Regression and Classification for Direction-of-Arrival Estimation with Convolutional Recurrent Neural Networks

使用CRNN进行DoA估计的回归和分类

*Zhenyu Tang, John D. Kanu, Kevin Hogan, Dinesh Manocha*

University of Maryland

zhy@cs.umd.edu, jdkanu@cs.umd.edu, khogan@cs.umd.edu, dm@cs.umd.edu

This work was supported in part by ARO grant W911NF-18-10313 and Intel. Project page <https://gamma.umd.edu/pro/speech/doa>

**https://github.com/RoyJames/doa-release**

# Abstract

We present a novel learning-based approach to **estimate the direction-of-arrival (DOA)** of a sound source using a convolutional recurrent neural network (CRNN) trained via regression on synthetic data and Cartesian labels.

We also describe an improved method to **generate synthetic data** to train the neural network using state-of-the-art sound propagation algorithms that model specular as well as diffuse reflections of sound.

We compare our model against three other CRNNs trained using different formulations of the same problem: classification on categorical labels, and regression on spherical coordinate labels.

In practice, our model achieves up to 43% decrease in angular error over prior methods. The use of diffuse reflection results in 34% and 41% reduction in angular prediction errors on LOCATA and SOFA datasets, respectively, over prior methods based on image-source methods. Our method results in an additional 3% error reduction over prior schemes that use classification networks, and we use 36% fewer network parameters.

Index Terms: speech recognition, sound propagation, direction of arrival estimation, data augmentation

我们提出一种新颖的基于学习的方法，通过使用对合成数据和笛卡尔标记进行回归训练的卷积递归神经网络（CRNN）来估计声源的到达方向（DOA）。

我们还描述了一种改进的方法，该方法可使用对声音的镜面反射和漫反射建模的最新声音传播算法来生成合成数据来训练神经网络。

我们将我们的模型与使用同一问题的不同公式训练的其他三个CRNN进行了比较：分类标签上的分类，球面坐标上的回归。

实际上，与现有方法相比，我们的模型可将角度误差降低多达43％。与基于图像源方法的现有方法相比，使用漫反射可使LOCATA和SOFA数据集上的角度预测误差分别减少34％和41％。与使用分类网络的现有方案相比，我们的方法可将错误减少3％，并且网络参数减少36％。

索引词：语音识别，声音传播，到达方向估计，数据增强

# Introduction

Estimating the direction-of-arrival (DOA) of sound sources has been an important problem in terms of analyzing multi-channel recordings [1, 2]. In these applications, the goal is to predict the azimuth and elevation angles of the sound source relative to the microphone, from a sound clip recorded in any multichannel setting. One of the simpler problems is the estimation of the DOA on the horizontal plane [3]. More complex problems include DOA estimation in three-dimensional space or the identification of both direction and distance of an audio source. Even more challenging problems correspond to performing these goals in noisy and reverberant environments.

就分析多通道录音而言，估计声源的到达方向（DOA）已经成为一个重要的问题[1、2]。在这些应用中，目标是根据任何多通道设置中记录的声音片段来预测声源相对于麦克风的方位角和仰角。最简单的问题之一是在水平面上估计DOA [3]。更复杂的问题包括三维空间中的DOA估计或音频源的方向和距离的标识。更具挑战性的问题对应于在嘈杂和混响环境中实现这些目标。

To analyze spatial information from sound recordings, at least two microphones with known relative positions must be used. In practice, various spatial recording formats including binaural, 5.1-channel, 7.1-channel, etc. have been applied to spatial audio related systems [4]. The Ambisonics format decomposes a sound field using a spherical harmonic function basis [5]. Compared with its alternatives, Ambisonics has the advantage of being hardware independent–it does not necessarily encode microphone specifications into the recording.

要分析录音中的空间信息，必须使用至少两个具有已知相对位置的麦克风。 实际上，包括双声道，5.1声道，7.1声道等在内的各种空间记录格式已应用于与空间音频相关的系统[4]。 Ambisonics格式使用球谐函数[5]分解声场。 与其他替代品相比，Ambisonics具有与硬件无关的优势–它不必将麦克风规格编码到录音中。

Recent work [6] has applied the Ambisonics format to DOA estimation and trained a CRNN classifier that yields more accurate predictions than a baseline approach using independent component analysis. While a regression formulation seems more natural for the problem of DOA estimation, some recent work [3] suggests a regression formulation may yield worse performance than that of the classification formulation for multilayer perceptrons.

arXiv:1904.08452v3 [cs.SD] 10 Jul 2019

In this work, we present a novel learning based approach for estimating DOA of a single sound source from ambisonic audio, building on an existing deep learning framework [6]. We present a CRNN which predicts DOA as a 3-D Cartesian vector. We introduce a method to generate synthetic data using geometric sound propagation that models specular and diffuse reflections, which results in up to 43% error reduction compared with image-source methods. We conduct a four-way comparison between the Cartesian regression network, two classification networks trained with cross-entropy loss, and a regression network trained using angular loss. Finally, we investigate results on two 3rd-party datasets: LOCATA [7] and SOFA [8], where our best model reduces angular prediction error by 43% compared to prior methods.

最近的工作[6]已将Ambisonics格式应用于DOA估计，并训练了一个CRNN分类器，该分类器比使用独立分量分析的基线方法产生的预测更准确。尽管对于DOA估计问题而言，回归公式似乎更自然，但一些近期的工作[3]表明，与多层感知器的分类公式相比，回归公式的性能可能较差。

在这项工作中，我们基于现有的深度学习框架[6]，提出了一种新颖的基于学习的方法，用于从歧义音频估计单个声源的DOA。我们提出了一种将DOA预测为3-D笛卡尔向量的CRNN。我们介绍了一种使用几何声音传播生成合成数据的方法，该方法对镜面反射和漫反射进行建模，与图像源方法相比，该方法最多可将错误减少43％。我们在笛卡尔回归网络，经过交叉熵损失训练的两个分类网络和经过角度损失训练的回归网络之间进行四向比较。最后，我们调查了两个第三方数据集的结果：LOCATA [7]和SOFA [8]，与以前的方法相比，我们的最佳模型将角度预测误差降低了43％。

Section 2 gives an overview of prior work. We propose our method in Section 3. Section 4 presents our results on two benchmarks and we conclude in Section 5.

# Related Work

## Overview

One classic approach to DOA estimation is to first determine the time delay of arrival (TDOA) between microphone array channels, which can be estimated by generalized cross correlation [9] or least squares [10]. The DOA can be computed from known TDOA and the array layout directly. Another approach is to use the signal subspace, as in the MUSIC algorithm [11]. With some restrictions on operating conditions, these techniques are very effective. However, they do not perform well in highly reverberant and noisy environments, or when the placement of signal sources is arbitrary [12]. More recently, researchers have applied modern machine learning techniques to speech DOA estimation with the goal of improving performance in noisy, realistic environments, which can be categorized into *classification* and *regression* networks. For both networks, angular error for a single example is proportional to the angular distance between the predicted and actual DOA.

DOA估计的一种经典方法是首先确定麦克风阵列通道之间的到达时间延迟（TDOA），可以通过广义互相关[9]或最小二乘[10]进行估计。 可以直接从已知的TDOA和阵列布局中计算出DOA。 另一种方法是使用信号子空间，如MUSIC算法[11]。 在操作条件上有一些限制，这些技术非常有效。 但是，它们在高度混响和嘈杂的环境中，或者当信号源的放置是任意的时，效果不佳[12]。 最近，研究人员将现代机器学习技术应用于语音DOA估计，目的是提高在嘈杂，现实环境中的性能，可以将其分类为分类和回归网络。 对于这两个网络，单个示例的角度误差与预测的和实际的DOA之间的角度距离成比例。

## Classification Formulations

In the classification formulation, DOA is encoded using a *categorical* representation: an approximately uniform mesh-grid defines the score for each of a finite set of possible categories, where each category corresponds to a unique region of the continuous DOA space. The mesh grid is defined by subdividing the DOA space at a given resolution. The DOA is decoded as the direction associated with the bin with highest score. The categorical formulation uses a discrete encoding, lending itself to a class-based formulation of DOA estimation. Generalized cross correlation (GCC) feature vectors of a microphone array input have been fed to a multilayer perceptron classifier, which predicts a DOA in one angular dimension [3] and show superior performance over the classic least square method [10] in both simulated and real rooms of various sizes. Perotin et al. [6] calculate acoustic intensity vectors using a first-order Ambisonics representation of audio. This representation serves as input to a CRNN, which predicts a DOA in two angular dimensions. Their CRNN yields more accurate predictions than a baseline approach using independent component analysis. CRNNs have also been used in [13] to identify the DOA for overlapping sound sources, in two angular dimensions.

在分类公式中，DOA使用分类表示进行编码：近似均匀的网格定义了有限类别的可能类别中每个类别的得分，其中每个类别对应于连续DOA空间的唯一区域。通过以给定的分辨率细分DOA空间来定义网格。将DOA解码为与得分最高的bin关联的方向。分类公式使用离散编码，从而使其适用于DOA估计的基于类的公式。麦克风阵列输入的广义互相关（GCC）特征向量已馈入多层感知器分类器，该分类器可预测一个角度维度的DOA [3]，并且在模拟和仿真中均表现出优于经典最小二乘法[10]的性能。各种大小的真实房间。 Perotin等 [6]使用音频的一阶Ambisonics表示来计算声强矢量。此表示用作CRNN的输入，它可以在两个角度维度上预测DOA。与使用独立成分分析的基准方法相比，它们的CRNN可以得出更准确的预测。在[13]中还使用了CRNN在两个角度维度上识别重叠声源的DOA。

## Regression Formulations

In the regression formulation, two representations of DOA are commonly used, which we refer to as *Cartesian* and *spherical*. With the Cartesian representation, DOA is encoded as a three dimensional vector in Cartesian (*x,y,z*) coordinates, pointing toward the source. With the spherical representation, DOA is encoded as a two-dimensional vector of azimuth (*θ*) and elevation (*φ*) angles. Both formulations encode DOA in continuous space, leading to a regression formulation of DOA estimation. In prior work, regression formulations have not shown superior empirical results for DOA estimation. Higher angular errors for regression than for classification is claimed in [3]. CNN regression has been used in [14] to estimate the Cartesian coordinates of a sound source in 3-D space. CRNN regression has been used in [13] and higher angular error is observed for DOA estimates for regression than for classification. Similar to this result, our experiments show a higher angular error for regression on spherical DOA than for classification. However, we discover a *lower* angular error for regression on Cartesian DOA, than for both classification on categorical DOA and regression on spherical DOA.

在回归公式中，通常使用DOA的两种表示形式，我们将其称为笛卡尔和球形。使用笛卡尔表示法，DOA被编码为笛卡尔（x，y，z）坐标中的三维向量，指向源。利用球形表示，DOA被编码为方位角（θ）和仰角（φ）的二维向量。两种公式都在连续空间中编码DOA，从而导致DOA估计的回归公式。在先前的工作中，回归公式并未显示出用于DOA估计的出色经验结果。在文献[3]中提出了比分类更高的回归角误差。 CNN回归已在[14]中用于估计3-D空间中声源的笛卡尔坐标。 CRNN回归已在[13]中使用，观察到的DOA估计的角度误差要比分类更高。与此结果类似，我们的实验显示，对于球形DOA而言，回归所产生的角度误差要大于分类所产生的角度误差。但是，我们发现笛卡尔DOA上回归的角度误差要比分类DOA上的分类和球面DOA的回归都低。

# Proposed Method

## Data Preparation

Our DOA estimation network also relies on a large amount of labeled training data. However, collecting Ambisonic recordings and manually labeling them for training is tedious and time-consuming. Therefore, in speech/audio related training, the common practice is to use image-source methods to generate synthetic impulse responses for augmenting the training data [15]. However, the distribution of synthetic data may not match that of the real data well enough, which can cause large generalization error when applying a synthetically trained DNN to real test data. To overcome this issue of domain mismatch, a more accurate approach for generating training data is needed.

我们的DOA估算网络还依赖于大量带标签的训练数据。 但是，收集Ambisonic录音并手动标记它们以进行培训既乏味又耗时。 因此，在语音/音频相关训练中，通常的做法是使用图像源方法来生成合成冲激响应，以增强训练数据[15]。 但是，合成数据的分布可能无法与真实数据充分匹配，这在将经过综合训练的DNN应用于真实测试数据时会导致较大的泛化误差。 为了克服域不匹配的问题，需要一种更精确的方法来生成训练数据。

Sound propagation methods compute the reflection and diffraction paths from the sound sources to a listener in the virtual environment. Image-source methods do not model sound scattering or diffuse reflections, which are important phenomenons in acoustic environments. We utilize the state-of-theart geometric sound propagation method [16, 17, 18] to generate synthetic data that is more accurate than image-source methods. Its benefit has also been observed in other speech tasks [19].

Following the suggested procedure in [6], we generated 42,000 rectangular room configurations with dimensions uniformly and independently sampled between 2*.*5*m*×2*.*5*m*×2*m* and 10*m*×10*m*×3*m*. Under each room configuration, we randomly populate three paired source-listener locations, both at least 0*.*5*m* away from walls. The geometric sound propagation method based on path tracing is used on each source-listener pair to generate its spatial room impulse response (SRIR). Then we convolve each SRIR with a randomly selected one-second clean speech sample from the Libri ASR corpus [20] to generate realistic reverberant speech recordings in Ambisonic format. Babble and speech shaped noise [21] are added to the convolved sound at signal-to-noise ratios (SNRs) following a normal distribution centered at 15*dB* with a standard deviation of 1*dB* as recommended by [22]. A short time fourier transform (STFT) is used to convert speech waveforms to spectrogram, and the features are extracted according to Section 3.2.

声音传播方法计算虚拟环境中从声源到听众的反射和衍射路径。图像源方法不能对声音散射或漫反射建模，这在声学环境中是很重要的现象。我们利用最先进的几何声音传播方法[16、17、18]来生成比图像源方法更准确的合成数据。在其他语音任务中也观察到了它的好处[19]。

按照[6]中建议的步骤，我们生成了42,000个矩形房间配置，其尺寸均匀且独立地在2.5m×2.5m×2m和10m×10m×3m之间采样。在每种房间配置下，我们随机填充三个成对的源侦听器位置，两个位置均距墙壁至少0.5m。在每个源侦听器对上使用基于路径跟踪的几何声音传播方法来生成其空间房间脉冲响应（SRIR）。然后，我们将每个SRIR与来自Libri ASR语料库的随机选择的一秒钟干净语音样本进行卷积[20]，以生成Ambisonic格式的逼真的混响语音录音。按照[22]的建议，按照以15dB为中心，正态分布为1dB的正态分布，以信噪比（SNR）将bble语和语音形噪声[21]添加到卷积声音中。短时傅立叶变换（STFT）用于将语音波形转换为频谱图，并根据3.2节提取特征。

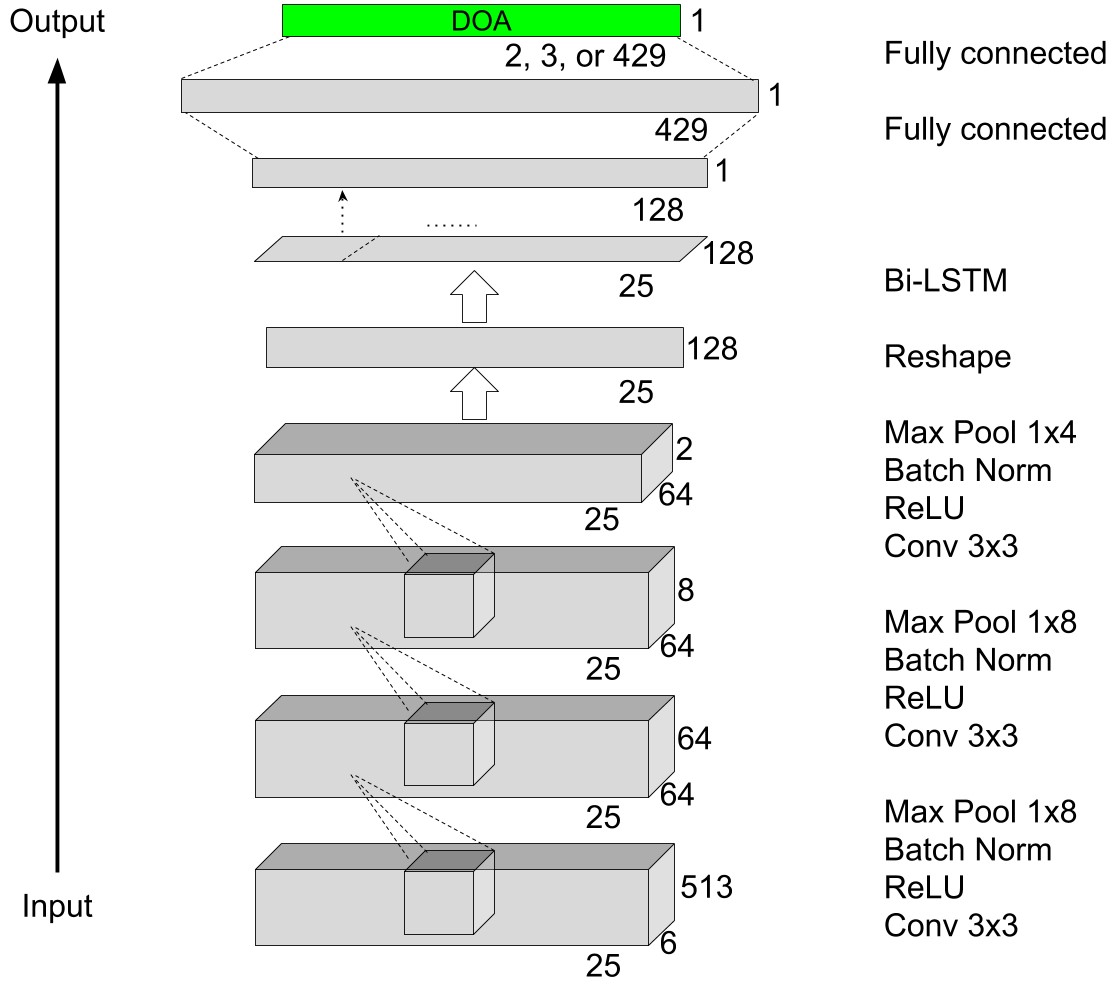


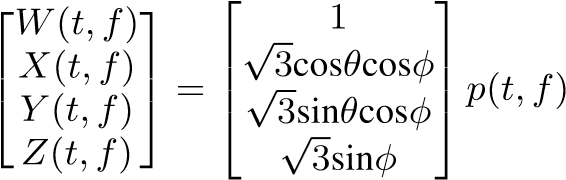
Figure 1: *General network architecture for each regression network and classification network. The dimensionality of the output vector (shown in green) is 2 for spherical formulation, 3 for Cartesian, and 429 for classification. Note that our implementation of the classifier is equivalent to the implementation in [6], but our regression networks differ in the size of the output layer, and use 36% fewer trainable parameters.*

图1：每个回归网络和分类网络的一般网络架构。 对于球形公式，输出矢量的维数（以绿色显示）为2，对于笛卡尔坐标系为3，对于分类为429。 请注意，我们对分类器的实现与[6]中的实现等效，但是我们的回归网络在输出层的大小上有所不同，并且使用的可训练参数减少了36％。

## Ambisonic Input Features

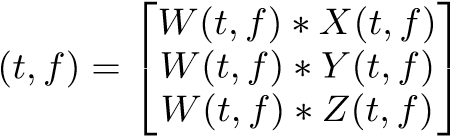
In theory, Ambisonics of an infinite number of bases can reproduce the recorded soundfield with no error. As a practical approximation, we use the first four bases/channels necessary for first-order ambisonics (FOA). The FOA channels are denoted by *W,X,Y,Z*, where *W* channel contains the zerothorder coefficients that represents the omnidirectional signal intensity, and *X,Y,Z* channels contain the first-order coefficients that encode direction modulated information. For a plane wave with azimuth *θ* and elevation *φ* creating a sound pressure *p*, the complex FOA components are:

从理论上讲，无穷多个音源的Ambisonics可以无错误地再现记录的声场。 作为实际的近似，我们使用一阶歧义（FOA）所需的前四个基数/通道。 FOA信道用W，X，Y，Z表示，其中W信道包含表示全向信号强度的零阶系数，X，Y，Z信道包含对方向调制信息进行编码的一阶系数。 对于具有方位角θ和仰角φ且产生声压p的平面波，复FOA分量为

*,* (1)

where *t* and *f* are time and frequency bins. We follow the approach in [6] to construct input features from raw FOA audio. The active and reactive intensity vectors are encoded as:

其中t和f是时间和频率块。 我们遵循[6]中的方法从原始FOA音频构造输入功能。 有功和无功强度矢量编码为：

**I** *,* (2)

**I***a*(*t,f*) = R{**I**(*t,f*)}*,***I***r*(*t,f*) = I{**I**(*t,f*)}*,* (3)

where R{·} and I{·} extract the real and imaginary components of a complex signal respectively. Both feature vectors are divided by to have a uniform range for deep neural network training.

其中R {·}和I {·}分别提取复数信号的实部和虚部。 两个特征向量均被除以以具有用于深度神经网络训练的统一范围。

## Output and Loss Formulation

We compare Cartesian (*x,y,z*) and Spherical (*θ,φ*) output representations in addition to the common Categorical representation in this work. A stacked CRNN formulation using Categorical outputs has been proposed in [6]. We use this network structure and derive a set of Categorical, Cartesian, and Spherical forms from it, differing only in the size of the output layer, our independent variable. Maintaining a high similarity among the network architecture enables us to conduct well controlled comparisons between output representations for DOA estimation. We visualize our network architecture for the following three output formulations in Fig. 1. Key differences between the formulations are summarized in Tab. 1.

除了这项工作中常见的分类表示，我们还比较了笛卡尔（x，y，z）和球形（θ，φ）的输出表示。 [6]中提出了使用分类输出的堆叠式CRNN公式。 我们使用这种网络结构，并从中得出一组分类，笛卡尔和球面形式，仅在输出层（我们的自变量）的大小上有所不同。 在网络体系结构之间保持高度相似性使我们能够在输出表示之间进行良好控制的比较，以进行DOA估计。 我们在图1中可视化了以下三个输出公式的网络体系结构。这些公式之间的主要区别总结在Tab 1。

### Categorical Outputs

We discretize the continuous DOA space into some number of possible categorical outputs. The angular resolution is chosen to be 10°, which results in 429 direction classes. Each training DOA label is assigned to the direction class that has the smallest angular difference from itself. We use this labeling to train a classifier that outputs a sigmoid vector of the direction class. The model is trained using cross-entropy loss.

我们将连续的DOA空间离散化为一些可能的分类输出。 角分辨率选择为10°，这将导致429个方向类别。 每个训练DOA标签都分配给与自身的角度差最小的方向类别。 我们使用这个标签来训练一个分类器，该分类器输出方向类的S型向量。 使用交叉熵损失训练模型。

### Cartesian Outputs

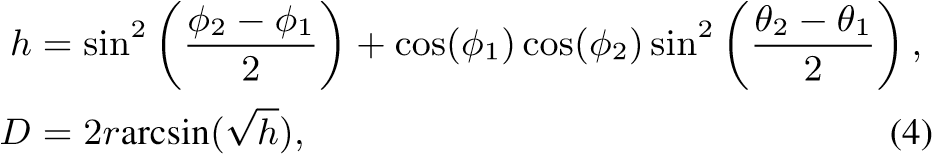
We define the output of the Cartesian network as a 3-D vector, representing DOA in Cartesian coordinates. Training labels are unit vectors in R3 pointing toward the source. We use meansquared error (MSE) as the loss to train our networks. Note that the output of the network is not constrained to lie on the unit sphere. As a result, a hypothetical output that is a large scalar multiple of the DOA label will result in a large loss, despite the perfect alignment of the output vector with the label. In practice, this property does not prevent our formulation from generating an accurate predictor, as shown in Section 4.

我们将笛卡尔网络的输出定义为3-D向量，以笛卡尔坐标表示DOA。 训练标签是R3中指向源的单位向量。 我们使用均方误差（MSE）作为训练网络的损失。 请注意，网络的输出不限于位于单位球面上。 结果，尽管输出矢量与标签完美对齐，但假设的输出是DOA标签的大标量倍数，将导致较大的损失。 实际上，此属性不会阻止我们的公式生成准确的预测变量，如第4节所示。

### Spherical Outputs

In contrast to the Cartesian formulation, the spherical formulation encodes DOA as a 2-D vector representation azimuth (*θ*) and elevation (*φ*) angles. This representation has only 2 degrees of freedom in 3-D, which means the Cartesian representation has added a redundant dimension to this learning problem. One issue with using this form is that the periodicity of spherical angles makes distance computation between two angles more complicated than in Cartesian coordinates. Conventional mean squared loss is discontinuous, and therefore non-differentiable, over predicted azimuth and elevation, which eliminates the guarantee of convergence of Gradient Descent. Instead, we compute the great-circle distance on a sphere’s surface using the haversine formula, which is differentiable:

与笛卡尔公式相反，球形公式将DOA编码为二维矢量表示方位角（θ）和仰角（φ）。 该表示形式在3-D中只有2个自由度，这意味着笛卡尔表示法为该学习问题添加了多余的维度。 使用这种形式的一个问题是球面角的周期性使两个角之间的距离计算比笛卡尔坐标中的复杂。 常规的均方损失在预测的方位角和高程上是不连续的，因此是不可微分的，从而消除了梯度下降收敛的保证。 取而代之的是，我们使用Haversine公式计算球体表面上的大圆距离，该公式是可微的：



where *r* is the radius of the sphere, and *D* is the great-circle distance between azimuth-elevation angles (*θ*1*,φ*1) and (*θ*2*,φ*2). By setting *r* = 1, we define *D* as our *Haversine loss*.

其中r是球体的半径，D是方位角（θ1，φ1）和（θ2，φ2）之间的大圆距离。 通过设置r = 1，我们将D定义为Haversine损失。

## Model Architecture and Training Procedure

As described in Section 3.2, the inputs to these networks are active and reactive intensity vectors. The size of the input is 6x25x513, containing 513 frequency bins (16kHz sample rate), for each 6-D intensity vector, computed at each of 25 frames.

如第3.2节所述，这些网络的输入是有功和无功强度矢量。 输入的大小为6x25x513，对于每个25维帧计算的每个6维强度矢量，包含513个频点（16kHz采样率）。

Table 1: *Three types of output representations studied in this paper. d is the number of classes into which a spherical surface is discretized. In our experiment, d* = 429*.*

表1：本文研究了三种类型的输出表示形式。 d是球面离散化的类的数量。 在我们的实验中，d = 429。

|  |  |  |  |
| --- | --- | --- | --- |
| Output Type | Activation | Dimension | Loss Function |
| Categorical | Sigmoid | R*d* | Cross-entropy |
| Cartesian | Linear | R3 | MSE |
| Spherical | Linear | R2 | Haversine |

The CRNN architecture is shown in Fig. 1. There are three convolutional layers, each consisting of 2-D convolution, a rectified linear unit (ReLU), batch normalization, and max pooling. The outputs of each of the three layers are 64x25x64, 64x25x8, and 64x25x2, respectively. The output of the last layer is flattened to a 128-D vector for each of the 25 frames. Each frame’s vector is fed into a two-layer bi-directional LSTM, and the output of the LSTM for each frame is fed through two timedistributed, fully-connected linear layers, generating a DOA estimate for each frame. As described in Section 3.3, we generate three forms of outputs depending on implementation: 2D azimuth-elevation angles, 3-D Cartesian coordinates, or 429 DOA classes. During training, losses are computed at each frame and error is backpropagated through the network. During evaluation, a single DOA estimate is taken as the uniformly weighted average of all estimates across frames.

CRNN体系结构如图1所示。共有三个卷积层，每个卷积层均由2-D卷积，整流线性单元（ReLU），批归一化和最大池化组成。 三层中每层的输出分别为64x25x64、64x25x8和64x25x2。 对于25个帧中的每个帧，最后一层的输出被展平为128-D矢量。 每帧的向量被馈送到两层双向LSTM，每帧LSTM的输出通过两个时间分布的，完全连接的线性层馈入，从而为每帧生成DOA估计。 如3.3节所述，我们根据实现方式生成三种输出形式：2D方位角仰角，3D笛卡尔坐标或429个DOA类。 在训练期间，会在每个帧上计算损失，并通过网络反向传播错误。 在评估期间，将单个DOA估计值作为所有帧中所有估计值的统一加权平均值。

# Experiment and Results

## Benchmarks

We evaluate each model on 1,189 samples from three staticsource microphone signals in the third-party sound localization and tracking (LOCATA) challenge dataset [7], and a dataset of ambisonic RIRs accompanying the Spatially Oriented Format for Acoustics (SOFA) convention [8]. No real RIRs are involved during the training phase. The noise and clean speech used for training and test are from different datasets. We use Eq. (4) as our angular error metric for visualization in Fig. 2.

Each signal in the LOCATA dataset is a real-world ambisonic speech recording with optically tracked azimuthelevation labels. In theory, ambisonic coefficients up to the 4th order can be captured by an Eigenmike microphone, but we only use its first order components.

From the SOFA dataset, we extract 225 SRIRs recorded in the Alte Pinakothek museum using Eigenmike®, Sennheiser AMBEO®, and SoundField®microphones. Positions and rotations of all loudspeaker and microphones are provided by measurements using laser meter and pointers. A reverberant test set is generated by convolving each SOFA SRIR with a random 1-second clip from the CMU Arctic speech databases [23]. Recorded background noise from the LOCATA dataset is added at a mean SNR level of 10*dB* with a standard deviation of 5*dB*.

我们评估了来自第三方声音定位和跟踪（LOCATA）挑战数据集中的三个静态源麦克风信号的1,189个样本中的每个模型，以及伴随着空间定向格式的声学（SOFA）约定[8]的歧义RIR数据集。 。在培训阶段不涉及实际的RIR。用于训练和测试的噪音和干净语音来自不同的数据集。我们使用等式。 （4）作为我们图2中可视化的角度误差度量。

LOCATA数据集中的每个信号都是具有光学跟踪的方位角标记的真实世界的混音语音记录。理论上，本征麦克麦克风可以捕获高达4阶的混响系数，但我们仅使用其一阶分量。

从SOFA数据集中，我们使用Eigenmike®，SennheiserAMBEO®和SoundField®麦克风提取在Alte Pinakothek博物馆中记录的225个SRIR。通过使用激光计和指示器进行测量，可以提供所有扬声器和麦克风的位置和旋转。通过将每个SOFA SRIR与来自CMU Arctic语音数据库的随机1秒片段进行卷积来生成混响测试集[23]。来自LOCATA数据集的已记录背景噪声以10dB的平均SNR电平添加，标准偏差为5dB。

## Results and Analysis

### LOCATA Dataset

|  |
| --- |
| (a) Recording 1 (b) Recording 2 (c) Recording 3  Figure 2: *Waveforms (top row) and angular tracking error (bottom row) for Recordings 1-3 in LOCATA Task 1. Shaded regions in the waveform indicate voice-active regions, while shaded regions in angular tracking error indicate the intersection of voice activity and regions containing predictions for all models. Each model must wait for a complete input window to make a prediction, hence the regions are not always identical between waveform and tracking error. Cartesian and Categorical models trained on our synthetic dataset achieve consistently lower tracking error compared with the classifier trained by Perotin et al. [6] and the MUSIC algorithm [11].*  图2：LOCATA任务1中的记录1-3的波形（顶部行）和角度跟踪误差（底部行）。波形中的阴影区域表示语音活动区域，而角度跟踪错误中的阴影区域表示语音活动的交集 以及包含所有模型预测的区域。 每个模型都必须等待完整的输入窗口做出预测，因此波形和跟踪误差之间的区域并不总是相同的。 与Perotin等人训练的分类器相比，在我们的综合数据集上训练的笛卡尔和分类模型实现了始终较低的跟踪误差。 [6]和MUSIC算法[11]。 |

We compute average error along the temporal axis, for each static-source signal in LOCATA Task 1. Estimates of DOA are generated for each frame in the microphone signal using a sliding window. The resulting estimates are interpolated to the timestamps provided in the LOCATA dataset. Prediction error is computed as the angular distance between the prediction and the ground-truth DOA. Angular tracking error is visualized in Fig. 2, for each predictor. Average angular error is computed over 234, 439, and 512 timestamps for Recordings 1, 2, and 3, respectively. Timestamps are selected to compute angular error if each algorithm makes a prediction for that timestamp, and the timestamp is located inside a voice-active region.

我们为LOCATA任务1中的每个静态源信号沿时间轴计算平均误差。使用滑动窗口为麦克风信号中的每个帧生成DOA估计。 结果估计值将插值到LOCATA数据集中提供的时间戳。 将预测误差计算为预测值与地面真实DOA之间的角距离。 对于每个预测变量，角跟踪误差在图2中可见。 对于记录1、2和3，分别在234、439和512个时间戳上计算了平均角度误差。 如果每种算法都对该时间戳做出预测，并且该时间戳位于语音活动区域内，则选择时间戳以计算角度误差。

To generate a prediction at frame *i* using a neural network, we feed in the sequence of 25 frames centered at *i*, and generate a sequence of 25 outputs of the network. If the output is Cartesian or spherical, we estimate DOA at *i* as the average of the outputs at each frame. If the output is a classification grid, we average the output grid over the frames, to produce a cumulative score for each DOA, and choose the DOA with highest score as the prediction. When running the MUSIC algorithm, we restrict it to 4-channel recordings, as well, for fair comparisons.

为了使用神经网络在第i帧处生成预测，我们输入以i为中心的25帧序列，并生成该网络的25个输出序列。 如果输出是笛卡尔或球形，则我们将i处的DOA估计为每帧输出的平均值。 如果输出是分类网格，则对帧上的输出网格求平均，以生成每个DOA的累积得分，然后选择得分最高的DOA作为预测。 当运行MUSIC算法时，为了公平比较，我们也将其限制为4通道录音。

Table 2: *Average angular tracking error within voice-active regions of LOCATA Task 1 recordings. Best performance in each column is highlighted in bold. All models are trained on data using specular and diffuse reflections in the geometric propagation algorithm, except for the MUSIC algorithm, which does not rely on training data. Perotin et al. [6] refers to the Categorical model trained on data generated by the image-source method.*

表2：LOCATA任务1录音的声音有效区域内的平均角度跟踪误差。 每列中的最佳性能以粗体突出显示。 除了不依赖训练数据的MUSIC算法外，所有模型都使用几何传播算法中的镜面反射和漫反射对数据进行训练。 Perotin等。 [6]指的是根据图像源方法生成的数据训练的分类模型。

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Recording 1 | Recording 2 | Recording 3 |
| MUSIC | 18.6° | 16.9° | 17.5° |
| Perotin et al. | 9.1° | 6.7° | 12.5° |
| Categorical | 9.3° | 6.3° | 3.2° |
| Cartesian | 8.5° | 5.8° | 6.8° |
| Spherical | 9.2° | 7.9° | 9.9° |

We observe that in Fig. 2 and Tab. 2, our Cartesian model achieves consistently lowest error on Recording 1 and 2, while our categorical model shows best performance in Recording 3. However, the Spherical model yields higher error than the other two models. We also observe that re-training the original model from Perotin et al. [6] using data generated by geometric method to augment data results in lower tracking error compared with the use of the image-source method.

我们在图2和Tab中观察到。 如图2所示，我们的笛卡尔模型在记录1和记录2上始终保持最低的误差，而我们的分类模型在记录3上表现出最好的性能。但是，球形模型产生的误差高于其他两个模型。 我们还观察到，重新训练了Perotin等人的原始模型。 [6]使用几何方法生成的数据来扩充数据，与使用图像源方法相比，可降低跟踪误差。

### SOFA Dataset

A larger scale test is performed using the SOFA dataset. MUSIC algorithm is not evaluated because this dataset does not provide microphone hardware configurations required by MUSIC. We compute the percentage of correctly predicted directions under error tolerances of 5°, 10°, and 15°, as well as each model’s average angular error on the whole dataset. It can be seen from Tab. 3 that our Cartesian model consistently achieves the best performance under each column, outperforming the baseline model by 43% in terms of average prediction error.

During our training procedure, we notice that each model is able to converge within dozens of epochs. However, the number of trainable parameters in regression models (i.e. Cartesian and Spherical) is only 64% of that in the classification model. This suggests that regression models tend to have a hypothesis set with lower complexity, which results in lower generalization error when tested on real data. We also tested the option of letting all models have approximately the same amount of trainable parameters, which has degraded the performance of the regression models. In conclusion, we are able to train a regression model that has superior performance over its corresponding classification model, although we do not observe obvious benefits in using the Spherical formulation.

使用SOFA数据集执行较大规模的测试。未评估MUSIC算法，因为此数据集未提供MUSIC所需的麦克风硬件配置。我们计算在5°，10°和15°的误差范围内正确预测的方向的百分比，以及每个模型在整个数据集中的平均角度误差。从Tab可以看到。 3我们的笛卡尔模型在每一列下始终获得最佳性能，就平均预测误差而言，其性能比基线模型高出43％。

在我们的训练过程中，我们注意到每种模型都可以在数十个纪元内收敛。但是，回归模型（即笛卡尔和球形）中可训练参数的数量仅为分类模型中可训练参数的数量的64％。这表明回归模型的假设集往往具有较低的复杂度，因此在对真实数据进行测试时，可以降低泛化误差。我们还测试了让所有模型具有大致相同数量的可训练参数的选项，这降低了回归模型的性能。总之，尽管我们没有发现使用球形公式的明显好处，但我们能够训练出性能优于其相应分类模型的回归模型。

Table 3: *Results on the SOFA dataset. First three columns show the percentage of DOA labels correctly predicted within error tolerances, followed by average angular errors, and %improvement on baseline. Best performance in each column is highlighted in bold.*

表3：SOFA数据集的结果。 前三列显示在误差容限内正确预测的DOA标签的百分比，其次是平均角度误差和基线改善百分比。 每列中的最佳性能以粗体突出显示。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | *<* 5° | *<* 10° | *<* 15° | Error | Improv. |
| Perotin et al. | 11.9% | 35.9% | 73.2% | 16.9° | - |
| Categorical | 24.4% | 58.2% | 88.7% | 9.96° | 41% |
| Cartesian | 24.4% | 66.3% | 88.2% | 9.68° | 43% |
| Spherical | 18.2% | 55.8% | 82.5% | 11.2° | 34% |

# Discussion and Future Work

In this paper, we demonstrate the benefits of using a geometric sound propagation simulator, as compared with image source methods for training DOA estimation networks, by reporting a higher accuracy on evaluation data. We evaluate the performance of a CRNN model in three output formulations: categorical, Cartesian, and spherical. We test them on two 3rd-party datasets and show that our Cartesian regression model achieves superior performance over classification and spherical models.

Evaluating classification models involves an additional factor: the resolution of the classification grid, which we kept fixed. Further, our work is limited to single-source localization problems, whereas in multi-source localization problems, classification models may have intrinsic advantages over regression models. Lastly, we restricted our simulation to extremely simple room settings to guarantee a fair comparison with the imagesource method. Furture work may involve experimentation on more complex room configurations.

在本文中，我们通过报告评估数据的更高准确性，证明了与用于训练DOA估计网络的图像源方法相比，使用几何声音传播模拟器的好处。我们在三种输出公式中评估了CRNN模型的性能：分类，笛卡尔和球形。我们在两个第三方数据集上对其进行了测试，结果表明我们的笛卡尔回归模型比分类模型和球形模型具有更高的性能。

评估分类模型涉及一个附加因素：分类网格的分辨率，我们保持不变。此外，我们的工作仅限于单源本地化问题，而在多源本地化问题中，分类模型可能比回归模型具有固有优势。最后，我们将模拟限制在极其简单的房间设置中，以确保与imagesource方法进行公平的比较。未来的工作可能涉及对更复杂的房间配置进行试验。

# References

1. C. Knapp and G. Carter, “The generalized correlation method for estimation of time delay,” *IEEE transactions on acoustics, speech, and signal processing*, vol. 24, no. 4, pp. 320–327, **1976**.

时延估计的广义相关方法

开发了最大似然（ML）估计器，用于确定,在存在不相关噪声的情况下, 在两个空间分离的传感器处接收到的信号之间,的时间延迟。 ML估计器可以实现为一对接收器预滤波器，后跟一个互相关器。相关器达到最大值的时间参数是延迟估计。

定性地，预滤波器的作用是加重传递到相关器的信号的信噪比（S / N）最高的频率，同时抑制噪声功率。通用的Eckart滤波器提供了相同类型的预滤波，可最大化相关器输出的S / N比。对于低信噪比，ML估计器显示为等效于Eckart预滤波。

1. M. S. Brandstein and H. F. Silverman, “A robust method for speech signal time-delay estimation in reverberant rooms,” in *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, **1997**, pp. 375–378.

混响室内语音信号时延估计的鲁棒方法

在存在多径信号的情况下，传统的时延估计器会出现明显的性能下降。 这限制了它们在混响箱中的应用，尤其是当感兴趣的信号是语音，并且可能无法在延时估计之前估计和补偿信道效应时。

本文详细介绍了一种替代方法，该方法将问题重新构建为相位数据的线性回归，然后通过最小化可靠的统计误差度量来估计时间延迟。 该技术显示较少受到房间混响效果的影响。 在一系列源位置和室内条件下进行了仿真，以说明所提出的时延估计方法相对于传统方法的实用性。

1. X. Xiao, S. Zhao, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, “A learning-based approach to direction of arrival estimation in noisy and reverberant environments,” in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, **2015**, pp. 2814–2818.

噪声和混响环境中, 基于学习的DoA估计方法

本文提出了一种基于学习的方法来处理麦克风阵列输入的到达方向估计（DOA）的任务。

诸如经典最小二乘（LS）方法之类的传统信号处理方法依赖于对信号模型的严格假设以及对到达时间延迟（time delay of arrival , TDOA）的准确估计。它们仅在相对干净的条件下才能很好地工作，但会受到噪音和混响失真的影响。

在本文中，我们提出了一种基于学习的方法，可以从大量模拟的噪声和混响麦克风阵列输入中学习，以进行可靠的DOA估计。具体来说，我们从广义互相关（GCC）向量中提取特征，并使用多层感知器神经网络来学习从此类特征到DOA的非线性映射。基于学习的方法的一个优势是，随着越来越多的训练数据变得可用，DOA估计将变得越来越准确。在模拟数据上的实验结果表明，所提出的基于学习的方法比最新的LS方法产生了更好的结果。与LS方法相比，对会议室中记录的真实数据的测试结果显示出改进的均方根误差（RMSE）。

https://github.com/singaxiong/SignalGraph

1. W. Zhang, P. Samarasinghe, H. Chen, and T. Abhayapala, “Surround by sound: A review of spatial audio recording and reproduction,” *Applied Sciences*, vol. 7, no. 5, p. 532, **2017**.

声音环绕：空间音频记录和再现的回顾

在本文中，将介绍各种用于空间音频的记录和再现技术的系统概述。 尽管双耳录音和渲染的设计类似于人类的双耳听觉系统，并专门为听众的两只耳朵再现声音，但使用大量麦克风和扬声器的声场录音和再现可以复制区域内的声学场景。 本文讨论了这两种根本不同类型的技术。 本文还简要回顾了最近流行的领域多区域复制。 本文最后讨论了该领域的现状和未解决的问题。

1. M. A. Gerzon, “Periphony: With-height sound reproduction,” *Journal of the Audio Engineering Society*, vol. 21, no. 1, pp. 2–10, **1973**.

伴奏：高音再现

可以通过实际的两声道，四声道和九声道系统来录制外音（围绕听众在垂直和水平方向上的声音再现）。

描述了19种不同系统的矩阵参数和麦克风技术，并给出了其他外围系统的设计程序。 讨论了幅度和能量方向分辨率，以及与当前仅限水平系统的兼容性。

1. L. Perotin, R. Serizel, E. Vincent, and A. Guerin, “Crnn-based´ joint azimuth and elevation localization with the ambisonics intensity vector,” in *IWAENC*, 2018.

基于Crnn的联合方位角和仰角定位以及歧义强度矢量

我们提出了基于堆叠卷积和递归神经网络（CRNN）的first-order Ambisonics（FOA）内容的源定位系统。 我们建议将FOA声强矢量用作CRNN的输入，它易于计算并且与到达声音的方向（DoA）紧密相关。 系统会估算方位角和仰角的点源的DoA。 我们对包括混响，噪声以及各种讲话者，麦克风方向在内的配置进行实验评估。 结果表明，与另一个最近的基于CRNN的系统相比，所提出的体系结构和输入允许网络在现实条件下返回准确的位置估计。

1. H. W. Lollmann, C. Evers, A. Schmidt, H. Mellmann, H. Bar-¨ fuss, P. A. Naylor, and W. Kellermann, “The LOCATA challenge data corpus for acoustic source localization and tracking,” in *IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM)*, Sheffield, UK, July **2018**.

声源定位和跟踪算法对于诸如个人助理，智能家居，电话会议系统，助听器或自治系统之类的广泛应用至关重要。为此已经提出了许多算法，但是到目前为止，还没有使用通用的数据库对其进行评估和比较。 IEEE-AASP声源定位和跟踪挑战（LOCATA）提供了新颖，全面的数据语料库，用于对声源定位和跟踪的最新算法进行客观基准测试。数据语料库包括六个任务，范围从具有静态麦克风阵列的单个静态声源的定位到具有移动麦克风阵列的多个移动扬声器的跟踪。它包含现实世界中的多通道录音，这些录音是通过助听器，集成在机器人头中的麦克风，在封闭的声学环境中的平面和球形麦克风阵列以及有关所涉及的阵列和由移动的人代表的声源的位置信息说话者或静态扬声器。

1. A. Perez-L´ opez and J. De Muynke, “Ambisonics directional room´ impulse response as a new convention of the spatially oriented format for acoustics,” in *Audio Engineering Society Convention 144*. Audio Engineering Society, 2018.

Ambisonics定向室的冲激响应是声学空间定向格式的新惯例

室内脉冲响应（RIR）测量是捕获给定空间声学特性的最常用方法之一。 与麦克风阵列一起使用时，RIR固有地包含方向信息。 由于对Ambisonics和虚拟现实音频的兴趣日益增长，新的球形麦克风阵列最近投放市场。 因此，已经公开发布了使用这种阵列测量的多个定向RIR（DRIR）数据库，称为Ambisonics DRIR。 但是，数据库之间没有格式共识。 为了改善互操作性，我们提出了Ambisonics DRIR的交换格式，作为一种新的空间定向声学格式（SOFA）约定。 作为一个用例，一些现有数据库已按照我们的建议进行了转换和发布。

1. C. Knapp and G. Carter, “The generalized correlation method for estimation of time delay,” in *IEEE Transactions on Acoustics, Speech, and Signal Processing, Volume 24 Issue 4*, 1976.

同引用1

1. Y. Huang, J. Benesty, G. Elko, and R. Mersereati, “Real-time passive source localization: a practical linear-correction least-squares approach,” in *IEEE Transactions on Speech and Audio Processing, Volume 9 Issue 8*, **2001**.

实时被动源定位：实用的线性校正最小二乘法

针对加性测量误差模型下的源定位问题，提出了一种线性校正最小二乘估计程序。 该方法可以在具有适度计算复杂性的实时系统中轻松实现，可在不假设噪声分布的先验知识的情况下产生有效的源位置估计器。 审查了现有的其他估计量，包括基于似然，球面交点，球面插值和二次校正最小二乘估计量，并比较了它们的复杂度，估计一致性和针对Cramer-Rao下限的效率。 数值研究表明，所提出的估计器在许多实际情况下的性能更好。

1. R. Schmidt, “Multiple emitter location and signal parameter estimation,” in *IEEE Transactions on Antennas and Propagation*, **1986**.

多发射器位置和信号参数估计

处理在传感器阵列上接收到的信号以确定发射器的位置引起了极大的兴趣，以至于在许多特殊情况的假设下都已进行了处理。一般问题考虑在任意协方差矩阵的噪声/干扰环境中具有任意位置和任意方向特性（增益/相位/极化）的传感器。

该报告首先涉及该问题的多发射器方面，其次涉及解决方案的普遍性。给出了多信号分类（MUSIC）算法的描述，该算法提供1）存在的入射波前数量的渐近无偏估计。 2）到达方向（DOA）（或发射器位置）； 3）入射波形之间的强度和互相关； 4）噪音/干扰强度。包括与基于最大似然（ML）和最大熵（ME）的方法以及常规波束形成的示例和比较。包含了将其用作按时间序列运行的多频估算器的示例。

1. J. DiBiase, H. Silverman, and M. Brandstein, “Robust localization in reverberant rooms,” in *Microphone Arrays, pp. 157-180*, **2001**.

混响室内的鲁棒定位方法

使用麦克风阵列进行讲话者定位作为一种自动跟踪方式被包含在任何通用语音捕获系统作为必要组件, 最近受到了广泛的关注。 几种算法方法可用于多通道数据的语音源定位。 本章总结了当前的领域，并对每种类型的一般优缺点进行了评论。 然后详细介绍了一种新的本地化方法。 通过利用现有方法的关键特征，与当今使用的传统定位技术相比，该新算法在声学条件（尤其是混响效果）方面表现出明显更强的鲁棒性。

1. S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, “Sound event localization and detection of overlapping sources using convolutional recurrent neural networks,” *IEEE Journal of Selected Topics in Signal Processing*, 2018.

使用卷积递归神经网络进行声音事件定位和重叠源检测

在本文中，我们提出了一种卷积递归神经网络，用于在三维（3D）空间中对多个重叠声音事件进行联合声音事件定位和检测（SELD）。

所提出的网络将一系列连续的频谱图时间帧作为输入，并将其并行映射到两个输出。作为第一输出，声音事件检测（SED）作为每个时间帧上的多标签分类任务执行，从而为所有声音事件类别产生时间活动。作为第二个输出，通过使用多输出回归估计每个声音事件类别的到达方向（DOA）的3D笛卡尔坐标来执行定位。所提出的方法能够将多个DOA与各自的声音事件标签相关联，并且相对于时间进一步跟踪该关联。所提出的方法将在每个音频通道上计算的频谱图的相位和幅度分量分别用作特征，从而避免了任何方法和特定于阵列的特征提取。在无回声，混响和真实场景中，对五个具有不同重叠声音事件的Ambisonic和两个圆形阵列格式数据集进行了评估。将该方法与两个SED，三个DOA估计和一个SELD基线进行了比较。结果表明，所提出的方法是通用的，适用于任何阵列结构，对看不见的DOA值，混响和低SNR场景均具有鲁棒性。与最佳基准相比，所提出的方法在整个数据集中估计的DOA数量始终保持较高的召回率。此外，对于大量重叠的声音事件，该召回率明显优于最佳基准方法。

1. J. M. Vera-Diaz, D. Pizarro, and J. Macias-Guarasa, “Towards end-to-end acoustic localization using deep learning: from audio signal to source position coordinates,” in *Sensors*, **2018**.

使用深度学习实现端到端声学定位：从音频信号到源位置坐标

本文提出了一种基于麦克风阵列并基于卷积神经网络（CNN）进行室内声源定位的新方法。据我们所知，所提出的解决方案是首次发表的工作，其中CNN被设计为使用原始音频信号作为输入信息来直接估计声源的三维位置，从而避免使用手工制作的音频特征。鉴于可用的本地化数据数量有限，我们在本文中提出了一种基于两个步骤的培训策略。我们首先使用从近距离语音记录生成的半合成数据训练网络，并在其中模拟从源传播到麦克风阵列的信号中的时间延迟和失真。然后，我们使用少量实际数据微调此网络。我们的实验结果表明，该策略能够产生可显着改善基于\ textit {SRP-PHAT}策略的现有定位方法的网络。此外，我们的实验表明，与其他方法相比，我们的CNN方法具有更好的抵抗说话者性别差异和不同窗口大小的能力。

1. T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, “A study on data augmentation of reverberant speech for robust speech recognition,” in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on. IEEE*, 2017, pp. 5220–5224.

混响语音数据增强技术用于鲁棒语音识别的研究

通过使用多条件训练数据，可以显着提高基于DNN的声学模型的环境鲁棒性。 但是，由于数据收集是一项昂贵的提议，因此对所需条件的模拟是一种经常采用的策略。 在本文中，我们详细介绍了用于远场ASR的数据增强方法。 我们将检查使用模拟房间脉冲响应（RIR）的影响，因为可能难以获得真实的RIR，并且还会增加点源噪声的影响。 我们发现，当添加点源噪声时，可以消除使用模拟RIR与实际RIR之间的性能差距。 进一步，我们表明，训练后的声学模型不仅在远距离通话场景中表现良好，而且在近距离通话场景中提供了更好的结果。 我们在几个LVCSR任务上评估了我们的方法，这些任务可以充分代表两种情况。

1. C. Schissler and D. Manocha, “Gsound: Interactive sound propagation for games,” in *Audio Engineering Society Conference: 41st International Conference: Audio for Games*. Audio Engineering Society, 2011.

我们提出了一种声音传播和渲染系统，用于为类似游戏的场景实时生成逼真的环境声学效果。 该系统使用射线跟踪对在任意反射深度处的听众可见的三角形进行采样。 然后使用基于射线的遮挡查询来验证从每个声源到听众的声音反射和衍射路径。 执行传播路径的帧到帧缓存以提高输出的一致性和准确性。 此外，我们提供了一个灵活的框架，在时间紧迫的情况下，它只占用一小部分CPU周期。 据我们所知，这是第一种可以为当前平台上的游戏生成逼真的声音和听觉化的实用方法。

1. C. Schissler, R. Mehra, and D. Manocha, “High-order diffraction and diffuse reflections for interactive sound propagation in large environments,” *ACM Transactions on Graphics (TOG)*, vol. 33, no. 4, p. 39, 2014.

高阶衍射和漫反射，用于在大型环境中进行交互式声音传播

我们提出了在大型虚拟环境中建模交互式漫反射和高阶衍射的新颖算法。 我们的公式基于基于射线的声音传播，可直接应用于复杂的几何数据集。 我们使用结合了光能传递和路径跟踪技术的增量方法来迭代计算漫反射。 我们还提出了用于与波长相关的简化和可见度图计算的算法，以在运行时加速高阶衍射。 整个系统可以在具有多个声源的大型动态场景中以交互速率生成合理的声音效果。 我们重点介绍了在复杂的室内和室外环境下的性能，并观察到与以前的方法相比性能提高了一个数量级。

1. C. Schissler and D. Manocha, “Interactive sound propagation and rendering for large multi-source scenes,” *ACM Transactions on Graphics (TOG)*, vol. 36, no. 1, p. 2, 2017.

大型多源场景的交互式声音传播和渲染

我们提出了一种在包含许多声源的大型动态环境中以交互速率生成合理的声学效果的方法。 我们的公式将基于听众的后向光线跟踪与声源聚类和混合音频渲染相结合，以处理复杂的场景。 我们提出了一种动态后期混响的新算法，该算法从听众对球形声源执行高阶射线跟踪。 通过聚类远处的声源并考虑相对可见性，我们实现了与声源数量成比例的亚线性缩放。 我们还描述了一种基于混合卷积的音频渲染技术，该技术可以以交互速率处理数十万条声音路径。 我们在多达200个声源的情况下，在许多室内和室外场景中演示了性能。 在实践中，我们的算法可以在多核PC上以交互速率计算超过50个反射阶数，并且与以前的几何声音传播算法相比，观察到的速度提高了5倍。

1. Z. Tang, L. Chen, B. Wu, D. Yu, and D. Manocha, “Improving reverberant speech training using diffuse acoustic simulation,” *arXiv preprint arXiv:1907.03988*, 2019.

使用扩散声学仿真改善混响语音训练

我们提出了一种高效且逼真的几何声学仿真方法，用于生成和增强与语音相关的机器学习任务中的训练数据。 我们基于物理的声学模拟方法能够在复杂的声学环境中对声音的遮挡，镜面反射和漫反射建模，而经典的图像方法只能在简单的房间设置中对镜面反射建模。 我们显示出，通过使用我们的综合训练数据，相同的神经网络在远场语音识别的真实测试集上获得了1.58％的显着性能提升，而在21％的关键词发现上却获得了显着的性能提升，而无需使用真实的冲激响应进行微调。

1. V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: an asr corpus based on public domain audio books,” in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 5206–5210.

Librispeech：基于公共领域有声读物的asr语料库

本文介绍了一种新的阅读英语语音语料库，适用于训练和评估语音识别系统。 LibriSpeech语料库来自LibriVox项目的有声读物，包含以16 kHz采样的1000个小时的语音。 我们已经免费提供了该语料库，以及单独准备的语言模型培训数据和预先构建的语言模型。 我们显示，在LibriSpeech上训练的声学模型比在WSJ本身上训练的模型在《华尔街日报》（WSJ）测试集中提供的错误率更低。 我们还发布了Kaldi脚本，使构建这些系统变得容易。

1. C. Valentini-Botinhao *et al.*, “Noisy speech database for training speech enhancement algorithms and tts models,” **2017**.

带噪语音数据库，用于训练语音增强算法和TTS模型

干净和带噪并行的语音数据库。该数据库旨在训练和测试以48kHz运行的语音增强方法。可以在与数据库相关的论文中找到更详细的描述。

对于28个说话者数据集，可以在以下论文找到详细信息：C. Valentini-Botinhao，X.Wang，S.Takaki和J. Yamagishi，“Speech Enhancement for a Noise-Robust Text-to-Speech Synthesis System using Deep Recurrent Neural Networks ”，Interspeech, 2016。

对于56位演讲者数据集：C. Valentini-Botinhao，X. Wang，S. Takaki和J. Yamagishi，“Investigating RNN-based speech enhancement methods for noise-robust Text-to-Speech”， Pro S.SSW, 2016.

用于创建嘈杂语音的一些噪声是从“需求”数据库获得的，可在此处找到：<http://parole.loria.fr/DEMAND/>

语音数据库是从CSTR VCTK语料库获得的，可在此处获取：<https://doi.org/10.7488/ds/1994>

用于创建此数据集的语音类和bable噪声文件位于以下位置：<http://homepages.inf.ed.ac.uk/cvbotinh/se/noises/>

1. S. Yin, C. Liu, Z. Zhang, Y. Lin, D. Wang, J. Tejedor, T. F. Zheng, and Y. Li, “Noisy training for deep neural networks in speech recognition,” *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, p. 2, **2015**.

用于语音识别的深度神经网络的噪声训练

语音识别中的深度神经网络的噪声训练深度神经网络（DNN）在语音识别方面取得了显著成功，部分归因于DNN模型在学习复杂的语音信号模式方面的灵活性。 但是，这种灵活性可能会导致严重的过度装配，从而在不利的声学条件（例如具有高环境噪声的声学条件）下导致性能下降。 我们提出了一种嘈杂的训练方法来解决此问题：通过有意和随机地将中等噪声注入训练数据中，可以学习更通用的DNN模型。 尽管这种“噪声注入”技术已经为神经计算社区所熟知，但尚未使用涉及高度复杂目标函数的DNN进行研究。 本文提出的实验证实，噪声训练方法对于DNN模型效果很好，并且可以为基于DNN的语音识别提供实质性的性能改进。

1. J. Kominek and A. W. Black, “The cmu arctic speech databases,” in *Fifth ISCA workshop on speech synthesis*, 2004.

北极语音数据库

CMU\_ARCTIC数据库是在卡内基梅隆大学语言技术学院构建的，具有语音平衡的美国英语单口语数据库，设计用于单元选择语音合成研究。

卡内基梅隆大学语言技术学院技术报告CMU-LTI-03-177可以提供有关数据库的结构和内容以及记录环境等的详细报告，也可以在此处获得。

该数据库包括从古腾堡计划（Project Gutenberg）的非版权文本中精心挑选的大约1150种语音。这些数据库包括美国英语的男性（bdl）和女性（slt）说话者（均具有出色的语音才能）以及其他有口音的说话者。

可从cmuarctic.data获得1132句子提示列表

分布包括16KHz波形和同步EGG信号。 CMU Sphinx使用基于FestVox的标记脚本进行了完全恐怖的标记。完整的可运行的节日声音包括在数据库发行版中，虽然可以通过改进标签等来制作更好的声音，但这些示例仍是示例。