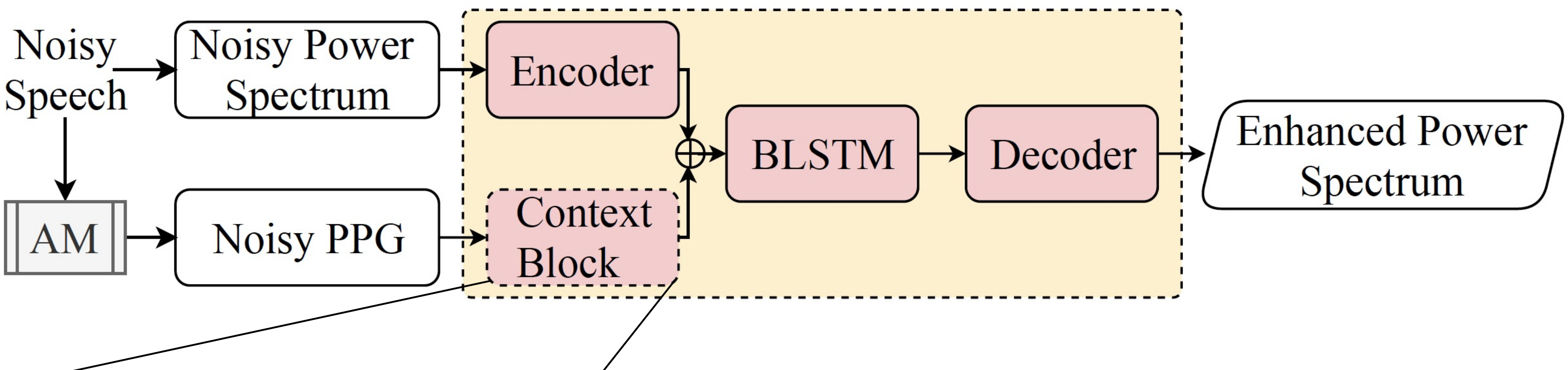
$$L_{SE} = \frac{1}{T}\frac{1}{F}\sum_{t=1}^{T}\sum_{f=1}^{F}\left|IRM(t,f) - \widehat{IRM}(t,f)\right|^2 \qquad where, IRM(t,f) = \frac{|S|^2(t,f)}{|S|^2(t,f) + |N|^2(t,f)}$$

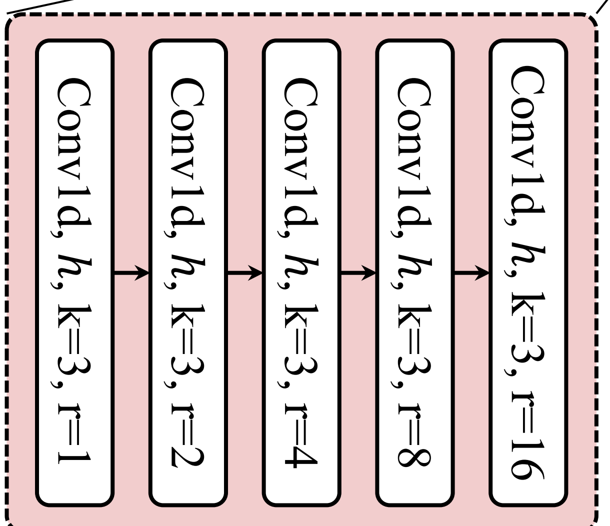
$$\left|\hat{S}\right|^2(t,f) = |Y|^2(t,f)\cdot\widehat{IRM}(t,f)$$

$|S|^2, |N|^2 and |Y|^2$ represent the power spectrum of clean speech, background noise and noisy speeches

Context Aggregation with dilated convolutions

PAN-b





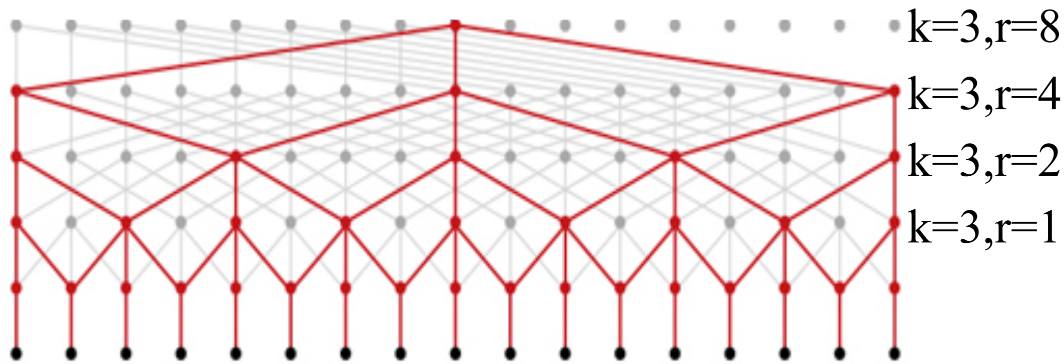
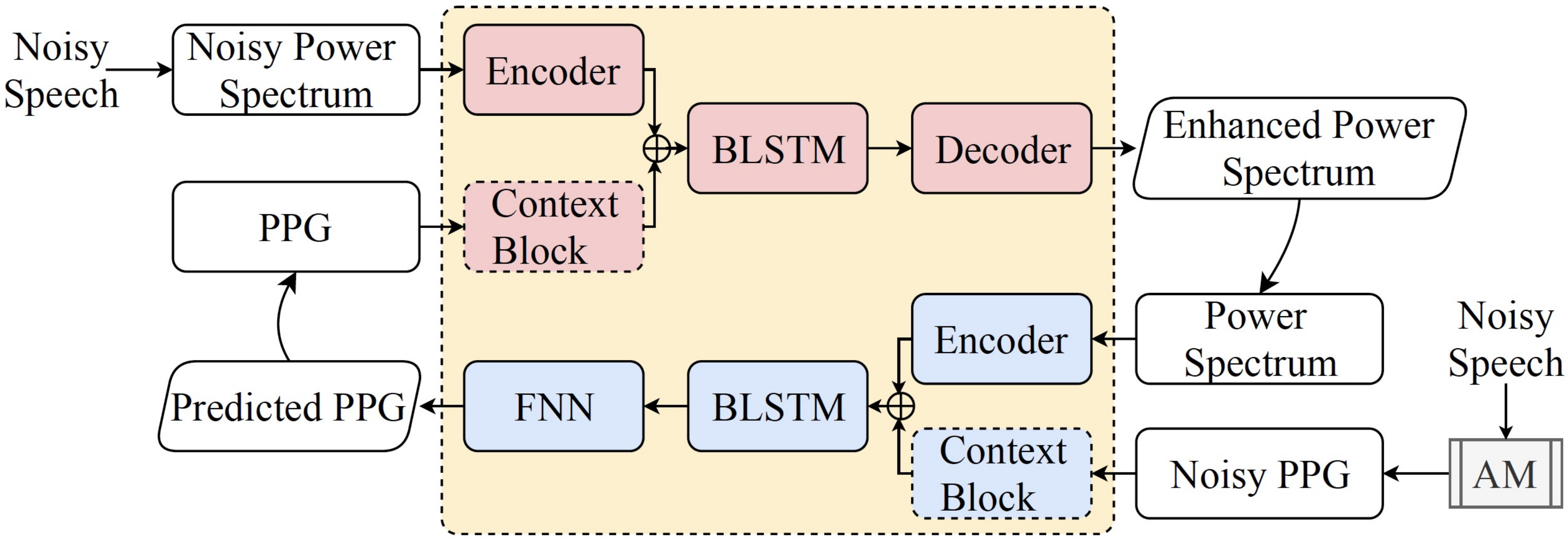


Fig. Dilated convolutions

Involving PPG prediction



PAN-b+PR

$$L_{PR} = KL(P||\hat{P}) = \sum_{t=1}^{T}\sum_{c=1}^{C} P(t,c)\log P(t,c) - P(t,c)\log \hat{P}(t,c)$$

$$L_{PAN} = L_{PR} + L_{SE}$$

Iterative prediction and training algorithm

Algorithm 1 The forward process of iterative training for PAN.

Input:

The power spectrum of noisy speech, $|X|^2$;

The PPG extracted from the noisy speech, Q;

Output:

The enhanced power spectrum, $|\hat{S}|^2$;

The predicted PPG, $\hat{P}$;

$\hat{P} \Leftarrow Q$;

$|\hat{S}|^2 = PAN(|X|^2, \hat{P}) \odot |X|^2$;

for $i = 1; i \leqslant N; i++$ **do**

    $\hat{P} = PR(|\hat{S}|^2, Q)$;

    $|\hat{S}|^2 = PAN(|X|^2, \hat{P}) \odot |X|^2$;

end for

return $|\hat{S}|^2, \hat{P}$;

Correcting noisy PPG (CNP) vs. predicting ground-truth PPG from scratch (PGP)

- In PAN-b+PR, the PPG predictor is trained to estimate the ground-truth PPG from scratch, which can be more inaccurate than the noisy PPG at the beginning of training. (PGP)
- We train the PPG predictor to learn how to correct the noisy PPG Q in log-scale (CNP):

$$P\log\hat{P} = \mathrm{PR}\left(\hat{S}, Q\right) + P\log Q$$

- Combining the loss function of PPG predictor L_{PR} and the above equation, we get:

$$\min L_{PR} = \min\left(KL(P||Q) - \sum \mathrm{PR}\left(\hat{S}, Q\right)\right)$$

Experimental settings

- Clean speech corpus
  - 98,991 mandarin utterances from 100 males and 100 females
  - 95 males and 95 females are randomly selected for training
  - 10 speakers are used for test.
- 5 noises from NOISEX-92 are used for training
  - factory1, Speech Shaped Noise, engine, optroom, babble
- 4 noises are used for test
  - buccaneer1, buccaneer2, factory2 from NOISEX-92
  - Cafeteria from DEMAND
- 4 SNR levels are trained and evaluated
  - -5, 0, 5, 10 dB

Experimental settings

- Optimizer: Adam with the learning rate of 0.001
- Acoustic model for PPG extraction:
  - A deep feed-forward sequential memory network (DeepFSMN) trained with a 5,000-hour mandarin speech dataset.
  - 10 DeepFSMN blocks with 512 hidden units in each block
  - The DeepFSMN is trained to model 244 senones by minimizing the cross-entropy (CE) loss.
- Evaluation settings:
  - Metrics: short-time objective intelligibility (STOI), perceptual evaluation of speech quality (PESQ) and character error rate (CER) of robust ASR
  - The ASR system is trained with a 20,000-hour mandarin speech dataset collected from 20 domains resulting in 5.71% CER on clean speech.

Experimental results

- The effect of phonetic information

Table 1: Effect of phonetic information in terms of STOI and PESQ on untrained noises and untrained speakers. The numbers represent the averages over the four test noises.

metrics	STOI (%)				PESQ			
SNR (dB)	-5	0	5	10	-5	0	5	10
unprocessed	64.39	75.38	85.29	92.15	1.01	1.32	1.71	2.10
CRN [5]	70.44	81.95	89.94	94.70	1.50	1.92	2.35	2.74
PAN-a (NP)	71.26	82.55	90.30	94.88	1.56	1.96	2.38	2.77
PAN-a (GP)	78.51	85.69	91.64	95.56	1.84	2.18	2.55	2.93

Experimental results

- Comparisons of different architectures and iterations

Table 2: Comparisons of different architectures and iterations in terms of average STOI, PESQ and PPG accuracy on test set.

metrics	STOI (in %)	PESQ	ACC (in %)
unprocessed	79.77	1.54	44.50
CRN [5]	84.26	2.13	-
PAN-a (N=0)	85.32	2.17	-
PAN-a+PR (PGP, N=1)	**85.64**	**2.21**	**49.95**
PAN-a+PR (PGP, N=2)	85.51	2.19	45.06
PAN-b (N=0)	85.86	2.20	-
PAN-b+PR (PGP, N=1)	**86.00**	**2.25**	**58.15**
PAN-b+PR (PGP, N=2)	85.94	2.21	54.01
PAN-b+PR (CNP, N=1)	**86.13**	**2.28**	**60.23**
PAN-b+PR (CNP, N=2)	85.30	2.18	56.60

Experimental results

- Correcting noisy PPG (CNP) vs. predicting ground-truth PPG from scratch (PGP)

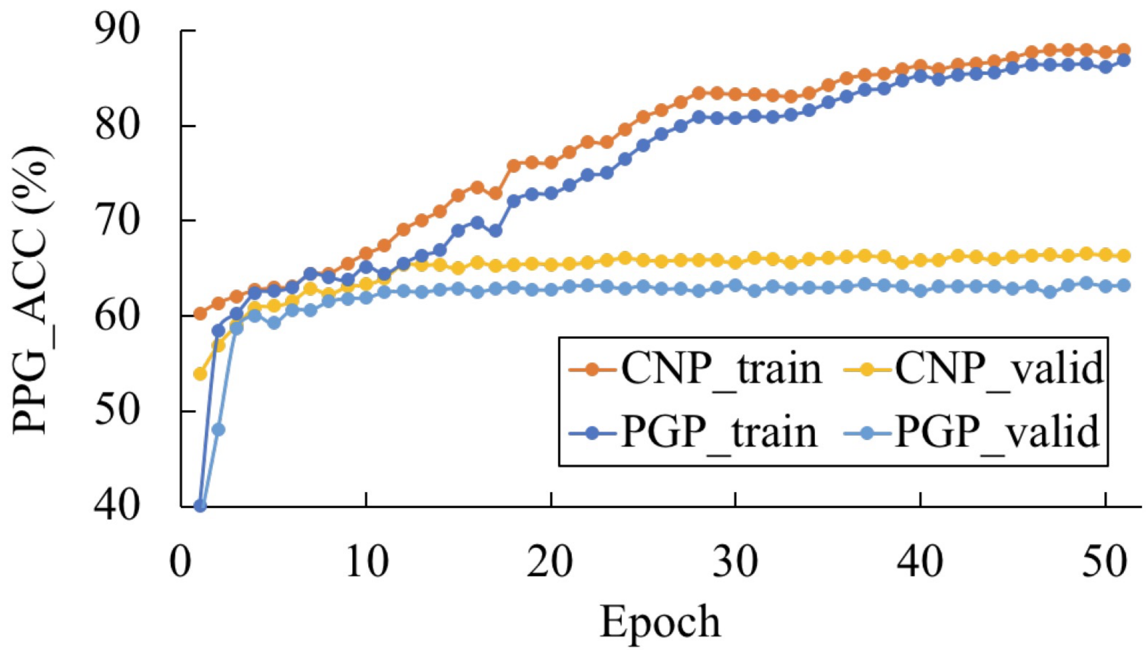


Fig. 2: The PPG_ACCs over training epochs for the CNP and PGP based predictor on the training and valid set.

Experimental results

- Independent front-end processing for ASR
  - Large-scale training with about 1,000 noises from MUSAN dataset
  - 10 untrained noises are used for test

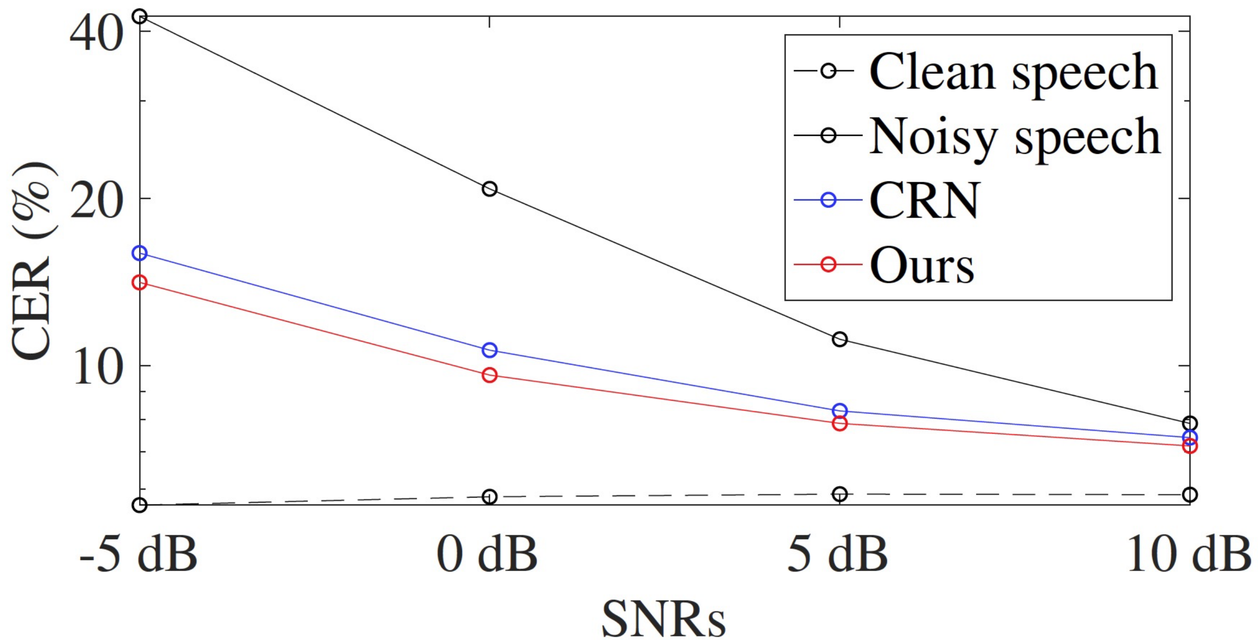


Fig. 3: The CERs of different models on the large-scale set.

Experimental results

- Independent front-end processing for ASR

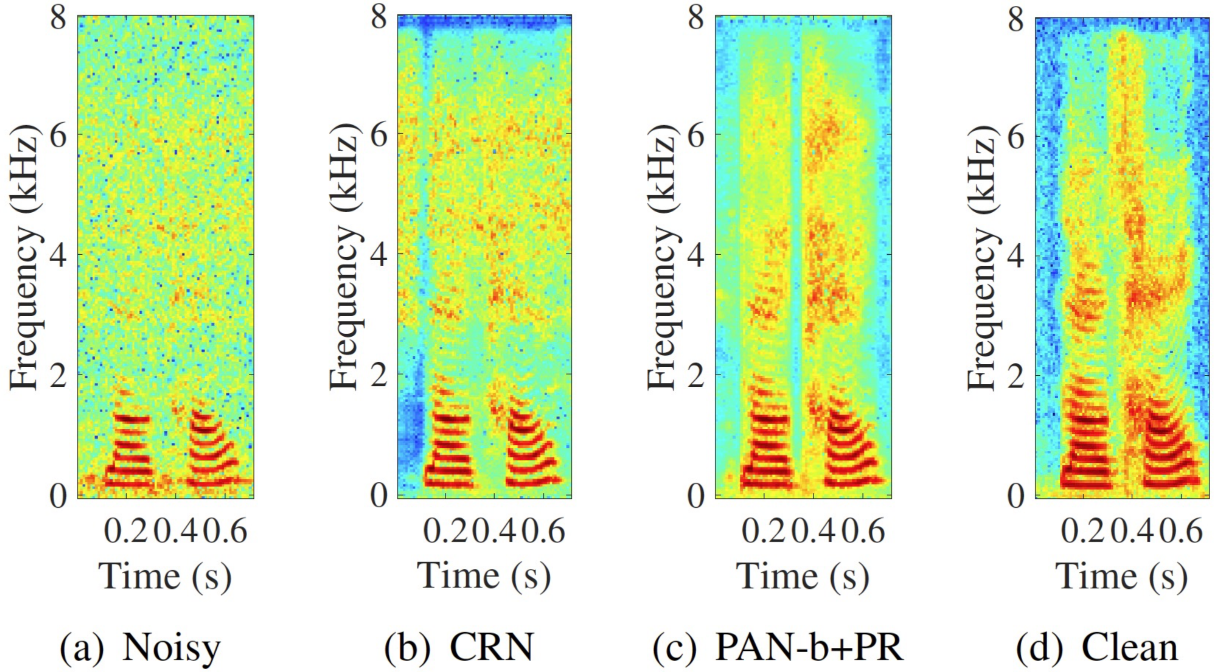


Fig. 4: The power spectrum of (a) noisy speech interfered by the ”buccaneer2” noise at 5dB, (b) enhanced speech by CRN, (c) enhanced speech by PAN-b+PR and (d) clean speech.

Conclusions

- We proposed the PAN to utilize the phonetic information for monaural speech enhancement.
- An iterative algorithm is proposed to train the PAN and PPG predictor.
- We find that correcting the noisy PPG is a better choice than predicting the ground-truth PPG from scratch.
- Experimental results show that utilizing the phonetic information can consistently improve the enhancement performance in terms of STOI, PESQ and CER.

Thanks for your attention!