Harmonic Attention for Monaural Speech Enhancement

Tianrui Wang, Weibin Zhu, Yingying Gao, Shilei Zhang, and Junlan Feng, *Fellow, IEEE*

***Abstract*—To further improve the quality of the enhanced speech, it is appealing that more profound articulatory and auditory knowledge should be introduced into the speech enhancement model. Among these, harmonics seriously affect speech timbre and play a crucial role in speech intelligibility. Especially in the frequency domain, harmonics appear as the local maximum peaks of energy, which could be expected to serve as anchors to recover the distorted speech. In this paper, an explicit modeling method, harmonic attention, is presented, patching the harmonics with the help of residual ones. In order to maintain the spectral structure of speech during the processing and to enable the network to support harmonic modeling, a harmonic attention-based progressive enhancement network (HAPNet) is applied, which gradually approaches clean speech with stacked modules of harmonic attention. In addition, to make enhanced speech more consistent with hearing, a loss function based on the loudness power compression (LC-SNR) is used, which measures both magnitude and phase values with appropriate auditory effects. The experimental visualization indicates that the harmonic attention can capture and recover the harmonics of speech. And the objective evaluations show that the presented HAPNet and LC-SNR outperform the referenced methods. Furthermore, the presented model trained on 100 hours of data achieves competitive results with the referenced models trained on 3000+ hours of data, and one trained on 500 hours of data yields the state-of-the-art performance.**

***Index Terms*—Monaural speech enhancement, harmonic, attention, loss function.**

I. Introduction

SPEECH enhancement (SE) aims to improve the intelligibility and the quality of degraded speech, which is used for applications such as remote conferences, hearing aids, and cell phones [1]. Many signal processing methods have been adopted in the past several decades, for example, spectral subtraction [2], subspace-method [3], statistical-based method [4], and Wiener filtering [5]. With the introduction of deep learning in recent years, speech enhancement has gradually been formulated as a supervised learning task. Compared to the traditional methods, mask-based [6], [7] and spectral-mapping-based [8], [9] networks significantly improve the intelligibility of the results. And, to avoid the confusion caused by the wrapped phase spectrum, early network-based enhancement methods with limited capacity have focused on the magnitude information and ignored the phase one for a long time [7], [10], [11], [12].

In recent years, applying phase information in the enhancement process has been verified for the improvement of the quality of the enhanced speech [13]. Time-domain [14], [15], [16], [17], [18] and complex frequency-domain [19], [20], [21], [22], [23], [24], [25], [26] methods have become the two mainstreams for modeling the information contained in the magnitude and phase spectra simultaneously.

The time-domain methods process the waveform to achieve signal-distortion ratio improvements [16]. However, the hearing evaluations are often degraded because it is difficult to filter out the redundant information for hearing without the help of the transform domain (e.g., loss function of magnitude) [17]. Moreover, weakly-structured waveform makes the time-domain models more abstract and uninterpretable [27].

The complex frequency-domain methods generally enhance the speech based on the complex ratio mask (CRM) [28]. [19] proposed a complex spectral convolutional recurrent network (CRN), where the magnitude and the phase information were expressed implicitly by the real and imaginary parts. Later, with the introduction of complex-valued operation [29], a deep complex convolution recurrent network (DCCRN) was proposed [20] and ranked first for the Interspeech 2020 Deep Noise Suppression challenge (DNS-Challenge) [30]. The mechanism of CRM becomes more explicit, as the distinguishability between the real and imaginary parts is maintained after the multi-layer processing. Afterward, a multi-stage-based model was submitted to the INTERSPEECH 2021 DNS challenge [31] with a superior perfomance, which consisted of a coarse processing module and a fine-tuning module [21], [22]. The forward inference can be interpreted as a Taylor approximation to the clean signal [23] and the residual structure can alleviate the gradient vanishing caused by the multi-stage modules [32].

In a sense, these above models are based on numerical methods to construct network architecture and to optimize training strategy. And the quality of the enhanced speech is still defective, especially in acoustically harsh environments. In order to improve the high fidelity and the auditory receptivity of recovered

Manuscript received 20 October 2022; revised 23 April 2023; accepted 24 May 2023. Date of publication 9 June 2023; date of current version 29 June 2023. This work was supported by Tianrui Wang during the super star program in China Mobile Research Institute. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Xiao-Lei Zhang. *(Corresponding author: Weibin Zhu.)*

Tianrui Wang and Weibin Zhu are with the Institute of Information Science and the Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing Jiaotong University, Beijing 100044, China (e-mail: 20120318@bjtu.edu.cn; wbzhu@bjtu.edu.cn).

Yingying Gao, Shilei Zhang, and Junlan Feng are with the AI Center, China Mobile Research Institute, Beijing 100032, China (e-mail: gaoyingying@chinamobile.com; zhangshilei@chinamobile.com; fengjunlan@chinamobile.com).

Digital Object Identifier 10.1109/TASLP.2023.3284522

speech, it is necessary to introduce articulatory and auditory mechanisms [33], [34].

In terms of articulatory mechanisms, it is important to turn the focus of the model to particular patterns of speech. FullSubNet [24], [25] took a time-frequency point and its adjacent time-frequency points as a sub-band unit to highlight the local time-frequency patterns. In [35], the speech signal was decomposed into excitation and vocal tract, and the two parts were enhanced separately. Among these articulatory features, harmonics affect the intelligibility and timbre of speech, which appears as comb-like structures in the spectrum, as peaks are located at integer multiples of the fundamental frequency with locally highest energy [36]. [37] found that the deep learning model tried to capture harmonics for speech enhancement. The locally high energy of harmonics makes them serve as anchors for recovering speech, which is noise-resistant and prioritized by gradient-based models. In addition, the time-frequency domain harmonic structure was shown to be effective for phase reconstruction [38]. However, it seems that the complete reliance on deep learning leads to an underutilization of the structural characteristics of harmonics. To explicitly adjust for harmonics, a harmonic gated compensation network (HGCN) was proposed in our previous work [39], [40], in which a harmonic location prediction module is adopted to adjust the spectrum. However, HGCN is sensitive to harmonic-like noise, so the harmonic prediction module must be connected behind a coarse enhancement module, which makes the model heavily limited by the performances of the coarse enhancement model, causing unstable results. Therefore, an explicit harmonic modeling strategy with robustness and efficiency is called for.

In terms of auditory mechanisms, it is important to make models aware of the characteristics of hearing perception. Mel scale [41] and equivalent rectangular bandwidth (ERB) mapping [42] are often used to compress high-frequency information in modeling. Compressing high-frequency information in line with the hearing can effectively represent the features of speech with less dimension [26], [43]. In addition, logarithmic and exponential compression of the magnitude are often implemented to simulate auditory loudness during feature processing [34], [44]. However, it seems more reasonable to introduce the auditory mechanism into the loss function than the model's feature to make the enhanced results consistent with human hearing. The model training is guided by introducing loudness spectrum, octave spectrum, and auditory masking effects into the loss function [45], [46]. Unfortunately, the auditory loudness mapping makes the loss function insensitive for phase, and the results is likely to be distorted by the involvement of too many auditory effects [45]. Therefore, a loss function that measures magnitude and phase simultaneously with appropriate auditory effects is required.

From the perspective of articulatory and auditory mechanisms, the following methods are presented in this paper:

- Harmonic attention is presented. Unlike conventional attention, which computes the correlation between two abstract features (query and key), harmonic attention module captures and patches harmonics based on the correlation between the spectrum and a high-resolution comb-pitch conversion matrix.
- A single-channel **h**armonic-**a**ttention-based **p**rogressive speech enhancement network is applied, which was referred to as the HAPNet. The progressive enhancement strategy preserves the structural patterns of the spectra during spectrum processing, which makes acoustic modeling practicable.
- A loudness power compression-based signal-to-noise ratio (LC-SNR) loss function is presented. Auditory compression is appropriately introduced into the loss function to achieve consistent results for hearing with less distorting.

Experiments based on the DNS-Challenge dataset [30], the Aurora dataset [47], and the collected music noises demonstrate that the presented methods not only outperform the referenced methods but also show satisfactory robustness to harmonic-like noise at a low SNR. And the visualization of the intermediate features and results show that harmonic attention module can capture the harmonic structure of speech and recover those harmonics submerged by noise. In addition, we conducted comparative experiments with another four advanced methods on the public DNS-Challenge testset. The results show that the presented methods yield state-of-the-art (SOTA) performance with less training data.

The paper is organized as follows. In Section II, the speech enhancement problem in the time-frequency domain is formulated. In Section III, the presented harmonic attention module and loss function are illustrated in detail. The experimental setup is introduced in Section IV. In Section V, experimental results and analyses are given. The discussion and conclusion of the research are provided in Section VI.

II. Problem Formulation

A. Signal Model

A noisy speech signal $\boldsymbol{x} \in \mathbb{R}^{T\times 1}$ can be formulated as:

$$\boldsymbol{x} = \boldsymbol{s} + \boldsymbol{n} \tag{1}$$

where $\{\boldsymbol{s}, \boldsymbol{n}\} \in \mathbb{R}^{T\times 1}$ denote clean and noise signals. T is the number of samples of the waveform. The signal in the time domain can be converted to the time-frequency (T-F) domain by the short-time Fourier transform (STFT),

$$\boldsymbol{X} = \boldsymbol{S} + \boldsymbol{N} \tag{2}$$

where $\{\boldsymbol{X}, \boldsymbol{S}, \boldsymbol{N}\} \in \mathbb{R}^{L\times 2\cdot F}$ are the complex spectra of the noisy, clean, and noise signal respectively. L and $2 \cdot F$ denote the number of time frames and frequency bins (real and imaginary part).

B. Mask-Based Problem Formulation

The complex frequency-domain speech enhancement aims to extract the target complex spectrum $\hat{\boldsymbol{S}} \in \mathbb{R}^{L\times 2\cdot F}$ from the noisy one,

$$\hat{\boldsymbol{S}} = \mathcal{F}(\boldsymbol{X}) \tag{3}$$

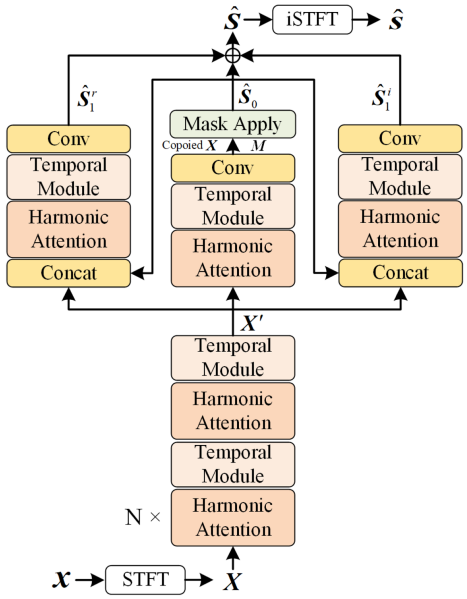

Fig. 1. Diagram of presented HAPNet, which consists of harmonic attention modules, temporal modules, and convolutions. The complex mask $\boldsymbol{M}$ and compensations $\hat{\boldsymbol{S}}_1^r, \hat{\boldsymbol{S}}_1^i$ are estimated to form the enhanced spectrum.

where $\mathcal{F}$ denotes the SE system. And the mask-based SE system $\mathcal{F}_M$ is widely used, which estimates a mask and then the mask is applied to the noisy spectrum to get the target one,

$$\boldsymbol{M} = \mathcal{F}_M(\boldsymbol{X}) \tag{4}$$

$$\hat{\boldsymbol{S}} = \mathcal{M}(\boldsymbol{X}, \boldsymbol{M}) \tag{5}$$

where $\boldsymbol{M} \in \mathbb{R}^{L\times 2\cdot F}$ denotes the estimated mask. $\mathcal{M}$ denotes the mask-applying method. Following the modeling strategy based on Taylor approximation [23], the mask-based methods are extended as,

$$\begin{aligned}\hat{\boldsymbol{S}} &= \mathcal{M}(\boldsymbol{X}, \boldsymbol{M}) + \sum_{q=1}^{+\infty} \frac{1}{q!} \frac{\partial^q \mathcal{M}(\boldsymbol{X}, \boldsymbol{M})}{\partial^q \boldsymbol{X}} \delta^q \\ &= \mathcal{M}(\boldsymbol{X}, \mathcal{F}_M(\boldsymbol{X})) + \sum_{q=1}^{+\infty} \frac{1}{q!} \mathcal{F}_q(\mathcal{M}(\boldsymbol{X}, \mathcal{F}_M(\boldsymbol{X}))) \\ &= \left\{ \hat{\boldsymbol{S}}_0^r + \sum_{q=1}^{+\infty} \frac{1}{q!} \hat{\boldsymbol{S}}_q^r, \hat{\boldsymbol{S}}_0^i + \sum_{q=1}^{+\infty} \frac{1}{q!} \hat{\boldsymbol{S}}_q^i \right\}\end{aligned} \tag{6}$$

where $\{\hat{\boldsymbol{S}}_0^r, \hat{\boldsymbol{S}}_0^i\} = \mathcal{M}(\boldsymbol{X}, \mathcal{F}_M(\boldsymbol{X}))$ is the Taylor-style first-order expansion. δ denotes the residual term of phase. $\mathcal{F}_q$ denotes the q-th DNN-based compensation module of a Taylor-style (multi-stage) based model, which estimates the real and imaginary compensation $\{\hat{\boldsymbol{S}}_q^r, \hat{\boldsymbol{S}}_q^i\}$.

III. Presented System

The **h**armonic-**a**ttention-based **p**rogressive speech enhancement network (HAPNet) is designed as shown in Fig. 1. It consists of stacked harmonic attention modules, temporal modules, and convolutions. The noisy waveform $\boldsymbol{x}$ is first transformed into time-frequency domain using the STFT. Then, several harmonic attention modules and two temporal modules are employed to recombine and refine the features, resulting in $\boldsymbol{X}' \in \mathbb{R}^{L\times C\times F}$, where C denotes the number of channels. Finally, following (6), the model estimates a complex mask $\boldsymbol{M} = \{\hat{\boldsymbol{M}}_0^r, \hat{\boldsymbol{M}}_0^i\}$ and the first-order compensations $\hat{\boldsymbol{S}}_1 = \{\hat{\boldsymbol{S}}_1^r, \hat{\boldsymbol{S}}_1^i\}$ for the real and imaginary parts (to make the model concise, only first-order compensation is used). In a nutshell, the whole forward calculation is formulated as,

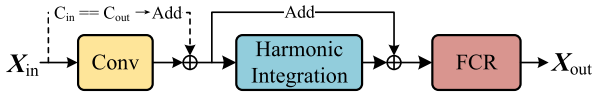

Fig. 2. Harmonic attention module block of HAPNet, which consists of convolution, harmonic integration, and frequency-channel recombination (FCR). The residual addition of convolution is used when the input channel is equal to the output one.

$$\hat{\boldsymbol{S}}_0 = |\boldsymbol{X}| \odot \tanh(|\boldsymbol{M}|) \odot e^{j(\boldsymbol{X}_{\text{phase}} + \boldsymbol{M}_{\text{phase}})} \tag{7}$$

$$\hat{\boldsymbol{S}} = \left\{ \hat{\boldsymbol{S}}_0^r + \hat{\boldsymbol{S}}_1^r, \hat{\boldsymbol{S}}_0^i + \hat{\boldsymbol{S}}_1^i \right\} \tag{8}$$

where $\hat{\boldsymbol{S}}_0 = \{\hat{\boldsymbol{S}}_0^r, \hat{\boldsymbol{S}}_0^i\}$ is the first part of (6). $|\cdot|$ and $\cdot_{\text{phase}}$ are the magnitude and phase. $\odot$ is the element-wise multiplication.

A. Harmonic Attention Module

Harmonic attention module is the key for HAPNet to captur, ande harmonics, which consists of three blocks connected in series, i.e., convolution, harmonic integration, and frequency-channel recombination (FCR), as shown in Fig. 2. The convolution roughly filters the input features and divides them in different channels, which consists of causal 2D convolution, batch normalization [48], and PReLU [49]. The residual addition will be used when the input channel number is equal to the output one after convolution. Due to the locality of convolution [50], the spectral structure is preserved after processing, which guarantees harmonic modeling.

1) Harmonic Integration: Considering the comb-like structure of the harmonics, their locations can be derived from those prominent and surviving harmonics, even though part of them are submerged by noise. The harmonic integration module first estimates the approximate distribution of harmonics based on a high-resolution comb-pitch conversion matrix, and then calculates the detailed adjustments by convolutions, as shown in Fig. 3.

A comb-pitch conversion matrix is designed to predict the pitch from the spectrum and to derive the corresponding harmonic distribution according to the predicted pitch. The pitch candidates are set first, and then the comb-like harmonics corresponding to each pitch candidate are modeled, resulting in a high-resolution comb-pitch conversion matrix $\boldsymbol{Q} \in \mathbb{R}^{N_c\times F}$, where N_c is the number of pitch candidates in the range of 60 to 420 Hz (normal pitch range of speech). The $\boldsymbol{Q}$ is designed as Algorithm 1 and its computing procedure is shown in Fig. 4. The pitch candidate f_c is taken from $[60/R]$ to $[420/R]$ in step R

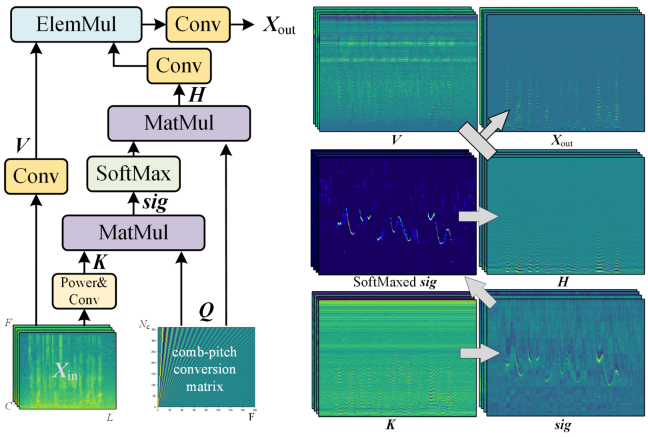

Fig. 3. Harmonic integration block of harmonic attention module. Some intermediate outputs of the third harmonic integration of the HAPNet (trained on 500 hours of data) are shown on the right. The distribution of harmonics $\boldsymbol{H}$ is estimated with the help of the artificial high-resolution comb-pitch conversion matrix $\boldsymbol{Q}$ (the pitch candidates resolution of $\boldsymbol{Q}$ in the figure is 1.0 Hz). And then, $\boldsymbol{H}$ is used to adjust the features.

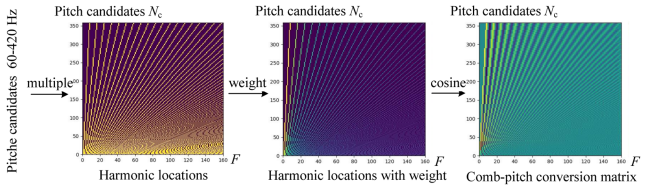

Fig. 4. The procedure for computing the comb-pitch conversion matrix. The harmonics locations are derived from the pitch candidates. The comb-pitch conversion matrix is obtained by weighting and modeling the peak-valley structure with cosine.

(resolution), where $[\cdot]$ is a rounding operation. The locations of harmonics are calculated from the 1st-order harmonic to the $\frac{sr}{(2\cdot R\cdot f_c)}$th-order one, where sr is the sampling rate. Then the weight $1/\sqrt{p}$ is used [51], where p denotes the order of harmonic, and the area between the current and the previous harmonics is filled with cosine curves, where the function linspace(a, b, c)[1] generates an arithmetic progression between a and b of length c.

To model the harmonics in the complex spectrum, the multi-head key $\boldsymbol{K} \in \mathbb{R}^{N_h\cdot C\times L\times F}$ is obtained by convolving the normalized [52] energy of input $\boldsymbol{X}_{\text{in}} \in \mathbb{R}^{C\times L\times F}$, where N_h is the number of heads. Then the significance spectrum $\boldsymbol{sig} \in \mathbb{R}^{N_h\cdot C\times L\times N_c}$ is obtained by multiplying $\boldsymbol{K}$ by $\boldsymbol{Q}$, which denotes the confidence that each candidate is the pitch,

$$\boldsymbol{K} = \text{conv}(\text{norm}(\boldsymbol{X}_{\text{in}}^2)) \tag{9}$$

$$\boldsymbol{sig} = \boldsymbol{K}\cdot\boldsymbol{Q}^\top \tag{10}$$

where norm denotes the layer normalization [52] in the frequency domain. Then the approximate distribution of harmonics

[1][Online]. Available: https://github.com/numpy/numpy

Algorithm 1: Comb-Pitch Conversion Matrix.

$\boldsymbol{Q} \leftarrow \boldsymbol{0} \in \mathbb{R}^{[360/R]\times F}$
for $f_c \leftarrow [60/R] \rightarrow [420/R]$ **do**
  $\text{loc}_{\text{last}} \leftarrow 0$; $\text{peak}_{\text{last}} \leftarrow 1$; $j \leftarrow f_c - [60/R]$
  for $p \leftarrow 1 \rightarrow [sr/(2\cdot R\cdot f_c)]$ **do**
    $\text{loc} \leftarrow [R\cdot f_c\cdot p\cdot F/(sr/2)]$ ▷ harmonic location
    $\text{peak} \leftarrow 1/\sqrt{p}$; $\boldsymbol{Q}_{j,\text{loc}} \leftarrow \text{peak}$ ▷ weight
    if $\text{loc}_{\text{last}} \neq 0$ **then** ▷ valley structures
      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**
          $\boldsymbol{Q}_{j,i+\text{loc}_{\text{last}}} = \boldsymbol{F}^{\cos}{}_i \cdot \boldsymbol{F}_i$
      else
        $\boldsymbol{Q}_{j,\text{loc}} \leftarrow \boldsymbol{Q}_{j,\text{loc}} - (\text{peak}_{\text{last}} + \text{peak})/2$
        $\boldsymbol{Q}_{j,\text{loc}_{\text{last}}} \leftarrow \boldsymbol{Q}_{j,\text{loc}_{\text{last}}} - (\text{peak}_{\text{last}} + \text{peak})/2$
    $\text{loc}_{\text{last}} \leftarrow \text{loc}$; $\text{peak}_{\text{last}} \leftarrow \text{peak}$

$\boldsymbol{H} \in \mathbb{R}^{N_h\cdot C\times L\times F}$ can be selected according to the softmax-based [53] normalized significance,

$$\boldsymbol{H} = \text{softmax}(\boldsymbol{sig})\cdot\boldsymbol{Q} \tag{11}$$

Finally the approximate distribution $\boldsymbol{H}$ is applied to the convoluted spectra $\boldsymbol{V} \in \mathbb{R}^{N_h\cdot C\times L\times F}$ and obtain the output $\boldsymbol{X}_{out} \in \mathbb{R}^{C\times L\times F}$ after a convolution as,

$$\boldsymbol{V} = \text{conv}(\boldsymbol{X}_{in}) \tag{12}$$

$$\boldsymbol{X}_{\text{out}} = \text{conv}\left(\boldsymbol{V}\odot\text{conv}(\boldsymbol{H})\right) \tag{13}$$

2) Frequency-Channel Recombination: The comb-like spectral structure is captured by the harmonic integration. But the interaction among the heads (channels) is absent. To improve the complementarity among the information captured by different filters, a frequency-channel recombination module (FCR) is employed, which consists of two stacked multi-head self-attention [54], one along with the channel (C-Attention) and the other one along with the frequency (F-Attention), as shown in Fig. 5.

The input $\boldsymbol{X}_{\text{in}} \in \mathbb{R}^{C\times L\times F}$ is firstly reshaped to $\boldsymbol{X}_{\text{c}} \in \mathbb{R}^{L\times C\times F}$. And three linear layers process the normalized $\boldsymbol{X}_{\text{c}}$ respectively to obtain {query $\boldsymbol{Q}_{\text{c}}$, key $\boldsymbol{K}_{\text{c}}$, value $\boldsymbol{V}_{\text{c}}$} $\in \mathbb{R}^{L\times N_{hc}\times C\times F'}$ for calculating the attention along with the channel dimension $\boldsymbol{A}_{\text{c}} \in \mathbb{R}^{L\times N_{hc}\times C\times C}$,

$$\boldsymbol{A}_{\text{c}} = \text{softmax}\left(\frac{\boldsymbol{Q}_{\text{c}}\cdot\boldsymbol{K}_{\text{c}}^\top}{\sqrt{d_c}}\right) \tag{14}$$

where $N_{hc}\times F' = F$ and d_c are the dimensions of query and key. $\boldsymbol{A}_{\text{c}}$ contains the inter-relationships among different channels, which is used to recombine the channels of $\boldsymbol{V}$ and then added the residual to obtain the output of the C-Attention,

$$\boldsymbol{X}_{\text{co}} = \boldsymbol{A}_{\text{c}}\cdot\boldsymbol{V}_{\text{c}} + \boldsymbol{X}_{\text{c}} \tag{15}$$

Then $\boldsymbol{X}_{\text{co}}$ is reshaped into $\boldsymbol{X}_{\text{f}} \in \mathbb{R}^{L\times F\times C}$ for calculating the attention along with the frequency dimension $\boldsymbol{A}_{\text{f}} \in$

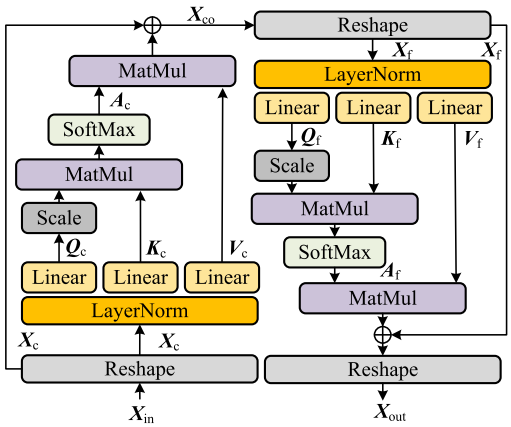

Fig. 5. The frequency-channel recombination (FCR) of harmonic attention module, which consists of two self-attentions, one along with the channel and the other along with the frequency.

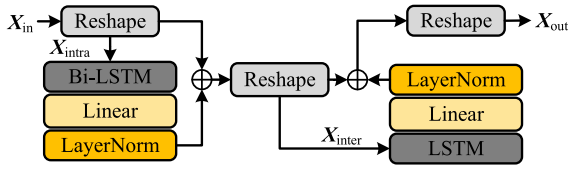

Fig. 6. Dual-path RNN. It consists of two RNNs that have recurrent connections in different dimensions. A bi-directional intra-RNN is applied to process local frequency-channel information. The inter-RNN is applied to capture the temporal dependency.

$\mathbb{R}^{L\times N_{hf}\times F\times F}$. The computation process is similar to C-Attention, and the output $\boldsymbol{X}_{\text{out}} \in \mathbb{R}^{C\times L\times F}$ of FCR is obtained.

B. Temporal Modules

Harmonic attention considers the frequency and channel information but lacks in temporal modeling. So, some powerful temporal modules need to be introduced to complement the absence of temporal information, such as LSTM-based DPRNN [55], [56] and time-domain multi-head self-Attention [37], [57], [58], [59], [60].

The DPRNN is shown in Fig. 6, the input $\boldsymbol{X}_{\text{in}} \in \mathbb{R}^{L\times C\times F}$ is reshaped to $\boldsymbol{X}_{\text{intra}} \in \mathbb{R}^{C\times L\times F}$, and a bi-directional intra-RNN is first applied to process the local frequency-channel information. Then the result of intra-RNN is reshaped to $\boldsymbol{X}_{\text{in}} \in \mathbb{R}^{L\times F\times C}$, and the inter-RNN is used to capture the temporal dependency. As the DPRNN condenses historical information into a memory vector, it is suitable for real-time tasks.

The processing of the time-domain multi-head self-attention is a self-attention in Fig. 5 along the time dimension. Although the causal matrix can make the attention causal, the localized information can limit the ability of attention. Hence, time-domain attention is more suitable for offline tasks with global information. And its performance is approximately proportional to the number of the looked frames.

The DPRNN and the time-domain self-attention are respectively inserted into HAPNet for evaluation.

C. Loss Function

Power compression is helpful for speech enhancement [44] and is commonly used in objective metrics. Inspired by this, we compress the predicted and referenced spectra as,

$$\begin{cases} \hat{\boldsymbol{S}}_{\text{com}} = |\hat{\boldsymbol{S}}| \odot (|\hat{\boldsymbol{S}}| + 1)^{\gamma-1} \odot e^{j\hat{\boldsymbol{S}}_{\text{phase}}} \\ \boldsymbol{S}_{\text{com}} = |\boldsymbol{S}| \odot (|\boldsymbol{S}| + 1)^{\gamma-1} \odot e^{j\boldsymbol{S}_{\text{phase}}} \end{cases} \tag{16}$$

where Zwicker power $\gamma \in [0.23, 0.27)$ is introduced from the loudness density [61]. And 1 is added to the spectrum before scaling to suppress the over-amplification of small values. Then, the time-domain scale-invariant signal-to-noise ratio (SI-SNR) [62] is migrated to the complex frequency spectrum, and the **l**oudness power **c**ompression based SNR (LC-SNR) can be calculated as,

$$\begin{cases} \boldsymbol{S}_{\text{t}} = (< \hat{\boldsymbol{S}}_{\text{com}}, \boldsymbol{S}_{\text{com}} > \cdot \boldsymbol{S}_{\text{com}})/||\boldsymbol{S}_{\text{com}}||^2 \\ \text{LC-SNR} = 10\log_{10}\left(||\boldsymbol{S}_{\text{t}}||^2/||\hat{\boldsymbol{S}}_{\text{com}} - \boldsymbol{S}_{\text{t}}||^2\right) \end{cases} \tag{17}$$

where $||\boldsymbol{S}||^2 =< \boldsymbol{S}, \boldsymbol{S} >$ represents the energy of the input, and the $\boldsymbol{S} \in \mathbb{R}^{T\times 2\cdot F}$ is reshaped to $\boldsymbol{S} \in \mathbb{R}^{1\times T\cdot 2\cdot F}$ before calculating.

IV. Experimental Setup

A. Datasets

To verify the performance of the presented methods, we conduct experiments on the data from the 2020 DNS-Challenge [30], which has 500 hours of speech from 2150 speakers and 180 hours of noise belonging to 150 classes.

For training, 72,000 5-seconds (100-hours) of noisy clips are generated. During the mixed process, a random clean utterance is mixed with a random noise under a random SNR. The SNR range is -5 dB to 15 dB with an interval of 1 dB. Then, The training data is divided into validation and training sets in a ratio of 1:4.

For testing, two kinds of noises (babble and train) are selected from Aurora [47], and they are mixed with 150 10-seconds of clean utterances under 5 SNRs (-6 dB, -3 dB, 0 dB, 3 dB, 6 dB) respectively, resulting in 1500 noisy-clean pairs. To investigate the robustness of models to harmonic-like noise, the noises randomly cut from 4 absolute music works are mixed with 150 utterances under 5 SNRs (-6 dB, -3 dB, 0 dB, 3 dB, 6 dB), resulting in 750 noisy-clean pairs. In addition, we also do a test on the DNS synthetic dataset [30]. All noisy or clean utterances are sampled at 16 kHz.

B. Model Setup

We compare the presented model with another four advanced models. CRN [19] is a complex spectral model. DCCRN [20] introduces complex-valued operations based on the CRN. HGCN+ [40] compensates the spectra according to the harmonic locations. MTFAA [43] implements the multi-scale time-frequency processing and the axial attention to model the signal. Among them, DCCRN, MTFAA, and HGCN+ have achieved impressive performance in DNS-Challenge [30], [63]. The setup of the referenced models and the presented HAPNet are detailed as

follows. Note that the configuration of the non-causal models in the experiments is the same as the causal one, except for the bidirectional RNN or the temporal attention without a causal mask.

CRN: The CRN maps the complex spectra using a convolutional recurrent network. The 32 ms Hanning window with a 25% overlap and 512-point STFT is used, resulting in 257-dimension spectral features. The real and imaginary parts are treated as two channels and concatenated together as the input. The CRN is an encoder-decoder structure. The encoder and decoder are comprised of 2D causal convolution (transpose convolution for decoder), batch normalization, and PReLU. The kernel size and the stride of convolution are (2,5) and (1,2). The channel numbers of the encoder and decoder are $\{16, 32, 64, 128, 128, 128\}$. Between the encoder and decoder, a 128-dimension LSTM followed by a 1024-dimension fully connected layer is used to model the temporal dependencies. A complex mask is estimated by the decoder. Then the enhanced result is obtained according to the (7).

DCCRN:[2] The framework of the DCCRN is similar to that of CRN. The 25 ms Hanning window with an overlap of 25% overlap and a 512-point STFT is used, resulting in 257-dimension spectral features. The corresponding modules of CRN are replaced by complex causal convolution, complex causal transpose convolution, complex batch normalization, and complex LSTM. The kernel size and the stride of convolution are (2,5) and (1,2). The channel numbers of the encoder and the decoder are $\{16, 32, 64, 128, 128, 128\}$. Two layers of the 128-dimension complex LSTM followed by a 2048-dimension fully connected layer are used between the encoder and the decoder. Since the complex-valued operation involves the real and imaginary parts, the encoder, the decoder, and the RNN-based temporal module have twice the parameters as that of CRN.

HGCN+: The wide-band version HGCN+ first roughly suppresses the noise using a CRN-based coarse enhancement module (CEM). And then, the harmonic locations are estimated by spectral integration. Finally, the spectrum estimated by CEM is compensated based on the harmonic locations. The 32 ms Hanning window with a 25% overlap and a 512-point STFT is used, resulting in 257-dimension spectral features. The kernel size, the stride, and the channel numbers of CEM are (2,5), (1,2), and $\{16, 32, 64, 128, 128, 128\}$. The configuration of convolution in CEM is the same as that in CRN, and other configurations remain the same as the original.

MTFAA:[3] Since the ability to denoise the noisy single-channel speech needs to be compared, the multi-channel full-band input of the MTFAA is replaced by the single-channel wide-band one, and the other settings remain the same as the original. The 32 ms Hanning window with a 25% overlap and a 512-point STFT is used, resulting in 257-dimension spectral features. The kernel size and the stride of the phase encoder module are (1,3) and (1,1). The output channel number of the phase encoder is 4. The (3,3) kernel size and six convolution blocks with dilations from 1 to 32 are used in the TF-Convolution module. And the causal and non-causal axial self-attentions are set separately for comparison.

HAPNet: The 20 ms Hanning window with a 50% overlap and a 320-point STFT is used, resulting in 161-dimension spectral features. The block number N in Fig. 1 is 4, and the channel numbers of harmonic attention modules are $\{12, 24, 24, 48, 48, 24, 12, 12\}$, where the first six belong to the main branch, and the last two belong to the compensation branch. The kernel size and the stride of convolution in the harmonic attention module in Fig. 2 are (2,3) and (1,1). And the kernel size, stride, and head number N_h of the harmonic integration are (1,3), (1,1), and 4. The N_{hc} and N_{hf} of FCR are set to 4 and 7. All modules with a residual addition do not change the dimensions of features. Four temporal modules are tested separately: causal DPRNN (HAPNet$_{\mathrm{D}}$), causal time-domain self-attention (HAPNet$_{\mathrm{A}}$), and the non-causal ones.

C. Loss Functions and Training Setup

To fairly measure the performance, the models in the model comparison experiments are trained based on a same loss function, SI-SNR. In addition, to evaluate the presented loss functions LC-SNR, some typical loss functions are used for comparisons, such as mean square error (MSE) [64], the SI-SNR, and the perceptual metric for speech quality evaluation (PMSQE) [45]. The optimizer is Adam [65]. And the initial learning rate is 0.001, which decays to an extent of 50% when the validation loss plateaus for three epochs, and the training is stopped if the loss plateaus for 15 epochs.

D. Evaluation Metrics

Since the subjective testing consumes a lot of resources, three exemplary objective methods are employed as the metrics. The perceptual evaluation of speech quality (PESQ) [66], which introduces the auditory masking effects and temporal alignment algorithms, is more consistent with human hearing but a little sluggish to the time delay differences. The short-time objective intelligibility measure (STOI) [46], which measures the correlation of estimated and referenced speeches based on the octave spectrum, is more sensitive to time delay differences. The scale-invariant signal-to-distortion ratio (SI-SDR) [67], which measures the waveform by a point-to-point SNR, is sensitive to signal distortion. As objective evaluation metrics, these three parameters are often combined to more comprehensively evaluate the quality of the enhanced speech.

V. Results and Analysis

Ablation studies are first conducted to demonstrate the necessity of each presented module, followed by SI-SNR based model comparisons to validate the performance of the HAPNet. After the model's architecture is determined, the presented loss function is tested. Finally, the comprehensive performance of the presented methods is validated.

[2][Online]. Available: https://github.com/huyanxin/DeepComplexCRN
[3][Online]. Available: https://github.com/echocatzh/MTFAA-Net

TABLE I
PERFORMANCE FOR DIFFERENT CONFIGURATIONS

N_h	R	Para.(M)	PESQ	STOI(%)	SI-SDR
1	1.0	1.20	2.917	96.84	19.60
2	1.0	1.32	2.954	96.96	19.71
4	1.0	1.68	**3.071**	**97.33**	19.99
5	1.0	1.92	3.017	97.18	**20.07**
4	0.5	1.68	3.054	97.29	20.03
4	2.0	1.68	2.974	97.03	19.76

Bold indicates the best score in each case.

TABLE II
PERFORMANCE OF MODELS WITH DIFFERENT SUB-MODULES

Model	Para.(M)	PESQ	STOI(%)	SI-SDR
HAPNet	1.68	**3.071**	**97.33**	**19.99**
- FCR	0.82	2.940	96.91	19.66
- Harmonic Integration	1.11	2.837	96.56	18.54
- FCR	0.24	2.787	96.18	18.26
+ C self-attention	1.44	2.908	96.85	19.34
+ F self-attention	1.44	2.956	97.06	19.69

-/+Means remove/add this module.

A. Ablation Studies

The ablation experiments are validated on the public test set of DNS-Challenge [30]. The number of the parameters (Para.), PESQ, STOI, and SI-SDR are used as the metrics for comparison. Note that the DPRNN is set as the temporal module in ablation experiments.

Harmonic integration is the most critical module of harmonic attention. So, some harmonic integrals with different parameters are compared, as shown in Table I. N_h denotes the number of heads for $\boldsymbol{K}$ and $\boldsymbol{V}$ of the harmonic integral, as shown in (9) and (12). And R denotes the frequency resolution of pitch candidates.

It can be seen from the results that the effect increases with the number of heads. Since the comb-pitch conversion matrix is fixed, the model can change the filtering targets of different heads to obtain the complementary features. And the multi-head can provide the conversion matrix with more diverse feature templates to capture the harmonics. Finally, excessive ability (too large N_h) will introduce bias to the time-domain metric (SI-SDR) due to using SI-SNR for training.

In addition, three comb-pitch conversion matrices with different frequency resolutions (0.5 Hz, 1 Hz, and 2 Hz) are tested to explore the necessity of the high-resolution candidates. The matrix with a resolution of 1 Hz shows the best result. We speculate that a small resolution causes redundant information, and a large one leads to vague high-order harmonics. Hence, an appropriate resolution improves the model's generalization while ensuring accurate harmonics.

Next, we investigate the contribution of each module of HAPNet. The harmonic integration and FCR are removed one by one. It can be seen from the results in Table II that the removal of harmonic integration and FCR significantly degrades the performance of the model. And, removing the FCR from the complete HAPNet causes more degradation than removing that from the HAPNet without harmonic integration, which proves that the harmonic structure captured by the harmonic integration facilitates the attention-based FCR to uncover the internal correlations of features. In addition, the self-attentions along the channel and the frequency (C/F self-attention) are used to replace the harmonic integration module for comparison. The results show that the harmonic integration module achieves more significant improvements (PESQ improved from 2.837 to 3.071), which indicates that the harmonic integration module models the feature more effectively and highlights the speech's acoustic structure for subsequent processing.

B. Referenced Model Comparison

Causal and non-causal models are tested on the test set that mixed with Babble and Train noises. PESQ and STOI are used as the metrics to evaluate the performance of the model, and the experimental results are shown in Table III.

1) Causal Models: From the results, it can be seen that HAPNet outperforms the referenced models among the causal methods. Compared to the DCCRN, the PESQ and STOI of HAPNet are improved by $\{0.34, 1.99\}$ with fewer parameters. It proves that modeling pattern within the feature can significantly alleviate the dependence of the model on the number of parameters. Compared to the HGCN+, PESQ and STOI are improved by $\{0.23, 1.11\}$ with one-fourth of the parameters. Since the harmonic prediction module in HGCN+ is not differentiable, the noise-sensitive integration module needs to follow a coarse enhancement module, which not only limits the performance of the model but also increases the parameters. The differentiable harmonic attention module of HAPNet overcomes this limitation. The causal MTFAA performed mediocrely in our experiments, and we speculate that there are two reasons for this. First, since the performance of time-domain attention is proportional to the number of frames, the short utterances of the training set limit the attention. Second, since the MTFAA is presented to handle full-band multichannel signals, some sub-modules lose their advantages in wideband single-channel experiments, such as the ERB scaling. However, the strong fitting ability of attention is also demonstrated by the fact that MTFAA performs better than DCCRN with fewer parameters. Similar results are obtained for HAPNet, where time-domain attention showed a slightly poorer performance than DPRNN with fewer parameters.

2) Non-Causal Models: Without the limitation of looking-ahead, the non-causal models perform better than the causal ones. It can be seen from the results that HAPNet performs best among the models with the RNN-based temporal module (HAPNet_D) and the models with the attention-based temporal module (HAPNet_A). In the CRN and DCCRN, the temporal module is placed between the Encoder and Decoder. The channel and frequency domains are mixed and then fed into LSTM. It is difficult to capture the temporal dependence with unstructured features. On the contrary, the harmonic attention module and the progressive enhancement strategy in HAPNet preserve the most feature structure during the intermediate processing. So, the improvement brought by non-causal modeling is more significant. In addition, a non-differentiable harmonic integration module

TABLE III
OBJECTIVE RESULT COMPARISONS AMONG DIFFERENT MODELS FOR THE TEST SET WITH BABBLE AND TRAIN NOISES

Model	Para. (M)	Cau.	Babble PESQ					Train PESQ					AVG	Babble STOI					Train STOI(%)					AVG
			-6dB	-3dB	0dB	3dB	6dB	-6dB	-3dB	0dB	3dB	6dB		-6dB	-3dB	0dB	3dB	6dB	-6dB	-3dB	0dB	3dB	6dB	
Noisy	-	-	1.08	1.12	1.19	1.29	1.45	1.17	1.29	1.45	1.67	1.95	1.37	66.38	74.18	81.18	86.97	91.41	89.19	92.96	95.62	97.37	98.47	87.37
CRN	1.72	√	1.26	1.45	1.70	2.00	2.35	2.04	2.38	2.71	3.00	3.28	2.22	76.59	84.09	89.48	93.23	95.72	94.40	96.43	97.73	98.56	99.08	92.53
DCCRN	3.67	√	1.31	1.52	1.80	2.14	2.50	2.19	2.53	2.85	3.13	3.38	2.34	79.41	86.34	91.05	94.20	96.44	95.15	96.88	98.01	98.73	99.18	93.54
HGCN+	7.60	√	1.39	1.63	1.94	2.28	2.62	2.35	2.67	2.96	3.22	3.44	2.45	82.32	88.27	92.35	95.00	96.75	95.85	97.28	98.24	98.88	99.27	94.42
MTFAA	2.19	√	1.38	1.60	1.87	2.19	2.53	2.26	2.58	2.88	3.16	3.40	2.38	81.50	87.48	91.67	94.52	96.42	95.50	97.03	98.05	98.74	99.18	94.01
HAPNet_D	1.68	√	**1.56**	**1.86**	**2.21**	**2.58**	**2.91**	**2.61**	**2.92**	**3.18**	**3.42**	**3.60**	**2.68**	**85.83**	**91.20**	**94.29**	**96.11**	**97.25**	**96.53**	**97.58**	**98.35**	**98.93**	**99.28**	**95.53**
HAPNet_A	**1.52**	√	1.50	1.79	2.14	2.50	2.82	2.53	2.85	2.12	3.36	3.56	2.52	83.95	90.11	93.72	95.78	97.06	96.29	97.43	98.26	98.84	99.23	95.07
CRN	2.10	×	1.32	1.53	1.80	2.13	2.48	2.09	2.42	2.76	3.06	3.32	2.29	78.35	85.49	90.59	93.98	96.15	94.83	96.67	97.87	98.64	99.14	93.17
DCCRN	4.99	×	1.37	1.63	1.94	2.28	2.63	2.29	2.61	2.91	3.19	3.42	2.43	81.72	87.70	92.41	94.82	96.82	95.55	97.28	98.22	98.85	99.26	94.26
HGCN+	9.30	×	1.49	1.79	2.13	2.45	2.77	2.53	2.81	3.05	3.32	3.56	2.59	85.21	90.33	93.57	95.66	97.19	96.45	97.62	98.45	98.97	99.29	95.27
MTFAA	2.19	×	1.53	1.78	2.08	2.39	2.70	2.43	2.72	2.99	3.25	3.47	2.53	85.31	90.10	93.40	95.62	97.13	96.31	97.54	98.39	98.95	99.29	95.20
HAPNet_D	1.67	×	**1.73**	**2.06**	**2.42**	**2.77**	**3.08**	**2.78**	**3.08**	**3.31**	**3.53**	**3.70**	**2.85**	**88.31**	**92.80**	**95.30**	**96.66**	**97.60**	**97.00**	**97.87**	**98.52**	**99.00**	**99.40**	**96.25**
HAPNet_A	**1.52**	×	1.59	1.88	2.22	2.57	2.91	2.59	2.91	3.20	3.44	3.64	2.69	86.71	91.75	94.65	96.31	97.38	96.67	97.69	98.43	98.96	99.39	95.79

CAU. denotes whether to use the causal setup.

TABLE IV
OBJECTIVE RESULT COMPARISONS AMONG DIFFERENT MODELS FOR THE TEST SET WITH HARMONIC-LIKE NOISES

Model	Cau.	PESQ					AVG	STOI					AVG	SI-SDR					AVG
		-6dB	-3dB	0dB	3dB	6dB		-6dB	-3dB	0dB	3dB	6dB		-6dB	-3dB	0dB	3dB	6dB	
Noisy	-	1.20	1.32	1.40	1.54	1.81	1.45	81.45	86.61	89.53	92.51	95.57	89.14	0.84	4.23	7.03	9.80	13.10	7.00
CRN	√	1.99	2.38	2.54	2.78	3.13	2.56	89.65	93.79	95.19	96.32	97.55	94.50	9.12	12.01	14.04	16.00	18.51	13.94
DCCRN	√	2.18	2.59	2.76	3.01	3.34	2.78	91.58	95.09	96.14	97.41	98.28	95.70	10.50	13.38	15.32	17.07	19.30	15.11
HGCN+	√	2.26	2.63	2.92	3.20	3.50	2.90	93.31	96.06	97.16	98.11	98.78	96.69	11.29	14.73	16.43	18.76	20.82	16.41
MTFAA	√	2.26	2.65	2.77	2.98	3.29	2.79	93.45	96.06	96.55	97.58	98.47	96.42	11.59	14.12	15.86	17.18	19.02	15.56
HAPNet_D	√	**2.73**	**3.05**	**3.17**	**3.36**	**3.60**	**3.18**	**96.24**	**97.52**	**97.99**	**98.60**	**99.07**	**97.88**	**15.26**	**17.36**	**18.75**	**20.11**	**22.22**	**18.74**
HAPNet_A	√	2.56	2.91	3.05	3.24	3.52	3.06	95.47	97.04	97.67	98.35	98.94	97.50	14.24	16.47	17.94	19.30	21.50	17.89
CRN	×	2.07	2.49	2.69	2.96	3.30	2.71	90.91	94.64	95.99	97.34	98.42	95.46	9.85	12.65	14.63	16.64	19.02	14.55
DCCRN	×	2.25	2.65	2.84	3.09	3.42	2.85	92.94	95.85	96.93	97.86	98.79	96.48	11.67	14.29	16.19	18.08	20.39	16.12
HGCN+	×	2.32	2.70	3.04	3.32	3.60	2.99	95.06	96.80	97.99	98.39	98.94	97.44	12.04	15.03	17.21	19.37	21.66	17.06
MTFAA	×	2.49	2.86	3.02	3.24	3.51	3.02	95.75	97.25	97.84	98.20	98.75	97.56	12.94	15.51	16.97	18.76	21.04	17.04
HAPNet_D	×	**3.09**	**3.37**	**3.49**	**3.67**	**3.87**	**3.50**	**97.52**	**98.37**	**98.68**	**99.09**	**99.42**	**98.61**	**17.49**	**19.54**	**20.81**	**22.31**	**24.33**	**20.89**
HAPNet_A	×	2.76	3.08	3.24	3.42	3.66	3.23	96.43	97.68	98.12	98.69	99.14	98.01	15.35	17.40	18.78	20.25	22.28	18.81

CAU. denotes whether to use the causal setup.

separates the two RNN-based temporal modules in the multi-stage based HGCN+, and the information impassability limits the improvement of non-causal HGCN+. The differentiability of harmonic attention module in HAPNet provides interoperability among temporal modules and brings significant improvements. Compared to the time-domain attention-based MTFAA, the improvement of non-causal modeling is more significant brought by HAPNet. We speculate that the highlighted harmonics can help the attention to capture the correlation among frames.

3) Parameter Comparison: It can be seen from the results that HAPNet achieves the best performance with the least parameters. Compared to the complex-valued operations in DCCRN, which involves the real and imaginary part modules for the implementation of the complex-valued process with twice the number of parameters, the harmonic attention module in HAPNet focuses on capturing internal patterns within features with fewer parameters. And the structure of the HAPNet is more straightforward than the multi-stage structure of the HGCN+, owing to the hot-pluggable harmonic attention module and the progressive enhancement strategy. Finally, HAPNet consists of the stacked harmonic attention module and the temporal module, which is simpler than MTFAA.

C. Robustness to Noise With Harmonic Structure

The noises with the harmonic structure are catastrophic for modeling harmonics. So, the robustness of models to music noise is evaluated experimentally. The performance of models is assessed by PESQ, STOI, and SI-SDR, as shown in Table IV.

From the results, it can be seen that the HAPNet outperforms the referenced models at all SNRs. Since the CRN, DCCRN, and MTFAA do not explicitly model harmonics, they perform mediocrely on the test set distorted by harmonic-like noises. Compared to CRN at 6 dB SNR, PESQ and STOI of HGCN+ are improved by $\{0.27, 1.03\}$ in Table III (Babble) and $\{0.37, 1.23\}$ in Table IV. More improvements are owed to the integration-based harmonic prediction in the HGCN+, which reinforces the higher-energy harmonic signal and suppresses the lower-energy ones. However, the HGCN+ is degraded in the case of signals with harmonic-like noise at low SNRs (The noise energy at -6 dB or -3 dB is greater than that of speech). The harmonic

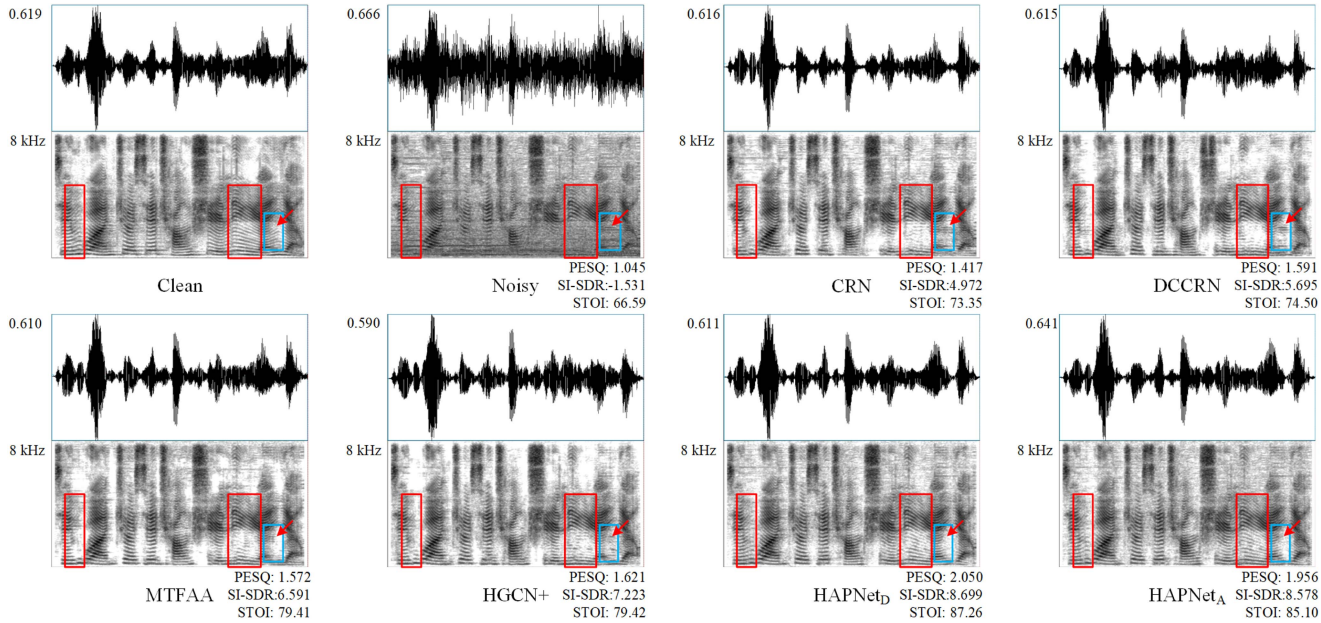

Fig. 7. Waveform and spectrum of a sample mixed with a harmonic-like noise at −6 dB SNR and the enhanced ones of HAPNet and referenced models (causal setup). And, the PESQ, SI-SDR, and STOI are shown in the lower right corner.

integration in the HGCN+ is rigid and unlearnable, which must follow a coarse enhancement module to prevent the trouble caused by harmonic-like noise. However, suppressing the noise at low SNR is difficult for the CRN-based coarse enhancement module with finite capability. The leaked harmonic-like noise makes the harmonic integration inaccurate, resulting in degraded performance. Compared to HGCN+, the learnable harmonic attention module of HAPNet captures harmonics and filters out the non-speech information simultaneously, which enables the model to extract speech well in the case of harmonic-like noise of low SNRs.

The pitches are correlated in a short time, so the improvement from non-causality is significant. The experimental results show that the improvement of non-causal HGCN+ compared to the causal one is limited in the low SNR due to the misled compensation module by inaccurate harmonic prediction. In contrast, the harmonics captured by harmonic attention module in HAPNet prompts the temporal dependencies to be uncovered; thus, the non-causal HAPNet achieves the most significant improvement compared to other models.

D. Visual Analysis

The harmonics are well-marked in the spectrum with high energy. We visualize the spectra enhanced by models to compare the capability to recover harmonics, as shown in Fig. 7. The mixture of music noise and engine noise is mixed with a clean speech at a −6 dB SNR. The causal models are used to enhance the degraded speech. The noisy spectrum shows that although the overall energy of the harmonics is high, they are still partially masked by the noise at a low SNR (colorful boxes in Fig. 7). Note that the energy of the high-frequency is relatively amplified due to the pre-emphasis and energy compression of the visualization tool. And low-energy high-frequency information affects human hearing less. PESQ, SI-SDR, and STOI are calculated to indicate speech quality.

As seen from the red boxes, it is difficult for non-harmonic-modeling models to capture and reconstruct those seriously distorted harmonic structures of speech. The harmonics in the spectrum enhanced by HGCN+ are better but still imperfect due to the interference of harmonic-like noise. Since the CRN, DCCRN, and MTFAA are only sensitive to high-energy regions, and the integral module in the HGCN+ is not resistant to harmonic-like noises, they mistakenly enhance the harmonic-like noises, as shown in the blue boxes.

It is clear from the spectra enhanced by HAPNet$_D$ and HAPNet$_A$ that the harmonic attention module is adequate for harmonic restoration. The distorted harmonics in the red box and most of the harmonics in the blue box are filled correctly. More complete harmonics result in a better quality (PESQ, SI-SDR, and STOI are much higher than referenced methods). In addition, the continuity of harmonics in the HAPNet$_D$-based enhanced spectrum is better than that in the HAPNet$_A$-based enhanced one. Considering that the RNN-based temporal module treats each frame equally, while the attention-based temporal module focuses only on frames with strong correlations, we speculate that the RNN-based temporal module is more suitable for causal speech enhancement, focusing on the short-time continuous information rather than the long-time information.

E. Comparison of Loss Functions

After the model structure is determined, the presented LC-SNR is compared with the widely used loss functions MSE,

TABLE V
TEST RESULTS OF HAPNET$_D$ TRAINED BY DIFFERENT LOSS FUNCTIONS

	PESQ	nPE	STOI	nST	SI-SDR(%)	nSI	CI
MSE	2.702	-0.534	96.399	-0.166	18.717	0.221	-0.160
PMSQE	**3.299**	**0.336**	96.477	-0.141	11.232	-1.149	-0.318
SI-SNR	3.071	0.003	97.331	0.129	19.993	0.454	0.196
LC-SNR	3.203	0.195	**97.484**	**0.178**	**20.100**	**0.474**	**0.282**

NPE, NST, and NSI denote the normalized PESQ, STOI, and SI-SDR.

SI-SNR, and PMSQE, as shown in Table V. Since some of the loss functions are strongly correlated to a particular metric (for example, PMSQE is a differentiable version of PESQ), and a good loss function should be able to uniformly improve all metrics, and the composite indicator (CI) is used as the primary referenced metric, which is calculated as follows,

$$\mathrm{CI} = \left[\sum_{\mathrm{Me}=\mathrm{Metric}}^{\mathrm{Metrics}} mean_m\left(\frac{\mathrm{Me} - mean_a(\mathrm{Me})}{std_a(\mathrm{Me})}\right)\right]/3 \quad (18)$$

where Metrics $= \{\mathrm{PESQ}, \mathrm{STOI}, \mathrm{SI\text{-}SDR}\}$, $mean_a(\mathrm{PESQ})$ and $std_a(\mathrm{PESQ})$ denote the average and standard deviation of PESQs from all models' results. $mean_m(\mathrm{PESQ})$ denotes the average of PESQs from the evaluated model's results, and the result in the normalized metrics nPE, nST, and nSI (PESQ, STOI, SI-SDR).

MSE is the most straightforward loss function that measures the distance between the enhanced signal and the referenced one. But the absence of auditory effects makes the results of MSE mediocre. SI-SNR calculates the SNR in the time domain, resulting a good overall performance but mediocre performance on PESQ. PMSQE is a differentiable version of PESQ. It improves the PESQ substantially, but the biased loudness-domain guidance leads to the severe degradation of the other metrics. The results show that LC-SNR performs well on all metrics and gets the highest composite indicator (0.282). Measuring the SNR on a complex spectrogram compressed by loudness power can simultaneously take into account auditory effects and numeric values of speech.

F. Comparison on Public Test Dataset

Models are compared on the public on-reverb test set of DNS-Challenge 2020 [30]. In addition to the models trained on the 100-hours dataset in our experiments, some excellent causal models are introduced. "DataSee" is the duration of data seen by the model during training, where "dynamic" indicates that the data is randomly generated during training. Note that our experiments trained HGCN+ with the same convolutional configuration as DCCRN based on SI-SNR. HAPNet$_{\mathrm{D(SI\text{-}SNR)}}$ and HAPNet$_{\mathrm{D(LC\text{-}SNR)}}$ are trained based on the SI-SNR and LC-SNR. The results in the Table VI show that the performance of HAPNet$_{\mathrm{D(SI\text{-}SNR)}}$ trained on data of 100 hours is comparable to that of the referenced models trained on data of 3000+ hours. There might be two reasons for this. First, compared to the speech (10 s or 30 s) used in other papers, the 100-hours data set used in our experiments is more efficient because of the 5 s duration leading to more samples with more noise-clean combinations. Second, HAPNet focuses on the patterns within the features, which enables itself to learn a more robust acoustic structure on a limited dataset, resulting in better results. In addition, HAPNet$_{\mathrm{D(LC\text{-}SNR)}}$ achieves the better results in both causal and non-causal models, which indicates that LC-SNR is desirable when the model is fixed. Finally, we trained the causal HAPNet$_{\mathrm{D(LC\text{-}SNR)}}$ based on a 500-hours dataset and obtained the SOTA performance.

Generally, the presented model and the loss function outperform the referenced method on the public test set with less training data and parameters.

TABLE VI
SYSTEM COMPARISON ON DNS-2020 SYNTHETIC TEST SET

Model	Cau.	Para.(M)	DataSee(H)	PESQ$_{\mathrm{WB}}$	PESQ$_{\mathrm{NB}}$	STOI(%)
Noisy	-	-	-	1.58	2.45	91.52
DCCRN [20]	√	3.67	dynamic	-	3.27	-
GaGNet [22]	√	5.94	300	3.17	3.56	97.13
FRCRN [60]	√	10.27	3000	**3.23**	3.60	**97.69**
HGCN+ [40]	√	5.29	3500	3.19	**3.65**	97.23
TaylorSENet [23]	√	5.40	3000	3.22	3.59	97.36
CRN	√	1.72	100	2.57	3.11	95.48
DCCRN	√	3.67	100	2.70	3.26	96.24
HGCN+	√	7.60	100	2.84	3.36	96.73
MTFAA	√	2.19	100	2.74	3.29	96.53
HAPNet$_{\mathrm{D(SI\text{-}SNR)}}$	√	1.67	100	3.07	3.53	97.33
HAPNet$_{\mathrm{A(SI\text{-}SNR)}}$	√	**1.52**	100	3.00	3.46	97.29
HAPNet$_{\mathrm{D(LC\text{-}SNR)}}$	√	1.67	100	**3.20**	**3.61**	**97.48**
HAPNet$_{\mathrm{A(LC\text{-}SNR)}}$	√	**1.52**	100	3.11	3.56	97.37
extra: HAPNet$_{\mathrm{D(LC\text{-}SNR)}}$	√	1.67	500	**3.33**	**3.72**	**97.90**
CRN	×	2.10	100	2.65	3.19	95.96
DCCRN	×	4.99	100	2.76	3.33	96.61
HGCN+	×	9.30	100	2.95	3.48	97.12
MTFAA	×	2.19	100	2.91	3.45	97.14
HAPNet$_{\mathrm{D(SI\text{-}SNR)}}$	×	1.67	100	3.24	3.67	97.89
HAPNet$_{\mathrm{A(SI\text{-}SNR)}}$	×	**1.52**	100	3.13	3.56	97.64
HAPNet$_{\mathrm{D(LC\text{-}SNR)}}$	×	1.67	100	**3.40**	**3.79**	**98.10**
HAPNet$_{\mathrm{A(LC\text{-}SNR)}}$	×	**1.52**	100	3.29	3.68	97.72

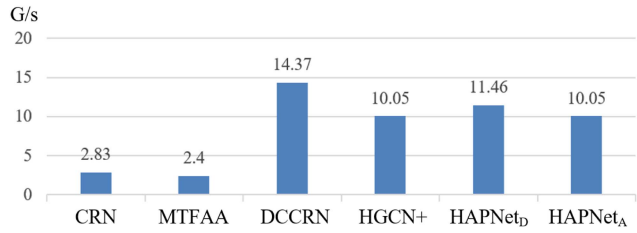

Fig. 8. The number of multiply-accumulate operations (MACs) of the presented and referenced models per second for wide-band signal.

G. Model Complexity

The number of parameters and the complexity of the model are essential during deployment. In addition to the parameters in the previous experiments, the opened tool OpCounter[4] is used to test the number of multiply-accumulate operations (MACs) of models. The results are shown in Fig. 8. Which shows that the computational complexity of HAPNet is acceptable compared

[4][Online]. Available: https://github.com/Lyken17/pytorch-OpCounter

to other models that have been submitted to the DNS-Challenge real-time speech enhancement track.

So far, it has been proved that HAPNet achieves satisfactory results in all aspects because of the excellent performance, small number of parameters, and acceptable complexity for the real-time processing.

VI. Discussion and Conclusion

Fitting complex distributions with more parameters is a prerequisite for the success of deep-learning-based methods. However, the redundant parametric modeling and the blinding introduction are not desirable. The articulatory and auditory mechanisms are the key to a better hearing quality with fewer parameters. The articulatory mechanism determines the time-frequency spectrum structure of speech and offers inspiration for modeling. On the other hand, the auditory mechanism determines the metric of the final result, which is appropriate for the loss function.

From the perspective of articulatory mechanisms, we chose the most prominent feature, comb-like harmonics. We present harmonic attention module, which explicitly recovers the speech components distorted by noise based on the special pattern of harmonics. The experimental results demonstrate that the compensation with the help of surviving harmonics brings more complete speech components and better auditory qualities for speech enhancement. And, the learnable integral process makes the model robust to harmonic-like noise. In addition, the preservation of feature structure is the key to introducing empirical articulatory mechanisms into the model, rendering the conventional Encoder-Decoder system unavailable; thus, a harmonic-attention progressive network (HAPNet) is applied. The complexity comparison experiments and the model parameters show that the empirical articulatory modeling effectively alleviates the computational burden caused by high-dimensional features. In conclusion, modeling feature structure and the progressive enhancement strategy are both essential to each other.

From the perspective of auditory characteristics, introducing too many auditory effects to the loss function will make it difficult to balance the model training, resulting in distortion of results. A good loss function should be able to improve multiple metrics synthetically. So, a loss function based on loudness power compression (LC-SNR) is used. The simultaneously measuring of magnitude and phase information with appropriate auditory effects leads to the superior performance of the model trained by LC-SNR with less distortion.

In summary, we design a harmonic attention module to capture the residual harmonics in the degraded speech. And a single-channel speech enhancement model, HAPNet, consisting of stacked harmonic attention module and temporal modules, is adopted to progressively recover the speech. In addition, a loss function based on the loudness power compression, named LC-SNR, is applied to persuade the enhanced speech to approximate the target speech both numerically and audibly. We conducted extensive experiments based on the DNS-Challenge dataset, the Aurora noise set, and some collected music noises. The experimental results prove that the presented methods outperform the previous advanced methods. On the public DNS-Challenge testset, the presented method yields SOTA performance with less training data.

In the future, the harsh-environment speech enhancement and target-speaker enhancement need to be further investigated. We will focus on the introduction of articulatory and auditory characteristics in deep learning and explore more effective and general methods for both human hearing and downstream tasks.

References

[1] J. Benesty, S. Makino, and J. Chen, *Speech Enhancement*. Berlin, Germany: Springer, 2006.

[2] K. Paliwal, K. Wójcicki, and B. Schwerin, "Single-channel speech enhancement using spectral subtraction in the short-time modulation domain," *Speech Commun.*, vol. 52, no. 5, pp. 450–475, 2010.

[3] F. Asano, S. Hayamizu, T. Yamada, and S. Nakamura, "Speech enhancement based on the subspace method," in *IEEE Trans. Speech Audio Process.*, vol. 8, no. 5, pp. 497–507, Sep. 2000.

[4] V. Sunnydayal, N. Sivaprasad, and T. K. Kumar, "A survey on statistical based single channel speech enhancement techniques," *Int. J. Intell. Syst. Appl.*, vol. 6, no. 12, pp. 69–85, 2014.

[5] I. Almajai and B. Milner, "Visually derived wiener filters for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 6, pp. 1642–1651, Aug. 2011.

[6] N. Roman, D. Wang, and G. J. Brown, "Speech segregation based on sound localization," *J. Acoust. Soc. Amer.*, vol. 114, no. 4, pp. 2236–2252, 2003.

[7] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1849–1858, Dec. 2014.

[8] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 1, pp. 7–19, Jan. 2015.

[9] Y. Zhao, D. Wang, B. Xu, and T. Zhang, "Monaural speech dereverberation using temporal convolutional networks with self attention," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1598–1607, 2020.

[10] D. Wang and J. Lim, "The unimportance of phase in speech enhancement," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 30, no. 4, pp. 679–681, Aug. 1982.

[11] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Proc. Interspeech*, 2013, pp. 436–440.

[12] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 65–68, Jan. 2014.

[13] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *Speech Commun.*, vol. 53, no. 4, pp. 465–494, 2011.

[14] S. Pascual, A. Bonafonte, and J. Serrà, "SeGAN: Speech enhancement generative adversarial network," in *Proc. Interspeech*, 2017, pp. 3642–3646.

[15] A. Pandey and D. Wang, "A new framework for CNN-based speech enhancement in the time domain," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 7, pp. 1179–1188, Jul. 2019.

[16] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.

[17] A. Défossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," in *Proc. Interspeech*, pp. 3291–3295, 2020.

[18] Z. Kong, W. Ping, A. Dantrey, and B. Catanzaro, "Speech denoising in the waveform domain with self-attention," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 7867–7871.

[19] K. Tan and D. Wang, "Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 6865–6869.

[20] Y. Hu et al., "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," in *Proc. Interspeech*, pp. 2472–2476. [Online]. Available: https://www.isca-speech.org/archive/interspeech_2020/hu20g_interspeech.html

[21] A. Li, W. Liu, X. Luo, C. Zheng, and X. Li, "ICASSP 2021 deep noise suppression challenge: Decoupling magnitude and phase optimization with a two-stage deep network," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 6628–6632.

[22] A. Li, C. Zheng, L. Zhang, and X. Li, "Glance and gaze: A collaborative learning framework for single-channel speech enhancement," *Appl. Acoust.*, vol. 187, 2022, Art. no. 108499.

[23] A. Li, S. You, G. Yu, C. Zheng, and X. Li, "Taylor, can you hear me now? A taylor-unfolding framework for monaural speech enhancement," in *Proc. Int. Joint Conf. Artif. Intell. Org.*, L. D. Raedt, Ed. Jul., 2022, pp. 4193–4200. [Online]. Available: https://www.ijcai.org/proceedings/2022/582

[24] X. Hao, X. Su, R. Horaud, and X. Li, "FullsubNet: A full-band and sub-band fusion model for real-time single-channel speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 6633–6637.

[25] J. Chen, Z. Wang, D. Tuo, Z. Wu, S. Kang, and H. Meng, "FullsubNet: Channel attention fullsubnet with complex spectrograms for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 7857–7861.

[26] H. Schroter, A. N. Escalante-B, T. Rosenkranz, and A. Maier, "DeepfilterNet: A low complexity speech enhancement framework for full-band audio based on deep filtering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 7407–7411.

[27] A. Li, W. Liu, C. Zheng, C. Fan, and X. Li, "Two heads are better than one: A two-stage complex spectral mapping approach for monaural speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1829–1843, 2021.

[28] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 3, pp. 483–492, Mar. 2016.

[29] C. Trabelsi et al., "Deep complex networks," in *Proc. Int. Conf. Learn. Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=H1T2hmZAb

[30] M. Strake, B. Defraene, K. Fluyt, W. Tirry, and T. Fingscheidt, "Interspeech 2020 deep noise suppression challenge: A fully convolutional recurrent network (FCRN) for joint dereverberation and denoising," in *Proc. Interspeech*, 2020, pp. 2467–2471.

[31] C. K. Reddy et al., "Interspeech 2021 deep noise suppression challenge," in *Proc. Interspeech*, 2021, pp. 2796–2800. [Online]. Available: https://www.isca-speech.org/archive/interspeech_2021/le21b_interspeech.html

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[33] V. Mitra, H. Nam, C. Y. Espy-Wilson, E. Saltzman, and L. Goldstein, "Articulatory information for noise robust speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 1913–1924, Sep. 2011.

[34] E. Plourde and B. Champagne, "Auditory-based spectral amplitude estimators for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 8, pp. 1614–1623, Nov. 2008.

[35] W. Jiang, Z. Liu, K. Yu, and F. Wen, "Speech enhancement with neural homomorphic synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 376–380.

[36] L. Welling and H. Ney, "Formant estimation for speech recognition," *IEEE Speech Audio Process.*, vol. 6, no. 1, pp. 36–48, Jan. 1998.

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

[38] 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. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 5560–5564.

[39] T. Wang, W. Zhu, Y. Gao, J. Feng, and S. Zhang, "HGCN: Harmonic gated compensation network for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 371–375.

[40] T. Wang, W. Zhu, Y. Gao, Y. Chen, J. Feng, and S. Zhang, "Harmonic gated compensation network plus for ICASSP 2022 DNS challenge," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 9286–9290.

[41] S. S. Stevens, J. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *J. Acoust. Soc. Amer.*, vol. 8, no. 3, pp. 185–190, 1937.

[42] B. C. Moore and B. R. Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," *J. Acoust. Soc. Amer.*, vol. 74, no. 3, pp. 750–753, 1983.

[43] G. Zhang, L. Yu, C. Wang, and J. Wei, "Multi-scale temporal frequency convolutional network with axial attention for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 9122–9126.

[44] A. Li, C. Zheng, R. Peng, and X. Li, "On the importance of power compression and phase estimation in monaural speech dereverberation," *JASA Exp. Lett.*, vol. 1, no. 1, 2021, Art. no. 014802.

[45] J. M. Martin-Doñas, A. M. Gomez, J. A. Gonzalez, and A. M. Peinado, "A deep learning loss function based on the perceptual evaluation of the speech quality," *IEEE Signal Process. Lett.*, vol. 25, no. 11, pp. 1680–1684, Nov. 2018.

[46] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 11, pp. 2009–2022, Nov. 2016.

[47] H.-G. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. ASR2000- Autom. Speech Recognit.: Challenges New Millenium ISCA Tut. Res. Workshop*, 2000, pp. 181–188. [Online]. Available: https://www.isca-speech.org/archive_open/asr2000/asr0_181.html

[48] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[49] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1026–1034.

[50] T. N. Sainath et al., "Deep convolutional neural networks for large-scale speech tasks," *Neural Netw.*, vol. 64, pp. 39–48, 2015.

[51] A. Camacho, *SWIPE: A Sawtooth Waveform Inspired Pitch Estimator for Speech and Music*. Univ. Florida Gainesville, 2007. [Online]. Available: https://d1wqtxts1xzle7.cloudfront.net/50152752/dissertation-libre.pdf?1478485226=&response-content-disposition=inline%3B+filename%3DA_sawtooth_waveform_inspired_pitch_estim.pdf&Expires=1687313133&Signature=OP2Z9GpCf095zaAZqO563O7fPShLXc9~Bp7sGfOm8MbOeSOHjROfcoM0jeJz43QlTGwViCfwHBjVP0rTyP5FEDSxTR3hH70KwrrIfqoIkzIC5F9ztGnev~7~3-bOrTH5N8pFG1GO0KQrzN6GiYaeGvBN0V0LHRBexVWjBKD5FgJTq9ZBisB26A0JDsuX~lMTDRBlVYyZm5Aic3yuq~LzKvATrlvEojp4~3eHIGe76k2qss0E9caedTHgumiLRARVxlz7hFUeoVb9iwEw6sWQt-9qCkHXwjqAo9KC-gjwppvYv37288Xoacepr5rTG5DsVa01ZDnD8VyzFN7DrelFSA__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA

[52] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," in *Proc. NIPS*, 2016. [Online]. Available: https://openreview.net/forum?id=BJLa_ZC9

[53] J. S. Bridle, "Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition," in *Neurocomputing*, Berlin, Germany: Springer, 1990, pp. 227–236. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-642-76153-9_28

[54] A. Vaswani et al., "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[55] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 46–50.

[56] X. Le, H. Chen, K. Chen, and J. Lu, "DPCRN: Dual-path convolution recurrent network for single channel speech enhancement," in *Proc. Interspeech*, 2021, pp. 2811–2815. [Online]. Available: https://www.isca-speech.org/archive/interspeech_2021/le21b_interspeech.html

[57] B. J. Borgström and M. S. Brandstein, "Speech enhancement via attention masking network (SEAMNET): An end-to-end system for joint suppression of noise and reverberation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 515–526, 2021.

[58] G. Yu, A. Li, C. Zheng, Y. Guo, Y. Wang, and H. Wang, "Dual-branch attention-in-attention transformer for single-channel speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 7847–7851.

[59] Y. Fu et al., "Uformer: A unet based dilated complex & real dual-path conformer network for simultaneous speech enhancement and dereverberation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 7417–7421.

[60] S. Zhao, B. Ma, K. N. Watcharasupat, and W.-S. Gan, "FRCRN: Boosting feature representation using frequency recurrence for monaural speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 9281–9285.

[61] E. Zwicker and B. Scharf, "A model of loudness summation," *Psychol. Rev.*, vol. 72, no. 1, pp. 3–26, 1965.
[62] Y. Isik, J. L. Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," in *Proc. Interspeech*, 2016, pp. 545–549. [Online]. Available: https://www.isca-speech.org/archive/interspeech_2016/isik16_interspeech.html
[63] H. Dubey et al., "ICASSP 2022 deep noise suppression challenge," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 9271–9275.
[64] L. Sun, J. Du, L. R. Dai, and C. H. Lee, "Multiple-target deep learning for LSTM-RNN based speech enhancement," in *Proc. Hands-Free Speech Commun. Microphone Arrays*, 2017, pp. 136–140.
[65] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Representations*, 2015. [Online]. Available: https://dblp.org/rec/journals/corr/KingmaB14.html?view=bibtex
[66] A.W. Rix, J.G. Beerends, M.P. Hollier, and A.P. Hekstra, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2001, vol. 2, pp. 749–752. [Online]. Available: https://ieeexplore.ieee.org/document/941023
[67] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 626–630.

Tianrui Wang received the B.E. degree in Internet of Things engineering from the North University of China, Taiyuan, China, in 2020. He is currently working toward the M.S. degree in information and communication engineering with the Institute of Information Science and the Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing Jiaotong University, Beijing, China. His research interests include speech enhancement and speech recognition.

Weibin Zhu received the Ph.D. degree from the Chinese Academy of Sciences, Beijing, China, in 1998. He is currently an Associate Professor with the Institute of Information Science and the Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing Jiaotong University, Beijing. During 1998–2005, he was a Research Staff Member with the IBM China Research Lab, Beijing. His research interests include acoustic model for speech recognition, prosody, and emotion model for speech synthesis, and deep learning for machine hearing.

Yingying Gao received the Ph.D. degree from the Institute of Information Science, Beijing Jiaotong University, Beijing, China, majoring in speech emotion generation modeling, in 2016. She is currently with the Artificial Intelligence and Intelligent Operation Center of China Mobile Research Institute, mainly engaged in speech recognition and end-to-end universal modeling.

Shilei Zhang received the B.S. and M.S. degrees in automation from Tianjin University, Tianjin, China, in 2001 and 2004, respectively, and the Ph.D. degree in speech processing from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2007. From 2007 to 2020, he was a Senior Researcher with IBM research - China. In 2021, he joins China Mobile Research Institute and leads speech team focusing on speech related research and application. He has authored or coauthored more than 50 papers in important academic conferences and journals, such as ICASSP, INTERSPEECH, IJCAI, and *Neural Networks*. His research interests maily include speech recognition, speech synthesis, speaker recognition, audio analysis, and multimodal recognition.

Junlan Feng (Fellow, IEEE) received the Ph.D. from the Chinese Academy of Sciences. She is/was a Chief Scientist of China Mobile Corp and the Board Chair of Linux Foundation Network. In 2001, she joined AT&T Labs Research as a Principal Researcher on speech recognition, language understanding and data mining till 2013. She has led the R&D Team on artificial intelligence with China Mobile since then. She has more than 100 technical journal and conference publications and more than 70 issued patents.