Innovative Directional Encoding in Speech Processing: Leveraging Spherical Harmonics Injection for Multi-Channel Speech Enhancement

Jiahui Pan, Pengjie Shen, Hui Zhang, Xueliang Zhang
College of Computer Science, Inner Mongolia University,
National & Local Joint Engineering Research Center of
Intelligent Information Processing Technology for Mongolian,
Inner Mongolia Key Laboratory of Mongolian Information Processing Technology
{panjiahui@mail.imu.edu.cn, {cszh,cszxl}@imu.edu.cn

Abstract

Multi-channel speech enhancement leverages multiple microphones to extract target speech signals amid background noise. Effectively utilizing directional cues is key for robust enhancement. While deep learning shows promise for multi-channel speech processing, most methods operate on short-time Fourier transform (STFT) coefficients directly. We propose using spherical harmonics transform (SHT) coefficients as auxiliary inputs to models. which concisely represent spatial distributions. SHT allows signals from varying numbers of microphones to be converted into coefficients of a consistent dimension. The proposed technique enables a single model to generalize to microphone arrays with varying configurations, rather than requiring a specialized model for each array layout. We present two architectures with SHT-based auxiliary inputs: parallel and serial. Specifically, the parallel model contains two encoders - one for STFT and another for SHT. By fusing both encoders' outputs in the decoder to estimate the enhanced STFT, it effectively incorporates spatial context. For the serial approach, we first apply SHT to the signals and then take STFT of the transformed signals as network inputs. Evaluations of the TIMIT dataset under fluctuating noise and reverberation demonstrate our model outperforms established benchmarks. Remarkably, these results are attained with reduced computations and parameters. Furthermore, experiments on the MS-SNSD dataset show the proposed method can enhance the generalization ability of networks. The source code is publicly accessible at https://github.com/Pandade1997/SH_injection.

1 Introduction

Multi-channel speech enhancement involves extracting a desired speech signal from noisy environments using data captured by multiple microphones. This technique is critical for applications including, but not limited to, video conferencing systems [Mehrotra *et al.*, 2011; Rao *et al.*, 2021], distant communication [Nguyen *et al.*, 2014; Tan *et al.*, 2019], and hearing aids [Doclo *et al.*, 2010; Nossier *et al.*, 2019]. Unlike single-channel methods that rely solely on spectral and temporal properties, multi-channel enhancement uniquely capitalizes on spatial information. By exploiting spatial cues like inter-channel differences, multi-channel systems can substantially improve speech clarity, background noise reduction, and overall listening experience compared to single-channel techniques. However, effectively integrating and processing spatial cues remains an open challenge.

The evolution of multi-channel speech enhancement techniques can be broadly categorized into two main approaches: traditional spatial filtering methods and deep learning techniques. Traditional methods focus on extracting signals based on differences in phase and timing across microphones, which struggle in noisy environments due to difficulties accurately estimating spatial information [Hoshuyama *et al.*, 1999; Souden *et al.*, 2009]. The second approach involves deep learning methods, which initially processed each channel independently [Wang and Wang, 2018] and later evolved into more advanced techniques such as merging deep neural networks with traditional beamforming [Li *et al.*, 2024] to combine the strengths of deep learning and array processing. However, the short-time Fourier transform (STFT) representation has difficulty expressing the spatial information of sound sources. Effectively incorporating spatial information remains a challenge. Independent per-channel processing fails to capture inter-channel dependencies and spatial relationships, which provide valuable context. Directly modeling full multi-channel spatial correlations is computationally infeasible. More efficient spatial feature extraction is needed to incorporate spatial information without excessive complexity and fully capitalize on spatial diversity in multi-channel scenarios.

Fortunately, spherical harmonic coefficients(SHCs) obtained via SHT provide a comprehensive spatial representation of soundfields [Kumar and Hegde, 2016; Varanasi *et al.*, 2020]. This spherical harmonic representation offers two key advantages for multi-channel speech enhancement: **Effective capture of spatial information [Lugasi and Rafaely, 2020]**: Unlike the STFT, SHT primarily captures the spatial distribution characteristics of soundfields. Grounded in spherical harmonics theory, the SHT discerns the spatial attributes of sig-

nals arriving from various directions and their inter-relations across microphone channels. Such spatial capture is crucial for multi-microphone speech enhancement. In a microphone array, this transform adeptly captures the spatial orientation of sounds, enabling more precise differentiation between target speech and background noise. **Enhanced spatial resolution [Rafaely, 2015]**: Spherical harmonics constitute a complete basis for functions defined on the spherical surface. Thus, any spherical function can be represented precisely as a linear combination of spherical harmonics. The SHT facilitates an accurate depiction of a sound field's spatial distribution. The order of the spherical harmonics determines the granularity of spatial feature capture. While lower orders delineate broad spatial patterns, higher orders characterize finer spatial nuances. By selecting a suitable order, the desired spatial resolution can be achieved. If the spatial information in these SHCs can be fully utilized, this may help compensate for the lack of descriptive spatial modeling in current mainstream approaches. Exploiting this could greatly improve the performance and robustness of multi-channel speech enhancement. Furthermore, SHT has the property of converting signals from varying numbers of microphones into coefficients of consistent dimension. This property can be utilized to address the problem of generalizing to varying numbers of microphones by formulating multi-channel speech enhancement independently of the number of input microphones. As a result, deeper neural networks can be employed for the task.

In this work, we propose a method to fully leverage spatial information using SHCs as auxiliary model inputs. We design two network variants that incorporate SHT as auxiliary inputs: parallel and serial approaches. Specifically, the parallel model contains two separate encoders - one that takes the STFT of the input signals, and another encoder that processes the SHCs obtained by applying SHT to the inputs. By fusing both encoders in the decoder to estimate the enhanced STFT, we effectively incorporate spatial context. For the serial approach, SHT is first applied to transform the multi-channel inputs into the spherical harmonic domain. Then, STFT is taken of the transformed signals represented as SHCs to obtain features. Unlike the parallel approach which processes STFT and SHT separately, the serial approach applies SHT before STFT so the network inputs encompass both spectral and spatial information from the start. This serial use of SHT then STFT aims to more deeply integrate spatial cues into feature extraction versus direct STFT features. Our contributions can be summarized as three-fold:

- We integrate SHT into deep learning methods to enable improved spatial processing for multi-channel speech enhancement.
- We introduce two novel network architectures: a parallel model that handles STFT coefficients and SHCs independently, and a serial model for jointly processing the combined spatial-spectral data.
- We demonstrate the proposed model achieves superior performance across various conditions, effectively generalizes to different microphone array configurations

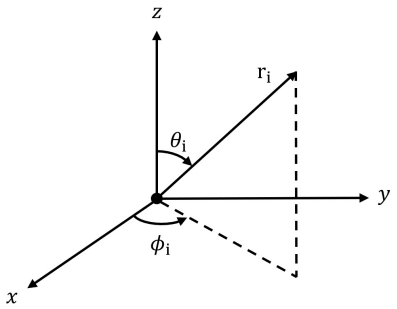

Figure 1: Defined spherical coordinate system.

2 Related Work

Traditional spatial filtering methods. Traditional approaches include spatial filtering methods such as delay-and-sum beamformer [Klemm *et al.*, 2008], minimum variance distortionless response (MVDR) beamformer [Souden *et al.*, 2009], super-directivity beamformer [Bitzer and Simmer, 2001], and others. These leverage phase and timing differences between microphones to preferentially extract signals from certain directions. Although these approaches can perform well, their performance depends on reliable estimation of spatial information, which can be challenging to accurately estimate in noisy conditions.

Deep learning techniques. Deep learning techniques have made significant progress in multi-channel speech enhancement recently. Earlier deep learning methods such as GCRN [Tan and Wang, 2019], DPCRN [Le *et al.*, 2021] focused on spectral mapping, processing each channel independently. To better preserve spatial cues, [Liu and Zhang, 2021] proposed the inplace gated convolutional recurrent neural network (IGCRN) which efficiently retains spatial information in each frequency bin without the downsampling and upsampling used in conventional CRNs. Later, [Tan *et al.*, 2022] introduced the concept of neural spectro-spatial filtering which jointly optimizes spectral and spatial filtering using a convolutional neural network with densely-connected blocks. This achieves significant gains over prior approaches for multi-microphone speech enhancement. More recent methods combine DNNs with traditional beamforming to better utilize spatial information. Examples like FasNet [Luo *et al.*, 2019], MIMO-Unet [Ren *et al.*, 2021], and EabNet [Li *et al.*, 2022] exploit complementary strengths of deep learning and array processing for state-of-the-art performance.

Spherical harmonic transformation. The spherical harmonic basis functions constitute a comprehensive set of basis functions in spherical coordinates, enabling the expression of any square-integrable function or star-shaped object in a mathematical closed form. Consequently, it is possible to decompose any sound field on the sphere into spherical harmonic components using the spherical harmonic transform, also known as the spherical Fourier transform, which employs the spherical harmonics. The primary applications in the field of speech involve sound field reconstruction and direction-of-arrival (DOA) estimation. [Xu *et al.*, 2021] utilized spherical harmonic decomposition and head-related impulse responses (HRIR) to establish a correlation

between spatial locations and received binaural audio signals. [Dwivedi *et al.*, 2022] introduced a low-complexity convolutional neural network-based method for DOA estimation in the spherical harmonic domain. [Weng *et al.*, 2023] put forward a solution to the problem of multi-source DOA estimation in challenging environments characterized by reverberation and noise, involving the utilization of a spherical harmonic domain beam-space MUSIC algorithm.

3 Methodology

We consider a set of microphone arrays of arbitrary configuration located at the origin of the Cartesian coordinates and composed of II omnidirectional microphones. Let $r_i = (r_i \cos\phi_i \sin\theta_i, r_i \sin\phi_i \sin\theta_i, r_i \cos\theta_i)^T$ denote the position of the i-th microphone of the array, where r_i represents the distance of the i-th microphone to the center of the array. The azimuth ϕ_i is measured counterclockwise from the x-axis, and the elevation angle θ_i is measured downward from the z-axis. The adopted spherical coordinate system is illustrated in Fig.1.

The array is assumed to be positioned in a reverberant sound field. According to the image method [Allen and Berkley, 1979], the sound pressure in a reverberant environment, generated by a single source in the far field, can be modeled as a sum of L significant plane waves produced by L image sources under free-field conditions. This can be assumed equivalently as L far-field sound sources generating plane waves propagating through space and picked up by the microphones. Let $\Psi_l = (\theta_l, \phi_l)$ denote the direction of propagation of the l-th sound source, and let $k_l = -(k \cos\phi_l \sin\theta_l, k \sin\phi_l \sin\theta_l, k \cos\theta_l)^T$ represent the wave number vector of the l-th plane wave.

3.1 Space Domain System Model

The signal received by the i-th microphone in the frequency domain can then be expressed as:

$$p_i(k) = \sum_{l=1}^{L} v_i(k, \Psi_l)\, s_l(k) + n_i(k), \tag{1}$$

where $v_i(k, \Psi_l)$ denotes the steering vector of the i-th microphone associated with the l-th plane wave. $s_l(k)$ is the complex amplitude of the l-th plane wave, and $n_i(k)$ is the noise received by the i-th microphone. The frequency domain received sound pressure model can be expressed in matrix form as:

$$p(k) = V(k, \Psi) s(k) + n(k), \tag{2}$$

where $V(k, \Psi)$ is the $I \times L$ dimensional direction matrix. $s(k) = [s_1(k), s_2(k), \ldots, s_L(k)]^T$ is the L dimensional source signal vector, $n(k) = [n_1(k), n_2(k), \ldots, n_I(k)]^T$ is the I dimensional zero-mean Gaussian white noise vector, and $n(k)$ is assumed to be uncorrelated with $s(k)$. By N-point STFT, the $p(k)$ in the T-F domain can be recorded as $P_i(t, f)$:

$$P_i(t, f) = H_i(t, f) X_i(t, f) + V_i(t, f), \tag{3}$$

where t represents frame index and f represents frequency bin obtained from STFT. $X_i(t, f)$ and $V_i(t, f)$ represent the target and noise component, respectively.

3.2 Spherical Harmonic Domain System Model

In this section, we describe the proposed method for modeling acoustic signals in the spherical harmonic domain through the utilization of the SHT. By calculating the coefficients of spherical harmonics, the received speech signal at a specific point on the sphere surface can be estimated. The spherical harmonics $Y_n^m(\theta, \phi)$ of order n ($n \in N$) and degree m ($m \in Z$ and $-n \le m \le n$) are defined as [Rafaely, 2015]:

$$Y_n^m(\theta, \phi) = \sqrt{\frac{(2n+1)}{4\pi} \frac{(n-m)!}{(n+m)!}} P_n^m(\cos\theta) e^{im\phi}, \tag{4}$$

where N is the truncated order of the soundfield, $(.)!$ is the factorial function, and P_n^m is the normalized associated Legendre polynomial. The spherical harmonic function $P_n^m(\cos\theta)$ captures the dependency on the elevation angle θ, while the complex exponential term $e^{im\phi}$ captures the dependency on the azimuth angle ϕ.

According to the Fourier acoustic principle, the sampled sound pressure $p(k, r)$ and its spherical harmonic domain representation $p_{nm}(k, r)$ at frequency k and angle (θ, ϕ) can be expressed as:

$$p(k, r) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} p_{nm}(k, r) Y_n^m(\theta, \phi). \tag{5}$$

As explained in [Rafaely, 2015], the coefficients p_{nm} diminish for k, r in a range smaller than N and can therefore be neglected for $n > N$. Hence, Eq.5 can be approximated for an appropriate finite order N:

$$p(k, r) \cong \sum_{n=0}^{N} \sum_{m=-n}^{n} p_{nm}(k, r) Y_n^m(\theta, \phi), \tag{6}$$

where N is the truncation order, $p(k, r)$ denotes the time-dependent amplitude of the sound pressure in free three-dimensional space, $p_{nm}(k, r)$ are the weights known as coefficients of the SHT, $k = 2\pi f/c$ is the wave number, f is the frequency, and c is the speed of sound in air. The coefficients $p_{nm}(k, r)$ are defined as [Rafaely, 2015]:

$$p_{nm}(k, r) = \int_0^{2\pi} \int_0^{\pi} p(k, \mathbf{r}) \left[Y_n^m(\theta, \phi)\right]^* \sin(\theta) \mathrm{d}\theta \mathrm{d}\phi, \tag{7}$$

where $(.)^*$ denotes complex conjugation. To satisfy the far-field condition, the distance d between the sound source and the center of the microphone array must exceed $8r^2 f/c$ [Meyer, 2001], where r is the array radius. This ensures negligible wavefront curvature effects. For $n \le N$, $p_{nm}(k, r)$ can be obtained as:

$$p_{nm}(k, r) \cong \frac{4\pi}{I} \sum_{i=1}^{I} p(k, \mathbf{r}_i) \left[Y_n^m(\theta_i, \phi_i)\right]^*, \tag{8}$$

where $\mathbf{r}_i = (r, \theta_i, \phi_i)$ is the location of the i-th physical microphone and I is the number of physical microphones.

3.3 Two variants of SH Injection

We propose a novel architecture that fully leverages spatial information by using SHCs as auxiliary inputs to the model. The key innovation is to introduce SHT for multi-channel speech enhancement to explore spatial clues. The microphone array signals are transformed into the spherical harmonic domain to obtain SHCs, $p_{nm}(k,r)$, up to order N, which compactly encode the spatial distributions.

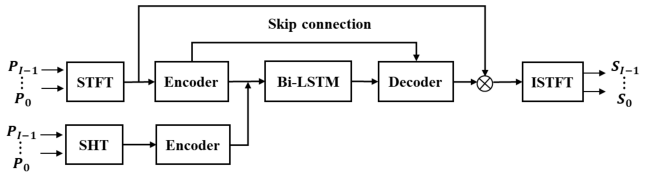

Figure 2: Proposed-parallel system.

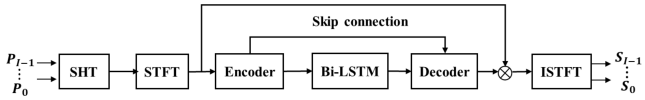

Figure 3: Proposed-serial system.

In proposed-parallel system, as is done in Fig. 2, the coefficients form SHT are then provided as auxiliary inputs to a dedicated spatial encoder, along with the spectrograms from the STFT fed to the main spectro-temporal encoder (as in IGCRN). Finally, the enhanced embeddings from both encoders are concatenated and passed to the decoder, which reconstructs the clean speech spectrogram. Using SHT provides orientation-invariant coefficients that concisely capture useful spatial properties and inject global contextual information about the soundfield to guide the model. The dual-encoder architecture enables joint modeling of spectral and spatial cues for improved speech enhancement.

In proposed-serial system, as is done in Fig. 3, the coefficients obtained from applying the SHT to the input audio signals are used as inputs to the STFT. This provides STFT coefficients that contain both spectral and spatial information from the beginning of the processing pipeline. The sequential application of SHT followed by STFT aims to deeply integrate spatial cues into the feature extraction process compared to directly using STFT features alone. By applying SHT first, spatial information from the spherical harmonics domain is incorporated into the spectral features computed by the subsequent STFT. This is intended to provide the neural network with inputs that integrate spatial cues more deeply than conventional STFT features. The algorithm outlining the proposed system is presented in Algorithm 1.

Algorithm 1 Algorithm for the proposed method.

Require:
A minibatch data $\{\mathbf{X}(t,f)_{mix}, \mathbf{p}_{nm}(k,r)_{mix}, \mathbf{s}(k)\}$. Where $\mathbf{X}(t,f)_{mix}$ is the result of STFT of mixed speech. $\mathbf{p}_{nm}(k,r)_{mix}$ denote the SHCs of order n and degree m, which are obtained by applying the SHT to the multi-channel mixed speech signal. $s(k)$ is the target speech. learning rate is μ_d.

Ensure:
The optimized proposed model

for number of training iterations **do**
  for m-th minibatch **do**
    if $proposed - parallel$ **then**
      $STFT_{out} = Encoder_{stft}(X(t,f)_{mix})$
      $SHT_{out} = Encoder_{sht}(p_{nm}(k,r)_{mix})$
      $LSTM_{out} = BiLSTM(STFT_{out}, SHT_{out})$
    else if $proposed - serial$ **then**
      $STFT_{out} = Encoder(STFT(p_{nm}(k,r)_{mix}))$
      $LSTM_{out} = BiLSTM(STFT_{out})$
    end if
    $est_{out} = Decoder(LSTM_{out})$
    $out = ISTFT(est_{out})$
    $Loss = MSE(out, s(k))$
    $\theta_d \leftarrow \theta_d - \mu_d \frac{\partial Loss}{\partial \theta_d}$
  end for
end for
return θ_d

4 Experiments

4.1 Datasets

To evaluate the performance and generalization of the proposed method, sound-field data was simulated using two datasets: TIMIT [Zue *et al.*, 1990] and MS-SNSD [Reddy *et al.*, 2019]. The simulation and evaluation process is described as follows:

TIMIT database. To evaluate the performance of the proposed system, the training data is generated by convolving multi-channel room impulse responses (RIRs) [Allen and Berkley, 1979] with speech signals from the TIMIT database. The TIMIT clips are divided into non-overlapping training, validation, and testing sets. Noise clips from the DNS-Challenge corpus are used for training and validation, while NOISEX-92 [Varga and Steeneken, 1993] and cafe noises from CHiME3 [Barker *et al.*, 2015] comprise the testing set. RIRs are generated using the image method based on a 9-microphone uniform circular array with radius 0.035 m, randomly positioned inside a $6 \times 5 \times 4\ \mathrm{m}^3$ room. The source-array distance is 1 m. RIRs are simulated with varying SNR (-6 to 6 dB) and RT60 (0.2 to 1 s) values. For the purpose of evaluation, we define three distinct SNR levels: -5dB, 0dB and 5dB. In addition, we explore five distinct T60 values, ranging from 0.2s to 0.6s, with intervals of 0.1s. This comprehensive configuration results in the generation of 350 pairs for each specific case.

MS-SNSD database. To assess the generalizability of the proposed system, we modified the MS-SNSD to employ gpuRIR's[Diaz-Guerra *et al.*, 2021] implementation of the image source method for room impulse response simulation. To simulate acoustic scenes, we selected a speech sample and a noise sample which were rescaled to random SNRs within the range of [-10, 10] dB and RT60 [0.2, 1.0]s. Sources were positioned randomly inside a room measuring $6 \times 5 \times 4\ \mathrm{m}^3$. The distance between sources and array was set to 1 m. The radius of the microphone array was 0.035 m. Microphone locations within the room were distributed uniformly at ran-

dom. Distinct testing and training sets were generated without overlap of speech or noise samples between sets. The number of microphones in the training and verification sets was 8, and the number of microphones in the test set was [4, 8, 12, 16]. The amount of data in the training set was 42,000 samples, the amount of data in the verification set was 4,200 samples, and the test set in each case contained 7,000 pieces of test data. Each scene simulation lasted two seconds.

4.2 Experiment Setup

Network Configuration

The proposed parallel and serial methods build upon the baseline IGCRN model [Liu and Zhang, 2021], which utilizes an encoder-decoder structure for multi-channel speech enhancement. For the Proposed-parallel system, the input features of the STFT and SHT are fed into two independent encoders. Each encoder consists of six cascaded 5×1 kernels inplace gated linear units (GLU), which are constructed using inplace convolutions as follows:

$$Y = ELU(BN(i\,\mathrm{Conv}(X) \otimes \mathrm{Sigmoid}(i\,\mathrm{Conv}(X)))), \quad (9)$$

where $ELU(.)$ and $Sigmoid(.)$ are activation functions, $BN(.)$ is batch normalization, iConv is inplace convolution, and $\otimes$ denotes element-wise multiplication. To achieve a similar computational cost and number of parameters as IGCRN, which has 64 GLUs, we set the number of input channels for each GLU here to 32. After the encoders, we concatenate their outputs along the channel dimension and feed them into a channel-wise LSTM to refine the spatial information. The decoder consists of six cascaded inplace transpose GLUs with 128 input channels per transpose GLU.

For the Proposed-serial systems, the coefficients obtained from applying the SHT to the input audio signals were used as inputs to the STFT. We set the number of input channels for each gated linear unit (GLU) here to 64.

Training Configuration

For the SHT, N = 4, resulting in 25 spherical harmonic functions: $Y_0^0(\theta,\phi)$, $Y_1^{-1}(\theta,\phi)$, $Y_1^0(\theta,\phi)$, $Y_1^1(\theta,\phi)$, $\cdots$, $Y_4^4(\theta,\phi)$. Then the complex value of each $Y_n^m(\theta_i,\phi_i)$ for the i-th microphone is specified. By employing (8) a set of $p_{nm}(k,r)$ is calculated which consists of 25 signals in the spherical harmonics domain.

All utterances had a frame length of 32 ms and a frame shift of 16 ms. The square-root Hann window was used as the analysis window. The sampling rate was 16 kHz. A 512-point discrete Fourier transform was used to extract complex STFT spectrograms. All models are trained using Adam optimizer with a fixed learning rate of 1e-3. If validation loss does not decrease for consecutive two epochs, the learning rate will be halved. All models are trained for 100 epochs.

4.3 Evaluation Metrics

In this study, the perceptual evaluation of speech quality (PESQ)[Rix *et al.*, 2001] and short-time objective intelligibility (STOI)[Taal *et al.*, 2011] metrics are used to objectively evaluate the enhancement performance of different models. PESQ assesses speech quality on a scale from -0.5 to 4.5, with higher scores indicating better quality. STOI gauges speech intelligibility on a scale of 0% to 100%, where higher scores reflect better intelligibility. Improved scores in both metrics reflect better performance.

4.4 Results and Discussions

In our experimental section, we carried out four main experiments. The first tested the overall performance of our method. The second, an ablation study, compared two model variants to identify the more effective one. The third assessed how the method performed with various microphone setups, checking its adaptability. The fourth extended these variants to more complex network architectures to show the method's generalizability. Together, these experiments comprehensively evaluated the method in different settings.

Performance of Proposed Structure

We compare against two baseline models: GCRN, which uses convolutional recurrent networks for complex spectral mapping, and IGCRN, which extends GCRN with inplace convolutions. We propose two variants of IGCRN:

- **Proposed-serial:** The coefficients obtained from applying the SHT to the input audio signals were used as inputs to the STFT, and then the STFT coefficients were fed into a single IGCRN model.
- **Proposed-parallel:** The SHT and STFT features are fed in parallel into two separate encoder branches of a dual-encoder IGCRN.

In Proposed-serial, the spatial and spectral features are combined into a single stream input to IGCRN. In Proposed-parallel, the SHT and STFT features are processed independently in dual encoder pathways before fusion. Both architectures augment the baseline with additional spherical harmonic spatial cues to enhance separation performance.

Experiments were conducted at -5dB, 0dB, and 5dB SNR levels, with 0.2s to 0.6s reverberation times. As shown in Figures 4 and 5, the results demonstrate the superiority of the two proposed models utilizing SHT over the baseline model without SHT. Specifically, at -5dB SNR, Proposed-parallel achieved an average PESQ score of 1.89, while Proposed-serial scored 1.79, compared to 1.74 for the baseline IGCRN. The STOI results followed a similar trend, with both proposed models surpassing the baseline STOI score. Overall, the results indicate that the proposed models can effectively incorporate and leverage the spatial cues from the SHT to achieve marked improvements in speech quality and intelligibility over the baseline model. By capturing the spatial information in the multi-channel input via spherical harmonics, the proposed models significantly outperform the baseline lacking this capability.

Ablation Study for Performance

We further analyze the performance difference between our Proposed-serial and Proposed-parallel models. At an SNR of -5dB, the mean gains observed are 0.1 for PESQ and 4.41% for STOI. As the SNR increases to 0dB, more substantial improvements are attained, with gains of 0.13 in PESQ and 2.69% in STOI. Finally, at the highest tested SNR level of 5dB, the enhancements over the unprocessed signals are 0.14 for PESQ and 1.41% for STOI, as shown in Figures

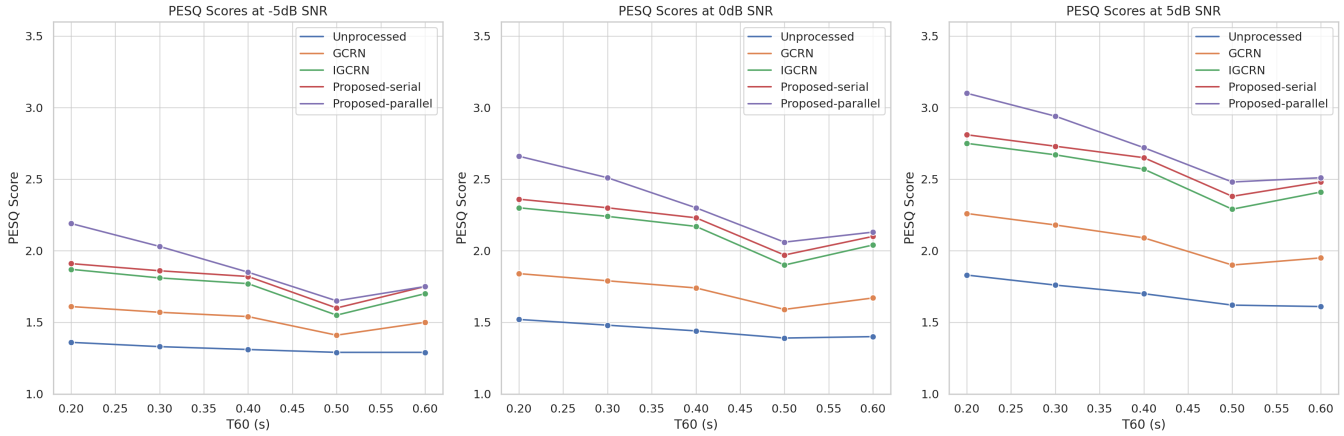

Figure 4: PESQ results comparing proposed models with baselines.

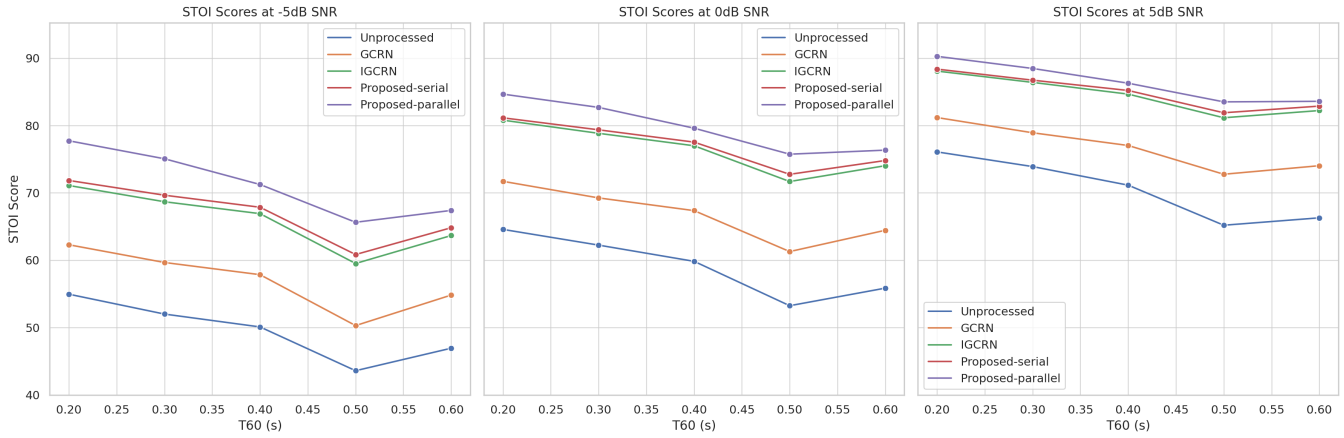

Figure 5: STOI results comparing proposed models with baselines.

4 and 5. The gains from parallel encoders remain consistent as we vary the noise and reverberation levels. This indicates that modeling the SHT and STFT features separately enables better representations to be learned, compared to serially processing the concatenated features. The network can train more specialized encoders when the inputs are independent. In contrast, the serial design forces the model to process the entire concatenated input in one encoder. This makes disentangling the spatial and spectrotemporal characteristics more challenging. The parallel approach does not have this constraint, allowing more robust joint representations to be formed after encoding. In summary, our ablation study demonstrates the superior performance achieved by modeling the SHT and STFT features in parallel architecture for multi-channel speech enhancement.

Performance of Generalization

By utilizing SHCs as auxiliary inputs, multi-channel speech signals can be transformed into a fixed dimension. This formulation allows designing multi-channel speech enhancement techniques independently of the microphone array size, thereby addressing the challenge of generalizing to different microphone configurations. This paper evaluates the proposed method's generalization ability under different numbers of microphones (specifically 4, 8, 12, and 16 microphones). Among these configurations, 8 microphones represent a matching condition, while the other three (4, 12, and 16 microphones) represent non-matching conditions. We use the CCAMS model [Casebeer *et al.*, 2021] as the baseline. The CCAMS model jointly trains a microphone selection mechanism and a speech enhancement network to perform multi-channel speech enhancement using an ad-hoc microphone array, without applying SHT to explore more comprehensive spatial representations. As shown in Figures 6 and 7, comparisons were made using the PESQ and STOI measures across four different microphone configurations. For PESQ, which reflects the perceived speech quality, the proposed method consistently outperformed both the unprocessed signals and the CCAMS model across all microphone configurations in both its serial and parallel formulations. Notably, the parallel version of the proposed method achieved a substantial increase in PESQ scores, particularly for MIC_12, indicating a significant improvement in perceived speech quality. Regarding STOI, which is indicative of speech intelligibility, the proposed method again demonstrated superior

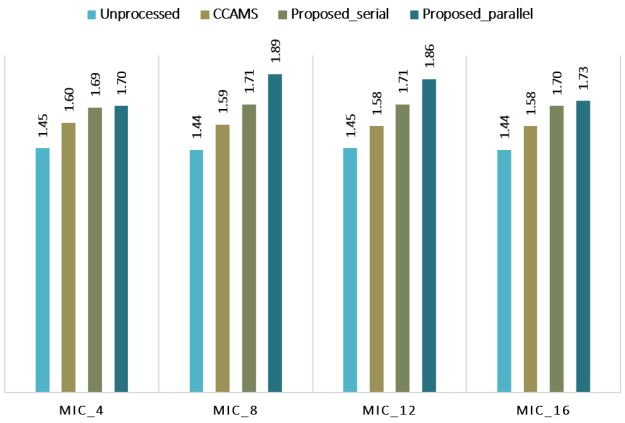

Figure 6: PESQ results comparing generalization in proposed models and baseline.

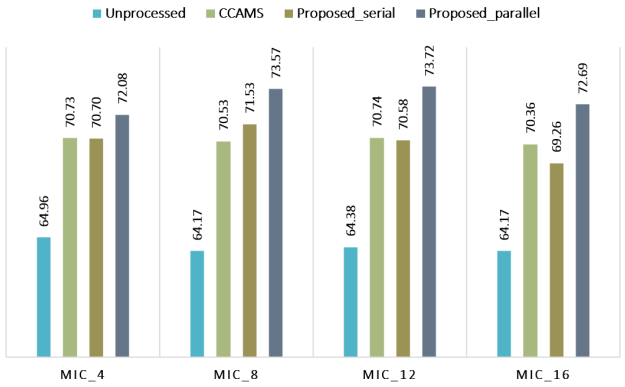

Figure 7: STOI results comparing generalization in proposed models and baseline.

performance over the unprocessed signals and the CCAMS model. The parallel version of the proposed method achieved the highest intelligibility scores, peaking at MIC_12. These results underscore the proposed method's ability to effectively enhance speech quality and intelligibility across varying numbers of microphones, highlighting its potential for robust application in diverse multi-channel speech processing scenarios.

Ablation Study for Generalization

In our research, we conducted ablation experiments to delve deeper into the potential of spherical harmonic information across different network architectures. We aimed to test two hypotheses: 1) Whether standardizing input dimensions through spherical harmonics would enable the use of more complex, deeper networks, thereby improving generalization, and 2) Whether our method could be effectively integrated with these deeper networks without affecting performance. Our baseline for comparison was the CCAMS model, which does not incorporate spherical harmonic information. We examined the performance of three distinct networks - IGCRN, ICCRN, and TFGrid [Wang *et al.*, 2023] - in serial and parallel configurations. As shown in Tables 1 and 2, our meticulously detailed PESQ and STOI scores for arrays of 4, 8, 12, and 16 microphones clearly showed that models using spherical harmonic information outperformed the baseline. Notably, the TFGrid network, especially in parallel, displayed a remarkable advantage over other models in all tested microphone configurations. This finding validated our first hypothesis by demonstrating the benefits of deeper networks when enabled by standardized spherical harmonic inputs. Furthermore, our second hypothesis was confirmed - the robust integration of our method with various network architectures was evidenced by consistent, notable performance improvements across models. This proves the flexibility and robustness of our approach, even when combined with complex networks.

Methods	MIC_4	MIC_8	MIC_12	MIC_16
Unprocessed	1.45	1.44	1.45	1.44
CCAMS	1.60	1.59	1.58	1.58
IGCRN_serial	1.69	1.71	1.71	1.70
ICCRN_serial	1.75	1.77	1.78	1.76
TFGrid_serial	2.23	2.34	2.36	2.31
IGCRN_parallel	1.70	1.89	1.86	1.73
ICCRN_parallel	1.76	1.79	1.79	1.75
TFGrid_parallel	**2.37**	**2.50**	**2.54**	**2.48**

Table 1: PESQ results for microphone number generalization.

Methods	MIC_4	MIC_8	MIC_12	MIC_16
Unprocessed	64.96	64.17	64.38	64.17
CCAMS	70.73	70.53	70.74	70.36
IGCRN_serial	70.70	71.53	70.58	69.26
ICCRN_serial	71.62	72.76	72.01	70.54
TFGrid_serial	81.10	82.74	82.22	80.51
IGCRN_parallel	71.54	77.17	76.75	73.52
ICCRN_parallel	72.08	73.57	73.72	72.69
TFGrid_parallel	**83.07**	**85.18**	**85.58**	**84.56**

Table 2: STOI results for microphone number generalization.

5 Conclusion

In this work, we propose a method to effectively utilize spatial information by incorporating SHT as an auxiliary input to models. We design two variants for incorporating SHT: parallel and serial approaches. Spherical harmonics provide a compact representation of spatial cues between microphones, improving multi-channel speech enhancement performance. By mapping speech signals to a fixed dimensional space via SHT, deeper networks can be employed to generalize to varying microphone numbers. Experiments demonstrate that both variants introducing SHT outperform established baselines with significantly reduced computations and parameters. By incorporating directional cues through SHT, our model effectively enhances speech. Ablation studies validate superior performance from modeling SHT and STFT features in parallel encoders rather than sequentially, highlighting benefits of joint spatial-spectral modeling. Moreover, our methodology demonstrates enhanced generalization across diverse microphone array configurations.

Acknowledgments

This work was partially supported by the China National Nature Science Foundation (No. 61876214, No. 61866030), the Innovation Research Project for Postgraduate Students at Inner Mongolia University, and the Fund for Supporting the Reform and Development of Local Universities (Disciplinary Construction).

References

[Allen and Berkley, 1979] Jont B Allen and David A Berkley. Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, 65(4):943–950, 1979.

[Barker *et al.*, 2015] Jon Barker, Ricard Marxer, Emmanuel Vincent, and Shinji Watanabe. The third 'chime'speech separation and recognition challenge: Dataset, task and baselines. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 504–511. IEEE, 2015.

[Bitzer and Simmer, 2001] Joerg Bitzer and K Uwe Simmer. Superdirective microphone arrays. In *Microphone arrays*, pages 19–38. Springer, 2001.

[Casebeer *et al.*, 2021] Jonah Casebeer, Jamshed Kaikaus, and Paris Smaragdis. Communication-cost aware microphone selection for neural speech enhancement with ad-hoc microphone arrays. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8438–8442. IEEE, 2021.

[Diaz-Guerra *et al.*, 2021] David Diaz-Guerra, Antonio Miguel, and Jose R Beltran. gpurir: A python library for room impulse response simulation with gpu acceleration. *Multimedia Tools and Applications*, 80:5653–5671, 2021.

[Doclo *et al.*, 2010] Simon Doclo, Sharon Gannot, Marc Moonen, and Ann Spriet. Acoustic beamforming for hearing aid applications. *Handbook on array processing and sensor networks*, pages 269–302, 2010.

[Dwivedi *et al.*, 2022] Priyadarshini Dwivedi, Raj Prakash Gohil, Gyanajyoti Routray, Vishnuvardhan Varanasiy, and Rajesh M Hegde. Joint doa estimation in spherical harmonics domain using low complexity cnn. In *2022 IEEE International Conference on Signal Processing and Communications (SPCOM)*, pages 1–5. IEEE, 2022.

[Hoshuyama *et al.*, 1999] Osamu Hoshuyama, Akihiko Sugiyama, and Akihiro Hirano. A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters. *IEEE Transactions on signal processing*, 47(10):2677–2684, 1999.

[Klemm *et al.*, 2008] M Klemm, IJ Craddock, JA Leendertz, A Preece, and R Benjamin. Improved delay-and-sum beamforming algorithm for breast cancer detection. *International Journal of Antennas and Propagation*, 2008, 2008.

[Kumar and Hegde, 2016] Lalan Kumar and Rajesh M Hegde. Near-field acoustic source localization and beamforming in spherical harmonics domain. *IEEE Transactions on Signal Processing*, 64(13):3351–3361, 2016.

[Le *et al.*, 2021] Xiaohuai Le, Hongsheng Chen, Kai Chen, and Jing Lu. Dpcrn: Dual-path convolution recurrent network for single channel speech enhancement. *arXiv preprint arXiv:2107.05429*, 2021.

[Li *et al.*, 2022] Andong Li, Wenzhe Liu, Chengshi Zheng, and Xiaodong Li. Embedding and beamforming: All-neural causal beamformer for multichannel speech enhancement. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6487–6491. IEEE, 2022.

[Li *et al.*, 2024] Andong Li, Guochen Yu, Zhongweiyang Xu, Cunhang Fan, Xiaodong Li, and Chengshi Zheng. Tabe: Decoupling spatial and spectral processing with taylor's unfolding method in the beamspace domain for multi-channel speech enhancement. *Information Fusion*, 101:101976, 2024.

[Liu and Zhang, 2021] Jinjiang Liu and Xueliang Zhang. Inplace gated convolutional recurrent neural network for dual-channel speech enhancement. *arXiv preprint arXiv:2107.11968*, 2021.

[Lugasi and Rafaely, 2020] Moti Lugasi and Boaz Rafaely. Speech enhancement using masking for binaural reproduction of ambisonics signals. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1767–1777, 2020.

[Luo *et al.*, 2019] Yi Luo, Cong Han, Nima Mesgarani, Enea Ceolini, and Shih-Chii Liu. Fasnet: Low-latency adaptive beamforming for multi-microphone audio processing. In *2019 IEEE automatic speech recognition and understanding workshop (ASRU)*, pages 260–267. IEEE, 2019.

[Mehrotra *et al.*, 2011] Sanjeev Mehrotra, Wei-ge Chen, Zhengyou Zhang, and Philip A Chou. Realistic audio in immersive video conferencing. In *2011 IEEE International Conference on Multimedia and Expo*, pages 1–4. IEEE, 2011.

[Meyer, 2001] Jens Meyer. Beamforming for a circular microphone array mounted on spherically shaped objects. *The Journal of the Acoustical Society of America*, 109(1):185–193, 2001.

[Nguyen *et al.*, 2014] Viet Anh Nguyen, Jiangbo Lu, Shengkui Zhao, Douglas L Jones, and Minh N Do. Teleimmersive audio-visual communication using commodity hardware [applications corner]. *IEEE Signal Processing Magazine*, 31(6):118–136, 2014.

[Nossier *et al.*, 2019] Soha A Nossier, MRM Rizk, Nancy Diaa Moussa, and Saleh el Shehaby. Enhanced smart hearing aid using deep neural networks. *Alexandria Engineering Journal*, 58(2):539–550, 2019.

[Rafaely, 2015] Boaz Rafaely. *Fundamentals of spherical array processing*, volume 8. Springer, 2015.

[Rao *et al.*, 2021] Wei Rao, Yihui Fu, Yanxin Hu, Xin Xu, Yvkai Jv, Jiangyu Han, Zhongjie Jiang, Lei Xie, Yannan Wang, Shinji Watanabe, et al. Conferencingspeech challenge: Towards far-field multi-channel speech enhancement for video conferencing. In *2021 IEEE Auto-*

matic Speech Recognition and Understanding Workshop (ASRU), pages 679–686. IEEE, 2021.

[Reddy *et al.*, 2019] Chandan KA Reddy, Ebrahim Beyrami, Jamie Pool, Ross Cutler, Sriram Srinivasan, and Johannes Gehrke. A scalable noisy speech dataset and online subjective test framework. *arXiv preprint arXiv:1909.08050*, 2019.

[Ren *et al.*, 2021] Xinlei Ren, Xu Zhang, Lianwu Chen, Xiguang Zheng, Chen Zhang, Liang Guo, and Bing Yu. A causal u-net based neural beamforming network for real-time multi-channel speech enhancement. In *Interspeech*, pages 1832–1836, 2021.

[Rix *et al.*, 2001] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, volume 2, pages 749–752. IEEE, 2001.

[Souden *et al.*, 2009] Mehrez Souden, Jacob Benesty, and Sofiene Affes. On optimal frequency-domain multichannel linear filtering for noise reduction. *IEEE Transactions on audio, speech, and language processing*, 18(2):260–276, 2009.

[Taal *et al.*, 2011] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2125–2136, 2011.

[Tan and Wang, 2019] Ke Tan and DeLiang Wang. Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:380–390, 2019.

[Tan *et al.*, 2019] Ke Tan, Xueliang Zhang, and DeLiang Wang. Real-time speech enhancement using an efficient convolutional recurrent network for dual-microphone mobile phones in close-talk scenarios. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5751–5755. IEEE, 2019.

[Tan *et al.*, 2022] Ke Tan, Zhong-Qiu Wang, and DeLiang Wang. Neural spectrospatial filtering. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:605–621, 2022.

[Varanasi *et al.*, 2020] Vishnuvardhan Varanasi, Harshit Gupta, and Rajesh M Hegde. A deep learning framework for robust doa estimation using spherical harmonic decomposition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1248–1259, 2020.

[Varga and Steeneken, 1993] Andrew Varga and Herman JM Steeneken. Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech communication*, 12(3):247–251, 1993.

[Wang and Wang, 2018] Zhong-Qiu Wang and DeLiang Wang. All-neural multi-channel speech enhancement. In *Interspeech*, pages 3234–3238, 2018.

[Wang *et al.*, 2023] Zhong-Qiu Wang, Samuele Cornell, Shukjae Choi, Younglo Lee, Byeong-Yeol Kim, and Shinji Watanabe. Tf-gridnet: Integrating full-and sub-band modeling for speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[Weng *et al.*, 2023] Liuqing Weng, Xiyu Song, Zhenghong Liu, Xiaojuan Liu, Haocheng Zhou, Hongbing Qiu, and Mei Wang. Doa estimation of indoor sound sources based on spherical harmonic domain beam-space music. *Symmetry*, 15(1):187, 2023.

[Xu *et al.*, 2021] Xudong Xu, Hang Zhou, Ziwei Liu, Bo Dai, Xiaogang Wang, and Dahua Lin. Visually informed binaural audio generation without binaural audios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15485–15494, 2021.

[Zue *et al.*, 1990] Victor Zue, Stephanie Seneff, and James Glass. Speech database development at mit: Timit and beyond. *Speech communication*, 9(4):351–356, 1990.