NAT3DSound: 3D Spatial Sound Field Synthesis with Multi-Modal Non-Autoregressive Transformer

Anonymous ACL submission

Abstract

3D spatial sound field synthesis takes the head-mounted audio signals and body poses as input and renders a 3D sound field around the center body, in which spatial audio can be inferred at any arbitrary position. To achieve this, a multi-modal system is required to spatialize input audio signals with the guidance of pose streams. Then, a numerical method is used to render the sound field with hundreds of spatial audio. However, there exist several challenges including hybrid audio input signals, cross-modal matching, and few data resources. To address them, we propose NAT3DSound, a unit-based framework for high-quality 3D spatial sound synthesis, which consists of 1) a general tokenization scheme for both dense and impulsive input audio signals, 2) a non-autoregressive transformer with multi-scale modality fusion for efficient cross-modal alignment and 3) a parallel sampling strategy for fast prediction. Furthermore, we investigate the feasibility of acoustic pre-training for low-resource learning in data-scarce scenarios. Extensive experiments and ablation studies demonstrate the effectiveness of NAT3DSound in terms of spatialization quality and generalization ability. Audio samples are available at http://NAT3DSound.github.io

1 Introduction

Virtual and augmented reality have gained significant attention in recent years, in which people will be digitized and interact with virtual environments. With the progress with computer vision and natural language processing, representing 3D human bodies (Bagautdinov et al., 2021; Zheng et al., 2023) and conducting interesting conversations (Touvron et al., 2023) becomes feasible over time.

For humans, perceiving the world is definitely multimodal, in which vision and hearing interact with each other inseparably. However, despite the impressive advances in visual representation of the 3D human body (Tian et al., 2023), its acoustic counterpart is often overlooked. Recently, Xudong et al. (2023) proposed the first 3D sound field modeling work to close the gap between visual and acoustic representations of human bodies. 3D spatial sound synthesis takes the egocentric audios from the head-mounted microphones and the body pose as the inputs and generates the 3D sound field, where the sounds at arbitrary positions can be inferred in the virtual space. To achieve this, they proposed a WaveNet-like multi-modal network to generate hundreds of spatial audios in a spherical space and then rendered the 3D sound field with a numerical approach (i.e. spherical harmonics (Samarasinghe et al., 2012)).

However, there still exist several open challenges to developing a high-quality 3D sound field synthesis system: 1) The hybrid composition of audio poses challenges for spectrogram-based optimization. The input audios include both the dense and structured signals like speech and the sparse and impulsive signals like body sounds. The spectrogram-based optimization tends to over-attenuate the impulsive sounds, bringing negative effects to audio spatialization. 2) Predicting a spatial sound that matches the body pose stream remains challenging, as the pose stream indicates the source of sound and its propagation. 3) The generalization ability of 3D sound field synthesis is not investigated.

In this work, we propose NAT3DSound, a high-quality 3D sound field synthesis model with unit-based non-autoregressive modeling. To tackle the hybrid composition of audio, we use a neural codec for general audio tokenization. Then, we build a non-autoregressive transformer with the proposed multi-scale audio-visual fusion for fine-grained multi-modal guidance. Also, a span-based parallel sampling is presented to improve the efficiency of our model. To alleviate the data scarcity in this task, we investigate low-resource learning by leveraging sufficient speech data for acoustic pre-training.

Experiments demonstrate that NAT3DSound generates audio samples with accurate spatialization effects and can generalize to novel persons, achieving state-of-the-art results in terms of spatialization quality and generalization ability.

Our contributions can be summarized as follows: 1) We propose NAT3DSound, the first unit-based non-autoregressive model for 3D sound field synthesis, which renders a field that the sound pressure at any position can be inferred. 2) We propose multi-scale cross-modal matching, including coarse text-visual and fine-grained audio-visual fusions to fully exploit the visual guidance. 3) We investigate the feasibility of speech-based pretraining for improving model performance and low-resource learning. 4) Experimental results reveal that NAT3DSound yields the best performance in terms of spatialization quality and generalization ability, and extensive ablation studies are conducted to validate the effectiveness of the proposed modules.

2 Related Work

2.1 Language Modeling in Audio

Recent audio generation works leverage the discrete representation of audio, which can be obtained by an encoder-decoder network. For audio tokenization, SoundStream (Zeghidour et al., 2021) is one of the first audio compression models, which uses a convolutional encoder and decoder network and performs residual vector quantization. Encodec (Défossez et al., 2022) follows the architecture of SoundStream and improves it with many modifications like a multi-scale STFT adversarial discriminator and a loss balancer to automatically tune the loss weights. Descript Audio Codec (DAC) (Kumar et al., 2023) also adopts a convolutional encoder-decoder architecture with adversarial learning, but proposes several new techniques like stable training without a sophisticated loss balancer. Based on the discrete representation of audio, AudioLM (Borsos et al., 2022) and MusicLM (Agostinelli et al., 2023) proposed a coarse-to-fine token generation paradigm, VALL-E (Wang et al., 2023) and SPEAR TTS (Kharitonov et al., 2023) are proposed to clone human voice with a short recording.

2.2 Non-autoregressive Audio Generation

Unlike autoregressive models (Oord et al., 2016; Wang et al., 2017; Kreuk et al., 2022; Yang et al., 2023), which use historical data as input to predict future values, non-autoregressive models do not depend on previous observations. For example, FastSpeech (Ren et al., 2019) learns a non-autoregressive student model from auto-regressive teacher models. Glow-TTS (Kim et al., 2020) uses normalizing flows and dynamic programming for text-to-speech. Inspired by diffusion model, many diffusion-based audio generation works are proposed (Yang et al., 2022; Liu et al., 2023; Huang et al., 2023b; Schneider et al., 2023; Huang et al., 2023a; Liu et al., 2022a). For example, Liu et al. (2023); Huang et al. (2023b) leveraged the diffusion model for text-to-audio by broadening its application to other tasks like image-to-audio. Schneider et al. (2023); Huang et al. (2023a) proposed addressing the text-to-music task with diffusion models. Recently, Ziv et al. (2024) proposed using a non-autoregressive transformer model with discrete acoustic representations and masked language modeling (Ghazvininejad et al., 2019). Our model draws inspiration from (Ziv et al., 2024) but proposes several novel modifications like multi-scale audio-visual fusion and span-based parallel sampling.

2.3 Audio-Visual Learning

In our real world, audio and visual are always connected. Many multi-modal self-supervised learning methods are proposed for better and more robust representation learning when compared to single-modal learning (Owens et al., 2016; Korbar et al., 2018; Morgado et al., 2020). Furthermore, audio-visual learning is applied to various applications like audio-visual navigation (Gan et al., 2020; Chen et al., 2020a,b), audio source location (Qian et al., 2020; Tian et al., 2021; Mo and Morgado, 2022), co-speech gesture generation (Yoon et al., 2019; Liu et al., 2022b; Zhu et al., 2023) and so on. Besides, our work is also related to visually-guided audio spatialization, which aims at transforming monaural audio into 3D audio with visual guidance. For example, Morgado et al. (2018); Li et al. (2018) synthesized the binaural audio by utilizing 360° videos; Gao and Grauman (2019); Zhou et al. (2020); Lu et al. (2019); Yang et al. (2020) achieved audio spatialization by learning the left and right channels. However, these methods only concern first-order ambisonics or binaural audio, and the whole 3D sound field is not modeled. Recently, Xudong et al. (2023) proposed using a 3D human pose along with the headset microphones

to generate the whole 3D sound field. They used a WaveNet-like (Oord et al., 2016; Richard et al., 2021) network architecture with a time-shifted L2 loss to handle the mixture of dense (e.g. speech) and impulsive (e.g. clapping and snapping) sounds. Our work follows the task definition of (Xudong et al., 2023), but propose a new multi-modal non-autoregressive model.

3 NAT3DSound

An overview of the proposed NAT3DSound is illustrated in Figure 1. NAT3DSound 1) uses the neural codec DAC (Kumar et al., 2023) to tokenize the hybrid source audio signals; 2) builds a non-autoregressive sequence-to-sequence model along with the proposed multi-scale audio-visual fusion module for audio spatialization; 3) generates the spatialized audios with a span-based parallel sampling strategy.

In this section, we first introduce the problem definition of 3D sound field synthesis. Then, we introduce the general audio tokenization for hybrid audio inputs. Based on acoustic tokens, we present a non-autoregressive model with the multi-scale audio-visual fusion module and span-based parallel sampling strategy. Finally, the generated spatial audios are used to render the 3D sound field with a numerical method. Also, we investigate speech pre-training to alleviate the data scarcity in 3D sound field synthesis.

3.1 Problem Definition

Let $\mathbf{a}_{1:T} = (a_1, a_2, ..., a_T)$ be an audio from the head-mounted microphones, where $a_t \in \mathbb{R}^C$ means the waveform at time t with the input channels C. Besides, let $\mathbf{p}_{1:S} = (p_1, p_2, ..., p_S)$ be the sampled 3D body pose stream, where $p_s \in \mathbb{R}^{J\times 3}$ means the 3D coordinates of J body joints at the frame s. For each paired audio-body sample $(\mathbf{a}_{1:T}, \mathbf{p}_{1:S})$, they share the same duration but are sampled at a different rate. For the soundbody dataset, the audio and pose are sampled at 48kHz and 30fps, respectively.

Our goal is to learn a mapping function $\mathcal{M}$ that gives the audio signal at any arbitrary position (x, y, z) in the 3D space with the input paired audio-body samples. However, it is impractical to record enough training data with densely populated microphones in the 3D space. Following the pipeline of (Xudong et al., 2023), the receiver microphones are constructed in a spherical space, and target data are distributed on the surface of this sphere. By taking the supervision from the spherical data, we can learn the mapping function with the polar coordinates (r, θ, ϕ), head-mounted audio signal $\mathbf{a}_{1:T}$ and the corresponding body pose $\mathbf{p}_{1:S}$:

$$\mathbf{s}_{1:T}(\theta_i, \phi_i) = \mathcal{M}(\theta_i, \phi_i, \mathbf{a}_{1:T}, \mathbf{p}_{1:S}), \quad (1)$$

where θ_i and ϕ_i indicates the azimuth and elevation of i-th microphone on the surface of the sphere. The radius r is equal for each target microphone (in fact, it is said that there is a very small difference) and is therefore omitted.

After learning the mapping from head-mounted audios and body poses to spatial audios, we encode the predicted spatial audios $\{\mathbf{s}_{1:T}(\theta_i, \phi_i)\}_1^N$ into the harmonic sound field coefficients, where N is the number of spatial audios, and the harmonic order is defined as $K = \lfloor\sqrt{N}\rfloor - 1$. The audio signal at any arbitrary spatial position $\mathbf{a}_{1:T}(x, y, z)$ can be rendered according to this numerical method. More details are presented in Section A.

3.2 Audio Tokenization

The recorded audio from the head-mounted microphones is hybrid, consisting of speech and non-speech body sounds like clapping. The speech is usually a structured and dense audio signal, while the non-speech bound sounds are sparse and impulsive, in which many high-frequency signals are introduced. This composite audio signal can pose a challenge for audio spatialization. Specifically, The time domain ℓ_2 loss, which has been demonstrated to be effective in audio spatialization (Richard et al., 2021), tends to over-suppress the impulsive sounds (Xudong et al., 2023). To address this problem, Xudong et al. (2023) proposed a time-delayed ℓ_2 loss, in which the loss is computed over a set of shift offsets. Different from them, we propose to leverage the neural codec to perform audio tokenization. The neural codec compresses the audio according to an encoder-decoder architecture and the residual vector quantization (Zeghidour et al., 2021; Défossez et al., 2022; Kumar et al., 2023). Formally, let E and D be the encoder and decoder of the neural codec, the input audio signal $\mathbf{a}_{1:T}$ can be converted into discrete units $\mathbf{u}_{1:T'} = (u_1, u_2, ..., u_{T'})$, where T' indicates the sequence length of tokenized audio, and each $u_t \in \mathbb{R}^{C\times Q}$ means the value of t-th acoustic token for each of the C input channels and Q codebooks.

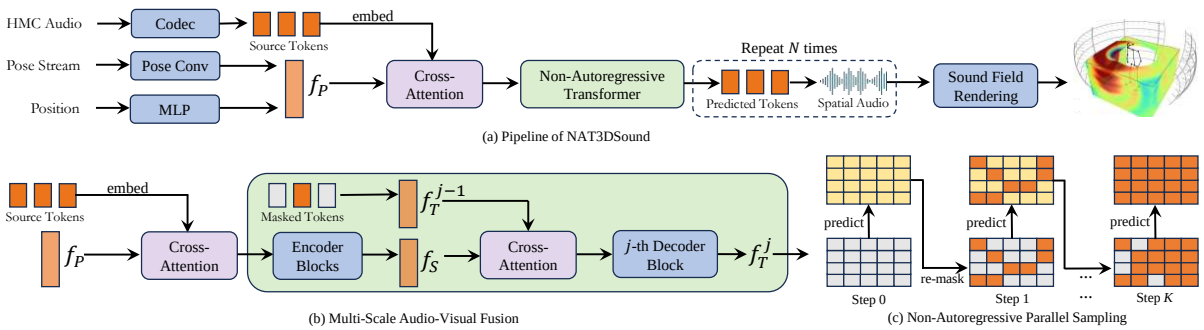

Figure 1: (Best viewed in color.) An overview of the proposed NAT3DSound model. (a) NAT3DSound takes the head-mounted microphone (HMC) audio, body pose stream, and spatial position as input, and non-autoregressively generates N spatial audio to render a 3D sound field. (b) NAT3DSound employs coarse fusion in source audio encoding and fine-grained fusion in target audio decoding. (c) NAT3DSound predicts the tokens across different codebooks in a parallel way.

3.3 Pose and Position Encoder

The body pose information is significant for audio spatialization since it provides where the sound begins and affects the sound propagation in the space. Also, the spatial audio also depends on its position (x_i, y_i, z_i), which can be mapped into a continuous Cartesian space as $(\cos\theta_i \sin\phi_i, \sin\theta_i \sin\phi_i, \cos\phi_i)$. Hence, we define two encoder networks $\mathcal{P}_1$ and $\mathcal{P}_2$ for body pose and target microphone positions, respectively. For the pose encoder $\mathcal{P}_1$, two temporal convolutional layers along with a ReLU activation layer are used to process the pose sequences, whose outputs are concatenated with the 3D body joint features and then mapped into the pose embeddings with an MLP layer. The position encoder $\mathcal{P}_2$ is simply defined as an MLP layer. Then, we concatenate the pose and position embeddings into a conditional feature f_p for guiding audio spatialization.

3.4 Non-Autoregressive Transformer

Masked Generative Modeling. In Sec. 3.2, we present audio tokenization to encode the mixed audio signal into discrete acoustic tokens. The mapping network $\mathcal{M}$ in Eq. (1) can be formulated as a sequence-to-sequence model with the external condition f_p. Auto-regressive models are the mainstay in sequence-to-sequence modeling, which predicts future tokens based on a combination of past tokens. However, this modeling paradigm introduces huge sampling latency, especially for long sequences. Here, we explore the non-autoregressive models, which have recently been shown to be effective in language, image, and audio generation with discrete representations (Ghazvininejad et al., 2019; Chang et al., 2022, 2023; Ziv et al., 2024). Given a source discrete audio representation $\mathbf{t}_{1:T'} = (t_1, ..., t_{T'})$ and the corresponding pose conditioning feature f_p, the mapping function predicts the target audio tokens in a non-autoregressive way:

$$\mathcal{M}(t_1, ..., t_{T'}|f_p, \mathbf{s}) = \prod_{k=1}^{K} \mathcal{M}(m(k) \odot \mathbf{t}|(1 - m(k)) \odot \mathbf{t}, f_p, \mathbf{s}), \tag{2}$$

where $k \in [1, K]$ indicates the decoding time step and $m(k) = (m_1(k), ..., m_{T'}(k))$ is a mask function. Let $\mathcal{I}_k$ be the masked indices for time step k, the mask function can be defined as:

$$m_i(k) \sim \mathbb{1}[i \in \mathcal{I}_k], \quad \text{where} \quad \mathcal{I}_k \sim \mathcal{U}(\{\mathcal{A} \subseteq \mathcal{I}_{k-1} : |\mathcal{A}| = \gamma(k) * T'\}). \tag{3}$$

In Eq. (3), $\gamma(k)$ is a mask schedule function to control the number of masked tokens in time step k. In practice, we use the cosine annealing function, which is defined as $\gamma(k) = \cos(\frac{\pi(k-1)}{2K})$. In inference, the generation begins with a fully masked sequence, in which the mask indices $\mathcal{I}_0 = \{1, T'\}$. Then, the model gradually predicts the tokens at the masked indices $\mathcal{I}_k$ with the unmasked (predicted) tokens. During training, we randomly select the masked positions and compute the cross-entropy loss.

Network Design. We use the transformer model to achieve the aforementioned masked generative modeling. Firstly, we encode the acoustic tokens from head-mounted microphones and pose features with an attention-based encoder $\mathcal{E}$. Then, the target spatial audio tokens are generated with the condition of input audio and pose. This non-autoregressive decoder $\mathcal{D}$ consists of multiple stacked attention blocks. Finally, the generated

tokens are converted into spatialized audio with the Codec. Since the target spatial audios share the same duration as the audios from head-mounted microphones, our model does not need an additional length predictor to predict the length of target sequence. In the following section, we elaborate on how to design an effective modality fusion approach for leveraging visual cues.

3.5 Multi-Scale Audio-Visual Fusion

The body pose stream can suggest where the body sound emanates from and influence its propagation. In other words, feature embedding of head-mounted microphone audio (source) and spatial audio (target) generation both require guidance from pose information. In this work, we propose a multi-scale audio-visual fusion (MAVF) module to fully exploit this visual cue. Specifically, we employ coarse fusion in source audio encoding and fine-grained fusion in target audio decoding (prediction). Given the acoustic token from head-mounted microphones $\mathbf{s}$ and pose embedding f_p, we use a cross-attention module $\mathcal{C}$ to extract the audio features by matching pose information:

$$\mathcal{C}(Q, K, V) = \text{softmax}(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}})V, \tag{4}$$

where Q is the learnable audio representation by taking token embedding and linear mapping on the input token $\mathbf{s}$ and K, V are the learnable pose representations obtained by mapping f_p. Then, a feed-forward layer is used to project the fused representation. Here, we slightly abuse K to indicate pose representation rather than the decoding steps in masked generative modeling.

This coarse audio-visual fusion is employed before the audio encoder. We empirically observe that this coarse fusion is enough for learning the pose-guided source audio representation, but on the decoding part, a fine-grained fusion strategy is preferred by the target audio generation (see Sec. **??**). For the decoding part, since the target spatial audio should preserve the semantic information from source audio, we propose to combine the audio-visual conditions with the decoder in a more fine-grained way. Concretely, the fused source inputs are forwarded to an encoder $\mathcal{E}$, whose output is defined as f_s. Let $f_t^{(j)}$ be the intermediate output of j-th attention block in decoder $\mathcal{D}$, we conduct conditioning on each of the attention blocks by residual connection:

$$f_t^{(j)} = f_t^{(j)} + \mathcal{C}(f_s, f_t^{(j)}), \tag{5}$$

where $\mathcal{C}$ is the cross-attention module defined in Eq. (4). It is worth noticing that the pose information is not explicitly modeled in the fine-grained fusion stage since it is already embedded in f_s.

3.6 Masking and Sampling Strategy.

Different from languages, adjacent acoustic tokens usually share similar information due to the receptive field of neural codec. Motivated by this, we follow the masking strategy of (Ziv et al., 2024), which uses non-overlapping spans of tokens as the base unit for masking scheme, rather than individual ones. Given a masking rate $\gamma(k)$, the number of spans N_k to be masked can be computed by:

$$1 - \frac{\binom{T'-L}{N_k}}{\binom{T'}{N_k}} \approx \gamma(k) \tag{6}$$

where L is the length of a single span. The above formulation expresses the expected masking rate when given a random placement of N_k spans of span length L over a sequence of length T'.

During inference, the model predicts the target tokens with K times for iterative refinement in a non-autoregressive manner. We use the prediction confidence at decoding iteration k as a scoring function and mask the least probable spans. After masking, our model predicts the masked spans with the condition of audio-visual fused representation and unmasked spans. However, recall that audio tokenization brings in multiple codebooks, which is usually solved by a coarse-to-fine prediction approach (Borsos et al., 2022; Ziv et al., 2024). In other words, each codebook is generated with the condition of previously generated codebooks. This is because for prompt-guided audio generation, the semantic and acoustic characteristics of the target audio are not explicitly cued and the codec conducts quantization in a coarse-to-fine manner. Different from them, we aim at modeling audio spatialization, in which the semantic and acoustic information is well-aligned with the proposed fine-grained guidance from source audio. Based on this, we find that predicting different codebooks in a parallel way is feasible for audio spatialization, which contributes to significant speedups compared to previous methods. After predicting the target audio tokens, we decode them into spatial audios with the neural codec DAC (Kumar et al., 2023).

3.7 Sound Field Rendering

Given all the spatial audios and their locations, the 3D sound field can be encoded in the form of

harmonic sound field coefficients (Samarasinghe et al., 2012) with a short-time Fourier transform. The maximum harmonic order $O = \lfloor \sqrt{N} \rfloor - 1$ is restricted by the number of spatial microphones. Then, the sound at any position in the space can be decoded with an inverse short-time Fourier transform based on encoded coefficients (Williams, 1999). More details can be found in Appendix A

4 Experiments

4.1 Dataset

We use the first 3D sound field dataset SoundingBodies (Xudong et al., 2023), which captures audio-pose pairs in an anechoic chamber with 5 Kinects for body pose tracking, 7 head-mounted microphones for gathering input audios and 345 microphones arranged on a sphere for collecting the spatialized audios. The sampling rates for pose video and recorded audio are 30fps and 48kHz, respectively. There are in total of 4.4 hours of body sounds and speech from eight participants. The participants performed various activities (32 types in total) to produce speech and body sounds like clapping, snapping, footsteps, and so on. In the published version, the eighth participant and the speech part of the seventh participant are not included. The original implementation split the sounds into clips of one second in length, which is not realistic in real-world applications. Here, we introduce two more challenging settings: 1) We randomly choose a participant except for the seventh one as the validation/test set, and the remaining participants are used for training. Since the test participant is unseen for training, we call it "Test-Unseen". 2) The validation and test set are randomly sampled from all participants to constitute all activities with no overlapping. The remaining data is the training set. We call it "Test-Seen".

4.2 Experimental Details

For tokenization, we use the descript audio codec (Kumar et al., 2023) with four codebooks, where each has a codebook size of 1024. We use the time-warping technique in (Xudong et al., 2023) to process the 7-channel audio inputs. For the transformer model, the hidden dimension for each attention block is 512 with 8 attention heads. We set 3 and 6 attention blocks for the encoding and decoding parts, respectively. We use AdamW optimizer to train our model with learning rate 5e-4, $\beta_1 = 0.9$, $\beta_2 = 0.95$, and a decoupled weight decay of 0.1. Following (Ziv et al., 2024), classifier-free guidance (Ho and Salimans, 2022) is used with a condition dropout of 0.3, and a guidance coefficient 10.0 annealed to 1.0. Since the lengths of speech data (150s or 300s) are too long, we crop speech data into multiple 15s segments. NAT3DSound requires about 5 days for pre-training and 3 days for fine-tuning with 2 Tesla V100 GPUs, respectively. For sampling, the span length is set as 60ms, and the decoding steps $K = 5$.

SoundingBoies (Xudong et al., 2023) is the first work on this topic, which adopts a WaveNet-like (Oord et al., 2016) network with shift-ℓ_2 and multiscale STFT losses. We conduct a comparison with SoundingBoies and its naive baseline: time warping. Also, we compare with the non-autoregressive audio generation models MagNet (Ziv et al., 2024) and Visual MagNet, which uses a direct cross-attention module to fusion the pose information. We use three metrics for evaluation: signal-distortion-ratio (SDR), angular error on the phase spectrogram, and ℓ_2 error on the amplitude spectrogram. The reported amplitude errors are amplified by a factor of 1000.

Collecting 3D spatial audio data is expensive and time-consuming and there is also limited relevant data available. Considering that the acoustic characteristics of different speakers share some commonalities, we propose to incorporate TTS data to alleviate data scarcity in 3D sound field synthesis. Specifically, we employ the same masked training for speech data to learn the general acoustic characteristics. After pre-training, the parameters of the encoder are initialized and the decoder is preserved. We demonstrate the effectiveness of pre-training in improving model performance (Sec. 4.3) and low-resource scenarios (Sec. 4.4).

4.3 Results

We report the comparison results with Test-Seen and Test-Unseen evaluation settings in Table 1 and 2, respectively. We can observe that 1) Our proposed realistic evaluation settings (Test-Seen and Test-Unseen) are more challenging than the baselines'. The baselines split the audio samples into 1-second clips and then selected the clips for testing, while we evaluated a whole section of audio. For example, SoundBodies get inferior performance on Test-Seen and Test-Unseen when compared to 1-second evaluation. Besides, Test-Unseen is harder than Test-Seen since the person on the test set has not appeared in the training set, which validates the

Method	Speech			Non-Speech		
	SDR(↑)	amplitude(↓)	phase(↓)	SDR(↑)	amplitude(↓)	phase(↓)
Time Warping*	-4.217	20.675	1.532	-0.016	1.031	0.505
SoundingBoies*	9.580	7.059	1.184	3.004	0.784	0.454
SoundingBoies	8.695	8.109	1.277	3.303	0.778	0.376
MagNet	6.776	10.173	1.331	1.780	0.835	0.382
Visual MagNet	7.972	8.932	1.300	3.070	0.783	0.383
NAT3DSound	9.533	7.530	**1.237**	3.434	0.763	**0.352**
NAT3DSound+pre.	**10.122**	**7.087**	1.249	**3.566**	**0.756**	0.354

Table 1: Results on the evaluation protocol of **Test-Seen**. "*" indicates the model is validated on a 1-second audio clip, while others are validated on a whole section of audio.

Method	Speech			Non-Speech		
	SDR(↑)	amplitude(↓)	phase(↓)	SDR(↑)	amplitude(↓)	phase(↓)
Time Warping*	-4.217	20.675	1.532	-0.016	1.031	0.505
SoundingBoies*	9.580	7.059	1.184	3.004	0.784	0.454
SoundingBoies	7.790	8.985	1.307	2.811	0.775	0.372
MagNet	5.871	10.252	1.338	2.329	0.810	0.376
Visual MagNet	7.194	9.626	1.311	2.065	0.820	0.374
NAT3DSound	9.118	8.117	1.269	2.970	0.753	0.336
NAT3DSound+pre.	**9.511**	**7.975**	**1.261**	**3.080**	**0.741**	**0.332**

Table 2: Results on the evaluation protocol of **Test-UnSeen**. "*" indicates the model is validated on a 1-second audio clip, while others are validated on a whole section of audio.

generalization ability of models. 2) Audio Spatialization requires careful cross-modal learning. The naive baseline Time Warping, which time-warps the input audio towards the receiver microphone, gains bad results across all tasks. 3) Our proposed NAT3DSound model achieves state-of-the-art performance on both evaluation settings compared to the first 3D sound field synthesis model SoundBodies, the advanced non-autoregressive audio generation model MagNet, and its variant Visual MagNet. 4) With acoustic pre-training, NAT3DSound+pre. outperforms NAT3DSound on both Test-Seen and Test-Unseen, which demonstrates the feasibility of leveraging sufficient speech data to improve audio spatialization. Equipped with acoustic pre-training, NAT3DSound achieves a new state-of-the-art: 10.122 and 3.566 SDRs on Test-Seen, 9.511 and 3.080 SDRs on Test-Unseen.

4.4 Low-resource Learning

We show the results of low-resource learning in Table 3. From the table, we can find that acoustic pre-training on sufficient TTS data can improve the performance in data-scarce scenarios. For example, even with the spatial data from a single person, our model gains 8.074 and 2.889 SDRs on speech and non-speech data, respectively. Despite the fact that the pre-training data is speech-only, they also improve the fine-tuning performance on non-speech data. A good start on the speech part helps the model optimization on hybrid spatial data. Besides, we can find that acoustic pre-training effectively reduces the amplitude error, which measures the mismatch in energy between ground truth. For example, pre-training reduces amplitude error from 10.690 to 8.636 on 1 person's speech section and 9.184 to 7.809 on 2 person's. For the angular phase error that mainly measures the time-domain alignment, the improvement from pre-training is limited.

4.5 Ablation Studies

We validate the effectiveness of our multi-scale audio-visual fusion and parallel sampling in the ablation studies, which are illustrated in Table 4.

For the ablation study on multi-scale audio-visual fusion, we can find that 1) Pose information is significant for audio spatialization since it sug-

Spatial Data	TTS Data	Speech			Non-Speech		
		SDR(↑)	amplitude(↓)	phase(↓)	SDR(↑)	amplitude(↓)	phase(↓)
1 person	0h	6.538	10.690	1.348	1.917	0.810	0.389
1 person	100h	8.074	8.636	1.281	2.889	0.749	0.359
2 person	0h	7.879	9.184	1.294	2.262	0.757	0.372
2 person	100h	9.134	7.809	1.262	2.919	0.727	0.357
full	100h	9.511	7.975	1.261	3.080	0.741	0.332

Table 3: Results of low-resource learning on **Test-Unseen**.

ID	Method	Speech			Non-Speech		
		SDR(↑)	amplitude(↓)	phase(↓)	SDR(↑)	amplitude(↓)	phase(↓)
Ablation study on multi-scale audio-visual fusion							
1'	no pose	6.899	10.127	1.280	2.845	0.799	0.380
2'	S:Coarse, T:Coarse	8.432	8.515	1.295	3.221	0.744	0.378
3'	S:Fine., T:Fine.	10.081	7.304	1.260	3.457	0.755	0.381
4'	S:Coarse, T:Fine.	10.122	7.087	1.249	3.566	0.756	0.354
Ablation study on parallel sampling							
5'	gradual sampling	10.438	7.017	1.256	3.489	0.722	0.372
6'	parallel sampling	10.122	7.087	1.249	3.566	0.756	0.354

Table 4: Results of ablation studies on **Test-Seen**. "S" indicates the modality fusion of source audio on the encoder part, while "T" indicates the modality fusion on the decoder part for target audio prediction.

gests where the body sound emanates from and influences its propagation. Without pose guidance, the model gains a clear performance drop (1' vs. others). 2) Fine-grained modality fusion on the decoder part is important since it provides both pose and source audio information to guide the prediction of target spatial audio. Fine-grained fusion clearly outperforms the coarse one. (2' vs. 4'). 3) A coarse fusion on the encoder part is enough to learn a great condition output from the source audio, the pose streams, and the target spatial position (3' vs. 4'). Since the performances of 3' and 4' are similar, we adopt a more concise design: we only conduct coarse fusion before the encoder.

For sampling strategy, we compare our parallel sampling with the gradual sampling adopted by previous works (Borsos et al., 2022; Ziv et al., 2024). Gradual sampling adopts a coarse-to-fine manner: each codebook is generated with the condition of previously generated codebooks, while parallel sampling predicts each codebook simultaneously. We can find the performance of parallel sampling is on par with the gradual sampling (5' vs. 6'). However, parallel sampling enjoys $4\times$ sampling speed-up compared to gradual sampling due to the four codebooks we used.

5 Conclusion

In this paper, we investigate a novel problem: 3D sound field synthesis, which takes body pose and head-mounted microphone signals as inputs and renders a 3D sound field that the audio at any position can be inferred. To achieve this, we propose NAT3DSound, the first unit-based model with hierarchical cross-modal modeling. We build a transformer architecture with masked generative modeling for audio spatialization. Further, we propose multi-scale audio-visual fusion to effectively integrate body pose stream and spatial position information, and a span-based parallel sampling strategy for efficiency. To alleviate the data scarcity, we propose acoustic pre-training with speech data to improve the performance on low-resource scenarios. Extensive experiments and ablation studies demonstrate the effectiveness of our model. We hope our NAT3DSound model could serve as a strong baseline for 3D sound field synthesis, and promote the development of AR and VR.

In future works, we plan to investigate better pre-training and fine-tuning schemes for the body sounds since there is limited research on this topic. Also, we plan to explore the audio propagation with general environments, not limited to the body pose.

6 Limitations and Potential Risks

Despite that our model achieves favorable performance on audio spatialization, it still has several limitations: 1) We build our model upon the pose-guided 3D sound field synthesis, in which the interaction with the surrounding environment (e.g. objects and other persons) and the environmental sounds (e.g. winds and rains) are neglected. 2) Our model relies on the neural codec for audio compression. There is a loss of information in this process, but this loss is small and can be further reduced with codec development. 3) Our model is not efficient enough to run on low-computing hardware like a consumer VR device.

NAT3DSound improves 3D sound generation, so it may help expand any impact that audio generation systems had on the broader world like copyright issues. We will add constraints and licenses when open-resourcing our code and pre-trained models.

References

Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. 2023. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Timur Bagautdinov, Chenglei Wu, Tomas Simon, Fabian Prada, Takaaki Shiratori, Shih-En Wei, Weipeng Xu, Yaser Sheikh, and Jason Saragih. 2021. Driving-signal aware full-body avatars. *ACM Transactions on Graphics (TOG)*, 40(4):1–17.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. 2022. Audiolm: a language modeling approach to audio generation. *arXiv preprint arXiv:2209.03143*.

Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. 2023. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. 2022. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11315–11325.

Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. 2020a. Soundspaces: Audio-visual navigation in 3d environments. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 17–36. Springer.

Changan Chen, Sagnik Majumder, Ziad Al-Halah, Ruohan Gao, Santhosh Kumar Ramakrishnan, and Kristen Grauman. 2020b. Learning to set waypoints for audio-visual navigation. *arXiv preprint arXiv:2008.09622*.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2022. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*.

Chuang Gan, Yiwei Zhang, Jiajun Wu, Boqing Gong, and Joshua B Tenenbaum. 2020. Look, listen, and act: Towards audio-visual embodied navigation. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9701–9707. IEEE.

Ruohan Gao and Kristen Grauman. 2019. 2.5 d visual sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 324–333.

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. 2019. Mask-predict: Parallel decoding of conditional masked language models. *arXiv preprint arXiv:1904.09324*.

Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Qingqing Huang, Daniel S Park, Tao Wang, Timo I Denk, Andy Ly, Nanxin Chen, Zhengdong Zhang, Zhishuai Zhang, Jiahui Yu, Christian Frank, et al. 2023a. Noise2music: Text-conditioned music generation with diffusion models. *arXiv preprint arXiv:2302.03917*.

Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. 2023b. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. *arXiv preprint arXiv:2301.12661*.

Eugene Kharitonov, Damien Vincent, Zalán Borsos, Raphaël Marinier, Sertan Girgin, Olivier Pietquin, Matt Sharifi, Marco Tagliasacchi, and Neil Zeghidour. 2023. Speak, read and prompt: High-fidelity text-to-speech with minimal supervision. *arXiv preprint arXiv:2302.03540*.

Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungroh Yoon. 2020. Glow-tts: A generative flow for text-to-speech via monotonic alignment search. *Advances in Neural Information Processing Systems*, 33:8067–8077.

Bruno Korbar, Du Tran, and Lorenzo Torresani. 2018. Cooperative learning of audio and video models from self-supervised synchronization. *Advances in Neural Information Processing Systems*, 31.

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. 2022. Audiogen: Textually guided audio generation. *arXiv preprint arXiv:2209.15352*.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. 2023. High-fidelity audio compression with improved rvqgan. *arXiv preprint arXiv:2306.06546*.

Dingzeyu Li, Timothy R Langlois, and Changxi Zheng. 2018. Scene-aware audio for 360 videos. *ACM Transactions on Graphics (TOG)*, 37(4):1–12.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. 2023. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*.

Jinglin Liu, Chengxi Li, Yi Ren, Feiyang Chen, and Zhou Zhao. 2022a. Diffsinger: Singing voice synthesis via shallow diffusion mechanism. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 11020–11028.

Xian Liu, Qianyi Wu, Hang Zhou, Yinghao Xu, Rui Qian, Xinyi Lin, Xiaowei Zhou, Wayne Wu, Bo Dai, and Bolei Zhou. 2022b. Learning hierarchical cross-modal association for co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10462–10472.

Yu-Ding Lu, Hsin-Ying Lee, Hung-Yu Tseng, and Ming-Hsuan Yang. 2019. Self-supervised audio spatialization with correspondence classifier. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 3347–3351. IEEE.

Shentong Mo and Pedro Morgado. 2022. Localizing visual sounds the easy way. In *European Conference on Computer Vision*, pages 218–234. Springer.

Pedro Morgado, Yi Li, and Nuno Nvasconcelos. 2020. Learning representations from audio-visual spatial alignment. *Advances in Neural Information Processing Systems*, 33:4733–4744.

Pedro Morgado, Nuno Nvasconcelos, Timothy Langlois, and Oliver Wang. 2018. Self-supervised generation of spatial audio for 360 video. *Advances in neural information processing systems*, 31.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba. 2016. Ambient sound provides supervision for visual learning. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 801–816. Springer.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Rui Qian, Di Hu, Heinrich Dinkel, Mengyue Wu, Ning Xu, and Weiyao Lin. 2020. Multiple sound sources localization from coarse to fine. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 292–308. Springer.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2019. Fastspeech: Fast, robust and controllable text to speech. *Advances in Neural Information Processing Systems*, 32.

Alexander Richard, Dejan Markovic, Israel D Gebru, Steven Krenn, Gladstone Alexander Butler, Fernando Torre, and Yaser Sheikh. 2021. Neural synthesis of binaural speech from mono audio. In *International Conference on Learning Representations*.

Prasanga N Samarasinghe, Thushara D Abhayapala, and Mark A Poletti. 2012. 3d spatial soundfield recording over large regions. In *IWAENC 2012; International Workshop on Acoustic Signal Enhancement*, pages 1–4. VDE.

Flavio Schneider, Ojasv Kamal, Zhijing Jin, and Bernhard Schölkopf. 2023. Mo\^ usai: Text-to-music generation with long-context latent diffusion. *arXiv preprint arXiv:2301.11757*.

Yapeng Tian, Di Hu, and Chenliang Xu. 2021. Cyclic co-learning of sounding object visual grounding and sound separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2745–2754.

Yating Tian, Hongwen Zhang, Yebin Liu, and Limin Wang. 2023. Recovering 3d human mesh from monocular images: A survey. *IEEE transactions on pattern analysis and machine intelligence*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. 2023. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*.

Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. 2017. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*.

E. G. Williams. 1999. Fourier acoustics. *Academic Press*.

XU Xudong, Dejan Markovic, Jacob Sandakly, Todd Keebler, Steven Krenn, and Alexander Richard. 2023. Sounding bodies: Modeling 3d spatial sound of humans using body pose and audio. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. 2019. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92). The Centre for Speech Technology Research (CSTR), University of Edinburgh. [sound].

D Yang, J Yu, H Wang, W Wang, C Weng, Y Zou, and D Diffsound Yu. 2022. Discrete diffusion model for text-to-sound generation. *arXiv preprint arXiv:2207.09983*.

Dongchao Yang, Jinchuan Tian, Xu Tan, Rongjie Huang, Songxiang Liu, Xuankai Chang, Jiatong Shi, Sheng Zhao, Jiang Bian, Xixin Wu, et al. 2023. Uniaudio: An audio foundation model toward universal audio generation. *arXiv preprint arXiv:2310.00704*.

Karren Yang, Bryan Russell, and Justin Salamon. 2020. Telling left from right: Learning spatial correspondence of sight and sound. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9932–9941.

Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2019. Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 4303–4309. IEEE.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. 2021. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507.

Zerong Zheng, Xiaochen Zhao, Hongwen Zhang, Boning Liu, and Yebin Liu. 2023. Avatarrex: Real-time expressive full-body avatars. *arXiv preprint arXiv:2305.04789*.

Hang Zhou, Xudong Xu, Dahua Lin, Xiaogang Wang, and Ziwei Liu. 2020. Sep-stereo: Visually guided stereophonic audio generation by associating source separation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 52–69. Springer.

Lingting Zhu, Xian Liu, Xuanyu Liu, Rui Qian, Ziwei Liu, and Lequan Yu. 2023. Taming diffusion models for audio-driven co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10544–10553.

Alon Ziv, Itai Gat, Gael Le Lan, Tal Remez, Felix Kreuk, Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2024. Masked audio generation using a single non-autoregressive transformer. *arXiv preprint arXiv:2401.04577*.

A Sound Field Rendering

Given all the spatial audios and their locations, the 3D sound field can be encoded in the form of harmonic sound field coefficients (Samarasinghe et al., 2012) with a short-time Fourier transform. The maximum harmonic order $O = \lfloor \sqrt{N} \rfloor - 1$ is restricted by the number of spatial microphones. Formally, we can use the harmonic sound field coefficients $\beta_{mn}(\tau, f)$ to express the target audio signal $\mathbf{t}_i$ captured by the i-th microphone at position (r_i, θ_i, ϕ_i):

$$\text{STFT}(\mathbf{s}_i) = \sum_{n=0}^{O} \sum_{m=-n}^{n} [\beta_{mn}(\tau, f) h_n(qr_i) Y_{nm}(\theta_i, \phi_i)], \tag{7}$$

where STFT is the short-time Fourier transform, τ and f are the time and frequency bins, respectively, $q = 2\pi f / v_{\text{sound}}$ is the wave number, v_{sound} is the sound speech, $h_n(qr_i)$ is the nth-order spherical Hankel function and $Y_{nm}(\theta_i, \phi_i)$ denotes the spherical harmonic of order n and degree m.

Then, we can obtain the encoded harmonic sound field coefficients $\beta_{mn}(\tau, f)$. Based on it, the sound at any position in the space $p(r, \theta, \phi, t)$ can be decoded with an inverse short-time Fourier transform (Williams, 1999):

$$p(r, \theta, \phi, t) = \text{iSTFT}\left(\sum_{n=0}^{O} \sum_{m=-n}^{n} \beta_{mn}(\tau, f) h_n(qr_i) Y_{nm}(\theta_i, \phi_i) \right), \tag{8}$$

where iSTFT denotes the inverse short-time Fourier transform.

B Datasets for Pre-training

In Table 5, we list the statistics of the datasets used for pre-training. We use these data for research only.

Dataset	Speech Time
VCTK (Yamagishi et al., 2019)	44 hours
CommonVoice-English v15.0 (Ardila et al., 2019)	3,437 hours
LibriSpeech-other (Panayotov et al., 2015)	500 hours

Table 5: Pre-training Dataset Statistics.