

TT-NET: DUAL-PATH TRANSFORMER BASED SOUND FIELD TRANSLATION IN THE SPHERICAL HARMONIC DOMAIN

Yiwen Wang, Zijian Lan, Xihong Wu, Tianshu Qu

Key Laboratory on Machine Perception (Ministry of Education)
School of Intelligence Science and Technology
Peking University, Beijing, China
{pku_wyw, qutianshu}@pku.edu.cn

ABSTRACT

In the current method for the sound field translation tasks based on spherical harmonic (SH) analysis, the solution based on the additive theorem usually faces the problem of singular values caused by large matrix condition numbers. The influence of different distances and frequencies of the spherical radial function on the stability of the translation matrix will affect the accuracy of the SH coefficients at the selected point. Due to the problems mentioned above, we propose a neural network scheme based on the dual-path transformer. More specifically, the dual-path network is constructed by the self-attention module along the two dimensions of the frequency and order axes. The transform-average-concatenate layer and upscaling layer are introduced in the network, which provides solutions for multiple sampling points and upscaling. Numerical simulation results indicate that both the working frequency range and the distance range of the translation are extended. More accurate higher-order SH coefficients are obtained with the proposed dual-path network.

Index Terms— Spherical harmonic analysis, dual-path transformer, translation matrix, sound field reproduction.

1. INTRODUCTION

With sound field recording, processing, and reproduction techniques, spatial audio enables the reconstruction of an acoustic environment [1]. The high-order ambisonics (HOA) can realize the expression of the three-dimensional sound field within a specific range [2]. The higher-order coefficients can be acquired through spherical microphone array (SMA) recordings [3]. Virtual reality applications demand additional translation degrees of freedom, often referred to as *six degrees of freedom* (6DoF).

A common way to implement 6DoF is to use multiple SMAs distributed in three-dimensional space. The commonly used method uses the mathematical properties of SH to convert the SH coefficients obtained at the sampling points into the higher-order SH coefficients at the selected point. In [4], Laborie *et al.* present the theoretical principle for estimating the SH coefficients, including the spatial sampling and encoding aspects. Based on the translation theorem, Samarasinghe *et al.* provide the solution for the global coefficients using higher-order microphones (HOM) [5]. In [6, 7], according to the category and location information of the sound source, Rafaely *et al.* summarize the translation equation corresponding to different situations. However, the numerical method is relatively complex and suffers from the problem of ill-conditioned matrices under higher-order conditions. Based on the idea of plane wave decomposition, Wang *et al.* propose a solution without using the translation theorem

[8]. Ueno *et al.* formulate an estimate of the harmonic coefficients based on infinite-order analysis by applying Bayes' theorem [9, 10]. Both methods have achieved better results in cylindrical coordinates, but the effect needs to be improved as the frequency increases. Furthermore, there is a lack of experimental validation for the case in the 3D space.

The ability of deep learning to model complex relationships between different representations has been applied to SH-based sound field problems recently. To achieve higher bandwidth SH coefficients and alleviate spatial aliasing problems, networks are used to model SH bases [11]. In the previous work, a U-Net-based generator is used to realize the upscaling of SH coefficients [12]. Given the limitations of the traditional numerical solution, a neural-network-based SH coefficient translation is proposed to achieve a more accurate SH representation at different distances, different frequency bands, and under complex sound source conditions in our work. The experimental results show that the proposed network-based method extends the working frequency range and obtains more precise translation coefficients under complex environmental conditions.

The rest of the paper is organized as follows: Section 2 introduces the theory of translation matrix in the SH domain, and Section 3 describes the proposed dual-path transformer model. Experimental setup and results are reported in Section 4 and Section 5. Finally, we conclude in Section 6.

2. THEORY OF TRANSLATION MATRIX

Consider the external far-field sound source situations. The sound pressure in the spherical coordinate system can be decomposed into the expansion of SH coefficients as

$$p(k, \mathbf{r}) = \sum_{n=0}^{\infty} j_n(kr) \sum_{m=-n}^n B_n^m Y_n^m(\theta, \phi), \quad (1)$$

where k is the wave number, $\mathbf{r} \equiv (r, \theta, \phi)$ is the spherical coordinate system denoted by elevation and azimuth angles, θ and ϕ , together with the radial distance r , $p(k, \mathbf{r})$ represents the sound pressure at \mathbf{r} , $j_n(kr)$ is the spherical Bessel function, $Y_n^m(\theta, \phi)$ is the basis function of SH, and B_n^m is the corresponding SH coefficient. Consider the translation from a global origin to a local translated origin $\mathbf{r}'' \equiv (r'', \theta'', \phi'')$, such that

$$\mathbf{r} = \mathbf{r}'' + \mathbf{r}', \quad (2)$$

where \mathbf{r} and $\mathbf{r}' \equiv (r', \theta', \phi')$ represent the position relative to the original global origin and the new coordinate center, respectively.

According to the addition theorems [13], the translation from the spherical Bessel functions to spherical Bessel functions is described as

$$j_n(kr)Y_n^m = \sum_{n'=0}^{\infty} \sum_{m'=-n'}^{n'} j_{n'}(kr')j_{n'}(kr')Y_{n'}^{m'}(\theta', \phi') \times \sum_{n''=0}^{\infty} j_{n''}(kr'')Y_{n''}^{m-m'}(\theta'', \phi'')C_{n'm'}^{nmn''}, \quad (3)$$

where $C_{n'm'}^{nmn''}$ contains the multiplication of two Wigner 3-j operators. See [6] for details.

For each translation position \mathbf{r}'' , the sound pressure is expressed by a set of SH coefficients with \mathbf{r}'' as the local coordinate center. Assuming that the SH coefficients of the Q local coordinate systems are truncated to order N'' , and the coefficients of the global origin are truncated to order N . According to the addition theorems, the relationship between the two types of coefficients can be established through the translation matrix of the SH coefficients. The derivation of this part is described in detail in [6]. The formula is expressed as

$$(b'')^T = T_{trans}b^T, \quad (4)$$

where $b'' = [B_{0(0)}^0, B_{1(0)}^{-1}, B_{1(0)}^0, \dots, B_{n(q)}^m, \dots, B_{N''(Q-1)}^{N''}]$ is a $1 \times (Q(N''+1)^2)$ vector, $B_{n(q)}^m$ represents the local spherical coefficient of the order n and degree m for the q -th local coordinate system, $b = [B_0^0, B_1^{-1}, B_1^0, \dots, B_n^m, \dots, B_N^N]$ is a $1 \times (N+1)^2$ vector, B_n^m represents the global coefficient. The shape of the translation matrix T_{trans} is $Q(N''+1)^2 \times (N+1)^2$. Each row element of the matrix is expanded according to the order of the global origin system.

3. PROPOSED METHOD

In the traditional method, the relationship of SH coefficients of the different coordinate centers is established through the translation matrix. Theoretically, the coefficients of the global origin can be realized by matrix inversion operation. However, in practice, due to the problem of Bessel nulls [5], the matrix has a large condition number, which seriously affects the results of numerical calculation. Ridge regression can alleviate the problem of matrix singularity to a certain extent [14], but in practical applications, different distance conditions bring errors to the coefficients at different orders. We propose a dual-path transformer-based SH coefficient translation network to solve the above problems in our work.

3.1. Model Description

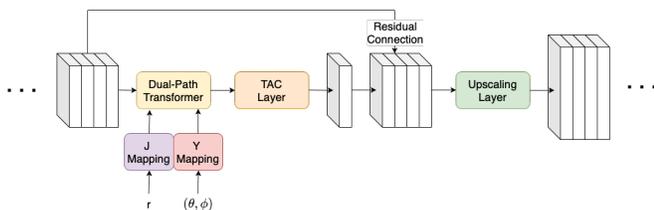


Fig. 1. Single-layer architecture of TT-Net.

A dual-path transformer-based network for translation named TT-Net is proposed. More specifically, we take a sample at one location as an example to illustrate our proposed network architecture in detail. One layer of the TT-Net is shown in Fig. 1. The input and output of the network are SH coefficients of different orders. In our work, the input feature of SH coefficients is a $K \times (N+1)^2$ matrix, where K is the total number of the frequency bins, and N is the order of the SH. The transformer mentioned in the paper refers to the encoder part. As mentioned in Eq. (3), for the SH translation, both radial spherical Bessel functions and angle-dependent SH functions contribute to the translation process. Due to the decoupling of distance and angle, two mapping networks named J and Y are constructed to replace the expression of $J_n(kr)$ and $Y_n^m(\theta, \phi)$, respectively. $Y_n^m(\theta, \phi)$ is only related to order and orientation, not frequency. Therefore, the same constraints are used for the Y network, the input is the angle information, and the output is the N -order vector after the reshaping operation. Similar structural constraints are applied to the J network so that the output dimension of the J network is consistent with the dimension of the spherical Bessel function, that is, $k \times (N+1)$. The output of the Dual-Path Transformer (DPT) remains the same shape as the input SH coefficients. The output of the DPT module serves as the input to a transform-average-concate (TAC) layer [15] and a fully connected (FC) layer. The TAC layer is used to integrate different SH coefficients, while the FC layer is used for upscaling the coefficients [16]. A residual connection is added between the TAC layer and the Upscaling Layer to help the training of the network [17].

The network is connected by the structure shown in Fig. 1. The order of the output of every single layer is larger than the input. The last layer of the network no longer uses the residual connection and upscaling, and the output of TAC from the last layer is used as the final output of the entire model.

3.2. Dual-path Transformer Module

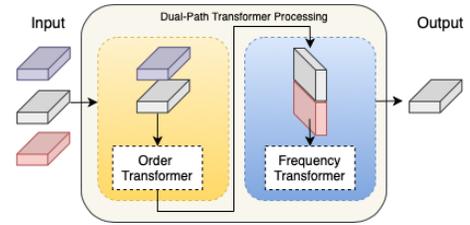


Fig. 2. Dual-path Transformer Module.

Dual-path transformer module has recently attracted much attention and has achieved good performance in speech separation [18, 19] and noise reduction [20]. The dual-path module used in our work is shown in Fig. 2. The input conditions represented by different colors are consistent with Fig. 1. Taking the spherical Bessel function as an example, for the same frequency bin k , $J_n(kr)$ of different orders is related to the SH coefficients of the same frequency bin. The coupling of coefficients and radial spherical functions at the same frequency bin is achieved through the self-attention mechanism. Compared with the standard encoder part in the transformer, a fully-connected layer is added to the last layer of the encoder to integrate the SH coefficient and the J function. It also plays a role in constraining the same input and output dimensions. The output of the intermediate layer is transposed and concatenated with the

Y function, and a similar self-attention operation is performed, followed by the FC layer.

4. EXPERIMENTAL SETUP

4.1. Datasets And Training Settings

During the translation process, information such as location, frequency, and the number of sampling points will affect the results. Therefore, different situations are taken into account in the simulation data. For our subsequent applications, we use the fourth-order SH coefficients as input, which means $N = 4$. The frequency bin ranges from 100Hz to 3000Hz with an interval of 100Hz, which means $K=30$. The coefficient of the global origin is $N'' = 8$. The distance between the set sampling point and the global origin is randomly sampled between $0.2m$ and $2.0m$. In our experiment, plane waves from 1 to 4 directions are randomly generated as the signal source. The amplitude of the signal is randomly chosen from 0.1 to 1.0. Since the low-frequency system noise brings difficulties to the solution of SH coefficients [21], noises of different signal-to-noise ratios (SNR) are added to enhance the noise immunity of the model. SNR varies from 10 to 30dB. The number of spatial sampling points is set from 4 to 10. The minimum setting is four sampling points to ensure the full rank of the translation matrix.

Networks are trained with mean squared error (MSE) loss. The order of the output of each dual-path transformer module is one greater than the input. Multi-head attention is introduced, and the number of heads is the same as the current order plus one. For stable learning, gradients are clipped to $[-1.0, 1.0]$. All models are trained distributedly with 2 TITAN RTX with batch size 32. The learning rate is initialized to be $3e-4$ and halved with Adam optimizer training [22]. For training stability, each data is normalized, and the number of data sampling points in each batch is ensured to be the same. The training data are trained sequentially from 10 to 4.

4.2. Evaluation Metrics

The evaluation is based on the similarity of the SH coefficients and the sound field. Euclidean distance metric (EDM) and cosine similarity (COSS) metrics are used to judge the difference between the recovered and the ideal coefficients, respectively, where EDM gives judgments from Euclidean space, and COSS provides judgments of structural similarity.

$$EDM = \frac{1}{N} \sum_{i=1}^N |\hat{y}_i - y_i|^2, \quad (5)$$

$$COSS = \frac{1}{N} \sum_{i=1}^N \frac{\hat{y}_i \cdot y_i}{|\hat{y}_i| \times |y_i|}, \quad (6)$$

N is the total number of test data, y is the SH coefficients of the global origin, and \hat{y} is the estimated result. Besides, signal-to-distortion ratio (SDR) is for evaluating the sound field based on the SH coefficients. SDR is defined as [9],

$$SDR = \frac{1}{N} \sum_{i=1}^N 10 \log_{10} \frac{\int_{r \in V} u_i(\mathbf{r})^2 d\mathbf{r}}{\int_{r \in V} |u_i(\mathbf{r}) - \hat{u}_i(\mathbf{r})|^2 d\mathbf{r}}, \quad (7)$$

where N is the total number of test data, u and \hat{u} represent the sound pressure reconstructed with the ideal SH coefficients and the estimated coefficients. SDR is calculated within a radius of $1.00m$ at $0.02m$ intervals. k representing frequency is omitted.

5. EVALUATION RESULTS AND DISCUSSION

5.1. Ablation Study

Table 1. Average results of 8 spatial sampling points.

Method	Strategy	COSS	EDM	SDR(dB)
LSM	Regularization	0.101	0.068	-0.074
TT-Net(1)	Lrg2Sml L2	0.334	0.058	0.827
TT-Net(2)	Lrg2Sml L2	0.528	0.049	1.357
TT-Net(4)	Lrg2Sml L2	0.732	0.043	2.037
TT-Net(4)	Lrg2Sml L1	0.472	0.052	1.397
TT-Net(4)	Sml2Lrg L2	0.411	0.056	0.793

The results of the ablation study are shown in Table 1. The test data uses eight randomly selected spatial sampling points in the 3D space with a $1m$ distance. The direction of the single sound source is randomly generated in the 3D space. The number of test datasets is 600. The result is the average metrics of all frequencies. The method of solving by inversion of the translation matrix in [7] is abbreviated as the least square method (LSM). One-layer and two-layer architectures shown in Fig. 1 are used for comparison to verify the optimal effect of increasing the dual-path module step by step. Correspondingly, the upscaling parts implement the mapping from order 4 to 8 or from other 4 to 6 with eight behind. The numbers in parentheses are the layers of the architecture. The number 4 indicates the scheme is that the order increases sequentially from 4 to 8. Lrg2Sml means that the network is trained with the number of spatial sampling points from 10 to 4, while Sml2Lrg is the opposite. L1 loss is chosen for comparison. The results show that Lrg2Sml helps the training. Results also show that the performance decreases if fewer DPT module is used. Although L1 loss function performs better in some regression tasks [23], this is not the case in our work. The final average results show that our proposed method is effective regarding SH coefficients and the recovered sound field.

5.2. Experiments Results

5.2.1. Results for different SNRs

The results for different SNRs are shown in Fig. 3(a). From left to right are the results of SDR, EDM, and COSS. Horizontal is the frequency information. The TT-Net(4) with the best results is used, and we use TT-Net for short. LSM is compared with our optimal scheme. From the SDR results, it can be seen that the performance of the LSM is degraded as the SNR decreases. In contrast, our method performs consistently on SDR. The metrics of EDM and COSS show that the results of the coefficients are in line with the previous discussion. The noises cause outliers to be added to the numerical solution results, which seriously degrades the results. On the contrary, our proposed method results in noise immunity.

5.2.2. Results for different distances

The results for different distances of spatial sampling points are shown in Fig. 3(b). Each test data uses eight randomly distributed sampling points on an equidistant sphere. The number of test samples per distance is 100. COSS shows that the traditional method works well at low frequencies. As the frequency increases, the results turn worse due to the increase of kr , which is related to the properties of spherical Bessel functions. The network method has

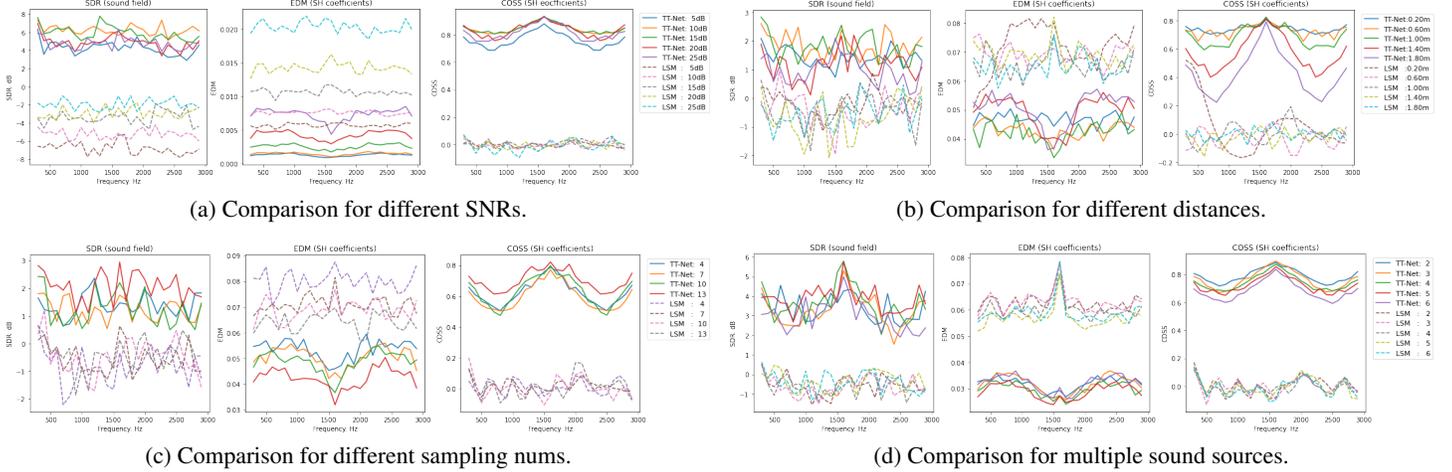


Fig. 3. Experiments results under different conditions.

stable performance under a shorter than $1.00m$ condition. However, the performance degrades under long-distance conditions, and there exists singularity in the frequency band below $1kHz$ and above $2kHz$. According to the analysis of the SDR results, the reconstructed sound field results remain the same, which indicates that the results of singular values are located in the higher-order part leading to little effect on the sound field near the origin.

5.2.3. Results for different nums of spatial sampling points

For experiments with different spatial sampling points, 4 to 13 spatial sampling points are selected, and the number of test samples for each condition is 100. We select four cases, with the number of 4, 7, 10, and 13 spatial sampling points for visualization. In LSM, using more spatial sampling points reduces the singularity of the translation matrix, which is reflected in all three metrics, as shown in Fig. 3(c). As the number increases, the network also has a consistent trend. Under the test conditions, the number outside the training set achieves better results. The results show that the proposed method can increase the stability as the number increases and confirms the TAC module’s role.

5.2.4. Results for multiple sound sources

The results for the multi-sources condition are shown in Fig. 3(d). In LSM, the influence of sound sources in different directions only affects the SH coefficients, which does not affect the solution of the translation matrix. Therefore, the traditional method is consistent. That is, the number of sound sources has little effect. The network is tested with 2 to 6 different numbers of sound sources. The results show that the SH coefficients of different numbers of sound sources tend to be consistent. The performance of our method does not degrade as the number of sound sources increases. The test sets with more than four sound sources are consistent with the other results on the metrics of EDM and COSS. It should be noted that there are specific differences in the results of the proposed method at different frequencies, which will be further analyzed in the follow-up work.

Fig. 4 visualizes an example of the sound pressure at 1000 Hz and 1800 Hz . The figure shows the sound pressure distribution on a horizontal plane of $2m \times 2m$. Two sound sources in this example are oriented at 0° and 225° . Sound pressures expanded by the output of

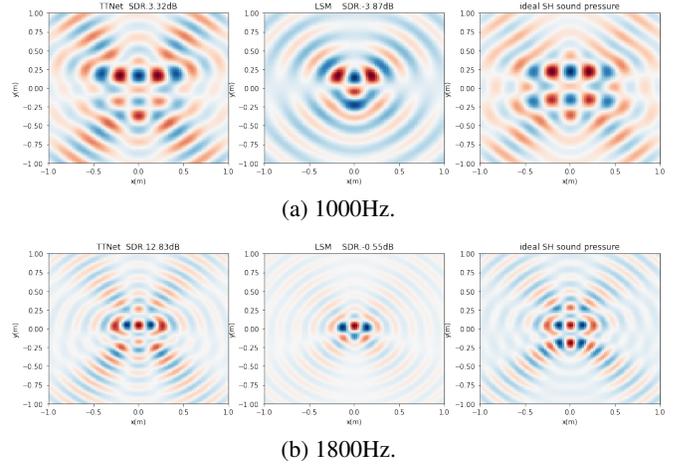


Fig. 4. Visualization of the sound pressure in the x-y plane.

TT-Net, LSM, and ideal SH coefficients are shown from left to right in each case. The results show that the network’s performance is stable and better under different frequency conditions.

6. CONCLUSION

We propose a sound field recording method based on a dual-path transformer network. The method applies to the translation of SH coefficients. This work continues our exploration of optimizing SH analysis using neural networks. The proposed method reduces the occurrence of singular solutions in solving SH coefficients. The simulation results of different frequency ranges, SNRs, and under more complex sound source conditions show that the method is more accurate than the traditional method to recover the SH coefficients. Under the same conditions, the proposed method brings a 3dB improvement in the SDR metrics. Detailed studies remain as future work for applications in real-world scenarios containing multiple scattering rigid balls. Future work will be tested in real scenarios to provide a feasible solution for the realization of 6DoF.

7. REFERENCES

- [1] M. Cobos, J. Ahrens, K. Kowalczyk, and A. Politis, "An overview of machine learning and other data-based methods for spatial audio capture, processing, and reproduction," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2022, no. 1, pp. 1–21, 2022.
- [2] M. A. Poletti, "Three-dimensional surround sound systems based on spherical harmonics," *Journal of the Audio Engineering Society*, vol. 53, no. 11, pp. 1004–1025, 2005.
- [3] T. D. Abhayapala and D. B. Ward, "Theory and design of high order sound field microphones using spherical microphone array," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. II–1949, IEEE, 2002.
- [4] A. Laborie, R. Bruno, and S. Montoya, "A new comprehensive approach of surround sound recording," in *Audio Engineering Society Convention 114*, Audio Engineering Society, 2003.
- [5] P. Samarasinghe, T. Abhayapala, and M. Poletti, "Wavefield analysis over large areas using distributed higher order microphones," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 3, pp. 647–658, 2014.
- [6] B. Rafaely, *Fundamentals of spherical array processing*, vol. 8. Springer, 2015.
- [7] T. Peleg and B. Rafaely, "Investigation of spherical loudspeaker arrays for local active control of sound," *The Journal of the Acoustical Society of America*, vol. 130, no. 4, pp. 1926–1935, 2011.
- [8] Y. Wang and K. Chen, "Translations of spherical harmonics expansion coefficients for a sound field using plane wave expansions," *The Journal of the Acoustical Society of America*, vol. 143, no. 6, pp. 3474–3478, 2018.
- [9] N. Ueno, S. Koyama, and H. Saruwatari, "Sound field recording using distributed microphones based on harmonic analysis of infinite order," *IEEE Signal Processing Letters*, vol. 25, no. 1, pp. 135–139, 2017.
- [10] M. Nakanishi, N. Ueno, S. Koyama, and H. Saruwatari, "Two-dimensional sound field recording with multiple circular microphone arrays considering multiple scattering," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 368–372, IEEE, 2019.
- [11] S. Gao, J. Lin, X. Wu, and T. Qu, "Sparse dnn model for frequency expanding of higher order ambisonics encoding process," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1124–1135, 2022.
- [12] Y. Wang, X. Wu, and T. Qu, "Up-wgan: Upscaling ambisonic sound scenes using wasserstein generative adversarial networks," in *Audio Engineering Society Convention 151*, Audio Engineering Society, 2022.
- [13] W. C. Chew, *Waves and fields in inhomogeneous media*, vol. 16. Inst of Electrical &, 1995.
- [14] G. C. McDonald, "Ridge regression," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 1, no. 1, pp. 93–100, 2009.
- [15] Y. Luo, Z. Chen, N. Mesgarani, and T. Yoshioka, "End-to-end microphone permutation and number invariant multi-channel speech separation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6394–6398, IEEE, 2020.
- [16] G. Routray, S. Basu, P. Baldev, and R. M. Hegde, "Deep-sound field analysis for upscaling ambisonic signals," in *EAA Spatial Audio Signal Processing Symposium*, pp. 1–6, 2019.
- [17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
- [18] J. Chen, Q. Mao, and D. Liu, "Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation," *arXiv preprint arXiv:2007.13975*, 2020.
- [19] K. Wang, B. He, and W.-P. Zhu, "Tstnn: Two-stage transformer based neural network for speech enhancement in the time domain," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7098–7102, IEEE, 2021.
- [20] F. Dang, H. Chen, and P. Zhang, "Dpt-fsnet: Dual-path transformer based full-band and sub-band fusion network for speech enhancement," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6857–6861, IEEE, 2022.
- [21] J. Lin, X. Wu, and T. Qu, "Anti spatial aliasing hoa encoding method based on aliasing projection matrix," in *2020 IEEE 3rd International Conference on Information Communication and Signal Processing (ICICSP)*, pp. 321–325, IEEE, 2020.
- [22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [23] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.