Mask processing in the time-frequency (T-F) domain through the neural network has been one of the mainstreams for single-channel speech enhancement. However, it is hard for most models to handle the situation when harmonics are partially masked by noise. To tackle this challenge, we propose a harmonic gated compensation network (HGCN). We design a high-resolution harmonic integral spectrum to improve the accuracy of harmonic locations prediction. Then we add voice activity detection (VAD) and voiced region detection (VRD) to the convolutional recurrent network (CRN) to filter harmonic locations. Finally, the harmonic gating mechanism is used to guide the compensation model to adjust the coarse results from CRN to obtain the refinedly enhanced results. Our experiments show HGCN achieves substantial gain over a number of advanced approaches in the community.

Index Terms— Speech Enhancement, Harmonic, Deep Learning, Pitch

2. PROPOSED HGCN

The overall diagram of the proposed system is shown in Fig. 1. It is mainly comprised of three parts, namely the coarse enhancement module (CEM), harmonic locations prediction module (HM), and gated harmonic compensation module (GHCM). CEM performs a coarse enhancement process on noisy speech. Then HM predicts harmonic locations based on the coarse result of the CEM. GHCM compensates for the coarse result based on the harmonic locations to get the refined result. Each module is described as follows.

2.1. Coarse enhancement module

A CRN [2] model is used to do the coarse enhancement process, which is an encoder-decoder architecture. Specifically, both the encoder and decoder are comprised of Batchnormalization (BN) [13], causal 2D convolution blocks (Causal-Conv) [14], and PReLU [15]. Between the encoder and the decoder, long short-term memory (LSTM) [16] is inserted to model the temporal dependencies. Additionally, skip connections are utilized to concatenate the output of each encoder layer to the input of the corresponding decoder layer (red line in Fig. 1). Time-domain waveform and T-F spectrum can be interconverted by STFT and inverse transform (iSTFT). In our model, both STFT and iSTFT are implemented by convolu-
Gated Harmonic Compensation

So, the input to the encoder is the noisy complex spectrum, denoted as \( S = \text{Cat}(S_r, S_i) \in \mathbb{R}^{T \times 2F} \), where \( S_r \) and \( S_i \) represent the real and imaginary parts of the spectrum respectively. And, we compress the input of the encoder with power exponent 0.23 as in [18]. The decoder predicts a complex ratio mask \( M = \text{Cat}(M_r, M_i) \in \mathbb{R}^{T \times 2F} \), where \( M_r \) and \( M_i \) represent the real and imaginary parts of mask. We use the mask applying scheme of DCCRN-E [3], which is called Mask Apply E in Fig. 1.

\[
S' = |S| \odot M_{\text{rms}} \odot e^{j(S_{\text{phase}}+M_{\text{phase}})} = (S_r^2 + S_i^2)^{0.5} \odot M_{\text{rms}} \odot e^{j[\arctan(S_r, S_i) + \arctan(M_r, M_i)]}
\]

(1)

where \( \odot \) denotes the element-wise multiplication operator. \( | \cdot | \) and \( (\cdot)_{\text{phase}} \) represent the magnitude and phase. \( M_{\text{rms}} = \tanh \{(M_r^2 + M_i^2)^{0.5}\} \) is the magnitude mask. \( \tanh \{ \cdot \} \) is the activation function proposed in [19]. \( C_A \) and \( C_B \) in Fig. 1 are introduced in the next section.

2.2. Harmonic locations prediction module

The enhanced result \( S' \) is first decoupled into \( |S'| \) and \( S'_{\text{phase}} \). HM will predict the harmonic locations based on the \( |S'| \).

There are peaks at integer multiples of the pitch and valleys at half-integer multiples, which are the characteristics of harmonics in the magnitude spectrum. Therefore, the pitch candidates can be set first, and the numerical integral of the multiple positions can be taken as the significance of each candidate. The candidate with the highest significance is the pitch \([9,20]\). So, the significance \( Q \) is calculated as,

\[
Q_{t,f} = \sum_{k=1}^{sr/f} \left( \frac{1}{\sqrt{k}} \cdot \log |S'_{t,kf}| - \frac{1}{\sqrt{k}} \cdot \log |S'_{t,(k-\frac{1}{2})f}| \right)
\]

(2)

where \( sr \) is half of the audio sample rate. \( f \) is the pitch candidate. And \( k \) denotes the multiple of the pitch.

For T-F models, 512 Fourier points are often used for audio with 16k sample rate. Since the frequency bandwidth is 31.25 Hz, few pitch candidates can be selected. To solve this problem, a high-resolution integral matrix \( U \) is designed as Algorithm 1 and Fig. 2 where \( \lceil \cdot \rceil \) is a rounding operation. We set the pitch candidates with a resolution of 0.1 Hz, and convert the multiple frequencies to the fixed spectral bins. A total of 3600 pitches in 60~420 Hz (normal pitch range of human) are taken as candidates. Then the Eq. (2) is improved to

\[
Q_t = \log |S'_r| \cdot U^T
\]

(3)

where \( Q_t \in \mathbb{R}^{1 \times 4200} \) denotes the pitch candidate significances of the \( t \)-th frame and the first 600 dimensions are 0. The candidate corresponding to the maximum value in \( Q_t \) is selected as the pitch, and the corresponding harmonic locations are used as the result \( R_H \in \mathbb{R}^{T \times F} \), where the harmonic locations are 1 and the non-harmonic locations are 0.

Fig. 1. Architecture of the proposed HGCN.

Fig. 2. High-resolution harmonic integral matrix \( U \).

The pitch and then harmonic locations for each frame can be predicted by Eq. (3), but in fact, there are no harmonics in non-speech and unvoiced frames, so we apply VAD and VRD to filter \( R_H \) (green and pinkish boxes in Fig. 4). In addition, the energy corresponding to the locations is low even if it’s harmonic (blue box in Fig. 4), which need to be filtered out. Therefore, the final harmonic gate is calculated as follows,

\[
\text{Gate} = R_{\text{VAD}} \odot R_{\text{VRD}} \odot R_A \odot R_H
\]

(4)

where \( R_{\text{VAD}} \in \mathbb{R}^{T \times 1} \) and \( R_{\text{VRD}} \in \mathbb{R}^{T \times 1} \) denote the speech activity frames and voiced frames respectively. \( R_A \in \mathbb{R}^{T \times F} \) denotes the non-low energy locations of speech. \( \mathbb{R}^{T \times 1} \) will be copied and expanded into \( \mathbb{R}^{T \times F} \).

Both VAD and VRD can be judged based on energy, so we design a speech energy detector to predict two non-low speech energy locations spectra \( R_A \) and \( R_B \) with different energy thresholds, where \( R_A \) is designed to filter out the lower energy locations of speech with a smaller threshold, and \( R_B \) is used for VAD and VRD with a larger threshold, which pays more attention to the locations with higher energy. Since the detector needs to be able to resist noise, we change the output channel number of the last CEM decoder to \( 2 + C_A + C_B \), where 2 is the channel number of complex ratio mask for speech enhancement. \( C_A \) and \( C_B \) are the channels number...
of the input \( X' \in \mathbb{R}^{T \times F \times C_{A/B}} \) for fully connected A (FC_A) and B (FC_B) respectively. FC_A and FC_B output 2-D (low-high) classification probabilities \( P_{t,f} = [p_t, p_f] \) for every T-F point \( P \in \mathbb{R}^{T \times F \times 2} \). And the category can be obtained by \( R_{t,f} = \text{argmax}(P_{t,f}) \), then we can obtain the results of the SED \( R_A \in \mathbb{R}^{T \times F} \) and \( R_B \in \mathbb{R}^{T \times F} \).

The labels for the SED are shown in Fig. [3]. We count the mean \( \mu \in \mathbb{R}^{F \times 1} \) of each bin in the logarithmic magnitude on the clean data \( |\hat{S}| = |\hat{S}_1|, \ldots, |\hat{S}_F| \in \mathbb{R}^{D \times T \times F} \), and standard deviation \( \sigma \in \mathbb{R}^{F \times 1} \) of bins are controlled according to different offset values \( \epsilon (\epsilon_A = 0 \text{ and } \epsilon_B = \frac{1}{2}) \), and the label for \( R_{A/B} \) is 1 if the logarithmic magnitude of clean is larger than \( \kappa \), 0 otherwise.

Then we can compute \( R_{\text{VAD}} \) and \( R_{\text{VRD}} \) based on \( R_B \).

\[
(R_{\text{VAD}})_t = \begin{cases} 
1 & \sum_{f=1}^{F} (R_B)_{t,f} > \epsilon \\
0 & \sum_{f=1}^{F} (R_B)_{t,f} \leq \epsilon 
\end{cases} \quad (5)
\]

\[
(R_{\text{VRD}})_t = \begin{cases} 
0 & H > L \\
1 & H \leq L 
\end{cases} \quad (6)
\]

where \( D \) represents the clip number of audio. \(|\hat{S}|\) represents the magnitude spectra. The energy thresholds \( \kappa = (\mu + \epsilon \cdot \sigma) \in \mathbb{R}^{F \times 1} \) of bins are controlled according to different offset values \( \epsilon (\epsilon_A = 0 \text{ and } \epsilon_B = \frac{1}{2}) \), and the label for \( R_{A/B} \) is 1 if the logarithmic magnitude of clean is larger than \( \kappa \), 0 otherwise.

Finally, we convert \( S'' \) into waveform by iSTFT.

### 3. EXPERIMENTS

#### 3.1. Dataset

We evaluate the HGNC on the DNS Challenge (INTER-SPEECH 2020) dataset [21]. This dataset includes 500 hours of clean speech from 2150 speakers. The noise dataset includes over 180 hours from 150 classes. For training, we generate 150 hours of noisy speech. The SNR is between 0 dB and 40 dB. And data is divided into training and validation set at 4 : 1. For testing, the SNR is between 0 dB and 20 dB. And the speech data in the testing doesn’t participate in the training or validation set, the noises are from [22]. A total of 2 hours of test audio are generated.

#### 3.2. Training setup and comparison methods

To ensure comparability, we train all models on our dataset with the same setup. The optimizer is Adam [23]. And the initial learning rate is 0.001, which will decay 50% when the validation loss plateau for 5 epochs and the training is stopped if loss plateau for 20 epochs. The kernel size and stride are (5, 2) and (2, 1). DCRN is utilized as the baseline.
system. And DCCRN \cite{1} is an improved version of DCRN, which ranked first in the Interspeech2020 DNS challenge real-time-track, so it’s utilized as the referenced system.

**DCRN**: The 32ms Hanning window with 25% overlap and 512-point STFT are used. The channel number of encoder and decoder is \{16, 32, 64, 128, 128, 128\}. And a 512-units FC layer after a 128-units LSTM is adopted.

**DCCRN**: The 25ms Hanning window with 25% overlap and 512-point STFT are used. The channel number is \{32, 64, 128, 256, 256, 256\}, and uses two layers complex LSTM with 128 units for real and imaginary parts respectively. And a dense with 1280 units is after the LSTM. And DCCRN looks ahead one frame in each decoder layer.

**HGCN(CEM+GHCM+HM)**: The parameter setting of CEM is the same as DCRN, except that the channel number of last decoder is changed to 22 (\(C_A = C_B = 10\)). Three GCBs are adopted, and their channel numbers and stride are \{8, 16, 8\} and \{1, 1\}. The \(\epsilon\) in Eq. (7) is set to 24. We designed the loss functions for \(S'\) and \(R_{A/B}\) of CEM, \(S''\) of GHCM, separately. For \(S'\), we use APC-SNR \cite{18}. For \(S''\), we use scale-invariant SNR (SI-SNR) \cite{24} and APC-SNR. For \(R_{A/B}\), we use Focal loss \cite{25}.

| Model          | RTF    | PESQ  | STOI(%) | SI-SDR(dB) |
|----------------|--------|-------|---------|------------|
| Noisy          | -      | 1.796 | 93.2    | 10.321     |
| DCRN           | 0.061  | 2.798 | 96.3    | 18.096     |
| DCCRN          | 0.263  | 2.887 | 96.7    | 18.845     |
| CEM            | 0.065  | 2.953 | 96.8    | 18.706     |
| +GHCM          | 0.099  | 3.018 | 97.0    | 18.897     |
| +HM            | 0.109  | 3.096 | 97.2    | 19.255     |

3.3. Experimental results and discussion

We compare the performance of HGCN with comparison methods on the test set, and three objective metrics are utilized in the experiments, namely wide band PESQ (PESQ), STOI, and SI-SDR, as shown in Table \cite{1}. To ensure the generality of the test set. We also did a test on DNS2020 synthetic test set, shown in Table \cite{2}. Compared with DCRN, the performance of the model has been gradually improved with the gradual addition of the CEM, GHCM, and HM modules.

The performance of CEM is improved compared to DCRN, which demonstrates the effectiveness of multi-task training \cite{26}, power compression, and loss function \cite{18}.

GHCM is added on the top of CEM, and only \(R_{A}\) is used as the gate. Although the performance of the model is improved on all indexes, the improvement ratio of CEM+GHCM on PESQ is greater than that on SI-SDR, even in DNS2020 test set, CEM+GHCM is higher than DCCRN on PESQ, but it is lower on SI-SDR. This is due to that the GHCM compensates for the magnitude and retains the phase of the coarse result. It further causes a slight mismatch between magnitude and phase, while PESQ and STOI only care about the magnitude, SI-SDR will be affected by both magnitude and phase. This is why we add SI-SNR to the loss function of \(S''\), otherwise, the effect will be worse.

HGCN (CEM+GHCM+HM) achieves the best results. We visualize the calculation process of the harmonic gate as shown in Fig. \cite{4}. We can observe that the HM can predict the exact harmonic locations, which can better guide the model to compensate for the magnitude spectrum.

Real Time Factor (RTF) is also tested on a machine with an Intel(R) Core(TM) i5-6200U CPU@2.30 GHz in a single thread (implemented by ONNX). We can observe that the proposed model brings better performance while maintaining good speed.

Table 2. System comparison on DNS-2020 synthetic test set.

| Model          | PESQ  | STOI(%) | SI-SDR(dB) |
|----------------|-------|---------|------------|
| Noisy          | 1.582 | 91.5    | 9.071      |
| DCRN           | 2.615 | 95.7    | 17.275     |
| DCCRN          | 2.711 | 96.0    | 17.967     |
| CEM            | 2.753 | 96.1    | 17.539     |
| +GHCM          | 2.812 | 96.3    | 18.414     |
| +HM            | 2.883 | 96.5    | 18.144     |

4. CONCLUSION

In this paper, to tackle the challenge of speech harmonics being partially masked by noise, a harmonic gated compensation network for monaural speech enhancement is proposed. First, we propose a high-resolution speech harmonic integral spectrum, which improves the accuracy of harmonic prediction by increasing the resolution of the predicted pitch. In addition, we design VAD and VRD to filter harmonic locations. Finally, the harmonic gating mechanism is used to guide the model to compensate for the coarse results from CRN to obtain the refinedly enhanced result. The experimental results show that the high-resolution harmonic integral spectrum can predict the harmonic locations accurately, and the HGCN performs better than referenced methods.

1https://github.com/huyanxin/DeepComplexCRN

4. CONCLUSION

In this paper, to tackle the challenge of speech harmonics being partially masked by noise, a harmonic gated compensation network for monaural speech enhancement is proposed. First, we propose a high-resolution speech harmonic integral spectrum, which improves the accuracy of harmonic prediction by increasing the resolution of the predicted pitch. In addition, we design VAD and VRD to filter harmonic locations. Finally, the harmonic gating mechanism is used to guide the model to compensate for the coarse results from CRN to obtain the refinedly enhanced result. The experimental results show that the high-resolution harmonic integral spectrum can predict the harmonic locations accurately, and the HGCN performs better than referenced methods.
5. REFERENCES

[1] Y. Luo and N. Mesgarani, “Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation,” IEEE/ACM Trans. Audio Speech Lang. Proc., vol. 27, no. 8, pp. 1256–1266, 2019.

[2] K. Tan and D. Wang, “Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement,” in Proc. of ICASSP. IEEE, 2019, pp. 6865–6869.

[3] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, “Dccrn: Deep complex convolution recurrent network for phase-aware speech enhancement,” arXiv preprint arXiv:2008.00264, 2020.

[4] A. Li, C. Zheng, L. Zhang, and X. Li, “Glance and gaze: A collaborative learning framework for single-channel speech enhancement,” arXiv preprint arXiv:2106.11789, 2021.

[5] W. Jin, X. Liu, M. S Scordilis, and L. Han, “Speech enhancement using harmonic emphasis and adaptive comb filtering,” IEEE transactions on audio, speech, and language processing, vol. 18, no. 2, pp. 356–368, 2009.

[6] D. Wang and J. Hansen, “Speech enhancement based on harmonic estimation combined with mmse to improve speech intelligibility for cochlear implant recipients.,” in INTERSPEECH, 2017, vol. 1, pp. 186–190.

[7] D. Yin, C. Luo, Z. Xiong, and W. Zeng, “Phasen: A phase-and-harmonics-aware speech enhancement network,” in Proc. of the AAAI Conference on Artificial Intelligence, 2020, vol. 34, pp. 9458–9465.

[8] Y. Wakabayashi, T. Fukumori, M. Nakayama, T. Nishiura, and Y. Yamashita, “Phase reconstruction method based on time-frequency domain harmonic structure for speech enhancement,” in Proc. of ICASSP. IEEE, 2017, pp. 5560–5564.

[9] A. Camacho, SWIPE: A sawtooth waveform inspired pitch estimator for speech and music, University of Florida Gainesville, 2007.

[10] M. Une and R. Miyazaki, “Musical-noise-free speech enhancement with low speech distortion by biased harmonic regeneration technique,” in Proc. of IWAENC. IEEE, 2018, pp. 31–35.

[11] Z. Ouyang, H. Yu, W. Zhu, and B. Champagne, “A deep neural network based harmonic noise model for speech enhancement,” in INTERSPEECH, 2018, pp. 3224–3228.

[12] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. Huang, “Free-form image inpainting with gated convolution,” in Proc. of ICCV, 2019, pp. 4470–4479.

[13] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in Proc. of The 32nd International Conference on Machine Learning, 2015, vol. 1, pp. 448–456.

[14] Wei J., Hu H., He Y., and Lu W., “Dilated causal convolution generative adversarial network end-to-end bone conduction speech blind enhancement method,” 2019.

[15] K. He, X. Zhang, S. Ren, and J. Sun, “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification,” in Proc. of ICCV, 2015, pp. 1026–1034.

[16] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[17] A. Li, C. Zheng, L. Zhang, and X. Li, “Glance and gaze: A collaborative learning framework for single-channel speech enhancement,” arXiv preprint arXiv:2106.11789, 2021.

[18] T. Wang and W. Zhu, “A deep learning loss function based on auditory power compression for speech enhancement,” arXiv preprint arXiv:2108.11877, 2021.

[19] W. Malfllet and W. Hereman, “The tanh method: I. exact solutions of nonlinear evolution and wave equations,” Physica Scripta, vol. 54, no. 6, pp. 563, 1996.

[20] M. R Schroeder, “Period histogram and product spectrum: New methods for fundamental-frequency measurement,” The Journal of the Acoustical Society of America, vol. 43, no. 4, pp. 829–834, 1968.

[21] C. K. A. Reddy, E. Beyrami, H. Dubey, V. Gopal, R. Cheng, R. Cutler, S. Matusveych, R. Aichner, A. Aazami, S. Braun, P. Rana, S. Srinivasan, and J. Gehrke, “The interspeech 2020 deep noise suppression challenge: Datasets, subjective speech quality and testing framework,” arXiv preprint arXiv:2001.08662, 2020.

[22] H. Hirsch and D. Pearce, “The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions,” in ASR2000-Automatic speech recognition: challenges for the new Millennium ISCA tutorial and research workshop (ITRW), 2000.

[23] D. P. Kingma and J. L. Ba, “Adam: A method for stochastic optimization,” in ICLR 2015 : International Conference on Learning Representations 2015, 2015.

[24] J. L. Roux, S. Wisdom, H. Erdogan, and J. R Hershey, “Sdr–half-baked or well done?,” in Proc. of ICASSP. IEEE, 2019, pp. 626–630.

[25] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” in Proc. of the IEEE international conference on computer vision, 2017, pp. 2980–2988.

[26] S. Lv, Y. Hu, S. Zhang, and L. Xie, “Dccrn+: Channel-wise subband dccrn with snr estimation for speech enhancement,” arXiv preprint arXiv:2106.08672, 2021.