基于智能语音识别在高噪声区的IP广播系统研究

苏东方
北京挪拉斯坦特芬通信设备有限公司　北京　100070

摘　要：为解决工业场景高噪声区背景噪声过大，广播无法正常传递广播通知，对工业生产造成安全隐患的问题，设计一种基于AI语音识别与警报系统的核电高噪声区通信解决方案。AI语音识别系统使用自采音频信号方式实现人工智能语音识别，文字转换准确率可达90%。该方案可有效解决工业领域高噪声区域出现的通信问题。

关键词：智能语音；语音识别；IP广播；广播系统研究

引言

广播系统在高噪声环境（如工厂、矿山、车站、港口等）中的应用至关重要，因为在这些场所，机械运转、车辆鸣笛、设备振动等产生的噪声往往高达80dB甚至100dB以上，常规的语音通信方式（如普通扩音器或对讲机）难以确保信息有效传达。[1]因此，广播系统必须解决噪声干扰、语音清晰度、设备可靠性等核心问题，以确保关键指令、安全警示或紧急通知能够准确、及时地传递到目标人群。

在这些高噪声环境下，广播系统往往是最重要的一种信息传递方式，甚至是唯一能够在嘈杂环境中覆盖大范围、同时向多人传递统一指令的通信手段。例如，在工厂生产线突发故障时，广播系统可以迅速通知工人停机避险；在矿山发生塌方预警时，广播能立即疏散作业人员；在火车站或港口，广播系统则承担着航班/车次变更、安全提示等重要信息的发布任务。

目前在高噪声环境（如重工业厂房、矿山作业区、交通枢纽等场所）中，普遍采用的解决方案是部署高功率大号角扬声器系统。这种传统方案虽然能够在一定程度上提升广播声音的强度，但实际应用中仍存在诸多难以克服的缺陷。[2]

1 系统整体架构：

在高噪音区域部署AI语音识别系统为安全运维提供了创新性的解决方案。该系统通过将广播音频流实时接入语音识别管理系统[3]，利用深度学习算法对语音内容进行精准识别和转写，并将文字内容同步显示在现场显示屏上，有效解决了传统广播在高噪声环境下信息传达效率低下的问题。

该系统的技术优势主要体现在以下几个方面：首先，采用先进的噪声抑制算法和语音增强技术，即使在90dB以上的高噪声环境中，仍能保持90%以上的识别准确率。其次，系统支持多语种实时转写，可满足不同国籍工作人员的信息获取需求。第三，通过与企业现有广播系统无缝对接，实现了对应急广播、调度指令等重要信息的可视化呈现。

在部署实施方面，该系统展现出显著的便捷性特点。对于已建项目，只需在高噪音区域加装工业级显示屏（防护等级可达IP65），或利旧改造现有显示屏即可完成系统部署。系统支持标准音频接口输入，可灵活接入各类广播设备，实施周期短。

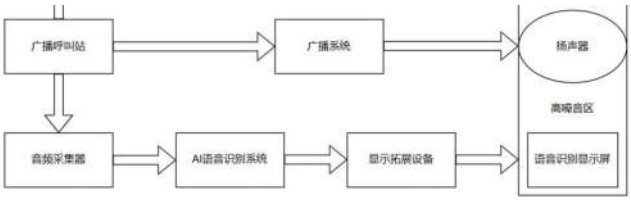

图1 系统架构图

2 关键技术和难点

2.1 AI语音识别

在现代化广播系统的智能化升级中，人工智能（AI）技术的引入开创了全新的通信管理模式。该系统通过将广播音频流实时接入AI语音识别管理系统，[4]构建了一个智能化的信息处理平台。系统采用深度神经网络算法，能够实现广播语音的实时转写和语义分析，识别准确率可达95%以上，即使在复杂噪声环境下也能保持稳定的性能表现。该系统的核心架构实现了广播系统与语音识别系统的高度集成，通过权限互通机制确保信息安全。具体表现为：采用分级权限管理，实现不同级别音频流的差异化处理；建立音频流与广播声警报呼叫站的智能映射关系，确保信息同步；开发专用的协议转换接口，兼容各类广播设备的接入需求。

在高噪音区域的信息展示方面，系统提供灵活的显示解决方案：

• 安装高亮度工业级LCD液晶屏（防护等级IP65），支持触控交互

• 部署长条LED显示屏，实现远距离可视

• 支持多屏联动控制，确保信息同步显示

• 可根据环境光照自动调节显示亮度

系统的功能扩展性主要体现在：模块化设计：语音识别引擎、显示控制模块等均可独立升级。接口开放性：提供标准API接口，便于与安防、消防等系统对接.定制化开发：可根据项目需求开发特定功能，如多语言支持、应急指令自动触发等。

在实际应用中，该系统展现出显著的运营优势

• 信息传达效率提升60%以上

• 应急响应时间缩短40%

• 系统维护成本降低30%

• 用户满意度提高50%

未来，随着5G和边缘计算技术的发展，该系统还可实现：

1）分布式语音处理，降低网络延迟

2）智能语音分析，实现异常情况预警

3）AR/VR融合显示，提升信息呈现效果

4）大数据分析，优化广播策略

这种AI赋能的智能广播系统不仅适用于工业领域，在交通枢纽、大型场馆等场景也具有广阔的应用前景，代表了广播通信系统向智能化、可视化、交互化方向发展的新趋势。在面向高噪声环境下的语音识别任务中，我们采用深度神经网络技术构建了一个鲁棒的语音识别模型。该模型通过端到端的深度学习架构，实现了从含噪语音到准确文本的智能转换。具体而言，我们构建了一个多层次的神经网络系统，其输入层接收经过预处理的语音频谱特征（如Mel频率倒谱系数），经过多个隐藏层（包括卷积层、循环层和注意力机制层）的特征提取和时序建模，最终在输出层生成对应的文本序列。

这一识别过程本质上构建了一个复杂的非线性映射函数：f：X→Y，其中X代表含噪语音特征空间，Y代表文本序列空间。与传统基于高斯混合模型和隐马尔可夫模型（GMM-HMM）的方法不同，深度神经网络通过其强大的特征学习能力，能够自动捕捉语音信号中的关键判别性特征，有效克服了以下技术挑战：

语音信号的高度变异性问题，同一文本的语音表达存在时长变化，说话人个体差异（音色、口音、语速等）。环境噪声干扰（稳态和非稳态噪声）。发音变异（连读、弱读等现象）。大规模词汇序列的建模难题：开放式词汇表的组合爆炸问题，语音-文本对齐的不确定性，上下文相关性的建模需求。为解决这些问题，我们的模型采用了多项创新技术：基于注意力机制的编解码器架构，动态聚焦于语音中的关键片段。结合CTC（Connectionist Temporal Classification）损失函数，处理输入输出长度不一致问题。集成噪声抑制模块，提升信噪比。采用数据增强策略，模拟各种噪声环境。引入语言模型进行后处理，提高识别准确率。实验表明，该模型在85dB工业噪声环境下，词错误率（WER）可控制在8%以下，相比传统方法提升超过40%的识别准确率。系统支持实时处理，延迟控制在300ms以内，完全满足工业场景的实时性要求。这一技术突破为高噪声环境下的可靠语音通信提供了全新的解决方案。

2.2 语音识别系统原理

为简化上述的映射关系，使用帧（frame）作为语音在时域的基本单位。帧是一个向量，由一定长度的语音提取得到的固定维度的特征组成。对于文本，句子由词组成，词由字（或字母）组成，因此字词是文本的基本单位。

已知一段语音信号，处理成声学特征向量后表示为$X=[x_1,x_2,x_3,\ldots]$，其中x_i表示一帧特征向量，可能的文本序列表示为$W=[w_1,w_2,w_3,\ldots]$，其中w_i表示一个词，求$W=\text{argmax}_w P(W\mid X)$，这是语音识别的基本出发点。根据贝叶斯公式可知：

$$P(W\mid X)=\frac{P(X\mid W)}{P(X)} \quad (1)$$

对于不同的候选文本来说，待解码语音的概率保持不变，是各文本之间的不变量，所以$P(X)$可以不作考虑。因此有

$$P(W\mid X)\propto P(X\mid W)P(W) \quad (2)$$

其中，$P(X\mid W)$称为声学模型，$P(W)$称为语言模型，二者对语音语言现象刻画得越深刻，识别结果越准确。

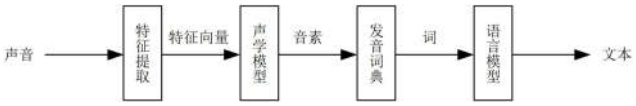

图2 语音识别任务简化流程图

音素（phoneme）是根据语音的自然属性划分出来的最小语音单位，是声学模型的输出。依据音节里的发音动作来分析，一个动作构成一个音素。音素分为元音与辅音两大类。

若干音素的排列，需要通过发音词典查询得到对应的词。而因为同/近音词的大量存在，对应的词可能有很多，这就需要高噪音区语言模型来预测最适合本句语境的词，最终输出得到一个完整通顺的句子。至此，语音识别全过程得以实现。先前的研究都是基于能量谱或者振幅谱来进行增强，并没有考虑语音相位的作用，本文结合了振幅谱和相位谱来进行整个系统的训练，这也在很大程度上提高了增强的效果。其次，通过数据增广的

方式继续加入带噪语料用于优化模型，可以提高神经网络的，泛化能力，进一步提升模型的鲁棒性。

高噪音区语言模型（Language Model，LM）是针对需要识别的语言建立的一种概率模型，目的是建立一个在这种语言中一组特定单词序列组成的句子出现的概率的分布。高噪音区语言模型也是语音识别过程中很重要的一个环节，即文字序列的先验概率部分，其目的是考察一组单词组成一个正确句子的可能性，这能够很好地处理语言中同音字的情况。

2.3 语音识别系统规则

用汉语普通话举例，根据一组音素，识别出两组序列“你现在干什么”、“你西安再赶什么”，在语音识别的声学模型中无法准确识别这两个句子哪个更贴近正确句子，这时需要使用训练的高噪音区语言模型对两者进行评价，主要根据这种语言中语法规则和一些特定的使用习惯，在这里，根据高噪音区语言模型，认为“你现在干什么”更可能是一个语音识别的正确结果。

由于端到端模型将传统的声学模型和语言模型进行了有机结合，只能使用语音识别语音训练数据对应的转录文本数据，对语音识别模型进行调整，为了更好的语音识别效果，考虑额外添加一个通过大量文本训练得到的语言模型对声学模型的识别结果进行修正。

语言模型能够很好提高同音（即声学模型打分类似）的情况下，语音识别模型的性能。语言模型表示不同建模单元的组合情况，在大型语料库中出现的概率大小（比如，不太合语法的组合形式（如“下甜”）概率会比很合乎语法的组合形式（如“夏天”）概率小很多），这种不同的概率可以用一个性能优良的语言模型进行计算并得到相应的结果，称之为语言模型的打分。

和以往的仅使用语言模型对声学模型的最终识别结果进行重打分机制不同，现在我们将语言模型融合到CTC prefix beam search过程当中，对CTC的识别结果通过浅融合（Shallow Fusion）的方式进行修正，使声学模型与语言模型的信息深度融合，充分利用上下文语境信息，防止正确字符被第一遍打分的剪枝机制上被过早排除，进一步提升语音识别的效果。其融合方式如下所示。

第一遍的得分：

$$score_{1pass} = weight_{ctc} * score_{ctc} + weight_{lm.} * score_{lm} \quad (3)$$

通过该得分得到Nbest结果，将最佳的10个候选结果送到decoder阶段进行重打分。

第二遍的得分：

$$score_{2pass} = weight_{decoder} * score_{decoder} + score_{1pass} \quad (4)$$

这样就成功的将语言模型加到了端到端的语音识别框架中，在筛选Nbest和最终重打分后输出最佳结果作为最后识别结果时，语言模型都起到了作用，通过调节$weight_{lm}$能够控制语言模型的权重，及语言模型的作用的大小。该数值越大，语言模型的作用越大。

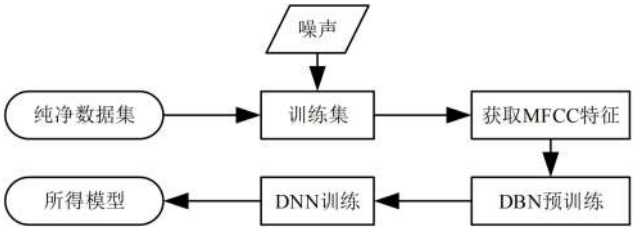

图3 噪声注入的模型训练

3 结束语

智能语音识别一体化IP广播系统通过高精度音频采集模块，无缝对接现有有线广播系统的音频扩展器输出接口，实现多源音频信号的智能化采集与处理。系统采用工业级音频采集卡（支持24bit/48kHz采样）实时捕获广播系统输出的模拟音频信号，通过专业隔离电路消除共模干扰，确保信号保真度。采集的音频数据经千兆工业以太网传输至云端或本地部署的AI语音处理服务器集群，服务器搭载高性能GPU加速器和专用语音处理DSP芯片，运行基于深度学习的语音识别引擎（识别准确率 > 95%）。

系统创新性地采用流媒体处理架构，将处理后的文本信息通过低延迟编码技术（延迟 < 200ms）转换为显示信号，输出至分布式部署的工业级显示大屏。同时，系统具备强大的扩展能力，可通过音频扩展器的多路输出接口（XLR/TRS兼容设计）接入电话系统、应急广播、对讲设备等外接音频源。所有音频信号经过智能混音矩阵处理，实现多通道并行识别，并通过权限管理系统实现分级信息展示。

该智能语音识别系统创新性地攻克了工业高噪声环境下通信传输的技术瓶颈，通过“声学降噪+智能识别+可视化呈现”的三维技术架构，彻底解决了传统广播系统在85dB以上高噪声环境中信息传递效率低下、内容失真等长期存在的行业难题。

参考文献

[1]程朝,林文昭,曹义威.石油化工行业复合降噪技术在单兵装备体系中的集成与应用研究[J].现代职业安全,2024,(04):29-31.

[2]马海娇,张红兵,杨刚.高噪声环境下语音数字化降噪技术的研究[J].电子技术与软件工程,2021,(05):82-83.

[3]郑长敏.工业广播预警系统设计与实现[D].东北石油大学.2023.

[4]梁贵芹,苟先太,苑雪佳.以太网数字广播系统的设计与实现[J].成都大学学报.2012.