文章编号:1671-4598(2023)07-0184-08 DOI:10.16526/j.cnki.11—4762/tp.2023.07.027 中图分类号:TP3;TH701 文献标识码:A

基于 Ca-GAN 增强的机坪管制指令识别方法研究

诸葛晶昌[1],胡宽博[1],杨新宇[1],吴 军[2]
(1. 中国民航大学 电子信息与自动化学院,天津 300300;2. 中国民航大学 航空工程学院,天津 300300)

摘要:我国枢纽机场长期处于繁忙状态,高负荷带来信息交互失真的风险,语音识别技术可用于辅助决策,然而管制语音特殊性及样本量局限性使传统深度学习技术难以直接应用于机坪管制领域;针对这一问题,提出了一种基于小样本学习的语音识别方法;首先提出数据增强方法,通过结合先验领域知识,构建基于数据生成策略组的生成对抗网络来增强声学模型识别能力来进一步提升模型效果;然后通过重构声学模型部分结构和参数;最后通过迁移学习方法将通用语音库中的声学建模特征应用到机坪管制语音指令的识别中;实验结果表明,该方法将字错率减少至 6.14%;该研究可应用于机场高级地面活动引导及控制系统中机坪管制语音指令的检测和识别,助力现代机场高质量运行。

关键词:智能交通;指令识别;语音识别;机坪管制;生成对抗网络

Research on Enhanced Apron Control Command Recognition Method Based on Ca-GAN

ZHUGE Jingchang[1], HU Kuanbo[1], YANG Xinyu[1], WU Jun[2]
(1. School of Electronic Information and Automation Civil Aviation University of China, Tianjin 300300, China;
2. School of Aeronautical Enginee ring, Civil Aviation University of China, Tianjin 300300, China)

Abstract: China's hub airports have been busy for a long time, and overload operation status brings the risk of distorted information interaction. Speech recognition technology is used to assist decision-making, however, the special characteristics of control speech and sample size limitations make it difficult to directly apply traditional deep learning techniques to the ramp control field. Aimed at this problem, a speech recognition method based on small sample learning is proposed. Firstly, a data enhancement method is proposed to further improve the model effect by combining a priori domain knowledge and constructing a generative adversarial network based on a data generation strategy group, and enhance the acoustic model recognition ability; Then, the partial structures and parameters of the acoustic model are reconstructed; Finally, the acoustic modeling features from a general speech library are applied to the recognition of ramp control speech commands by a migration learning method. The experimental results show that the method reduces the word error rate to 6.14%. This research is applied to the detection and recognition of the ramp control speech commands in advanced ground activity guidance and control systems in airports, which can help modern airports operate with high quality.

Keywords: intelligent transportation; instruction identification; speech recognition; apron control; generative adversarial networks

0 引言

中国民航不断发展,航班数量的增加,导致机场愈加繁忙,特别是国内枢纽机场大多已经接近最大容量限制。国际机场理事会预测,到 2040 年中国民航的航班数量将占世界民航的 16.1%,可以预见,我国大型机场特别是枢纽机场将长期保持高负荷的运行状态。机坪管制从人员、车辆、设备、信息、环境等方面开展飞机地面保障和运行调度工作,目前机场飞机地面运行普遍采用管制员人工语音调度的方式进行,机场飞行区运行负荷的不断增加,为管制人员带来巨大压力,导致管制员人为因素造成的异常事件呈上升趋势。因此,管制员语音指令的准确识别已经成为实现机场地面运行辅助决策、资源调配、预测预警的重要环节,是提升场面运行指挥技术保障能力的有效手段,有利于深化“平安民航”建设,为机场飞行区场面运行提供安全保障。

自动语音识别技术(ASR, automatic speech recognition)已经在空中交通管制(ATC, air traffic control)领域应用,如在空中交通管制领域使用深度神经网络(DNN, deep neural network)与双向长短时记忆网络(BiLSTM, bi-directional long short-term memory)进行语音识别[1],将自动语音识别技术应用到空中交通管制领域构建的封闭式跑道运行预防装置[2]。深度学习方法在机场运行流程中的应

收稿日期:2022-11-04; **修回日期**:2022-12-08。
基金项目:国家自然科学基金(52005500);中央高校基本科研业务费中国民航大学专项(3122019047)。
作者简介:诸葛晶昌(1981-),男,天津人,博士,副教授,主要从事航空器地面运行支持方向的研究。
引用格式:诸葛晶昌,胡宽博,杨新宇,等.基于 Ca-GAN 增强的机坪管制指令识别方法研究[J].计算机测量与控制,2023,31(7):184-191,198.

用也是大势所趋[3]，既提升了机场运行效率，更是为机场安全提供了保障[4]。卷积神经网络（CNN，convolutional neural networks）在语音识别领域中的应用经典案例为 CLDNN（CLDNN，convolutional，long short-term memory，fully connected deep neural networks）[5]，其通过将 CNN、LSTM、DNN 连接在一起，通过三者互补性，以提升模型识别性能。科大讯飞公司也提出了一种新式 CNN 结构，即深度全序列卷积神经网络（DFCNN，deep fully convolutional neural network），直接将音频转化为图像进行处理，在保留音频在时频域的信息方面表现出色。同时与 CTC（connectionist temporal classification）可以很好地结合[6]。语言模型方面，A Vaswani 等提出了基于纯注意力机制的 Transformer 模型，并论证了其在自然语言处理和计算机视觉领域的优越性，且受到领域内学者的一致认可。最初将 Transformer 模型应用到语音领域的 Dong[7] 等人也提出 Speech-Transformer 模型，这种 Sequence-to-Sequence 模型正好被用于解决语音识别中的分类问题。同时为了解决训练中数据量不足而引发的问题，赵凯琳[8]等，张一珂[9]等都提出了数据增强的策略。另一方面，生成对抗网络（GAN，generative adversarial networks）[10] 的出现使得进行无监督的数据增强策略更为可行，但原始 GAN 存在部分缺陷，如两元化的“极小极大博弈”[11]，在缺少损失函数的约束下，难以持续进行，容易陷入最稳状态，无法生成新的样本。随着条件生成对抗网络（CGAN，conditional generative adversarial nets）[12]及各种变种 GAN 网络如拉普拉斯生成对抗网络（LAPGAN，laplacian pyramid of adversarial networks）[13]等的提出，在各种约束下，GAN 的部分缺陷逐渐被弥补。

由于管制员机坪管制语音指令区别于标准普通话的特殊性，现有的语音识别模式方案无法发挥最佳效果，限制了其在机坪管制领域的应用。机坪管制指令特点在于：1）指令简短规范，信息密度大；2）受通话环境影响，和周围噪声干扰、通信干扰和管制双方通话习惯等有关。因此，适用于机坪管制员语音的识别方法的研究显得至关重要。

本文参考李响 [14] 等提出了基于生成联合深度卷积神经网络（G-DFCNN，generator-deep convolutional neural network）结构的语音识别方法，实现了机坪管制指令的准确识别，并依据小样本学习方法增强了识别模型的准确性和鲁棒性。本方案提出适用于管制指令音频识别的数据增强方法，构建基于增强策略组的级联生成对抗网络来生成虚假样本参与训练，通过改进 DFCNN 网络结构用以提升声学特征的匹配度，使用 Transformer 模型搭建语言模型，以弥补语音识别中最常用的 N-Gram 语言模型只能关注连续词的缺点。最后通过迁移学习方法实现对声学建模单元的高效利用以提升语音识别的准确性。

1　数据增强策略组

考虑到机坪管制指令的小样本集问题，本文通过数据增强方法实现对小样本的扩充。在图像数据增强领域，常用的数据增强方法有尺度变换和像素变换或是直方图均衡化以及调整白平衡等。而在语音方面则可以将音频的语谱图当作图像来对待，语谱图的两个维度分别代表音频的时间和频率，语谱图中的颜色深浅则代表语音的强弱，正对应了图片尺度特征和灰度，因此可以查询指定时间和频率的能量分布。本文数据增强方法包括利用如 SamplePairing 和 Mix-up 等批次化处理的增强技术获得新的样本，这些模板化生成策略处理得到的新音频中声学特征信息都被部分保留，随着变换尺度的增加，部分新特征也会逐渐失真。或通过有规律的破坏完整的信息链，迫使卷积网络学习或猜测更深层次的内容。如神经网络中加入 Dropout 的操作，或自然语言处理领域中的掩码思想。另一方面，通过生成对抗网络来生成虚假样本参与训练也是一种有效的手段。参考 LAPGAN 和深度语音增强生成对抗网络（DSEGAN，deep speech enhancement GAN）的链式生成器思想，我们构建了基于混合增强策略组的 GAN 模型，通过将音频的各个尺度特征当作图像中的残差特征进行提取和生成，最后通过与原音频进行级联以生成新数据。

生成网络结构如图 1 所示。通过 3 种增强方法生成训练所需的虚假样本参与训练。

1.1　预处理

1.1.1　分帧加窗

机坪管制指令的识别可以看作序列到序列的分类问题，音频单采样点所蕴含的信息密度远低于拼音或音素所蕴含的信息密度，为增大音频单帧的信息密度，更好地匹配标签，同时应对采样中的随机信号的干扰，采取分帧加窗是将时变语音信号处理成短时平稳信号，用于之后的特征提取。由于读取到的语音指令语音信号表示为时域排序的离散一维数组，而单采样点所蕴含的特征信息不足，且包含随机信号特征，需要联合前后点经由变换得到平稳信号以摒除随机信号特征的影响，同时能使得每一帧信息密度增大。因此每帧信息一般维持 10～30 ms 内，以保持语音信号的特性基本不变。对于一段语音进行加窗操作时设置的参

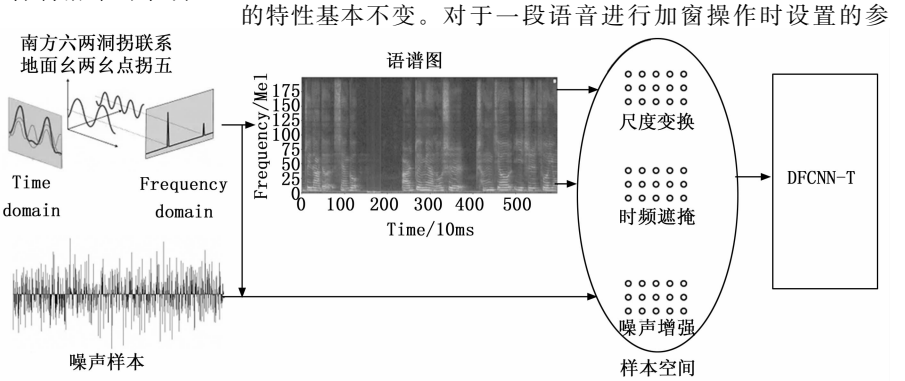

图 1　数据增强策略组结构

数包括帧长、帧移以及重叠时间。因此本文设置帧长为 25 ms，每次时移 10 ms，其中会有 15 ms 的重叠区域，以这种方式来防止帧与帧之间丢失重要的信息。同时为了减少信息的泄露，所以在加窗时选择海明窗：

$$W(n,\alpha) = (1-\alpha) - \alpha \cdot \cos(2\pi \cdot \frac{n}{N-1}) \quad (1)$$

其中：α 一般取 0.46，N 为窗口大小，$0 \leqslant n \leqslant N-1$。

1.1.2 MFCC 特征

为了全面体现管制指令语音信号特征，除了时域特征外本文选取了梅尔倒谱系数（mel-frequency cepstral coefficients，MFCC）作为频域特征，MFCC 是在分帧加窗后经由 FFT 变换获得其在频域上的特征，再对各帧频谱取模平方得到的。

$$X_a(k) = \sum_{N=0}^{N-1} x(n) \cdot e^{-j2\pi nk/n}, 0 \leqslant k \leqslant N \quad (2)$$

其中：$x(n)$ 为输入信号，N 表示傅里叶变换点数。

经过梅尔带通滤波器组对频谱进行平滑化，并消除谐波的作用，突显原先语音的共振峰。由于 MFCC 特征随维度增加到一定程度后，蕴含的信息将会变少，本文中实验部分提取 32 维特征向量，采用 64 个滤波器组成梅尔带通滤波器组。最后经由对数运算得到当前帧的对数能量。

1.2 增强策略组

1.2.1 噪声增强

噪声增强有多种形式，如加性噪声即在增强过的语音信号上添加一个或多个随机噪声段，增强随机系数倍数后经由加性或乘性方法加入原噪声，用以提升语音识别方法的鲁棒性。

加性噪声是为了减少人声信息在音频中所占比，同样的，可以通过加强人声的降噪处理进一步推动神经网络识别噪声、学习到人声信息，增加系统鲁棒性。混响增强方法基于 ISM（image source method）方法，通过模拟封闭空间中的各个反射面的特性，将声源视作光源，随着在空间中各个反射面的扩散与折射，呈现出真实情景中封闭空间中的混响效果。

1.2.2 时域增强

1）时移变换：

仅沿着时间轴随机移动语音信号，不改变信号的其他属性。对应语谱图的横向平移。在实际的管制语音音频中，包含语音特征的音频序列在起始和结束时可能产生一定范围的空白，因此采用时移增强，将信息段沿着时间轴按一定比例滚动，以模拟真实场景中的起始、结尾环境并创造新的音频样本数据，增加方法鲁棒性。

2）音速变换：

用以模拟不同管制员语速的快慢，在实际机场塔台管制指令中，语速比日常对话要快上很多，同时语句本身包含信息量大，但不同管制员语速并不能完全相同，通过音速的变换可以模拟其中的影响。在卷积网络中通过多层卷积后的跨越长时的声学特征也可以汇聚到一张图层中，因此，该方法作为策略组中时间维度的展缩工具。

3）音高变换：

改变音频升降幅度，通模拟通信过程中音频幅值各个情况下的高低变化，如远近场的变化、输入输出设备音量变化等，使神经网络忽略音高所带来的音频特征的差异。

音高变换后语谱图从高频区到低频区语谱强度都有所提升。虽然不会改变音频中有效信息的占比，但可以一定程度上消除声学特征匹配过程中音高带来的差异。

1.2.3 语谱遮掩

在时间维度随机抽去一段数据，形成类似于 Dropout 的操作，隐藏的内容不能超过一定阈值，否则样本与标签的对应将会冲突，卷积神经网络通过猜测可以得出隐含信息。频域的遮掩则类似于去噪，根据研究表明，人声通常集中在一定范围频段，虽然男声、女声、齿音、鼻音等分布范围都不相同，但其中信息基本包含在一定区间。如女性发声中 1.6～3.6 kHz 影响音色的质量。在频域遮掩的过程中，如果将连续的有效信息遮掩掉，则会导致识别效果下降。本文将导致识别效果明显降低的批次舍弃，同时通过取 loss 趋于平缓后的识别结果均值作为识别结果以降低误差。由于卷积神经网络中图像的输入为两个有效维度即图片的尺寸变换应保留自身特征，因此图像数据增强方法中部分多维特征变换方法可能会大幅减少输入语音的固有特征所占比例，破坏已经提取的特征信息链，使其产生失真，无法提升识别性能。因此频谱遮掩方法中需要设置阈值来遏制连续域内的多次遮掩。

$$W_1 = \sum_{i=0}^{n} l_i \quad 0 < W_1 \leqslant t \cdot \varepsilon \quad (3)$$

$$W_2 = \sum_{i=0}^{n} l_i \quad 0 < W_2 \leqslant t \cdot \varepsilon \quad (4)$$

其中：W_1、W_2 分别为时频域遮掩总量，n 为遮掩数量，l_i 为遮掩范围，t 为样本时间长度，f 为样本特征频率范围，ε 为遮掩系数。实验结果表明阈值为时频域尺度的 20%以内时效果最佳。遮掩效果如图 2～3 所示。

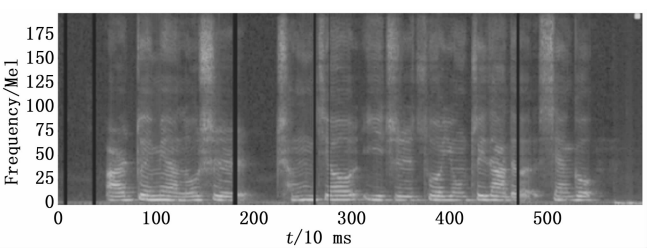

图 2 时域遮掩后的语音信号语谱图

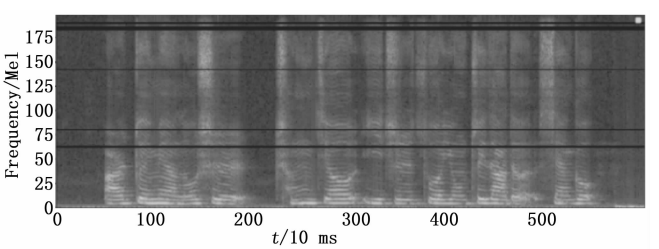

图 3 频域遮掩后的语音信号语谱图

2　Ca-GAN

GAN 方法的提出，为小样本学习领域带来了新的思路。通过生成器域鉴别器的二元博弈，我们可以得到能够产出大量虚假样本的生成器。在一些难以获得大量标注数据的领域起到了至关重要的作用。而如何判断数据的真实分布与虚假分布则是 GAN 训练中需要解决的问题。

2.1　模型结构

Quan 等[15]证明了通过在 GAN 原始生成器后进一步增加附加的生成器能够获得更好的图像构建效果，而 Huy Phan 等[16]则在语音增强领域提出了 DSEGAN，通过使用多级生成器增强映射 $G=G_1\rightarrow G_2\rightarrow\cdots\rightarrow G_N$ 来实现对声学特征的重建。

LAPGAN 也采取了类似的思想，LAPGAN 最先用于计算机视觉领域，由 Facebook 等人提出，通过金字塔式的多层下采样获取图像的各级特征，并训练鉴别器进行判断。区别于原始 GAN 的两元极端思想，LAPGAN 引入了条件信息，只通过各级残差特征与原始图像的结合创造出新的样本。

生成器采样过程为：

$$\tilde{I}_k=\tilde{u}(\tilde{I}_{k+1})+\tilde{h}_k \tag{5}$$

$$\tilde{h}_k=G_k(z_k,\tilde{u}(\tilde{I}_{k+1}))=\tilde{I}_k-\tilde{u}(I_{k+1}) \tag{6}$$

其中：$\tilde{I}_k$ 为第 k 次下采样的图像，由 $k+1$ 层生成，$\tilde{u}()$ 代表下采样操作，$\tilde{h}$ 代表拉普拉斯金字塔 k 层系数（由相邻层的差值构造）。

本文采用类似结构来构建所需 Generator，相关结构如图 4 所示。

Generator 生成样本流程如图 4 所示。本文设置三类金字塔层，分别为加性噪声层、时频特征层和频谱特征层，将以增强策略组的形式展示。不同于 LAPGAN 方法的对于图像的分解式特征金字塔，Generator 通过将原始样本通过时频和频谱等增强变换方法逐层构建自己的特征金字塔。

其中，单层 Generator 结构如图 5 所示。

相对的，鉴别器结构如图 5（b）所示，参考 DCGAN 和 SEGAN 中的形式，生成器和鉴别器由多个卷积块组成，每个卷积块包含二维卷积/反卷积、batchnormization 层和相应的激活函数。在生成器中为了保护负数域中的数据特征不失真而采用了 tan$h$ 激活函数。而在鉴别器中则采用 Leaky _ relu 函数。

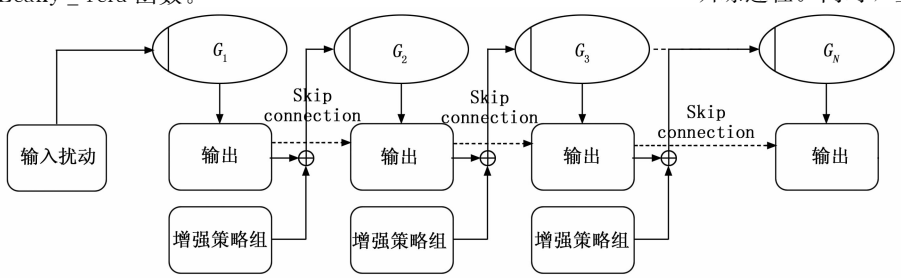

图 4　Ca-GAN 中的生成器结构

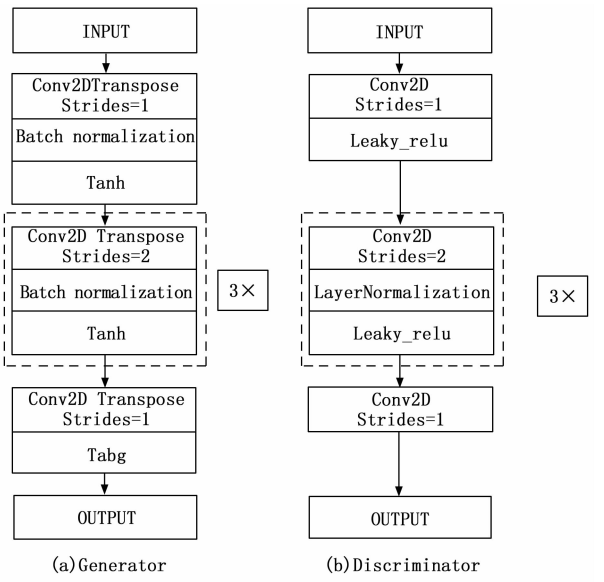

图 5　单个生成器和鉴别器的基础结构

2.2　GAN 训练

为了迫使级联的生成器链能够学习到真实数据分布之间的特征，鉴别器的任务是对真实分布与虚假分布的差异。然而在实际的训练过程中，经常会出现各种问题，如鉴别器识别的效果太差，难以约束；或是鉴别器的效果太好导致生成器的更新停滞等。为了减少这些问题的出现，评价的标准采用 Wasserstein 距离[17]，相较于 Jensen-Shannon 散度[18]和 Kullback-Leibler 散度[19]因此通过二元博弈的公式如下：

$$\min_G\max_{D\in L}\underset{x\sim P_r}{E}[D(x)]-\underset{z\sim P_g}{E}[G(z)] \tag{7}$$

而在 WGAN 的基础上，为了得到更稳定的 loss 输出[20]，将传统的权重裁剪方法变换为梯度惩罚的方法[21]，即增加一个正则项作为约束：

$$L_{GP}=\underset{x\sim P_r}{E}[f(x)]-\underset{\tilde{x}\sim P_g}{E}[f(\tilde{x})]+k\underset{\tilde{x}\sim P_{\tilde{x}}}{E}[(\|\nabla f(\tilde{x})\|_2-1)^2] \tag{8}$$

因此二元博弈公式也可以表示为：

$$\min_G\max_{D}\underset{G(z)\sim P_g}{E}[D(G(z))]-\underset{x\sim P_r}{E}[D(x)]-k\underset{x\sim P_{\tilde{x}}}{E}[(\|\nabla f(\tilde{x})\|^p] \tag{9}$$

通过 WGAN-GP[22]的训练方法，可以实现更加稳定的训练过程。同时，生成器的质量也会有所提升。

3　基线模型设计

3.1　声学模型

在机坪管制语音指令的识别中，本文选用深度全序列卷积神经网络（DFCNN，deep fully convolutional neural network）来实现声学特征处理及训练，利用其在时间和空间上的平移不变性卷积来克服语音信号本身的多样性。

卷积神经网络的平移不变性和卷积

采样过程非常适用于语音识别的研究中。在卷积神经网络中，卷积层是利用多个卷积核滤波器对原始的图像进行卷积操作来提取多个抽象特征，而在语音识别过程中通过将语音转换为图像的形式而进行计算。池化层对卷积层进行池化处理，使提取的特征更加紧凑并减少神经元个数。DFCNN 的特点在于直接将音频的语谱图作为神经网络的输入，相比于其他网络模型，这一特点保留了更多的音频特征信息。传统 DFCNN 结构如图 6 所示。

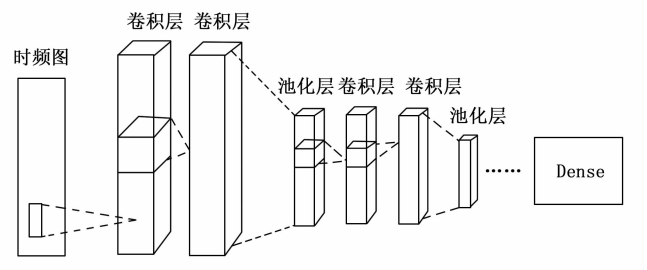

图 6 经典 DFCNN 模型结构

本文通过改进 DFCNN 模型来实现对音频信息的提取和处理。由于音频被转化为语谱图的形式，应用于图像数据增强的方法在音频处理上更便于理解和学习。其通过大量的卷积层和池化层提取了音频在时间和频率两个维度的特征，音频在经过提取后的数据不仅能够真实地表达当前帧特征信息，而且在相当长的时间维度的上的相关性也可以轻易地体现，模拟了循环递归网络的一部分特性，改进后结构如图 7 所示。

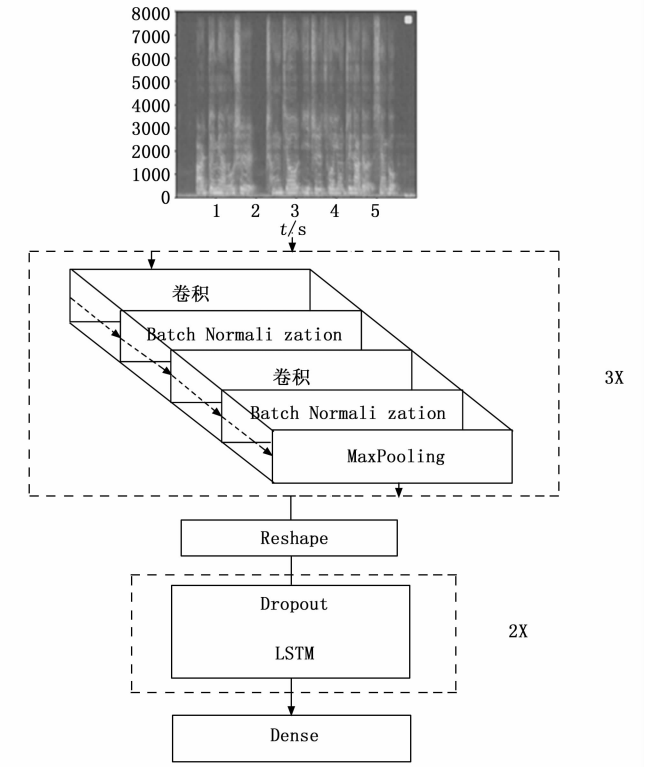

图 7 改进 DFCNN 模型结构

本文通过改进其网络结构，以追求最优效果。通过调整其中神经网络的层数，来减少参与运算的参数，减小运算负荷，增加 Dropout 层防止过拟合，采用批标准化（BN，batch normalization）层增加网络泛化能力。同时增加 LSTM 层对扁平化低维数据进行运算提取音频中的时域信息，从而得到更好的识别效果。

3.2 语言模型

在机坪管制指令识别中，由于指令本身具有字义固定、无混淆发音等特点，语言模型并不能发挥太多作用，但在迁移学习方法中，由于引入了大量同音字、多音字等容易混淆的字符，如果缺少语言模型，将会导致语音识别准确率大幅下降。因此本文设置语言模型，根据声学模型的结果给出概率最大的汉字序列，以实现声学序列到汉字序列的分类。N-gram 模型是语音识别中常用的语言模型，但其仅关注前一个字符而引入的有限的局部文本信息，很难有效地发现孤立的识别错误，如同音字替换错误。

而本文采用的 Transformer 语言模型不仅能通过编码学习到顺序信息，同时也基于自注意力机制寻找输入输出序列的最优匹配。Transformer 模型通过注意力机制、编码解码、残差前馈网络和线性化等特点解决了传统神经网络算法中的缺陷，如根据卷积神经网络思想，结合多头注意力机制，实现了并行的运算，加快了运算的速度。Transformer 方法实际是正是为了理解输入和输出序列之间对应的关系，包含两个主要模块：编码模块和解码模块，通过编码器对时间序列进行编码处理，Transformer 结合编码器的当前的输出和上一时刻的输出来生成下一时间步长的输出，通过这一流程可以出色地表达出序列的时域相关性，从而解决迁移学习所带来的语义问题。

3.2.1 编码器层

编码器层分为 6 层，由 6 个编码单元组成，但其相互之间不会共享权值。每个编码单元包含一个多头注意力通道和前馈通道，该前馈通道包含矩阵线性变化 Linear 层和 Relu 非线性激活处理，每一个子层之后都会接一个残差连接和归一化层。其中，残差连接层避免了梯度消失的问题，而归一化层通常采用 BN 层。BN 的作用在于对网络层中每一小批数据进行归一化处理，防止多层前向计算后的数据偏差过大，造成梯度方面出现问题。

在多头自注意力层、求和与归一化层、前馈神经网络这 3 个不同层的结合下，最终得到编码器的输出。

输入前需要进行位置编码是由于 Transformer 不包含递归和卷积，因此序列的顺序信息无法得到利用，但通过位置编码字符向量嵌入和字符位置向量嵌入可以实现将位置信息作为输入传输到网络中。在编码器和的解码器堆栈底部需要嵌入位置编码，本文采用不同频率的正弦和余弦函数来进行编码，表示为：

$$PE(pos,2i)=\sin\left(\frac{pos}{10000^{2i/d_{model}}}\right) \tag{10}$$

$$PE(pos,2i+1)=\cos\left(\frac{pos}{10000^{2i/d_{model}}}\right) \tag{11}$$

其中：*pos* 表示位置，$2i$ 代表维数。位置编码中不同维度对应着不同的正弦信号。

注意力函数一般分为乘性和加性两类，虽然乘性函数和加性函数理论上复杂度相同，但实际应用中乘性函数一般计算速度更快一些，空间利用效率更高。缩放点积注意力机制如图 8 所示。

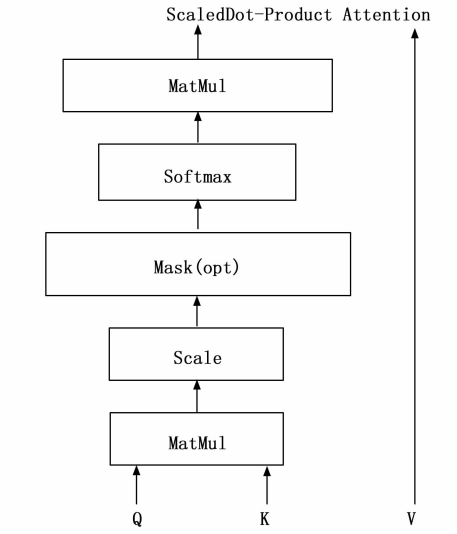

图 8　缩放点乘注意力机制

多头注意力机制并非直接将 Q（qurey）、K（key）、V（value）输入网络，而是通过多个不同的线性变换，将 Q、K、V 进行投影，然后将最终得出的注意力结果进行拼接，这种操作能使 Transformer 模型在不同表示子空间中的不同位置共同关注信息。

3.2.2　解码器层

解码器层是由 6 个解码器组成，与编码器的结构类似，但比编码器多了一个掩盖多头自注意力层。这个层包括了第一层掩盖多头自注意力层和第二层多头自注意力层。Transformer 模型为自回归模型，在预测过程中，查询 Q 自于上一层解码器输出，而键 K 和值 V 自于编码器，编码器可以并行计算，解码器需要分布出结果。通过最终的线性变换，可以得到最高概率的汉字字符索引，从而得到输出。

4　实验结果与分析

4.1　实验平台

本文是在 Ubuntu18.04 系统下，基于 Tensorflow 2.X、Keras 2.X 框架下完成的。

硬件环境为：CPU i7-10700 八核处理器、GPU RTX3090、64 G 内存。

4.2　数据集

1）开源语音数据集，包括 Thchs30，Aishell-1 和 Google Speech Command，其中 Thchs30 和 Aishell-1 为中文长句数据集，Google Speech Command 为英文短语型数据集。清华大学发布的 Thchs30 包含超过 10 000 个超过 30 小时的语音文件，这些文件内容主要由文章和诗歌组成。AISHELL-1 中文语音数据集，包含约 178 小时的开源数据。Google Speech Command 由 TensorFlow and AIY 发布。它包含 65 000 个简短的有声句子。每个段包含一个语音命令。总共有大约 30 种不同的语音命令。

2）民航专业数据集：

该语料库由空管专业人员录制的语音与裁剪标注后的实际管制员语音组成，包含机场管制员与各个航司飞行员通话的内容，并包含已标注的对应文本序列和音素信息。

本数据库的建立参考《空中交通无线电通话用语》和《CCAR93-R5 民用航空空中交通管理规则》根据规范要求，发音速度保持适中，在发送需要被记录的信息时会适当降低语速，单词发音清楚、音量平稳、正常语调，采样频率 16 000 Hz，采样大小为 16 bits，共 1 200 条，此数据集分为纯中文和中英混合两部分，划分为训练集、验证集、测试集三部分，比例为 0.8，0.1，0.1。

区别于普通话语音数据库，管制指令语音数据库中关键词重复度相对较高，仅有约 200 个声学建模单元，其规模远小于普通话语音数据库中的声学建模单元，但相对的字义密度大，因此受噪声干扰明显。因此在数据增强方法下语音识别效果更能得到较好的表现。

用于增强模型验证的离线噪声数据集由三方面组成：

1）实际空管塔台及驾驶舱语音中捕捉。在实际管制通话过程中，部分塔台或驾驶舱由于工作环境的影响，会出现不同程度的噪音干扰正常的交流通话。

2）人为制造通信系统中常见的噪声如高斯白噪声、均匀白噪声、随机噪声等。

3）来源于公开的噪声库包括来自 Google Speech Command 的噪声和来自 noiseX-92 的噪声。Google Speech Command 中包含 6 种噪声，如“白噪声”、“运动自行车”等。noiseX－92 来自信号处理信息库（SPIB），包含 15 种噪声，例如“粉红噪声”，“工厂地板噪声 1”，“军用车辆噪声”等。

噪音将基于指定的音频文件，响度将均衡。将噪声应用于由超参数 α［0.00，0.10］控制的不同尺度的干净数据，以模拟 ASR 实验中信噪比水平［5 dB，25 dB］的范围。

此外，混响实验参数如下：在产生混响时，我们参考了图像源（ISM，image source method）方法。在基础的设计中，在适当的封闭空间（例如［6，6，3］）被作为基本实验环境，默认为长方体。声源和麦克风的坐标随机出现在虚拟房屋的中间区域，例如声源（例如［4，4，1.5］）、麦克风组（例如［2，2，1］、［2，2.1，1］）。声源与麦克风在水平方向上的坐标距离大多为 2 米左右。在生成过程中，原始音频以 16 000 Hz 采样，最大反射次数为最大值的 1/3，墙壁材质使用默认材质“hard _ surface”。此时，混响效果明显，音频的主观声音感知有所变化。room impulse response（RIR）[23] 通过 pyroomacoustics[24] 实现。

4.3 经典数据增强

本文实验评价指标选用语音识别中常用的字错率(WER, word error rate),即需要替换 S、删除 D 或插入 I 的字符数除以标签集对应的词序列的总个数。

$$WER = \frac{S+D+I}{N} \cdot 100\% \tag{12}$$

为验证语音数据增强的可行性与有效性,将原始数据集与数据增强处理后的数据集进行对比分析。结果如表 1 所示。

表 1 数据增强下的识别结果

数据增强	模型结构	WER/%
无增强	DFCNN *	10.61
噪声增强	DFCNN *	9.29
时移增强	DFCNN *	9.73
音速增强	DFCNN *	10.47
音高增强	DFCNN *	10.76
时域遮掩	DFCNN *	9.29
频域遮掩	DFCNN *	6.93

从表 1 可以看出,数据增强带来了正确率的提升,但部分尺度变换类的处理方法如时域的调整和时域的遮掩并没有增加新的特征,因此带来的提升并不大,但相对的,另一部分则相当于有监督地创造了新特征,扩充了样本容量,因此效果明显。同时数据增强策略组中的各方法的混合施加,也为识别结果带来了提升。其中时域增强为音速、音高和时移增强效果随机混合施加于样本特征后得到的最优识别效果,语谱遮掩为时域、频域和时频域 3 种遮掩中的最优识别效果。

4.4 Ca-GAN

结合 DSEGAN 思想,通过噪声增强、时域增强、频谱遮掩等增强策略组构建级联生成对抗网络利进行数据生成,放入鉴别网络中进行识别来进行二元博弈训练,通过 Wasserstein 距离评估真实分布与虚假分布之间的差异。结果如表 2 所示。

表 2 策略组数据增强结果

数据增强	模型结构	WER/%
无增强	DFCNN *	10.61
SEGAN	DFCNN *	7.83
DCGAN	DFCNN *	8.23
Ca-GAN	DFCNN *	6.14

如表 2 结果所示,通过 Ca-GAN 生成数据后,相较于纯净的基线模型,基于 Ca-GAN 的增强策略明显降低了字错率且效果优于单层生成器的 DCGAN 与基于二维输入的 SEGAN。同时,相较于纯粹的模板化的数据增强方式,基于 GAN 的生成方式效果更好。

4.5 基线模型

经多次试验,调整出最佳模型参数得出最优识别结果为字错率 10.61%。输入三维数组,其中第二维为 mfcc 特征维度,取 32 维,采用 ctc-loss 作为估计依据,以 Adam 优化算法创建优化器,最后经由 softmax 函数进行归一和评分。在调整好学习率、batch _ size、初始化函数、正则化和 Dropout 等参数后得到当前最优模型。试验采用十折交叉验证,即将数据分为十份,各份依次充当测试集,其余分为训练、验证集,以增加试验可靠性。并在无数据增强情况下进行如下对比试验以验证方法可行性,结果如表 3 所示,其中 DFCNN * 为改进后的声学模型。

表 3 各模型语音识别结果

数据集	模型结构	WER/%
专业数据	DFCNN	11.26
专业数据	GRU	15.23
专业数据	BILSTM	14.43
专业数据	DFCNN *	10.61

在 Transformer 模型参数中,设置隐藏节点数为 512,将多头注意力数设置为 8,经多次尝试,以默认 6 组编码解码器层为优。

同时,为了防止过拟合,将 dropout 层参数设置为 0.2,并采取标签柔滑化(Label Smothing)通过降低正确分类样本的置信度,提升模型的自适应能力来防止过拟合:

$$Y = y(1-\mathcal{L}) + u \cdot \mathcal{L} \tag{13}$$

其中:Y 为处理后的样本标签,$\mathcal{L}$ 为平滑因子,y 为原始数据,u 为$\mathcal{L}$的相关系数。

Transformer 模型损失函数设置为:

$$loss = -[Y\log p + (1-Y)\log(1-p)] \tag{14}$$

其中:p 为预测分数。

为对比观察 Transformer 模型的效果,采用相同声学模型,分别结合两种数据集和有无语言模型进行对比。通用数据集则采用了 Aishell-1、Thchs30 这几个开源数据集进行测试。结果如表 4 所示。

表 4 语言模型效果

样本	模型结构	WER/%
专业数据	DFCNN *	11.11
专业数据	DFCNN * -T	10.61
通用数据	DFCNN *	20.84
通用数据	DFCNN * -T	16.38

从表 4 可以看出,无 Transformer 模型精度要稍逊一筹,主要归因于 Transformer 中注意力机制考虑到了信息的连续性和空间分布,学习到了字符在整句中的相对位置以及上下文连续性。由于管制指令的特殊性,在发音、用词方面尽量避免混淆,语言模型在管制指令的识别中效果并不明显。但可以看出在日常对话识别过程中,由于多音字、相似词汇等的干扰,声学模型识别出的音素或拼音并不能很好地对应上正确汉字,本文采用拼音为建模单元,在日常对话情况下,语言模型将字错率降低了 4.46%。

4.6　迁移学习

本文迁移学习预训练采用的数据集为 Thch30、Aishell-1，实验中先对公开数据集进行训练，在到达一定效果后，通过冻结模型前部分层参数，仅训练后一部分全连接层参数，来达到将通用样本中声学建模单元的迁移，结果如表 5 所示。

表 5　迁移学习效果

迁移学习应用	模型结构	WER/%
否	DFCNN＊-T	10.61
是	DFCNN＊-T	8.32

表中可以看到，混合数据集训练后用于迁移学习，字错率明显降低，专用数据集中部分字符与通用数据集中字符重合，虽然同时引入了多音字符的干扰，但在语言模型的匹配下，字错率依然能够减少 2.29%。将训练好的最优模型与参与迁移学习的最优模型进行对比，可以看出迁移学习后的模型效果更好，相比于原始模型，迁移学习能够学习到更多声学特征，语音识别的效果也会更佳。

5　结束语

本文针对管制指令语音识别存在的问题，提出了生成联合深度卷积网络的结构，依据空中交通管理规范，建立了机坪管制指令语音数据库，构建了基于改进 DFCNN 和 Transformer 的语音指令识别模型。为了解决样本不足的问题，通过小样本学习中基于数据合成的数据增强方法依据来对数据进行扩充，即将音频语谱图当作标准图像进行尺度变换和时频遮掩，能够防止数据被简单复制而影响实验结果。本文设置了数据生成策略组并构建了级联生成对抗网络对数据进行混合增强以针对机坪管制指令的特点进行实验以提升识别方法鲁棒性，其中频谱遮掩方法效果显著，将字错率降至 6.14%，明显优于原始模型方法。另一方面，通过迁移学习方法将通用样本中的声学建模特征应用到小样本的学习中，对照组实验结果显示，迁移学习方法将字错率减少至 8.32%。实验结果表明，本文方法效果显著，机坪管制指令语音识别字错率降低至 6.14%，证明本文方法的有效性，本文方法将有望应用于机场高级地面活动引导及控制系统中机坪管制语音指令的检测和识别，实现机坪管制决策支持，助力现代机场高质量运行。

参考文献：

[1] ŠMiDL L, ŠVEC J, PRAZAK A, et al. Semi-supervised training of DNN-based acoustic model for ATC speech recognitio- n [C] // International Conference on Speech and Computer. Leipzig, Germany, 2018, 646－655.

[2] CHEN S, KOPALD H. The closed runway operation prevention device: Applying automatic speech recognition technology for aviation safety [C] //Eleventh USA/Europe Air Traffic Management Research and Development Seminar (ATM2015), 2015: 1－10.

[3] SRINIVASAMURTHY A, MOTLICEK P, SINGH M, et al. Iterative learning of speech recognition models for air traffic control [C] //Proceedings of Interspeech 2018, 2018: 3519－3523.

[4] LIN Y, LI Q, YANG B, et al. Improving speech recognition models with small samples for air traffic control systems [J]. Neurocomputing, 2021, 445 (20): 287－297.

[5] SAINATH T N, VINYALS O, SENIOR A, et al. Convolutional, Long Short-Term Memory, fully connected Deep Neural Networks [C] // ICASSP 2015－2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2015.

[6] 甘振业，周世华，曾　浩，等. 基于 DFCNN-CTC 端到端的藏族学生普通话发音偏误检测 [J]. 西北师范大学学报（自然科学版），2020，56 (5)：49－53，108.

[7] DONG L, SHUANG X, BO X. Speech-Transformer: A No-Recurrence Sequence-to-Sequence Model for Speech Recognition [C] // ICASSP 2018－2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Seoul, South Korea, 2018, 5884－5888.

[8] 赵凯琳，靳小龙，王元卓. 小样本学习研究综述 [J]. 软件学报，2021，32 (2)：349－369.

[9] 张一珂，张鹏远，颜永红. 基于对抗训练策略的语言模型数据增强技术 [J]. 自动化学报，2018，44 (5)：891－900.

[10] GOODFELLOW IAN, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets [C] //Advances in Neural Information Processing Systems, 2014: 2672－2680.

[11] GOODFELLOW I, MIRZA M, COURVILLE A, et al. Multi-prediction deep boltzmann machines [C] //In Advances in Neural Information Processing Systems, 2013: 548－556.

[12] MIRZA M, OSINDERO S. Conditional Generative Adversarial Nets [J]. Computer Science, 2014: 2672－2680.

[13] DENTON E L, CHINTALA S, FERGUS R. Deep Generative Image Models using a Laplacian Pyramid of Adversarial Networks [C] //Advances in neural information processing systems, 2015: 1486－1494.

[14] 李　响，李国正，邓明君，等. 基于语音频谱图像特征的人体疲劳检测方法 [J]. 仪器仪表学报，2021，42 (2)：123－132.

[15] QUAN T M, NGUYEN-DUC T, JEONG W K. Compressed sensing MRI reconstruction using a generative adversarial network with a cyclic loss [J]. IEEE transactions on medical imaging, 2018, 37 (6): 1488－1497.

[16] PHAN H, MCLOUGHLIN I V, PHAM L, et al. Improving GANs for speech enhancement [J]. IEEE Signal Processing Letters, 2020, 27: 1700－1704.

[17] WENG L. From gan to wgan [J]. arXiv preprint arXiv: 1904.08994, 2019.

[18] NIELSEN F. On a generalization of the Jensen-Shannon divergence and the Jensen-Shannon centroid [J]. Entropy, 2020, 22 (2): 221.

（下转第 198 页）