ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9747647

HARMONIC GATED COMPENSATION NETWORK PLUS FOR ICASSP 2022 DNS CHALLENGE

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

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

ABSTRACT

The harmonic structure of speech is resistant to noise, but the harmonics may still be partially masked by noise. Therefore, we previously proposed a harmonic gated compensation network (HGCN) to predict the full harmonic locations based on the unmasked harmonics and process the result of a coarse enhancement module to recover the masked harmonics. In addition, the auditory loudness loss function is used to train the network. For the DNS Challenge, we update HGCN with the following aspects, resulting in HGCN+. First, a high-band module is employed to help the model handle full-band signals. Second, cosine is used to model the harmonic structure more accurately. Then, the dual-path encoder and dual-path rnn (DPRNN) are introduced to take full advantage of the features. Finally, a gated residual linear structure replaces the gated convolution in the compensation module to increase the receptive field of frequency. The experimental results show that each updated module brings performance improvement to the model. HGCN+ also outperforms the referenced models on both wide-band and full-band test sets.

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

1. INTRODUCTION

Speech enhancement (SE) aims to improve speech quality. Many researchers introduce intuitive signal processing ideas to deep learning [1, 2]. [3] explores an SE model combining signal processing and deep learning, but simple structure limits its performance. [4] introduces the complex operations, which improve the performance but with less auditory characteristic. [5, 6] introduce the auditory feature, but spectra with wide frequency bandwidth degrade the performance. [7] verifies the effectiveness of the hearing pipeline structure. [8] verifies the necessity of modeling the harmonic. Inspired by these works, we proposed the HGCN for SE [9]. In HGCN, a high-resolution harmonic integration algorithm is proposed to predict the harmonic locations. Then the locations are used as a gate to help the subsequent module compensate for the result of the coarse enhancement module to obtain a refined result. To make the enhancement more consistent with human hearing, we previously proposed a loss function based on auditory loudness power compression (APC-SNR) [10]. Both HGCN and APC-SNR have proved their effectiveness for SE.

Compared to the HGCN, each module of the HGCN+ is updated and the HGCN+ can handle full-band (FB, 0~24 KHz) signals. Since the wide-band (WB, 0~8 KHz) is more likely to contain high energies, tonalities and long sustained sounds, while the high-band (HB, 8~24 KHz) tends to have low energies, noise and rapidly decaying sounds [11], the HB and WB spectra are modeled separately [5]. The HB spectrum is enhanced by a lightweight NSNet [2] and the WB spectrum is enhanced by an HGCN that is updated in the following aspects. 1) The dual-path encoder and DPRNN [12, 13] are introduced to take full advantage of the features. 2) Cosine is adopted to model the harmonic peak-valley structure, and the voiced region detection (VRD) is judged based on the harmonic integration significance. 3) The gated convolution is replaced by a residual gated structure comprised of linear layers and Gated Recurrent Units (GRUs) [14] to increase the receptive field of frequency. Since we model the WB and HB spectra separately, HGCN+ can h andle both WB and FB signals without resampling. Experimental results show that HGCN+ outperforms the referenced methods on test sets.

2. PROPOSED HGCN+

The overall diagram of the HGCN+ is as shown in Fig. 1, which is comprised of four parts, namely the high-band module (HBM), coarse enhancement module (CEM), harmonic locations prediction module (HM), and gated harmonic compensation module (GHCM). The noisy signal is split into WB and HB spectra after a short-time Fourier transform (STFT). The HBM enhances the HB spectrum. And the WB spectrum is firstly passed to CEM to obtain a coarse result. Subsequently, HM predicts harmonic locations based on the coarse result. Then, GHCM compensates for the coarse result based on the harmonic location gates to get the refined WB result. Finally, the enhanced WB and HB spectra are concatenated and converted to waveform by an inverse STFT (iSTFT).

⋆Work done during internship at China Mobile Research Institute.

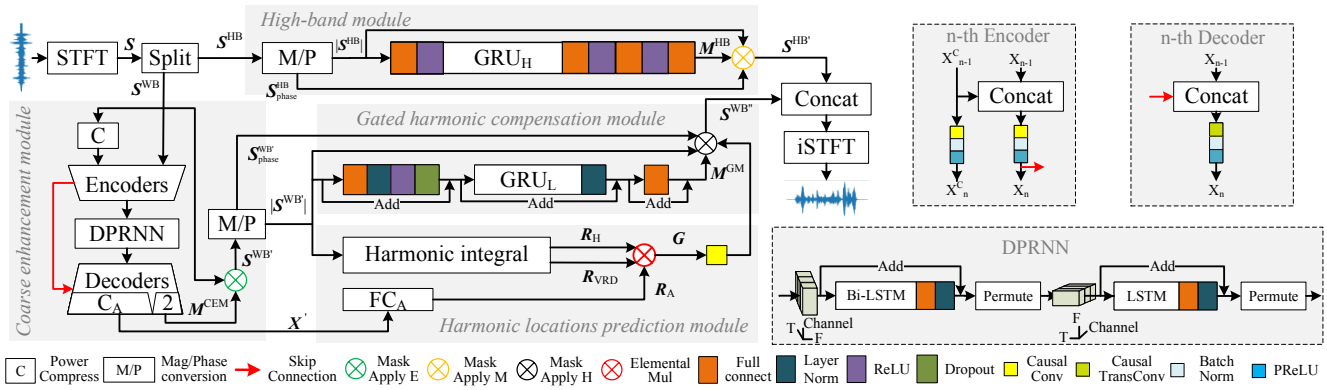


Fig. 1. Architecture of the proposed HGCN+.

2.1. High-band model

HB spectra contain less speech information, and elaborate operation on them could bring a computational burden, so a lightweight magnitude module [2] is used. The HB spectrum $\boldsymbol{S}^{\text{HB}} = \text{Cat}(\boldsymbol{S}_r^{\text{HB}}, \boldsymbol{S}_i^{\text{HB}}) \in \mathbb{R}^{T\times 2F}$ is firstly converted to $|\boldsymbol{S}^{\text{HB}}|$ and $\boldsymbol{S}_{\text{phase}}^{\text{HB}}$, where $|\boldsymbol{S}^{\text{HB}}| = (\boldsymbol{S}_r^{\text{HB}\,2} + \boldsymbol{S}_i^{\text{HB}\,2})^{0.5}$ and $\boldsymbol{S}_{\text{phase}}^{\text{HB}} = \arctan(\boldsymbol{S}_i^{\text{HB}}, \boldsymbol{S}_r^{\text{HB}})$ are the magnitude and phase, $\boldsymbol{S}_r^{\text{HB}}$ and $\boldsymbol{S}_i^{\text{HB}}$ are the real and imaginary parts, T and F denote the number of frames and the STFT bins respectively. The magnitude is used as input to predict the mask $\boldsymbol{M}^{\text{HB}}$. And HBM consists of fully connecting (FC), GRUs and ReLU [15], as shown in Fig. 1. Finally, the HB enhanced result $\boldsymbol{S}^{\text{HB}'}$ is obtained by the Mask Apply M as follow,

$$\boldsymbol{S}^{\text{HB}'} = |\boldsymbol{S}^{\text{HB}}| \odot \sigma(\boldsymbol{M}^{\text{HB}}) \odot e^{j\boldsymbol{S}_{\text{phase}}^{\text{HB}}} \tag{1}$$

where $\odot$ is the element-wise multiplication, $\sigma(\cdot)$ is sigmoid.

2.2. Coarse enhancement module

The noise with higher energy than speech is catastrophic for HM. So CEM is designed to roughly suppress the noise of WB spectra, which is an encoder-decoder structure. Before being input to the encoder, the WB spectrum $\boldsymbol{S}^{\text{WB}} = \text{Cat}(\boldsymbol{S}_r^{\text{WB}}, \boldsymbol{S}_i^{\text{WB}})$ is compressed by a power of 0.23 and is input to the dual-path encoder together with the original spectrum, as shown in Fig. 1. Both the encoder and decoder are comprised of 2D causal convolution, batch normalization [16], and PReLU [17]. Between the encoder and decoder, DPRNN [12, 13] is inserted to model the multidimensional dependencies and Skip Connection concatenates the output of each encoder to the input of the corresponding decoder (red line in Fig. 1). The output channel of the last decoder is $(\text{C}_\text{A} + 2)$, where 2 is the estimated mask of CEM $\boldsymbol{M}^{\text{CEM}} = \text{Cat}(\boldsymbol{M}_r^{\text{CEM}}, \boldsymbol{M}_i^{\text{CEM}})$, C_A is introduced in the next section. The Mask Apply E of CEM is as follows,

$$\boldsymbol{S}^{\text{WB}'} = |\boldsymbol{S}^{\text{WB}}| \odot \tanh(|\boldsymbol{M}^{\text{CEM}}|) \odot e^{j(\boldsymbol{S}_{\text{phase}}^{\text{WB}} + \boldsymbol{M}_{\text{phase}}^{\text{CEM}})} \tag{2}$$

2.3. Harmonic locations prediction module

The harmonics that are masked by noise can be deduced from the unmasked harmonics (red box in Fig. 3). [18,19] proposed to model the peak-valley structure of the harmonics on spectra to detect the pitch. The pitch candidates are set first, and the integral of the multiple locations is taken as the significance Q_{t,f_c} of each candidate f_c,

$$Q_{t,f_c} = \sum_{k=1}^{8000/f_c} \left(\frac{1}{\sqrt{k}} \cdot |S_{t,kf_c}^{\text{WB}'}|^{0.5} - \frac{1}{\sqrt{k}} |S_{t,(k-\frac{1}{2})f_c}^{\text{WB}'}|^{0.5}\right) \tag{3}$$

where k denotes the multiple of the pitch. The candidate with the highest significance is regarded as the pitch.

To detect the pitch with fine-resolution based on STFT bins with wide-bandwidth, we proposed a high-resolution harmonic integration matrix $\boldsymbol{U}$ in HGCN, which sets candidates in 60~420 Hz (normal pitch range of speech) with a resolution of 0.1 Hz. In this paper, the peak-valley modeling is improved by cosine function, and the $\boldsymbol{U}$ is designed as Algorithm 1 and Fig. 2(a), where $[\cdot]$ is a rounding operation, $\text{linspace}(a, b, c)$ generates an arithmetic progression between a and b of length c. Then the Eq. (3) is updated to,

$$\boldsymbol{Q}_t = |\boldsymbol{S}_t^{\text{WB}'}|^{0.5} \cdot \boldsymbol{U}^\top \tag{4}$$

where $\boldsymbol{Q}_t \in \mathbb{R}^{1\times 3600}$ denotes the pitch candidate significances of the t-th frame. The candidate corresponding to the maximum value in $\boldsymbol{Q}_t$ is regarded as the pitch, and then the corresponding harmonic peak-valley structure is deduced based on the pitch as $\boldsymbol{R}_\text{H} \in \mathbb{R}^{T\times F}$, as shown in Fig. 3.

There is no harmonic in unvoiced and silent frames (orange box in Fig. 3), so we apply the voiced region detector (VRD) to filter $\boldsymbol{R}_\text{H}$. In addition, the energy is low even if it's harmonic (gray box in Fig. 3), which needs to be filtered out. Therefore, the final harmonic gate $\boldsymbol{G}$ is calculated as follows,

$$\boldsymbol{G} = \boldsymbol{R}_{\text{VRD}} \odot \boldsymbol{R}_\text{A} \odot \boldsymbol{R}_\text{H} \tag{5}$$

where $\boldsymbol{R}_\text{A} \in \mathbb{R}^{T\times F}$ denotes the non-low energy locations of speech and is detected by a speech energy detector (SED). Since SED needs to resist noise, we change the output channel number of the last CEM decoder to $(2 + \text{C}_\text{A})$, and C_A is the channels number of the input $\boldsymbol{X}' \in \mathbb{R}^{T\times F\times \text{C}_\text{A}}$ for an FC layer (FC_A) to get a 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 $\boldsymbol{R}_\text{A}$ is obtained by $\boldsymbol{R}_{t,f} = \text{argmax}(P_{t,f})$.

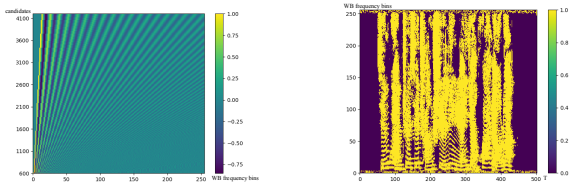
The SED is designed to filter out the low energy parts, so we generate labels for SED based on energy. The mean of

each bin in the clean logarithmic magnitude is counted as $\boldsymbol{\mu} = \sum_{t=1}^{T} \log|\dot{\boldsymbol{S}}_t|/T$, where $|\dot{\boldsymbol{S}}|$ represents the clean magnitude. And the label is 1 if the logarithmic magnitude of clean is larger than $\boldsymbol{\mu} \in \mathbb{R}^{F\times 1}$, 0 otherwise, as shown in Fig. 2(b).

The significances of the voiced frames are higher than that of the unvoiced and silent frames in the integral spectrum $\boldsymbol{Q}$, as shown in Fig. 3. So the VRD is designed as,

$$(\boldsymbol{R}_{\text{VRD}})_t = I(\max(\boldsymbol{Q}_t) > (\alpha \cdot \xi)) \tag{6}$$

where $\max(\cdot)$ denotes the maximum value in the vector. If the input is true, $I(\cdot)$ outputs 1. ξ is the moving average which is updated as $\xi_{\text{new}} = 0.9\xi_{\text{old}} + 0.1\sum_{t=1}^{T}\max(\boldsymbol{Q}_t)/T$. α is the scale factor (α is 0.4 in our experiments).



(a) (b)

Fig. 2. (a) High-resolution harmonic integral matrix $\boldsymbol{U}$. (b) Labels for speech energy detection.

2.4. Gated harmonic compensation module

To recover the masked harmonics with the help of the unmasked harmonics, the module needs to have a wide receptive field of frequency, so a gated residual structure consisting of FCs and GRUs is adopted to compensate the coarsely processed spectra according to the harmonic locations $\boldsymbol{G}$, as shown in Fig. 1. And the harmonic gated compensation mask applying (Mask Apply H) is used as,

$$\boldsymbol{S}^{\text{WB}''} = \left[1 + \text{CC}(\boldsymbol{G}) \odot \sigma(\boldsymbol{M}^{\text{GM}})\right] \odot |\boldsymbol{S}^{\text{WB}'}| \odot e^{j\boldsymbol{S}^{\text{WB}'}_{\text{phase}}} \tag{7}$$

where $\boldsymbol{M}^{\text{GM}}$ is the estimated mask of GHCM. Since some harmonic peaks maybe masked by noise, the magnitude that needs to be compensated is not a complete harmonic structure, a causal convolution (CC) is used to process the gate.

Finally, $\boldsymbol{S}^{\text{WB}''}$ and $\boldsymbol{S}^{\text{HB}'}$ are concatenated into FB complex spectra and converted to the waveform by iSTFT.

2.5. Loss function

The magnitude loss is used to model the HB magnitude. To make the WB enhancement results consistent with the human hearing, we introduce the time-domain SI-SNR to measure the complex spectrum compressed by the auditory loudness power exponent $\boldsymbol{\gamma}$ [10]. In addition, FocalLoss [20] is used as a loss function for the SED. The loss function of the whole model is defined as,

$$L = L_{\text{HB}} + L_{\text{APC}}^{\boldsymbol{S}^{\text{WB}'}} + L_{\text{APC}}^{\boldsymbol{S}^{\text{WB}''}} + L_{\text{focal}} \tag{8}$$

$$L_{\text{HB}} = \left\| |\boldsymbol{S}^{\text{HB}'}| - |\dot{\boldsymbol{S}}^{\text{HB}}| \right\|^2 + \left\| \log|\boldsymbol{S}^{\text{HB}'}| - \log|\dot{\boldsymbol{S}}^{\text{HB}}| \right\|^2 \tag{9}$$

$$L_{\text{focal}} = -\alpha(1 - P_{t,f})^{\beta} \cdot \log P_{t,f} \tag{10}$$

Algorithm 1 Integral matrix

$\boldsymbol{U} \leftarrow \boldsymbol{0} \in \mathbb{R}^{3600\times F}$
for $f_c \leftarrow 600 \rightarrow 4200$ **do**
  $\text{loc}_{\text{last}} \leftarrow 0$; $\text{peak}_{\text{last}} \leftarrow 1$; $j \leftarrow f_c - 600$
  for $k \leftarrow 1 \rightarrow [8000/(0.1 \cdot f_c)]$ **do**
    $\text{loc} \leftarrow [0.1 \cdot f_c \cdot k \cdot F/8000]$
    $\text{peak} \leftarrow 1/\sqrt{k}$; $\boldsymbol{U}_{j,\text{loc}} \leftarrow \text{peak}$
    if $\text{loc} - \text{loc}_{\text{last}} > 1$ **then**
      $\text{num}_{\text{iner}} = \text{loc} - \text{loc}_{\text{last}}$
      $\boldsymbol{F}^{\cos} \leftarrow \cos(\text{linspace}(0, 2\pi, \text{num}_{\text{iner}}))$
      $\boldsymbol{F} \leftarrow \text{linspace}(\text{peak}_{\text{last}}, \text{peak}, \text{num}_{\text{iner}})$
      for $i \leftarrow 1 \rightarrow \text{num}_{\text{iner}}$ **do**
        $\boldsymbol{U}_{j,i+\text{loc}_{\text{last}}} = \boldsymbol{F}_i^{\cos} \cdot \boldsymbol{F}_i$
    else
      $\boldsymbol{U}_{j,\text{loc}} \leftarrow \boldsymbol{U}_{j,\text{loc}} - (\text{peak}_{\text{last}} + \text{peak})/2$
      $\boldsymbol{U}_{j,\text{loc}_{\text{last}}} \leftarrow \boldsymbol{U}_{j,\text{loc}_{\text{last}}} - (\text{peak}_{\text{last}} + \text{peak})/2$
    $\text{loc}_{\text{last}} \leftarrow \text{loc}$; $\text{peak}_{\text{last}} \leftarrow \text{peak}$

$$\begin{cases} \boldsymbol{S}_{\text{C}}^{\text{WB}'} & = |\boldsymbol{S}^{\text{WB}'}| \odot (|\boldsymbol{S}^{\text{WB}'}| + 1)^{\frac{\gamma-1}{2}} \odot e^{j\boldsymbol{S}^{\text{WB}'}_{\text{phase}}} \\ \dot{\boldsymbol{S}}_{\text{C}}^{\text{WB}} & = |\dot{\boldsymbol{S}}^{\text{WB}}| \odot (|\dot{\boldsymbol{S}}^{\text{WB}}| + 1)^{\frac{\gamma-1}{2}} \odot e^{j\dot{\boldsymbol{S}}^{\text{WB}}_{\text{phase}}} \\ \boldsymbol{S}_{\text{t}} & = (< \boldsymbol{S}_{\text{C}}^{\text{WB}'}, \dot{\boldsymbol{S}}_{\text{C}}^{\text{WB}} > \cdot \dot{\boldsymbol{S}}_{\text{C}}^{\text{WB}})/||\dot{\boldsymbol{S}}_{\text{C}}^{\text{WB}}||^2 \\ L_{\text{APC}}^{\boldsymbol{S}^{\text{WB}'}} & = 10\log_{10}\left(||\boldsymbol{S}_{\text{t}}||^2/||\boldsymbol{S}_{\text{C}}^{\text{WB}'} - \boldsymbol{S}_{\text{t}}||^2\right) \end{cases} \tag{11}$$

where $L_{\text{APC}}^{\boldsymbol{S}^{\text{WB}'}}$ and $L_{\text{APC}}^{\boldsymbol{S}^{\text{WB}''}}$ are the loss of CEM and GHCM respectively, and they are calculated as Eq. (11). $||\cdot||$ denotes the 2-norm of the vector. We set α=1 and β=2 in experiment.

3. EXPERIMENTS

3.1. Dataset

To evaluate the performance of the updated WB part of HGCN+, we generate 100 hours of WB data with signal-to-noise ratios (SNR) ranging in 0~40 dB using the speech and noise provided by INTERSPEECH 2020 DNS Challenge (DNS-2020) [21]. We divide the data into training and validation set by 4:1. For testing, we generate 540 noisy-clean pairs with 3 SNRs (-5 dB, 0 dB, 5 dB) using the noise and speech that didn't appear in the training or validation set.

For the ICASSP 2022 DNS Challenge (DNS-2022) [22], we generate a FB set with 3500 hours duration using provided 181 hours of noise, read-English, Russian, French, and Spanish data, and the SNR ranges -5~25 dB. In addition, half of the utterances are convolved with random synthetic and real room impulse responses (RIRs) before being mixed. We divide the data into training and validation by 4:1.

3.2. Training setup and comparison methods

For the WB experiments, we use HGCN (CEM+GHCM+HM) as the reference and introduce the improvements of each module step by step for comparison (CEM$^+$, GHCM$^+$, HM$^+$, $^+$

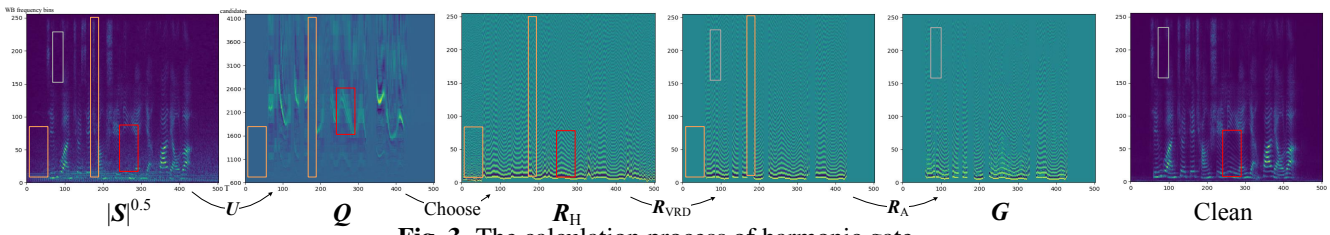


Fig. 3. The calculation process of harmonic gate.

denotes the updated version). The 32 ms hanning window with 25% overlap and 512-point STFT is used. The kernel size is $(5, 2)$. The channel number and stride of encoder and decoder are $\{12, 24, 48, 64, 96, 96\}$ and $(2, 1)$. For HGCN, a 512-units FC after 3-layer 128-units LSTM is adopted. The channel number and stride of GHCM are $\{8, 16, 8\}$ and $(1, 1)$ and the out channel of last decoder is 22 ($C_A = C_B = 10$). Since the GHCM$^+$ and the DPRNN are residual, the hidden dimension is the same as the input. The optimizer is Adam [23]. And the initial learning rate is 0.001, which decays 50% when the valid loss plateaus for 5 epochs, and training is stopped if loss plateaus for 20 epochs.

For HGCN+ in DNS-2022, the 32 ms Hanning window with 25% overlap and 1536-point STFT is used. The kernel size and stride are $(5, 2)$ and $(2, 1)$. The channel numbers of the encoder and decoder are $\{12, 24, 48, 64, 96, 96\}$ and the channel number of last decoder is 6 ($C_A = 4$). The hidden cell of the HBM is 256 and 2-layer GRU is used. HGCN+ is trained for 24 epochs on 3500 hours of data with a learning rate of 0.000125. The audio speed is adjusted in 0.9~1.1 during training. The number of parameter (Para.) is 5.29 M.

3.3. Experimental results and discussion

The Real-Time-Factor (RTF) of WB and FB models both are tested on a machine with an Intel(R) Core(TM) i5-6200U CPU@2.30 GHz in a single thread (implemented by ONNX).

Table 1. System comparison on the wide-band test set.

Model	Para. (M)	RTF	PESQ-WB				STOI(%)			
			-5dB	0dB	5dB	AVG	-5dB	0dB	5dB	AVG
Noisy	-	-	1.11	1.20	1.37	1.23	72.6	81.2	88.2	80.7
HGCN	**0.93**	**0.11**	1.59	1.95	2.37	1.97	80.2	88.2	93.6	87.4
+CEM$^+$	3.60	0.17	1.62	2.00	2.44	2.02	82.8	90.3	94.5	89.1
+GHCM$^+$	4.12	0.16	1.62	2.01	2.47	2.03	83.0	90.3	94.5	89.2
+HM$^+$	4.11	0.14	**1.65**	**2.04**	**2.48**	**2.06**	**83.6**	**90.8**	**94.8**	**89.7**

To evaluate the performance of WB models (100 hours training data), two metrics are utilized, namely PESQ (PESQ-WB, PESQ-NB (narrow-band, 0~4 KHz)) and STOI [24, 25]. The comparison based on WB set is shown in Table 1. It can be seen that the dual-path encoder and DPRNN in CEM$^+$ bring an improvement in terms of feature utilization and frequency dependencies modeling with more arithmetic complexity. GHCM$^+$ improves the performance because the linear module has a wider receptive field with less computational complexity than the convolution. In HM$^+$, the significance-based VRD reduces the computation of CEM, and the introduction of cosine makes gates more accurate and instructive.

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

Model	Para.(M)	PESQ-WB	PESQ-NB	STOI(%)
Noisy	-	1.58	2.45	91.5
DCCRN [4]	**3.67**	-	3.27	-
GaGNet [7]	5.94	3.17	3.56	97.1
HGCN+	5.29	**3.19**	**3.65**	**97.2**

Table 3. Challenge results for track 1 on DNS-2022.

Model	Para.(M)	SIG	BAK	OVRL	WAcc	Final Score
NSNet2 [2]	6.17	3.62	3.93	3.26	0.63	0.60
HGCN+	**5.29**	**4.01**	**4.55**	**3.81**	**0.65**	**0.68**

We evaluate the HGCN+ (3500 hours training data) on the DNS-2020 and the DNS-2022 test sets. Because HGCN+ processes the HB and WB spectra separately, and the WB spectrum contains more speech information, the model has good performance on the WB test set can bring gains to the processing of FB signal, as shown in Table 2. In DNS-2022, ITU-T P.835 framework [26] and Word Accuracy (WAcc) are used to evaluate the speech quality, and HGCN+ outperforms the baseline [2] with less parameters in all metrics, as shown in Table 3. By audiometric analysis, we also found that non-speech sound, such as breath, sob, etc. is usually regarded as noise and filtered out, which results in incomplete human voice but better speech and the BAK MOS which is relatively larger than SIG MOS. In addition, our method makes a confusing pitch choice in the case of multi-speaker and degrades the performance also. The RTF of HGCN+ in processing the FB signal is 0.16 and it consumes 5.22 ms per frame. The frame size and overlap are 32 ms and 25%. Since the model is causal without looking forward, the latency is 32+8=40 ms, which satisfies the requirements of the challenge.

4. CONCLUSION

In this paper, we improve each module of our HGCN. First, we model the harmonic integration by cosine and propose a significance-based VRD to predict the harmonic locations efficiently. Second, we introduce the dual-path encoder, DPRNN, and residual linear structure to CEM and GHCM to enhance model performance. Finally, we add a high-band module to help the model handle the FB signal, resulting in HGCN+. HGCN+ outperforms the referenced models on the DNS-2020 and DNS-2022 test sets.

5. REFERENCES

[1] K. Tan and D. Wang, "Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement," in *IEEE ICASSP*, 2019, pp. 6865–6869.

[2] S. Braun and I. Tashev, "Data augmentation and loss normalization for deep noise suppression," in *International Conference on Speech and Computer*. Springer, 2020, pp. 79–86.

[3] JM Valin, "A hybrid DSP/deep learning approach to real-time full-band speech enhancement," in *IEEE MMSP*, 2018, pp. 1–5.

[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] S. Lv, Y. Fu, M. Xing, J. Sun, L. Xie, J. Huang, Y. Wang, and T. Yu, "S-dccrn: Super wide band dccrn with learnable complex feature for speech enhancement," *arXiv preprint arXiv:2111.08387*, 2021.

[6] H. Schröter, T. Rosenkranz, A. Maier, et al., "DeepFilterNet: A low complexity speech enhancement framework for full-band audio based on deep filtering," *arXiv preprint arXiv:2110.05588*, 2021.

[7] 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.

[8] D. Yin, C. Luo, Z. Xiong, and W. Zeng, "PHASEN: A phase-and-harmonics-aware speech enhancement network," in *AAAI*, 2020, vol. 34, pp. 9458–9465.

[9] T. Wang, W. Zhu, Y. Gao, J. Feng, and S. Zhang, "HGCN: Harmonic gated compensation network for speech enhancement," *arXiv preprint arXiv:2201.12755*, 2022.

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

[11] N. Takahashi and Y. Mitsufuji, "Multi-scale multi-band densenets for audio source separation," in *IEEE WASPAA*, 2017, pp. 21–25.

[12] X. Le, H. Chen, K. Chen, and J. Lu, "DPCRN: Dual-path convolution recurrent network for single channel speech enhancement," in *ISCA INTERSPEECH*, 2021, pp. 2811–2815.

[13] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation," in *IEEE ICASSP*, 2020, pp. 46–50.

[14] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[15] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2011, pp. 315–323.

[16] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating deep network training by reducing internal covariate shift," in *IMLS ICML*, 2015, vol. 1, pp. 448–456.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Delving Deep into Rectifiers: Surpassing human-level performance on imageNet classification," in *IEEE ICCV*, 2015, pp. 1026–1034.

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

[19] 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.

[20] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *IEEE ICCV*, 2017, pp. 2980–2988.

[21] 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.

[22] R. Cutler A. Aazami S. Matusevych S. Braun S. Eskimez M. Thakker T. Yoshioka H. Gamper R. Aichner H. Dubey, V. Gopal, "ICASSP 2022 Deep Noise Suppression Challenge," in *IEEE ICASSP*, 2022, pp. 46–50.

[23] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

[24] A. W Rix, M. P Hollier, A. P Hekstra, and J. G Beerends, "Perceptual evaluation of speech quality (PESQ) the new itu standard for end-to-end speech quality assessment part i–time-delay compensation," *Journal of the Audio Engineering Society*, vol. 50, no. 10, pp. 755–764, 2002.

[25] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE TASLP*, vol. 19, no. 7, pp. 2125–2136, 2011.

[26] B. Naderi and R. Cutler, "A crowdsourcing extension of the ITU-T recommendation P.835 with validation," *arXiv e-prints*, pp. arXiv–2010, 2020.