HGCN: HARMONIC GATED COMPENSATION NETWORK FOR SPEECH ENHANCEMENT

Tianrui Wang[⋆†]*, Weibin Zhu*[⋆]*, Yingying Gao*[†]*, Junlan Feng*[†]*, Shilei Zhang*[†]

[⋆] Institute of Information Science, Beijing Jiaotong University, Beijing, China
[†] China Mobile Research Institute, Beijing, China

ABSTRACT

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

1. INTRODUCTION

Speech enhancement aims to improve speech quality by using various algorithms. In recent years, deep learning methods have been applied and achieved promising results in this area. These models could be divided into two main categories, time-domain (T) models and time-frequency domain (T-F) models. T models process the waveform directly to obtain the target speech [1, 2]. T-F models precess the spectrum after the short-time fast Fourier transform (STFT) [3–5]. Generally speaking, for speech enhancement, it's the T-F structure of speech that is enhanced. In some sense, the processing of the comb harmonic structure of speech constitutes the basis of T-F models [6, 7]. However, in the case of low SNR, the harmonic structure may be masked severely by noise. [8] constructs a frequency domain transformation structure to capture harmonic correlations. [9] borrows harmonic enhancement to reconstruct phase. But neither of them explicitly considers the reconstruction of harmonics.

In principle, the harmonics can be obtained directly from the pitch, while the pitch can be obtained via spectral integral [10]. Since the harmonic structure will seldom be completely masked on the spectra even if speech is seriously corrupted by noise, the pitch predicted by spectral integral is reliable. However, **i)** the frequency resolution after STFT in the deep learning model is fixed and low, which causes the prediction of the pitch by the former spectral integral to be less accurate. And **ii)** the magnitude values that need to be compensated in the harmonic locations are difficult to be obtained [11, 12].

In this paper, a harmonic gated compensation network (HGCN) is proposed. To tackle challenge **i)**, we increase the resolution of the pitch candidates and propose a high-resolution harmonic integral spectrum. To tackle challenge **ii)**, we design a gated [13] compensation module to adjust the magnitude of harmonic. In addition, we design a speech energy detector (SED) to do VAD and VRD, which are used to filter harmonic locations. The experimental results show that each sub-module brings a performance improvement, and the proposed method performs better than referenced ones.

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 [3] model is used to do the coarse enhancement process, which is an encoder-decoder architecture. Specifically, both the encoder and decoder are comprised of Batchnormlization (BN) [14], causal 2D convolution blocks (CausalConv) [15], and PReLU [16]. Between the encoder and the decoder, long short-term memory (LSTM) [17] 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-

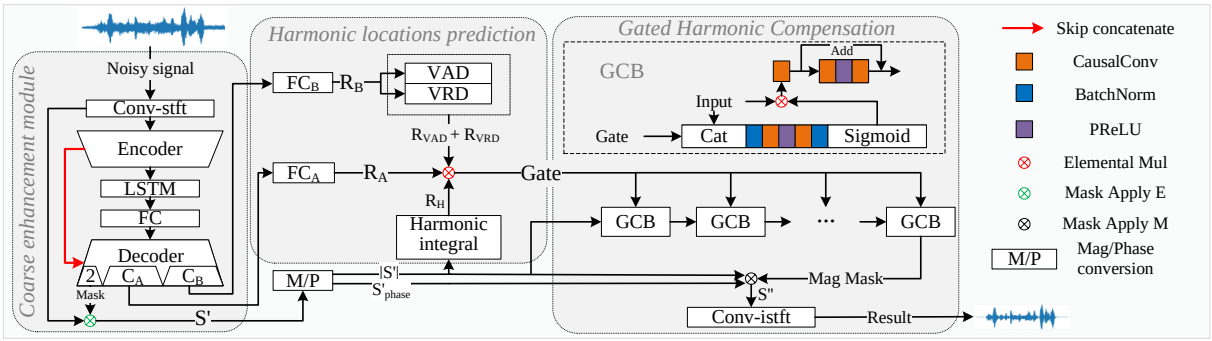

Fig. 1. Architecture of the proposed HGCN.

tion [18]. So, the input to the encoder is the noisy complex spectrum, denoted as $\boldsymbol{S} = \text{Cat}(\boldsymbol{S}_r, \boldsymbol{S}_i) \in \mathbb{R}^{T\times 2F}$, where $\boldsymbol{S}_r$ and $\boldsymbol{S}_i$ 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 [19]. The decoder predicts a complex ratio mask $\boldsymbol{M} = \text{Cat}(\boldsymbol{M}_r, \boldsymbol{M}_i) \in \mathbb{R}^{T\times 2F}$, where $\boldsymbol{M}_r$ and $\boldsymbol{M}_i$ represent the real and imaginary parts of mask. We use the mask applying scheme of DCCRN-E [4], which is called Mask Apply E in Fig. 1,

$$\begin{aligned}\boldsymbol{S}^{'} &= |\boldsymbol{S}| \odot \boldsymbol{M}_m \odot e^{j(\boldsymbol{S}_{\text{phase}}+\boldsymbol{M}_{\text{phase}})} \\ &= (\boldsymbol{S}_r^2 + \boldsymbol{S}_i^2)^{0.5} \odot \boldsymbol{M}_m \odot e^{j[\arctan(\boldsymbol{S}_i,\boldsymbol{S}_r)+\arctan(\boldsymbol{M}_i,\boldsymbol{M}_r)]}\end{aligned} \tag{1}$$

where $\odot$ denotes the element-wise multiplication operator. $|\cdot|$ and $(\cdot)_{\text{phase}}$ represent the magnitude and phase. $\boldsymbol{M}_m = \tanh\{(\boldsymbol{M}_r^2 + \boldsymbol{M}_i^2)^{0.5}\}$ is the magnitude mask. $\tanh\{\cdot\}$ is the activation function proposed in [20]. C_A and C_B in Fig. 1 are introduced in the next section.

2.2. Harmonic locations prediction module

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

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 [10, 21]. So, the significance $\boldsymbol{Q}$ is calculated as,

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

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 31.25Hz, few pitch candidates can be selected. To solve this problem, a high-resolution integral matrix $\boldsymbol{U}$ is designed as Algorithm 1 and Fig. 2, where $[\cdot]$ is a rounding operation. We set the pitch candidates with a resolution of 0.1Hz, and convert the multiple frequencies to the fixed spectral bins. A total of 3600 pitches in 60~420Hz (normal pitch range of human) are taken as candidates. Then the Eq. (2) is improved to

$$\boldsymbol{Q}_t = \log|\boldsymbol{S}^{'}_t| \cdot \boldsymbol{U}^\top \tag{3}$$

where $\boldsymbol{Q}_t \in \mathbb{R}^{1\times 4200}$ 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 $\boldsymbol{Q}_t$ is selected as the pitch, and the corresponding harmonic locations are used as the result $\text{R}_\text{H} \in \mathbb{R}^{T\times F}$, where the harmonic locations are 1 and the non-harmonic locations are 0.

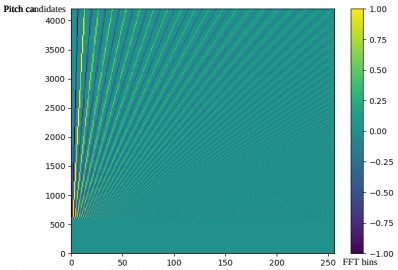

Fig. 2. High-resolution harmonic integral matrix $\boldsymbol{U}$.

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,

$$\text{Gate} = \text{R}_\text{VAD} \odot \text{R}_\text{VRD} \odot \text{R}_\text{A} \odot \text{R}_\text{H} \tag{4}$$

where $\text{R}_\text{VAD} \in \mathbb{R}^{T\times 1}$ and $\text{R}_\text{VRD} \in \mathbb{R}^{T\times 1}$ denote the speech activity frames and voiced frames respectively. $\text{R}_\text{A} \in \mathbb{R}^{T\times F}$ 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 $(2 + \text{C}_\text{A} + \text{C}_\text{B})$, 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 $\boldsymbol{X}^{'} \in \mathbb{R}^{T\times F\times \mathrm{C_{A/B}}}$ for fully connected A ($\mathrm{FC_A}$) and B ($\mathrm{FC_B}$) respectively. $\mathrm{FC_A}$ and $\mathrm{FC_B}$ output 2-D (low-high) classification probabilities $P_{t,f} = [p_0, p_1]$ for every T-F point $\boldsymbol{P} \in \mathbb{R}^{T\times F\times 2}$. And the category can be obtained by $\mathrm{R}_{t,f} = \mathrm{argmax}(P_{t,f})$, then we can obtain the results of the SED $\mathrm{R_A} \in \mathbb{R}^{T\times F}$ and $\mathrm{R_B} \in \mathbb{R}^{T\times F}$.

The labels for the SED are shown in Fig. 3. We count the mean $\boldsymbol{\mu} \in \mathbb{R}^{F\times 1}$ of each bin in the logarithmic magnitude on the clean datas $|\dot{\mathbb{S}}| = [|\dot{\boldsymbol{S}}|_1, \cdots, |\dot{\boldsymbol{S}}|_d] \in \mathbb{R}^{D\times T\times F}$, and standard deviation $\boldsymbol{\sigma} \in \mathbb{R}^{Q\times 1}$ of means,

$$\boldsymbol{\mu} = \sum_{i=1}^{D}\left[\left(\sum_{t=1}^{T}\log|\dot{\mathbb{S}}|_{i,t}\right)/T\right]/D \tag{5}$$

$$\boldsymbol{\sigma} = \sqrt{\sum_{i=1}^{D}\left[\left(\sum_{t=1}^{T}\log|\dot{\mathbb{S}}|_{i,t}\right)/T - \boldsymbol{\mu}\right]^2/D} \tag{6}$$

where D represents the clip number of audio. $|\dot{\boldsymbol{S}}|$ represents the magnitude spectra. The energy thresholds $\boldsymbol{\kappa} = (\boldsymbol{\mu} + \varepsilon \cdot \boldsymbol{\sigma}) \in \mathbb{R}^{F\times 1}$ of bins are controlled according to different offset values ε ($\varepsilon_{\mathrm{A}} = 0$ and $\varepsilon_{\mathrm{B}} = \frac{4}{3}$), and the label for $\mathrm{R_{A/B}}$ is 1 if the logarithmic magnitude of clean is larger than $\boldsymbol{\kappa}$, 0 otherwise.

Then we can compute $\mathrm{R_{VAD}}$ and $\mathrm{R_{VRD}}$ based on $\mathrm{R_B}$,

$$(\mathrm{R_{VAD}})_t = \begin{cases} 1 & ,\sum_{f=1}^{F}(\mathrm{R_B})_{t,f} > \epsilon \\ 0 & ,\sum_{f=1}^{F}(\mathrm{R_B})_{t,f} \le \epsilon \end{cases} \tag{7}$$

$$(\mathrm{R_{VRD}})_t = \begin{cases} 0 & ,\mathrm{H} > \mathrm{L} \\ 1 & ,\mathrm{H} \le \mathrm{L} \end{cases} \tag{8}$$

where $\mathrm{H} = \sum_{f=F/2}^{F}(\mathrm{R_B})_{t,f}$ and $\mathrm{L} = \sum_{f=1}^{F/2}(\mathrm{R_B})_{t,f}$ denote the number of selected speech points in high and low frequency respectively. ϵ is the threshold for VAD.

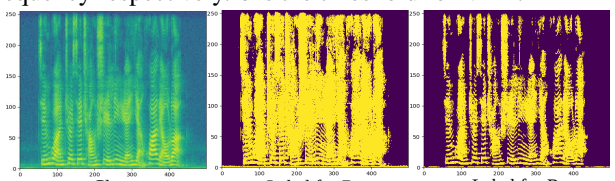

Fig. 3. Labels for speech energy detection. $\mathrm{R_A}$ is to filter out the lower energy locations with a smaller threshold, $\mathrm{R_B}$ focus on the higher energy locations with a larger threshold.

2.3. Gated harmonic compensation module

A gated mechanism [13] is used to guide the model to compensate for the coarse result $\boldsymbol{S}^{'}$ of CEM. The GHCM is composed of multiple gated compensation blocks (GCB) in series to predict the magnitude compensation mask, where GCB is composed of gated convolution layer (GCL) and residual convolution (RC). The input $\boldsymbol{X}_{\mathrm{in}}$ of the first GCB is $|\boldsymbol{S}^{'}|$, and the subsequent input is the output of the previous one.

The GCB introduce the gate mechanism during convolution. As shown in Fig. 1, we first obtain an attention map $\boldsymbol{\alpha} \in \mathbb{R}^{T\times F}$ by concatenating gate Eq. (4) and the input feature in channel followed by $\mathrm{CB}_{1\times 1}$,

$$\boldsymbol{\alpha} = \mathrm{sigmoid}\left(\mathrm{CB}_{1\times 1}(\mathrm{Cat}(\mathrm{Gate}, \boldsymbol{X}_{\mathrm{in}}))\right) \tag{9}$$

Algorithm 1 Integral matrix

$\boldsymbol{U} \leftarrow \boldsymbol{0} \in \mathbb{R}^{4200\times F}$
for $f \leftarrow 600 \rightarrow 4200$ **do**
  last_index $\leftarrow 0$
  for $k \leftarrow 1 \rightarrow [\mathrm{sr}/f]$ **do**
    index $\leftarrow [f \cdot k \cdot F/\mathrm{sr}]$
    $\boldsymbol{U}_{f,\mathrm{index}} \leftarrow \boldsymbol{U}_{f,\mathrm{index}} + (1/\sqrt{k})$
    if index − last_index > 1 **then**
      $i \leftarrow [(\mathrm{index} - \mathrm{last\_index})/2]$
      if $(\mathrm{index} - \mathrm{last\_index}) \mod 2 \neq 0$ **then**
        $\boldsymbol{U}_{f,i} \leftarrow \boldsymbol{U}_{f,i} - 1/(2\sqrt{k})$
        $\boldsymbol{U}_{f,i+1} \leftarrow \boldsymbol{U}_{f,i+1} - 1/(2\sqrt{k})$
      else
        $\boldsymbol{U}_{f,i} \leftarrow \boldsymbol{U}_{f,i} - 1/\sqrt{k}$
    else
      $\boldsymbol{U}_{f,\mathrm{index}} \leftarrow \boldsymbol{U}_{f,\mathrm{index}} - 1/(2\sqrt{k})$
      $\boldsymbol{U}_{f,\mathrm{last\_index}} \leftarrow \boldsymbol{U}_{f,\mathrm{last\_index}} - 1/(2\sqrt{k})$

where $\mathrm{CB}_{1\times 1}$ is comprised of BN, CausalConv, and PReLU. Secondly, GCL applies the $\boldsymbol{\alpha}$ to the magnitude as $\tilde{\boldsymbol{X}} = \boldsymbol{X}_{\mathrm{in}} \odot \boldsymbol{\alpha}$, then fed $\tilde{\boldsymbol{X}}$ into a convolutional layer. Finally, an RC follows the GCL and does a compensation process.

In GCBs, PReLU is used as the activation function, except for the last block which uses sigmoid to predict the compensation magnitude mask $\boldsymbol{M}_{\mathrm{GHCM}} \in \mathbb{R}^{T\times F}$. The magnitude mask applying is used and call it Mask Apply M,

$$\boldsymbol{S}^{''} = (|\boldsymbol{S}^{'}| + \boldsymbol{M}_{\mathrm{GHCM}} \odot |\boldsymbol{S}^{'}|) \odot e^{j\boldsymbol{S}^{'}_{\mathrm{phase}}} \tag{10}$$

Finally, we convert $\boldsymbol{S}^{''}$ into waveform by iSTFT.

3. EXPERIMENTS

3.1. Dataset

We evaluate the HGCN on the DNS Challenge (INTERSPEECH 2020) dataset [22]. 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 0dB and 40dB. And data is divided into training and validation set at $4:1$. For testing, the SNR is between 0dB and 20dB. And the speech data in the testing doesn't participate in the training or validation set, the noises are from the ESC-50 [23]. 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 [24]. 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

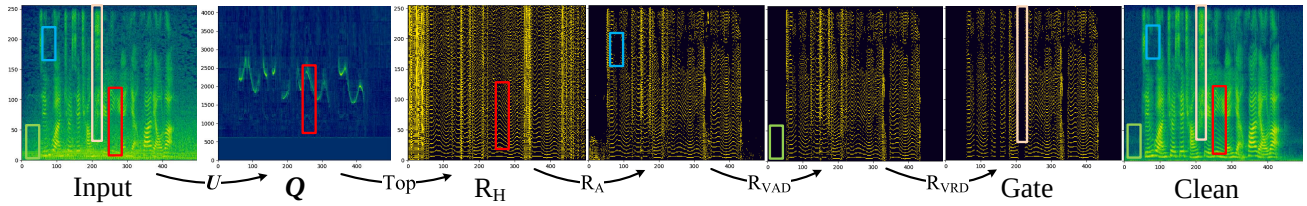

Fig. 4. The calculation process of harmonic gate.

system. And DCCRN [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 are $\{8, 16, 8\}$. The ϵ 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 [19]. For $S^{''}$, we use scale-invariant SNR (SI-SNR) [25] and APC-SNR. For $R_{A/B}$, we use Focal loss [26].

Table 1. System comparison on the test set.

Model	RTF	PESQ	STOI(%)	SI-SDR(dB)
Noisy	-	1.796	0.932	10.321
DCRN	**0.061**	2.798	0.963	18.096
DCCRN	0.263	2.887	0.967	18.845
CEM	0.065	2.953	0.968	18.706
+GHCM	0.099	3.018	0.970	18.897
+HM	0.109	**3.096**	**0.972**	**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 1. To ensure the generality of the test set. We also did a test on DNS2020 synthetic test set, shown in Table 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 [27], power compression, and loss function [19].

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

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. 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. 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	0.915	9.071
DCRN	2.615	0.957	17.275
DCCRN	2.711	0.960	17.967
CEM	2.753	0.961	17.539
+GHCM	2.812	0.963	17.841
+HM	**2.883**	**0.965**	**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 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] A. Défossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain.," in *Interspeech 2020*, 2020, pp. 3291–3295.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

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

[18] R. Gu, J. Wu, S. Zhang, L. Chen, Y. Xu, M. Yu, D. Su, Y. Zou, and D. Yu, "End-to-end multi-channel speech separation," *arXiv preprint arXiv:1905.06286*, 2019.

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

[20] W. Malfliet and W. Hereman, "The tanh method: I. exact solutions of nonlinear evolution and wave equations," *Physica Scripta*, vol. 54, no. 6, pp. 563, 1996.

[21] 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.

[22] C. K. A. Reddy, E. Beyrami, H. Dubey, V. Gopal, R. Cheng, R. Cutler, S. Matusevych, 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.

[23] K. J. Piczak, "ESC: Dataset for Environmental Sound Classification," in *Proc. of the 23rd Annual ACM Conference on Multimedia*. pp. 1015–1018, ACM Press.

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

[25] 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.

[26] 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.

[27] 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.