学校代码：10126　　学号：31709050
分 类 号：TP183　　编号：

论文题目

噪声环境下的语音关键词检测

学　院：计算机学院
专　业：计算机技术
研究方向：智能信息处理
姓　名：谷悦
指导教师：张学良 教授

2019年5月31日

原 创 性 声 明

本人声明：所呈交的学位论文是本人在导师的指导下进行的研究工作及取得的研究成果。除本文已经注明引用的内容外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得内蒙古大学及其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

学位论文作者签名：谷振 指导教师签名：张学良

日 期：2019-5-13 日 期：2019-5-13

在学期间研究成果使用承诺书

本学位论文作者完全了解学校有关保留、使用学位论文的规定，即：内蒙古大学有权将学位论文的全部内容或部分保留并向国家有关机构、部门送交学位论文的复印件和磁盘，允许编入有关数据库进行检索，也可以采用影印、缩印或其他复制手段保存、汇编学位论文。为保护学院和导师的知识产权，作者在学期间取得的研究成果属于内蒙古大学。作者今后使用涉及在学期间主要研究内容或研究成果，须征得内蒙古大学就读期间导师的同意；若用于发表论文，版权单位必须署名为内蒙古大学方可投稿或公开发表。

学位论文作者签名：谷振 指导教师签名：张学良

日 期：2019-5-13 日 期：2019-5-13

噪声环境下的语音关键词检测

摘 要

随着智能家居、智能手机和自动化设备的快速发展，基于语音技术的人机交互变得越来越流行，例如谷歌公司的Google Now，微软公司的Cortana、亚马逊公司的Alexa和苹果公司的Siri已变得十分流行。为了实现免手持的语音识别体验，语音识别系统需要持续不断地监听特定的唤醒词语来开始语音识别任务，这个过程通常被称为关键词检测（Keyword Detection，KWD）或关键词识别（Keyword Spotting，KWS）。考虑到目前很多设备计算资源受限并且大都使用电池作为能量供应，这要求关键词检测必须满足小内存占用和低能量消耗的要求。

在现实世界的环境中，噪声干扰不可避免，噪声鲁棒性对于关键词识别任务而言至关重要。为了提高关键词检测系统的鲁棒性，通用的方法是在系统前端增加一个语音增强模型。本文为提高关键词检测模型的鲁棒性做了三个方面的尝试。首先，本文将预训练的语音增强模型和关键词检测模型连接起来形成一个更复杂的系统。整个模型使用联合训练的方法，因此关键词检测系统包含的语言学信息可以通过反向传播的方法传递给增强模型。第二，本文提出了一种新的卷积循环神经网络，这种网络结构需要参数量和计算量更少，并且不会降低增强模型的性能，极大程度上满足了小内存、低功耗的设备部署需求。最后，为了进一步提升系统的性能，设计了特征转换模块，将输入特征从能量谱转换为梅尔谱，有效地减少了计算量，更适合关键词检测系统。

本文采用的基于联合训练框架的CNN-MelCRN32关键词检测系统在测试集上的准确率为93.17%，与带噪训练的（基于Multi-condition训练方法）基线系统相比相对提升64.2%，显著地提高了关键词系统的噪声鲁棒性。

关键词：鲁棒性关键词识别 关键词检测 语音增强 卷积循环神经网络 联合训练

KEYWORD DETECTION IN NOISY ENVIRONMENTS

ABSTRACT

With the proliferation of smart homes and mobile and automotive devices, speech-based human-machine interaction becomes prevailing, e.g., in Google Now, Microsoft Cortana, Amazon Alexa, and Apple Siri. To achieve hands-free speech recognition, the system continuously listens for specific wake-up keywords, a process often called keyword detection (KWD) or keyword spotting (KWS) , to start speech recognition. From the perspective of practice, the keyword detection systems typically run on the small-footprint device with low power consumption.

Robustness against noise is critical for keyword detection systems in the real-world environments. To improve the robustness, a speech enhancement frontend is involved. This thesis attempts to improve the robustness of the keyword detection model in three aspects. Firstly, instead of treating the speech enhancement as separated preprocessing before the keyword detection system, in this study, the pre-trained speech enhancement frontend and the convolutional neural networks (CNNs) based keyword detection system are concatenated, where a feature transformation block is used to transform the output from enhancement frontend into the keyword detection system's input. The whole model is trained jointly, thus the linguistic and other useful information from the keyword detection system can be back-propagated to the enhancement frontend to improve its performance. Secondly, to fit the small-footprint requirement of on-device deploying, a novel convolution recurrent network is proposed, which needs fewer parameters and computation and does not degrade performance. Finally, by changing the input features from the power spectrum to Mel spectrum, less computation and better performance are obtained. our experimental results demonstrate that the proposed method significantly improve the KWS system with respect to noise robustness.

The proposed model, joint trained CNN-CRN32 achieves an accuracy of 93. 17% under noisy conditions, which is 64.2% higher than the baseline trained Multi-conditional data. The proposed method significantly improve the robustness of the keyword detection systems.

KEYWORDS: Robust Keyword Detection, Speech Enhancement, Convolutional Recurrent Neural Network, Joint-Training

目 录

图目录

表目录

第一章 引言

1.1 研究背景及意义

语音是人与人交互最主要的方式，人类可以用简短的语音表达出丰富的信息，这种特性是其它交互方式（如图像等）所不具备的，因而语音是有可能在未来广泛使用的人机交互方式。由于语音信号高度抽象，长久以来语音信号的研究与处理一直是一个艰深的问题。近些年随着深度学习的快速发展，人机语音交互得到了迅猛
发展，机器逐渐能够“听见”人们的话语，并进一步作出回应。目前，基于语音交流的人机交互方式已经进入人们的日常生活中，在移动设备、智能家居、机器人等领域迅速发展，使得人与机器、人与人之间的沟通交流更加的简单便捷。关于语音信号的研究根据不同的应用需求细分出很多领域，像是自动语音识别、语音合成、说话人确认、语音增强等 * MERGEFORMAT [1]领域。其中，关键词检测（Keyword Detection, 关键词检测或 Spoken Term Detection，STD）,也称为关键词识别（Keyword Spotting,
KWS），是一种需要在连续的语音流中识别出预定义词的技术。相较于自动语音识别任务，关键词检测技术的目标不是完整的一句话或一段话，而是设定的特定词。

语音关键词检测技术具有广泛的应用场景。最初提出这项技术主要是为了监测电话通讯中的敏感词和实时语音服务监测中的特殊词。在电话监控中，使用语音关键词检测技术对电话通话进行实时监控，发现敏感的特殊词。而在实时语

音服务的应用场景中，例如语音自动分机接驳服务、语音查询系统等，仅仅需要检测出关键词就可以完成系统功能。随着移动互联网时代的到来，语音关键词检测技术有了更加丰富的应用场景，例如智能家居、智能手机，在这些应用场景中语音关键词检测技术成为人机交互的语音接口。语音关键词检测技术可以持续不断的监听特定的关键词，用户仅仅需要说出预设关键词就可以唤醒设备开始工作，从而为用户提供免手持的语音识别体验。由于智能设备硬件资源和和电源供应等方面受限，相较于自动语音识别系统而言，关键词检测系统受到以下限制：

1、关键词检测模型需要以一种非常节能高效的方式运行，模型计算资源需求小。

2、关键词检测模型大部分的输入是静音或者背景噪声而不是语音，并且，系统大部分输入的语音是与预设的关键词无关的，因此系统的虚警率（False Alarm rate, FAR）和错误拒绝率（False Rejection Rate, FRR）必须要小。

当前，关键词检测系统在相对安静的环境中表现良好，例如最新的研究 * MERGEFORMAT [2]表明他们能够在谷歌语音命令数据集 * MERGEFORMAT [3]上达到95%左右的准确率，并且模型所需的计算资源较少。然而，真实环境中的噪声对于关键词检测系统而言仍然是一个的挑战。关键词检测系统在噪声环境中识别性能有所下降，因此系统需要对不同环境下和不同响度的噪声都具有鲁棒性。系统处理噪声的能力非常重要，因为这个能力决定了用户是否能在距离设备比较远、带有一定噪声和混响的情况下，顺畅地进行免手持的人机交互。

鲁棒性关键词检测技术具有十分重要的意义。在真实生活场景中，噪声是无处不在的。具有噪声鲁棒性的关键词检测系统能够为用户提供更加稳定的服务。在不同的环境中，鲁棒性关键词检测系统都能够准确响应用户指令，同时也为下一步语音多轮对话提供服务。

1.2 相关研究工作

1.2.1 关键词检测研究现状与分析

关键词检测技术的探索起始于上个世纪70年代，到现在已经持续了几十年。关键词检测系统研究方法大致分为三种：基于模板匹配的方法、基于关键词-补白（keyword/filler）模型的方法和基于大词表连续语音识别（Large Vocabulary Continuous Cpeech Recognition, LVCSR）的方法。基于动态时间规划（Dynamic Time Warping, DTW）的模板匹配方法是在20世纪80年代提出的，这种方法将待求解的全局最优问题转换为每一步的局部最优问题[7]，目前基于DTW或DTW变形的方法（segmenta-DTW）仍处在主流研究行列中[8] * MERGEFORMAT [9] * MERGEFORMAT [10]。基于模板匹配的方法具有模型尺寸小、计算量少的优点，但是准确率不如其他两种方法。基于keyword/filler模型的关键词-补白模型的方法会为关键词和非关键词（补白）分别建立模型，通过使用维特比解码算法来判断输入的音频中是否出现了关键词[11][12]。这种方法需要大量含有关键词和非关键词的训练数据，并且对于新加入的关键词不够灵活，需要使用含有新关键词的训练数据重新训练模型。第三种方法是基于LVCSR的，这种方法是源于语音识别任务，但与语音识别任务不同的是，关键词检测系统的目标为孤立词，因而基于LVCSR的关键词检测系统的解码网络只需要包含keyword和filler的路径，因而解码网络比语音识别的解码网络小，解码速度比较快[13] * MERGEFORMAT [14] * MERGEFORMAT [15]。但是关键词检测系统必须满足移动手持设备低功耗、低内存占用的要求，关键词检测系统一般只能利用智能设备自身的硬件资源完成识别任务，不像语音识别任务能利用远程云端的资源，因而基于LVCSR的关键词检测系统应用场景比较有限。

自2006年以来，深度学习成为机器学习的一个重要领域，在图像、语音领域都取得了巨大的成功。基于深度学习的声学模型建模已经逐渐取代了基于高斯混合模型的声学模型，并且深入语音的各大领域。在关键词检测领域，一种端到端的模式逐渐成为主流，将待识别的语音直接输入到模型中，直接输出待识别的关键词或者非关键词。基于端到端的关键词检测系统通常包括三个部分：特征提取模块、神经网络模块和输出后验得分的计算模块[16] * MERGEFORMAT[17]。在识别阶段，关键词检测系统首先对输入语音按帧提取特征，之后将特征送入训练好的神经网络模型中，输出各个关键词和非关键词的后验概率，最后对后验概率以一定的窗长进行平滑，平滑后的后验得分如果超过预先设定的阈值或者选取平滑后多个关键词中最大的后验得分，就认为识别出了某个关键词。这种基于端到端的关键词检测系统，与其他传统方法比较，具有实现简单、模型尺寸小、计算量少并且准确率高的优点。本文的研究所使用的基线系统就是基于卷积神经网络的端到端的模型，这种模型最早由谷歌的Sainath Tara等人[34]提出。

关键词检测系统已经在现实生活中广泛使用，例如各种智能设备的唤醒功能。关键词检测是与实际结合非常紧密、极其影响用户体验的任务，因而市场对关键词检测系统的要求也更加的严苛，需要在模型性能、模型复杂度等诸多方面达到一个平衡。目前，评价一个关键词检测系统的性能主要有以下指标：准确率、ROC曲线（Receiver operating characteristic curve, ROC）等。其中，准确率是评估
关键词检测系统分类准确率（关键词和非关键词）的指标，ROC综合考虑了系统的错误拒绝率（False rejection rate, FRR）和错误接收率（False alarm rate,
FAR），曲线越偏向零点代表性能越好。由于不同的设备部署限制，除了评估关键词检测模型性能外还需要评估模型计算资源需求情况，通常将模型大小和模型计算复杂度作为参考指标。总之，关键词检测
系统需要运行在计算资源受限的设备上，同时具有较好的性能表现，因而关键词检测系统必须在计算复杂度和模型性能之间进行平衡。

1.2.2 语音增强发展历史与研究现状

近年来，语音关键词检测系统已经应用到人们的日常生活中，然而关键词检测系统在实际使用中总是受到各种各样的干扰，比如背景噪声、房间混响以及信道不匹配等问题。语音增强（Speech Enhancement,
SE）技术正是要将感兴趣的语音关键词从受干扰的混合语音中提取出来，或者抑制甚至消除干扰。语音增强可以作为关键词检测系统的预处理环节。因此，从提高语音关键词检测系统鲁棒性的角度看，语音增强技术十分重要。

语音增强按照通道数量可分为单通道增强和多通道增强，其中单通道语音增强技术最具挑战性，也最具实用价值的。从应用的角度看，它具有部署简单、对运行设备要求低的优点，适合低计算资源的关键词检测任务。单通道语音增强方法包括传统的信号处理方法和掩膜估计方法，本文中使用的是基于掩膜估计的有监督学习的语音增强方法。基于掩膜估计的语音增强方法源自人类听觉的
掩蔽效应，以掩膜估计作为目标，使用估计出的掩膜对带噪语音进行掩蔽，从而增强语音质量。有监督学习的掩膜估计方法将语音增强看作回归问题，训练学习器从带噪语音预测人工构造的理想掩膜。

(1) 早期有监督学习模型

最早用于解决语音增强问题的有监督学习模型是上世纪80年代的多层感知器（Multilayer Perceptron，MLP）。Tamura * MERGEFORMAT [18]和Waibel * MERGEFORMAT [19]
提出使用多层感知器从短时加窗的带噪语音波形预测纯净语音波形。由于当时的计算能力有限，这个多层感知器模型非常小，同时实验数据规模也很小，与基于信号处理的方法相比没有体现出优势。另外一个比较早的有监督学习的语音增强方法是在1994年提出的 * MERGEFORMAT [20]
，作者使用多层感知器预测纯净语音的对数能量谱，同样由于计算能力的限制，使用的特征简单，语音和噪声的变化相对较少，模型也相对简单。

后来，基于有监督学习的语音增强方法从计算听觉场景分析中借鉴了掩膜估计，不再直接对纯净语音的波形或者能量谱进行预测，这种结构保留的预测目标要比直接的频谱估计有效 * MERGEFORMAT [21]
。在单声道语音增强方面比较早的是Seltzer等人的研究 * MERGEFORMAT [22]
，在双声道方面比较早的是Roman和Wang等人的研究 * MERGEFORMAT [23]
。Seltzer提出的方法是针对自动语音识别模型中纯净语音特征被噪声掩盖的问题，他们使用贝叶斯分类器来判别时频单元被噪声掩盖的可能性。为解决这一问题使用了很多特征，其中包括梳状滤波器比、自相关函数尖峰比以及频谱平坦度等。从对自动语音识别模型性能提升的角度看，掩膜估计要比谱减类方法和直接频谱估计类方法效果好很多。较早的技术主要尝试去估计理想二值掩膜（Ideal Binary
Mask，IBM），而Tchorz和Kollmeier提出局部信噪比掩膜，他们使用幅度调制谱作为输入，训练一个多层感知器来对每个频带的局部信噪比进行预测 * MERGEFORMAT [24]。

总的来说，早期的有监督语音增强方法将语音增强看作二值分类问题或者回归问题，使用多层感知器作为分类器预测理想二值掩膜，或者使用回归模型估计局部信噪比。

(2) 支持向量机与混合高斯模型

2009年Kim将混合高斯模型（Gaussian Mixture
Model，GMM）引入语音增强任务，他使用混合高斯模型对梅尔频域上的理想二值掩膜进行建模，该方法使用幅度调制谱及其一阶差分、二阶差分作为特征，在低信噪比的场景下进行训练 * MERGEFORMAT [25]
。值得注意的是，Kim的方法是首个对正常听力者来说，依旧能够提高语音可懂度（Short-Time Objective Intelligibility，STOI）的方法。但缺点是该方法只有在噪声类型匹配的条件下才能够生效，其泛化能力十分有限。后续的改进多是针对泛化能力进行的，包括引入自适应性方法，以及使用对噪声更鲁棒的语音特征。

(3) 深度神经网络

尽管支持向量机取得了很大的成功，但在应对线性不可分问题时引入的核技巧计算复杂度太高。针对支持向量机的这个问题，研究者们提出使用深度神经网络（Deep Neural Network，DNN）代替或者辅助支持向量机。

在文献 * MERGEFORMAT [26]
中，Wang使用深度神经网络代替支持向量机作为分类器，通过使用多个说话人多种噪声构造的数据进行训练，得到了非常大的效果提升。该方法使用由Group
LASSO得到的组合互补特征作为输入，以理想二值掩膜作为训练目标，即使在噪声不匹配的情况下依然提升了语音可懂度。

除特征外，对深度学习模型来说，模型的学习目标是非常重要的，在语音增强任务中有多种训练目标可供使用，包括理想二值掩膜、理想比率掩膜、能量谱、对数能量谱等。同时损失函数也有多种选择，包括均方误差函数、交叉熵函数以及命中率与误检率的差（HIT-False Alarm，HIT-FA）等。Wang对这些训练目标和损失函数进行了评价 * MERGEFORMAT [27]
，他们发现通常情况下，掩膜可以较好的提高语音可懂度，而对提高语音质量来说，掩膜稍微好于直接估计频谱；回归问题中均方误差函数的效果比较好，二分类问题中HIT-
FA作为损失函数可以得到较高的语音可懂度，而交叉熵函数可以得到较高的语音质量（Perceptual Evaluation of Speech Quality，PESQ）。对于保留掩膜中的结构化信息，结合非负矩阵分解，文献 * MERGEFORMAT [28]
中提出将理想比率掩膜进行分解得到掩膜的基，然后训练深度神经网络对基的系数进行预测，这样的训练目标更具鲁棒性。

语音是一种时序信号，其时序信息对于语音增强也至关重要。在文献 * MERGEFORMAT [29]
中，深度神经网络作为一种非线性特征提取器，将线性不可分语音特征映射为更线性可分的特征，之后结合条件随机场对语音的时序性进行建模，从而预测纯净语音的掩膜。Huang分别使用深度神经网和递归神经网（Recurrent Neural Network，RNN）作为分类器，他的实验发现这两者模型的差别很小 * MERGEFORMAT [30]。在文献 * MERGEFORMAT [31]中，长短时记忆网 * MERGEFORMAT [32]（Long Short-Term Memory，LSTM）作为模型对梅尔域的频谱进行了估计，之后将梅尔频谱转换为比率掩膜并对带噪语音进行掩蔽，这个方法比传统的语音增强方法效果要好，但是没有跟最近的有监督学习方法进行对比。

1.2.3 鲁棒的关键词检测技术的研究现状

近几年，鲁棒的关键词检测的研究逐步得到研究者的重视。为了提高关键词检测系统的噪声鲁棒性，应用最常见、最广泛的方法是多环境训练策略（Multi-condition Training） * MERGEFORMAT [33][34] * MERGEFORMAT [35]
，这种方法是使用不同噪声环境下的带噪声频数据来训练神经网络。然而，多环境训练的方法为了得到较好的性能通常需要训练大型的模型，大型的模型很难部署在资源受限的设备上。

随着深度学习的快速发展，基于深度学习语音增强技术取得了重大进步。在自动语音识别（Automatic speech recognition,
ASR）领域，语音增强技术被引入进行语音识别的前端处理工作，作用是在语音识别模型识别之前增强带噪的语音，从而使识别模型对背景噪声更具鲁棒性。鉴于语音增强技术在ASR任务上的优秀表现，在关键词检测系统中引入语音增强技术的表现也可以预见
。通过使用前端增强技术，在将特征传递给关键词检测系统之前，过滤掉目标语音中的干扰，从而显著提高关键词检测系统的识别效果。2018年腾讯公司Yu Meng等人提出基于文本依赖的鲁棒关键词检测方法 * MERGEFORMAT [36]，显著提升了关键词检测系统在噪声环境下的模型性能，然而这种基于双向长短时记忆神经网络（Bi-directional Long Short-Term Memory,
BiLSTM）的方法需要较多的参数和计算量，不适宜部署在计算资源受限的设备上。另外，谷歌提出基于麦克风阵列的鲁棒关键词检测系统 * MERGEFORMAT [37]
具有良好的抗噪能力，但是多麦克风的方法应用场景受限，并且成本高于单通道语音增强方法。

1.3论文的主要研究内容及组织结构

1.3.1 研究内容

本文为计算资源受限的关键词检测系统设计了一种低内存占用、低功耗的语音增强方法。对比BiLSTM增强模型，我们提出的模型性能相当甚至更好，并且只需要少量的参数和计算量。主要研究内容如下：

1、考虑到语音增强和关键词检测不是两个完全独立的任务，它们可以互相影响，本文首次在关键词检测领域应用联合训练策略，将两个模型连接在一起得到一个更大更深的模型，之后进行联合优化来提高模型的噪声鲁棒能力。文中对比了四种训练策略：多环境训练策略、直接语音增强前端策略和再训练（Retraining）策略、联合训练（Joint-training）策略。

2、为满足关键词检测系统计算资源受限的要求，本文设计了一种基于卷积循环神经网络（convolutional recurrent network,
CRN）的小型语音增强前端模型，对比了基于BiLSTM的语音增强前端模型。进一步地为计算资源更加少的智能设备设计了压缩的语音增强模型。最后，本文在性能和计算资源占用两个方面分别对比了提出的模型和基于BiLSTM的增强模型。

3、常用的增强特征能量谱并不适用于鲁棒关键词检测系统，本文提出了基于梅尔域的特征即梅尔谱，并设计了特征转换模块。文中对比了这两种不同的谱特征，并进行分析。

1.3.2 组织结构

论文内容按如下章节划分为五个部分，具体结构如下：

第一章是引言，主要论述了关键词检测和语音增强的相关概念及研究意义，简单介绍了鲁棒关键词检测的研究现状，并对本项工作的主要研究内容和各个章节的内容进行了说明。

第二章介绍了关键词检测模型的原理和系统结构，对特征提取、声学模型、后验概率平滑处理三个方面进行了详细描述，并建立了基线系统，阐述了性能评价指标和实验结果。

第三章描述了基于深度学习的有监督语音增强方法，详细描述了有监督语音增强系统的原理，常用的训练目标、模型训练策略，构建了基于BiLSTM的语音增强模型。在噪声匹配和不匹配的测试集上对比了我们提出的联合优化策略与其他两种常用策略之间的性能差异。

第四章描述了本文提出的基于CRN的语音增强模型和梅尔域增强方法，阐述了基于能量谱和梅尔谱两种特征域的CRN网络结构，并介绍了压缩模型。最后，对比了本实验中的全部模型，从模型类型、特征域、音节长度敏感度、训练策略、噪声泛化五个方面进行分析。

最后对全文工作进行总结，并且对未来改进工作进行简述。 Equation Section (Next)Equation Chapter (Next) Section 1

第二章 关键词检测系统

在本文的研究中，关键词检测系统参考TensorFlow reference基线系统进行设计，基于端到端模式并丢弃了语言模型，使用深度学习框架Pytorch实现。关键词检测整体结构如图2-1所示。整个系统框架包括三个主要部分：特征提取模块、基于卷积神经网络（Convolutional Neural Networks, CNN）的声学模型、后验概率处理模块。特征提取模块进行声音活动检测（Voice Activity Detection, VAD）并对每一帧音频计算特征向量，之后将预设帧数的帧级别特征向量按帧堆叠成特征谱图，送入卷积神经网络中。我们训练CNN声学模型来预测每一个标签（label）的后验概率，这些标签代表一个关键词或者说是关键词的一部分。后验概率处理模块会将CNN的输出进行平滑，对每一个标签计算一个平滑窗内的置信度，置信度分数最高的即为预测的关键词或者非关键词。

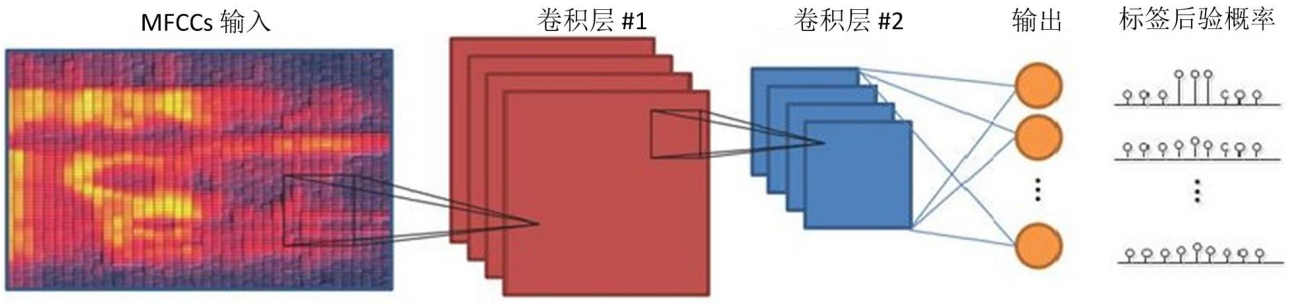

图2. 1 关键词检测系统框架图
Figure 2.1 Framework of keyword detection system

2.1 特征提取

语音信号承载着重要的语义信息，特征提取的目的就是将语音信号的重要信息用数字向量的形式表示。特征提取前首先将语音信号的模拟信号转变为数字信号，之后对数字信号进行特征提取，将经过特征提取后输出的声学特征向量作为声学模型的输入。目前，常用的语音识别声学特征有感知线性预测(Perceptual Linear Predictive, PLP)[38]、滤波器组（Fourier-transform-based log Filter-bank, Fbank） * MERGEFORMAT [39]、梅尔频率倒谱系数(Mel Frequency Cepstrum Coefficient, MFCC)[40]等。其中Fbank和MFCC使用最为广泛，本文所采用的是Fbank特征和MFCC特征。

MFCC特征考虑了人类听觉感知只聚焦在某些特定频率的特性，语音信号的MFCC特征提取过程如图2.2所示。

首先对音频信号进行数模转换，之后对音频信号进行预加重、分帧、加窗处理。预加重操作的目的是提高语音中高频的部分，使得信号在低频到高频的整个频谱变得平坦。语音信号具有短时平稳性，即10-30 ms内可以认为语音信号的统计学性质近似不变，因而称一短段语音信号为一帧，从而语音信号被划分为多帧信号。分帧时，为了避免丢失信息，采取重叠分段的方法，一帧的时长为帧长，相邻两帧的起始位置时间差为帧移。加窗操作是指将语音信号与窗函数相乘，方便之后做傅里叶变换。本实验中，帧长为30 ms，帧移为10 ms，窗函数使用汉明窗。经过预处理后，对语音信号作快速傅里叶变换得到频谱，之后对频谱取模平方后可以得到信号功率谱。梅尔滤波器组用一组梅尔频率上线性分布的三角窗滤波器对功率谱进行卷积滤波，并求取对数。最后用离散余弦变换算法对上一步结果进行计算，去除各维信号的相关性，即可得到梅尔倒谱特征。为了进一步提高系统的识别性能，会对MFCC特征参数计算一阶差分参数（Delta）和二阶差分参数（Delta-Delta）。最终可得13维MFCC特征及其一阶二阶差分，加上对数能量特征，共40维特征。在本研究中，关键词检测系统的特征是40维的MFCC特征。

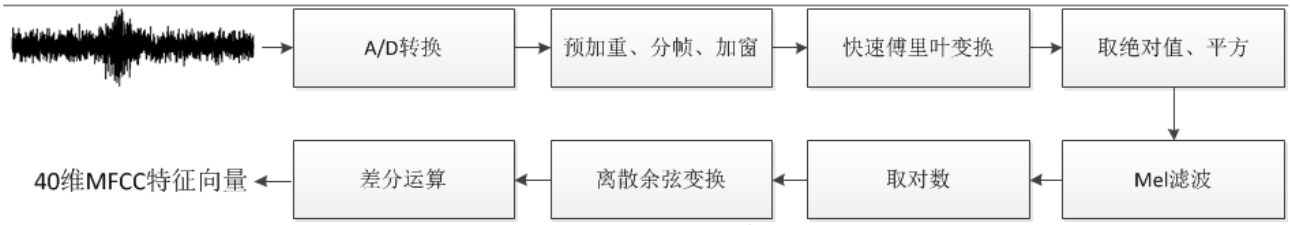

图2．2 MFCC特征提取流程图

Figure 2.2 Flow diagram of MFCC feature extraction

2.2 基于卷积神经网络的声学模型

声学模型是关键词检测任务的核心部分，选择合适的声学模型是构建一个准确可靠并且计算资源占用少的关键词检测系统的重要一步。早期，基于HMM（Hidden Markov Model,
HMM）模型的声学模型是关键词检测任务声学模型的主流选择。然而，随着智能手机等移动交互设备的兴起，对于关键词检测模型的计算资源消耗有了更严格的要求。当前，基于深度神经网络的端到端的声学模型 * MERGEFORMAT [2] * MERGEFORMAT [5] * MERGEFORMAT [34]
逐渐取代了基于HMM的声学模型，这类模型不需要解码部分，实现较为容易，并且模型的输出为每一个标签的后验概率，经过后验概率处理模块平滑处理后给出每一个标签的置信度得分。相较于基于深度神经网络（Deep Neural Networks,
DNN）的声学模型，CNN模型能更好的学习语音频谱中的时频关系，并且需要更少的计算资源；相较于序列模型，CNN比较简单，分类效果也可接受。

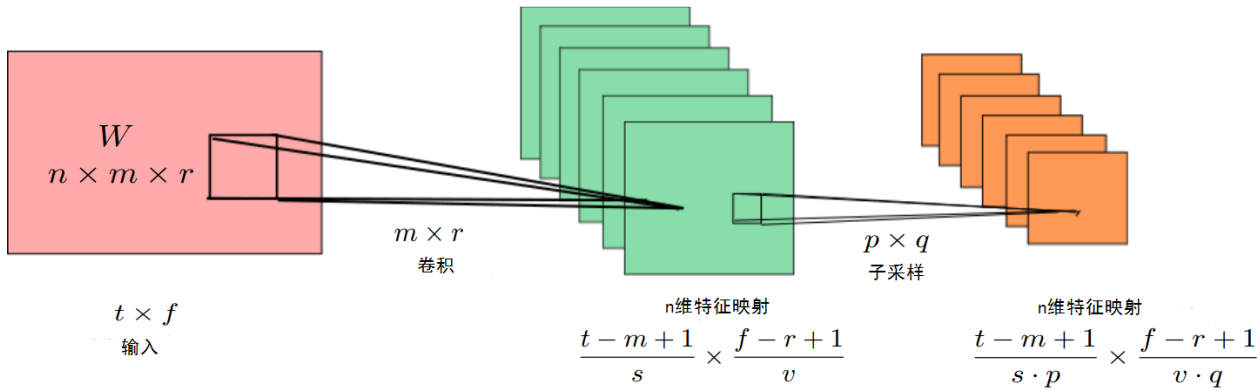

图 2．3 卷积神经网络

Figure 2.3 Convolutional neural network

一个简单的CNN结构如图2.3所示，图中展示了一层卷积层和一层池化层。网络的输入信号为t*f的特征向量，其中t代表时间维度，f代表频率维度。卷积核大小为m*r（m<=t, r<=f）。s代表时间轴的长度，v代表频率轴的长度。经过卷积后得到n个特征映射，对这些特征映射进行降采样，池化窗大小为p*q。根据不同的任务判断是否需要池化，在本文的关键词检测模型中没有进行池化操作。

基于卷积神经网络的声学模型，最后一层为softmax层，其输出的每一个结点对应一个关键词标签或者非关键词的标签，输出值为某关键词或非关键词（未知词）的后验概率估计值。在我们的关键词检测系统中，CNN的输出层有12个结点对应12种标签，分别对应10种关键词、“非关键词”以及静音。原始的后验概率估计值带有噪声，因而需要对后验概率估计进行平滑，之后计算一个平滑窗内
的平滑置信度，比较这12种标签经过平滑处理的置信度打分，最大置信度分数对应的标签就是系统预测的关键词或“非关键词”。

2.3 后验概率处理

基于CNN的声学模型在输出层给出每个标签的后验概率，这个后验概率值是每一帧音频对应标签的后验概率。声学模型直接输出的后验概率有很多噪声，因而需要在一个固定时长的平滑窗

内对后验概率值进行平滑。平滑公式为：

(2.1)

其中，　表示第i个标签，j表示第j帧，　　表示平滑后的后验概率，

表示平滑窗第一帧对应的下标。

CNN输出了计算每个标签的帧级别置信度，需要将一定帧数的后验概率组合起来，计算出一个大小为
滑动窗内的后验概率值即置信度分数。置信度计算公式为：

(2.2)

计算第j帧之前一个滑动窗大小的置信度，其中，
。如果是单个关键字，当置信度超过预先设定的阈值时，模型认为这段音频包含关键词；如果是多个关键词和非关键词分类，选取最大的置信度对应的标签，模型认为这段音频包含该关键词或非关键词。

2.4 语音关键词检测基线系统

2.4.1 基线系统介绍

在设计实验时，我们参考了谷歌研究关键词检测任务，对12类不同的标签进行分类，12类标签如下：yes、no、up、down、left、right、on、off、stop、go、unknown和静音，其中unknown表示“非关键词”所有与前面10个关键词不同的语音都被认为是非关键词。训练数据和谷歌一致，采用谷歌开源的数据集Speech Commands Dataset，模型参考谷歌提出的CNN结构 * MERGEFORMAT [34]。

本文的数据集采用的是谷歌2018年4月发布的开源数据集Speech Commands Dataset，第二版与2017年8月份发布的第一版相比，第二版的数据集包含了更多类型的关键词，是一个专门为英文关键词检测任务制作的数据集。Speech Commands Dataset数据集包含了105，829个时长为1秒的单个单词的音频和6种背景噪声（包括高斯白噪声、粉红噪声、厨房场景噪声、跑步场景噪声、骑行自行车噪声、动物叫声噪声），这些音频由上千个志愿者录制并且谷歌进行筛选、标注，具体收集过程详情见参考文献 * MERGEFORMAT [41]。

Speech Commands Dataset包含35种关键词，我们的实验使用了上述十种关键词作为待识别的对象，其余backward、tree、three等25种关键词我们用作测试使用。我们按照8:1:1的比例将纯净的数据集分为训练集、验证集和测试集，将这些语音分别以信噪比-3、0、3、6dB和噪声混合。噪声使用Speech Commands Dataset自带的6种噪声，噪声时长达一分钟，因此合成时我们从噪声时长中随机选择1秒和纯净语音混合，得到的带噪语音数量是纯净语音数量的24倍，在训练时，还需要在训练集和验证集中加入一倍的纯净语音。在没有混响的条件下，语音传播符合加性噪声假设，因此采用人工合成的数据与真实数据有较好的一致性，使用人工合成的数据训练的增强模型在真实场景下依然能够很好地保持其性能。

为了更加公平准确地的测试模型，我们还在测试集中加上另外25种在训练集中未出现的关键词，以同样的方式混合噪声。最终测试集包含21万句音频，其中10种待检测关键词和25种干扰关键词的比例为1:1.3。为了进一步测试模型的泛化能力，我们还设计了噪声不匹配的测试集，使用了100种模型从未见过的噪声和纯净的测试集语音混合，噪声不匹配的测试集包含约360万句音频。所有音频在进一步处理前都降采样到16 kHz。

基线系统的特征采用的是40维的MFCC声学特征，窗长为30 ms，帧移为10 ms。选择30 ms窗长的原因是想要获取更丰富的上下文信息。训练基线系统的损失函数是交叉熵，交叉熵在多分类问题中应用非常广泛。优化器使用的是Adam优化器，学习率设置为1*e-4。基线系统的网络参数设置如表2.1所示。

表 2. 1 基线系统网络参数设置信息

Table 2.1 Configuration information of the baseline network parameters

类型	m	r	n	s	v	p	q
conv	20	8	64	1	1	-	-
max-pool	-	-	64	2	2	2	2
conv	10	4	64	1	1	-	-
linear	-	-	12	-	-	-	-
softmax	-	-	12	-	-	-	-

基线系统共有2层卷积层和一层全连接输出层，m*r代表卷积核大小，n为通道数，s和v表示步长，p*q为池化核大小。模型的参数量和计算复杂度与最终输出分类类别数相关，在本研究中，基线系统的参数量为493.7K，计算复杂度为95.9M次乘加。

训练基线系统时，我们直接使用带噪的音频训练CNN，这种训练方式叫做多环境训练（Multi-condition Training）。多环境训练策略是提高模型噪声鲁棒能力应用最广泛的一种方式，训练完成的模型本身具有很强的鲁棒能力。在之后的实验中，基于多环境训练策略的基线系统作为我们提出方法的对比对象。

2.4.2性能评价指标和性能测试结果

目前，在关键词检测领域主要是以分类准确率(Accuracy)、分类结果混淆矩阵(Confusion Matrix)、ROC曲线（Receiver operating characteristic curve , ROC）作为评价指标来衡量模型的识别性能。分类准确率并不能完全衡量模型的性能，例如模型将多少正例划分为反例。分类结果混淆矩阵比较全面的展示了模型的性能，分类结果混淆矩阵如表2.2所示。

表 2. 2 分类结果混淆矩阵

Table 2.2 Confusion matrix of the classification result

真实情况	预测结果	
	正例	反例
正例	TP（真正例）	FN（假反例）
反例	FP（假正例）	TN（真反例）

对于二分类问题，可将样例的真实类别与预测类别的组合划分为真正例（True Positive, TP）、假正例（False Positive, FP）、真反例（True Negative, TN）、假反例（False Negative, FN）四种情形，四者之和为样例总数。分类准确率可以表示为：

(2.3)

在实际应用中，关键词检测的任务比较看重模型将正确的预设关键词和非关键词识别出来的能力，例如在关键词的应用之一唤醒任务上，我们希望当我们说出唤醒词唤醒设备时，它能准确被唤醒，这叫做“真正例率”（True Positive Rate, TPR）或召回率（Recall Rate），可以表示为：

(2.4)

当我们说一些发音近似唤醒词但不是唤醒词的语句时，它能够不被唤醒，这叫做“假正例率”（False Positive Rate, FPR）或虚警率（False Alarm Rate, FAR），表示为：

(2.5)

ROC曲线可以综合描述这两种能力。ROC曲线源于“二战”中用于敌机检测的雷达信号分析技术，二十世纪六七十年代开始被用于一些心理学、医学检测应用中，此后被引入机器学习领域。显示ROC曲线的图称为“ROC图”，图2.4为一个ROC曲线的示意图。

ROC曲线的横轴为FPR，纵轴为TPR，因而曲线越“偏向”左上角表示性能越好。若一条ROC曲线被另一条ROC曲线完全“包住”，则可断言后者ROC曲线对应的识别器性能更优；若两条ROC曲线发生交叉，则无法轻易判断两者对应的识别器孰优孰劣。此时比较合理的判断依据是比较ROC曲线下的面积，即AUC（Area Under ROC Curve），取值范围为[0, 1],图2.4中AUC面积为0.79。黄线和绿色虚线的交点对应的阈值是等错误率(Equal Error Rate , EER)，表示 ，是一个折衷的值。

在本文中，基线系统的性能评估指标使用准确率、ROC曲线、AUC和EER。但在关键词领域，与常见的ROC曲线不同，ROC曲线的纵坐标为 ，即错误拒绝率（False Rejection Rate, FRR）。因此在我们的实验评估中，关键词检测系统的ROC曲线越“偏向”左下角，AUC和EER越小表示性能越好。

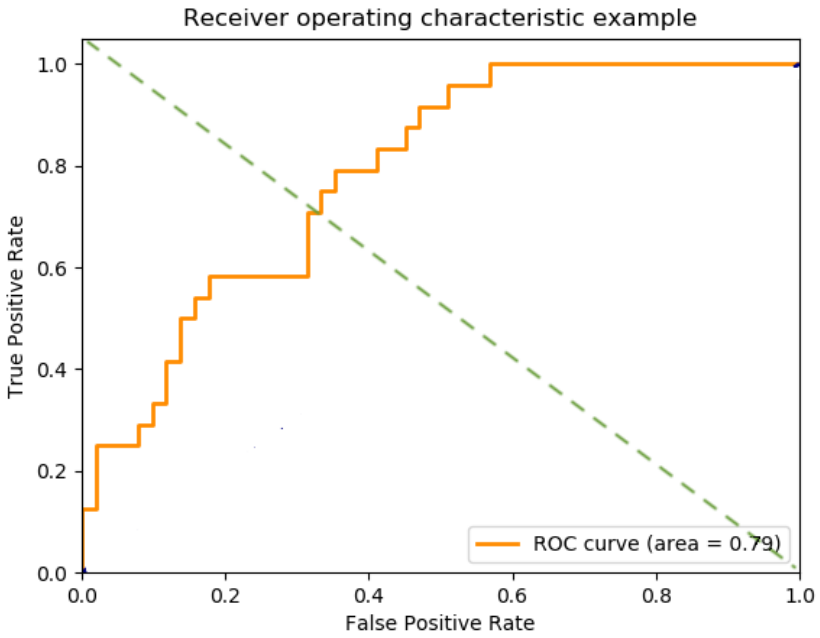

图 2. 图_2 * ARABIC 4 ROC曲线样例

Figure 2.4 An emaple of ROC

我们在测试集上测试了带噪训练和纯净语音训练的识别模型，结果如表2.3和图2.5所示。noise_baseline表示带噪语音训练的模型，而pure_baseline表示纯净语音训练的模型，从表2.3中可以直观的看到，背景噪声对关键词检测任务形成很大干扰。纯净语音训练的模型在纯净语音测试集的准确率约为90%，但在带噪测试集上的识别准确率急剧下降至72.46%。基于多环境训练策略的noise_baseline模型，对噪声有一定的鲁棒能力，相对于纯净训练的模型，其准确率提升了约6%，从图2.5中也可以发现noise_baseline模型优于pure_baseline模型，因此本文基线模型采用noise_baseline。noise_baseline在噪声匹配的测试集上准确率为80.89%，在噪声不匹配的测试集上准确率为68.81%，我们提出的方法会在此基础上进一步提高关键词检测系统对噪声的鲁棒能力，提高对未在训练集中出现的噪声泛化能力。

表 2. 3 基线系统实验结果

Table 2.3 The result of baseline experiments

模型	结果		
	Accuracy(%)	AUC(%)	EER(%)
noise_baseline	80.89	1.99	7.28
pure_baseline	72.46	6.05	13.68

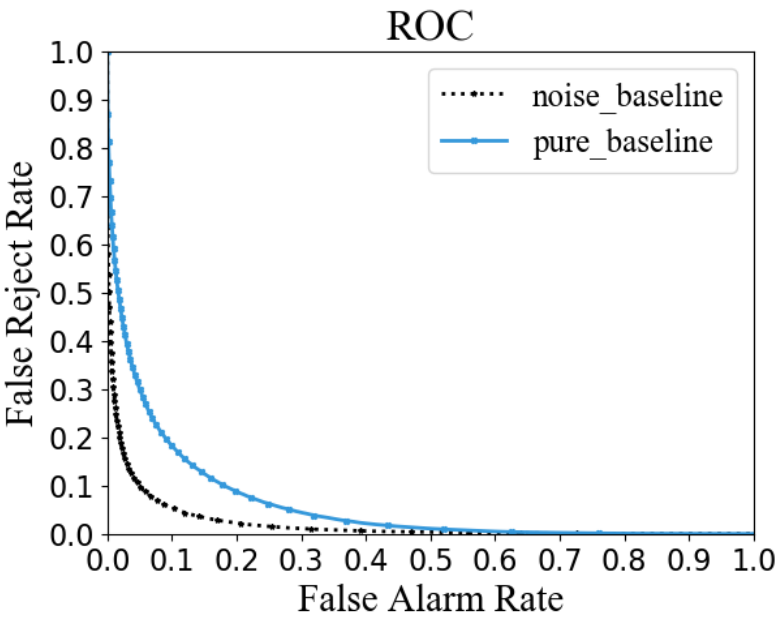

图 2. 5 基线系统的ROC曲线
Figure 2.5 ROC curve of the baseline

2.5 本章小结

本章主要对关键词检测系统进行了概述，对如何构建一个完整的关键词检测系统进行了细致描述。特征提取部分主要介绍了分帧加窗等预处理操作和MFCC特征提取过程。声学模型模块主要介绍了基于CNN的计算过程。之后对如何计算标签的置信度得分进行了描述。此外，介绍了主流的语音关键词检测系统性能评估指标。最后，构建了基于CNN的关键词检测基线系统，并对比了纯净语音训练策略和多环境训练策略，确定基于多环境训练策略的模型为基线模型。

第三章 基于深度学习的有监督的语音增强方法

3.1 有监督语音增强方法

语音增强问题可根据通道数分为单通道增强技术以及多通道增强技术，本文主要研究了单通道增强方法。单通道语音增强方法包括传统的信号处理方法以及掩膜估计方法，其中基于有监督学习的掩膜估计方法是近几年兴起的。掩膜估计的语音增强方法源自计算听觉场景分析，以掩膜估计作为目标，使用估计出的掩膜对带噪语音进行掩蔽，从而增强语音质量。有监督学习的方法将
基于掩膜估计的语音增强看作回归问题，根据带噪语音对人工构造的理想掩膜进行学习，从语音特征中学习语音、发声人和背景噪声之间的模式。

语音增强的目的是构建一个可靠的语音增强模型，在识别任务的前端对噪声进行过滤。按照机器学习的研究框架，模型包括训练和测试两个阶段，如图3.
1所示。语音增强系统的输入是带噪的混合语音，输出目标是各种时频掩膜或者目标语音，不管输出是时频掩膜
还是目标语音，必须预先知道和噪声混合的目标语音（纯净语音），因而必须预先人工合成带噪语音而无法使用实际带噪语音。将目标语音和噪声按照一定的信噪比可以得到混合语音。在训练阶段，目标语音作为模型训练的标签，目标语音和混合语音提取特征后送入模型中进行训练，在本文中，模型的训练目标是隐式理想比率掩膜，隐式理想比率掩膜中蕴含着人声和噪声的掩蔽关系。
在测试阶段，混合语音提取特征后送入增强模型，和增强模型预测的隐式掩膜相乘后可以得到降噪的特征，进而再送入识别模型进行识别任务。

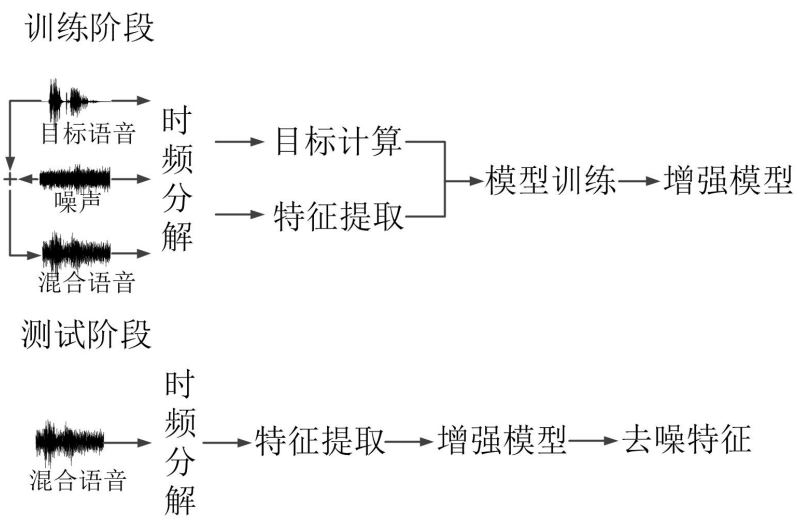

图3. 1 基于深度学习的语音增强结构图

Figure 3.1 A block diagram of the speech enhancement system which is based on deep learning

时频分解可以将输入的一维时域语音信号分解成二维的时间-频率信号，时间-
频率域同时域一样是表示语音信号的方式，与语音信号本身包含相同的信息，因而没有将时频分解归类在特征提取模块中。特征提取模块的输入是时间-
频率域的语音信号，输出是提取的特征。近年来，语音增强研究中使用的特征随着学习器或者说增强模型的能力不断增强，呈现由复杂到简单的趋势。早期
使用信号处理方式增强带噪语音时，由于学习器的能力不足或者当时计算机算力受限，往往需要设计极其复杂的语音信号特征提取算法。目前，较新的工作更倾向于使用简单的特征，尤其是在基于深度学习的有监督的语音增强任务中，短时傅里叶变换（Short Time Fourier Transform,
STFT）的能量谱特征受到研究者们的青睐。实际上，STFT是语音的时间-频率域表示，因而可以认为是
直接使用原始语音数据作为输入特征。在不同的识别任务中，对特征的要求不一样，还需要结合具体的任务加以分析。

3.2 训练目标

在有监督的语音增强模型训练中，掩膜是常用的训练目标。掩膜是一个非常重要的概念，它能够将带噪语音的频谱转换为纯净语音的频谱。它定义在短时频谱上，设纯净语音的连续波形信号为 ，其离散采样得到的序列为 ；设带噪语音的连续波形为 ，其离散采样得到的序列为

。短时傅里叶变换操作符STFT为如下过程：

1、将离散信号　　进行分帧，按照每窗长（　　）为一帧，帧之间重叠为

2、对分帧后的信号进行加窗处理，得到加窗后的信号

3、对每一帧加窗后的信号进行离散傅里叶变换：

(3.1)

4、将每帧的傅里叶谱按照顺序排在一起构成短时傅里叶谱

，其幅度为　　，其相位为　　，能量谱为　　。

逆短时傅里叶变换操作符iSTFT为如下过程：

1、对每一帧的傅里叶谱进行逆傅里叶变换：

(3.2)

2、对每一帧的波形进行加窗，得到加窗后的单帧波形

3、将加窗后的每一帧进行重叠相加操作，重叠的部分相加，不重叠的部分拼接，得到合成的波形

。

对于掩膜　　和带噪语音的频谱　　，使用如下的公式可以获得纯净语音频谱　　的估计

(3.3)

其中　　表示矩阵点乘。使用纯净语音频谱的估计

进行逆短时傅里叶变换即可得到对于纯净语音波形的估计　　。

现有的掩膜类型很多 * MERGEFORMAT [42]，其中幅度谱类的掩膜有理想二值掩膜（Ideal Binary Mask, IBM）、理想比率掩膜（Ideal Ratio Mask, IRM）、类维纳滤波掩膜（Wiener Like Mask）以及频谱振幅掩膜（Spectral Magnitude Mask, SMM）。

理想二值掩膜是最先开始使用的训练目标，公式表示如下：

(3.4)

在一个时频单元中，如果信噪比（Signal-to-Noise Ratio, SNR）超过预设的阈值，则在这个时频单元中IBM值为1，否则，IBM值为0，没有特殊情况的话，阈值一般设置为零。最优化准则为最大化SNR，其中　　。

理想比率掩膜，又被译为理想浮值掩膜，是目前应用最广泛的训练目标，它表示目标语音能量在混合的语音和噪声中所占的比重，可以看作是一个“软”的理想二值掩膜，常与损失函数均方误差配合，定义为：

(3.5)

最优化准则为最大化SNR，其中 。类维纳滤波掩膜和理想比率掩膜相似，当理想比率掩膜公式中β为2时，类维纳滤波掩膜和理想比率掩膜只差一个根号，最优化准则同理想比率掩膜一样定义如下：

(3.6)

频谱振幅掩膜表示目标语音能量在带噪语音中所占的比重，取值范围为0到正无穷，定义为：

(3.7)

相较于理想浮值掩膜，频谱振幅掩膜更难估计。最优化准则是要准确的估计 ，其中 。

以上介绍的掩膜可以根据刻画对象不同分为两大类：基于时频掩膜的（masking-based）和基于频谱映射（mapping-based）的。基于时频掩膜的训练目标刻画的是目标语音与干扰噪声之间的关系，即刻画掩膜；而基于频谱映射的训练目标刻画的是目标语音的性质，即刻画纯净语音。后一类掩膜可以对识别系统进行整体优化，从而为系统带来额外的性能提升，提高系统的背景噪声鲁棒能力。

3.3 模型训练策略

在第二章中，我们简单介绍了多环境训练策略，随即直接使用带噪语音训练关键词检测模型。这种训练方法使用简单，应用非常广泛，并且基于多环境训练策略的关键词检测系统本身就具备一定的背景噪声鲁棒能力。为了在此基础上继续提高模型在噪声环境中的识别性能，我们从鲁棒性自动语音识别领域借鉴并尝试了多种训练策略，目的是为了使语音增强模型和语音关键词检测模型互相影响，互相优化。在噪声环境下的语音关键词研究中，我们首次调研了多种训练策略，并针对关键词检测任务本身的特性，对训练策略进行了改进。

第一种策略是分别训练好语音增强模型和关键词检测模型，即固定住语音增强模型和关键词检测模型，在测试阶段直接将增强模型放在关键词检测模型的前端，使得带噪语音先经过增强模型进行降噪，之后将纯净的语音特征送入关键词检测模型进行识别。在本文中，我们把这种模型简称为SE+KWD。SE+KWD策略能够提高识别器的识别性能，但是有一个巨大的缺陷，语音增强模型的训练目标几乎都是以最大化信噪比为准则的，在这一准则下语音的可懂度和语音真实性无法得到保证，可能导致语音失真，反而使得语音关键词检测模型无法准确识别。

考虑到SE+KWD策略的缺陷，第二种策略是使用经过增强模型降噪之后的特征训练和测试关键词检测模型。首先先训练好一个语音增强模型，随即固定住增强模型，在关键词检测模型的训练和测试阶段，都直接将训练好的语音增强模型放在语音关键词检测前端，语音关键词检测模型的输入特征是降噪之后的语音特征。本文中，我们称这种训练策略为Retraning+KWD，即再训练识别器。Retraining+KWD策略优于SE+KWD策略，但是Retraining+KWD策略太过依赖于语音增强模型的性能。

语音增强模型和语音关键词检测模型并不是两个完全独立的任务，两个模型应该互相影响互相优化。第三种策略是联合训练，即将两个模型联合起来生成更大更深的模型，通过反向传播算法联合训练两个模型，语音关键词检测模型中的语言学和其他有用的信息可以传送给语音增强模型，使两个模型彼此优化。在本文中，我们称这种训练策略为Joint+KWD，在我们的所有试验结果中，联合优化都展示了远远优于其他训练策略的性能。与多任务的训练不同的是，我们有针对性地只使用关键词检测模型的损失函数作为优化的目标，具体的Joint+KWD的细节会在之后的实验介绍中阐述。

3.4 性能评估

为了方便、客观地对不同模型增强后的语音进行度量和比较，使用短时客观可懂度打分（Short-Time Objective Intelligibility score，STOI）来客观度量语音的可懂度 * MERGEFORMAT [43]，STOI表示了纯净语音和增强语音在短时时域包络上的相关性，同时表现出了与人类可懂度打分很强的相关性。使用语音质量感知评估（Perceptual Evaluation of Speech Quality，PESQ）来客观度量语音质量 * MERGEFORMAT [44]，同STOI一样PESQ也是通过对比增强后的语音与纯净语音获得的，STOI的得分范围为

，PESQ的得分范围为 。

除了语音可懂度与语音质量的评价指标外，信噪比的提升也是语音增强一个比较重要的评价指标，Wang * MERGEFORMAT [45]
提出在不同的掩膜作为增强的目标时，信噪比的计算应当使用基于目标的信噪比计算公式，比如当IRM作为学习目标时，应当使用理想的IRM对带噪语音的短时幅度谱进行掩膜，并结合带噪语音的相位信息合成纯净语音，以这样合成的纯净语音作为计算信噪比时的对比目标，而不是直接使用纯净语音作为对比目标，对于不同的掩膜使用不同的合成语音作为计算信噪比时的对比目标。Pascual的模型是一种End-To-End的模型，它直接得出对于纯净语音波形的预测，因此需要直接使用传统的信噪比计算方法，而我们的模型在学习时同时考虑幅度信息和相位信息，但是没有理想的掩膜作为对比目标来计算信噪比。尽管如此，这三种模型理想情况下（预测误差为0），都有能力得到纯净语音，所以我们直接使用传统的信噪比计算方式来计算信噪比，即使用纯净语音与剩余噪声（distortion）进行对比：

(3.8)

在每一种信噪比下对比模型性能。

在将增强模型应用于关键词检测等识别任务时，除了上述指标外，通常会将增强模型为识别模型带来的性能提升作为重要的评估指标。

3.5 基于Joint-BiLSTM的鲁棒关键词检测系统

3.5.1 系统框架

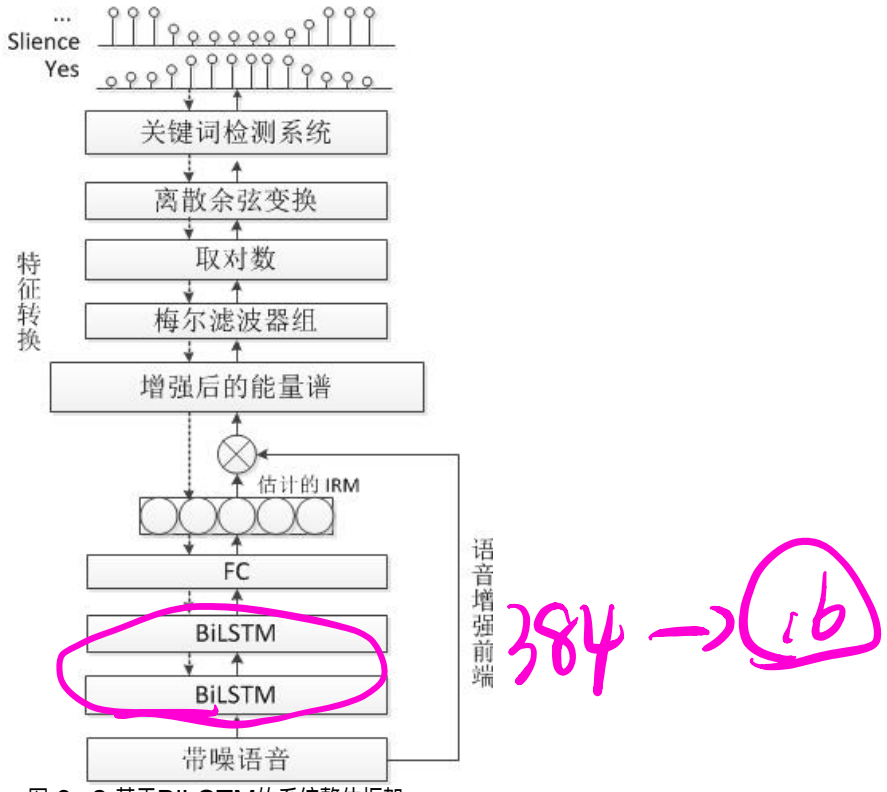

图 3. 2 基于BiLSTM的系统整体框架
Figure 3.2 Schematic diagram of the system based BiLSTM

整个系统框架如图3.
2所示。在系统中主要有三个部分：语音增强前端、特征转换模块、关键词检测模型，这三个部分连接在一起形成了一个更大更深的模型。在这个结构上，我们可以尝试不同的训练策略。我们把增强后的能量谱（幅度谱）特征转换为MFCC特征单独作为一个模块，在这个模块中可以替换成不同的特征转换方式。语音增强前端可以通过计算时频掩膜重构没有噪声的能量谱
。图中的实线表示神经网络前向计算的信息流动过程，虚线表示反向传播时，神经网络梯度信息的流动过程，例如在联合训练时，梯度信息可以从关键词检测模型传递给语音增强模型，使得增强模型训练出的时频掩膜更容易被识别模型识别。

3.5.2 实验设置

关键词检测系统的输入特征为MFCC，输出为某一个类别的后验概率，通过计算平滑窗内的置信度得分来“推断”关键词类别。识别模型使用的网络结构为第二章中提到的CNN模型，增强模型为2层BiLSTM，每层384个节点。训练CNN的损失函数是交叉熵，交叉熵在多分类问题中应用非常广泛。训练BiLSTM模型的损失函数为均方误差，优化器使用的是Adam优化器，学习率设置为1*e-4。Batch size设置为256。

实验使用了三种训练策略进行训练，第一种策略是在测试阶段直接将增强模型放置在关键词检测模型前端。首先分别用同一份带噪的语音数据训练CNN识别模型和BiLSTM增强模型，在测试阶段连在一起，记为BiLSTM+KWD模型。其中BiLSTM模型的损失函数如下：

(3.9)

为目标掩膜，
为BiLSTM模型预测的掩膜，t和f分别表示时间和频率，T表示总帧数，F表示时频单元总数。在这种训练策略中，计算的是显式掩膜的均方误差。CNN模型的损失函数为交叉熵，多分类的交叉熵公式如下：

(3.10)

m为一个batch中的样本总数，n为类别总数，y为真实类别标签，　　为识别模型预测类别标签。

第二种策略是利用降噪后的声学特征重新训练CNN识别模型，在本研究中记为BiLSTM+Retraining。首先训练BiLSTM增强模型，将训练好的增强模型接在识别模型前端，重新训练并测试CNN识别模型的性能，两个模型的损失函数同前一种策略保持一致。

第三种策略是联合训练，将两个模型结合成一个更大的模型，利用反向传播将语言学信息和其他有用的信息从CNN识别模型传递到BiLSTM增强模型中。在本研究中，记为BiLSTM+Joint模型。联合训练的损失函数设置与上述两种策略不同。一般来说，常见的联合训练有两种方式：

1. 将两个模型结合在一起，损失函数设置为均方误差损失和交叉熵损失之和，他们之间的比值由一个参数来调节，公式如下：

(3.11)

Loss是整个大模型的优化目标，　　是识别模型的损失函数，
是增强模型的损失函数。这种方式增强模型和识别模型都是从随机初始化开始训练。

2.

第二种方式需要将识别模型和增强模型分别进行预训练，再将两个模型融合在一起，整个模型的损失函数不再使用增强模型的均方误差损失，只是用识别模型的交叉熵损失函数。这种方式，更具识别目标导向性，因为训练增强模型的目的就是为了提高识别模型的性能。本文所使用的联合训练框架采用这种融合方式。

3.5.3 实验结果与分析

本节实验对比基线系统和加上增强模型的系统，还在此基础上对比了几种训练策略。在噪声匹配的测试集上结果如表3.1所示。

表 3. 1 测试集实验结果

Table 3.1 The result of experiments

模型	结果		
	Accuracy(%)	AUC(%)	EER(%)
noise_baseline	80.89	1.99	7.28
BiLSTM+KWD	87.64	1.30	6.66
BiLSTM+Retraining	90.18	1.17	5.92
BiLSTM+Joint	91.64	1.10	6.15

从表中可以看出，我们训练的增强模型显著提升了关键词检测模型的识别性能，大大提高了关键词检测模型的背景噪声鲁棒能力。对比基线系统noise_baseline模型和直接增强模型的BiLSTM+KWD，识别模型的准确率绝对提升了将近7%，从AUC和EER指标上我们也能看出明显的提升。对比几种训练策略，可以得出联合训练策略优于其它两种策略的结论，比基线系统绝对提升接近11%，相对提升56%，极大地提升了关键词检测系统的识别率。如图3.3所示，从ROC曲线中我们能更明显的看出四种模型之间的性能差距。

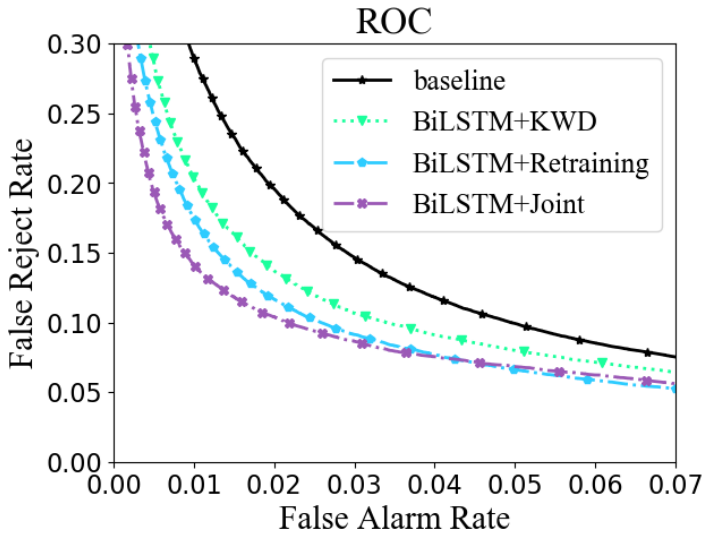

图 3. 3 测试集ROC曲线

Figure 3.3 The ROC curve of experiments

实际生活中，我们希望关键词检测模型能够在较低的误唤醒率（虚警率）的情况下尽可能降低错误拒绝率，从图3.3中可以看出，BiLSTM+Joint模型误唤醒率百分之二以下时显著优于其它模型，更加突出了联合训练策略的优势。

尽管基于BiLSTM的增强模型具有很好的降噪效果，但是几乎没有使用价值，因为在关键词检测任务中，模型的计算资源需求受到严格的限制，在提升识别性能的同时，必须以最小的计算资源消耗为前提。表3.2中列出了模型的参数量和计算复杂度。

表 3. 2 基线系统和BiLSTM的参数量以及计算复杂度

Table 3.2 The number of parameters and multiplies used for the baseline and BiLSTM

模型	计算资源占用评估	
	参数量	计算复杂度
noise_baseline	493.7K	95.87M
BiLSTM	5661.0k	432.7M

从表3.2可以看出，BiLSTM模型所需的参数量是基线系统的十几倍，计算复杂度也多了3倍，性能提升所带来的计算代价太过昂贵，这是当前任何移动智能设备都不能接受的。针对这个缺陷，在第四章中，我们提出了一种低资源占用的语音增强方法，能够保持和BiLSTM模型一样的甚至更好的增强性能，同时，参数量和计算量都要比BiLSTM显著降低。

3.6 本章小结

本章我们介绍了基于深度学习的有监督的语音增强系统流程和研究方法，介绍了监督学习的训练目标、识别模型和增强模型结合的多种训练策略，并讨论了常用的性能评估指标。在最后一小节中，我们对比了基线系统和在前端加上降噪模块两类
模型，增强模型可以显著提高识别模型的性能；除此之外，我们还对比了几种训练策略，三种训练策略都显著提升了识别模型的性能，其中以联合训练策略优化效果最好。最后我们讨论了基于BiLSTM增强模型的缺陷。

第四章 基于CRN的低资源需求的鲁棒关键词检测系统

4.1 低资源占用的卷积循环神经网络

对于一个具有实用价值的鲁棒关键词检测系统，应该满足以下要求：较好的人声和噪声泛化能力、尽可能低的错误拒绝率和虚警率以及较高的实时性、较低的计算资源消耗。一个处理噪声泛化能力的非常简单有效的方法是使用不同的噪声进行训练 * MERGEFORMAT [46]
，即多环境训练。同样地，也可以在训练数据中加入多个说话人的数据来提升模型对说话人的鲁棒性。但是经研究发现DNN没办法追踪某个特定的说话人，因为DNN是每一帧独立判断的，不包含前后文的信息 * MERGEFORMAT [47]。最近的研究发现 * MERGEFORMAT [48]
，为了利用上下文信息，将语音增强表示为序列到序列的映射会更好，即使用RNNs捕捉上下文信息。

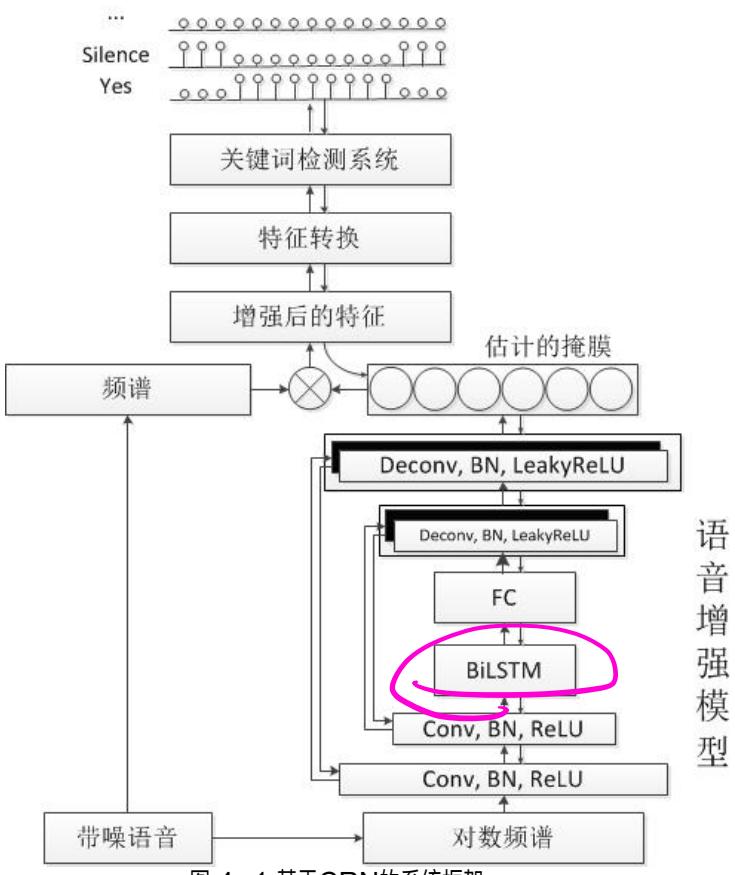

图 4. 1 基于CRN的系统框架

Figure 4.1 Schematic diagram of the system based CRN

在上一章中，我们使用BiLSTM模型得到了很好的增强效果，但是需要较多的计算资源，系统的延迟也比较高。受到CRN相关研究 * MERGEFORMAT [49]
的启发，我们设计了一种新颖的CRN结构，在CRN结构中组合了CNN编码器和长短时记忆层。在设计时考虑到计算资源占用的问题，我们设计的CRN非常简单。除此之外，我们还研究了特征转换模块，并对比了能量域特征和梅尔域特征。在实验中，我们跟基线系统做了对比，并且对比了三种训练策略。系统整体结构如图4.1所示。

框架图右下角即本文提出的CRN结构，它采用CNN进行编码解码，解码器之后是一层sigmoid输出层。编码器是堆叠了几层卷积层（两种特征域的系统结构卷积层数目不同），从原始输入中提取高层次的特征信息，解码器的结构跟编码器相对应，这保证了系统的输出跟输入有相同的维度。这种卷积编解码的机制使得我们可以把能量谱当作一张图片来对待。编码器每一层的激活函数我们使用的是修正线性单元（Rectified Linear Unit,
ReLU），解码器每一层的激活函数我们使用的是带泄露修正线性单元（Leaky Rectified Linear Unit, Leaky ReLU），在编码部分使用ReLU是为了得到尽可能少的编码，在解码部分选择使用Leaky ReLU的原因是Leaky ReLU相较于ReLU而言，其梯度处处不为零，从而能够更好的训练编码部分。在激活函数之前都使用了批正则化（Batch Normalization, BN）。卷积核的数量设置具体可见参数表4.1。

在卷积编码解码结构中无法利用上下文信息，而为了追踪目标说话人，利用上下文信息非常重要，因此引入了长短时记忆记忆网络。长短时记忆网络是循环神经网络的一种，目前已被应用在语音识别和音频分类领域，并取得了巨大成功。为了考虑语音的时序变化，我们在编码器和解码器之间加入一层双向LSTM。在本文中，双向LSTM由如下公式定义：

(4.1)

(4.2)

(4.3)

(4.4)

(4.5)

(4.6)

其中， 、 、 、 分别表示输入、块输入、记忆细胞和t时刻的隐层激活函数。 和 表示权重和偏置量。 表示sigmoid非线性激活， 表示向量元素相乘。

为了使编码器的输出维度符合LSTM的输入维度，我们压平了编码器输出的频率和通道维数，得到了一个序列的特征向量。LSTM的输出序列会重新改变维度来符合解码器的输入维度。通过将CNNs和BiLSTM组合，我们设计的CRN既包含了CNN强大的特征提取能力又包含了RNNs的时序建模能力。

4.2 能量谱特征和梅尔谱特征

系统中的理想比率掩膜定义在不同的时频域上。一般来说，在语音增强领域能量谱是一种常见的选择，但在鲁棒关键词检测的系统中，我们认为有更好的选择。在我们提出的关键词系统中，语音增强模型的输出会被送入关键词检测模型中，而关键词检测模型使用的特征是MFCC特征，因而能量谱中的频带被整合并通过梅尔滤波器组来提取MFCC特征。这意味着能量谱中的信息被压缩了。

因此在能量谱上做增强没有必要而且效率较低。另一方面，梅尔谱能够直接转换为MFCC，所以我们提出在梅尔谱上预测理想比率掩膜。带噪语音频谱Y、纯净语音频谱S和噪声频谱都定义在梅尔域上，理想比率掩膜的公式为：

(4.7)

为了从频谱中提取MFCC特征，我们设计了特征转换模块，如图4.2所示。将梅尔谱转换为MFCC需要先进行离散余弦变换（Discrete Cosine Transformation, DCT）。为了进行对比，我们也构建了基于能量域频谱的语音增强模型，首先将增强后的能量谱特征送入梅尔滤波器组，再进行离散余弦变换。注意梅尔滤波器组滤波和DCT变换都可以用矩阵乘法实现，进一步可以在神经网络中表示为线性的网络层。因此，包含特征转换模块，所有的神经网络层都可以使用反向传播算法进行训练。

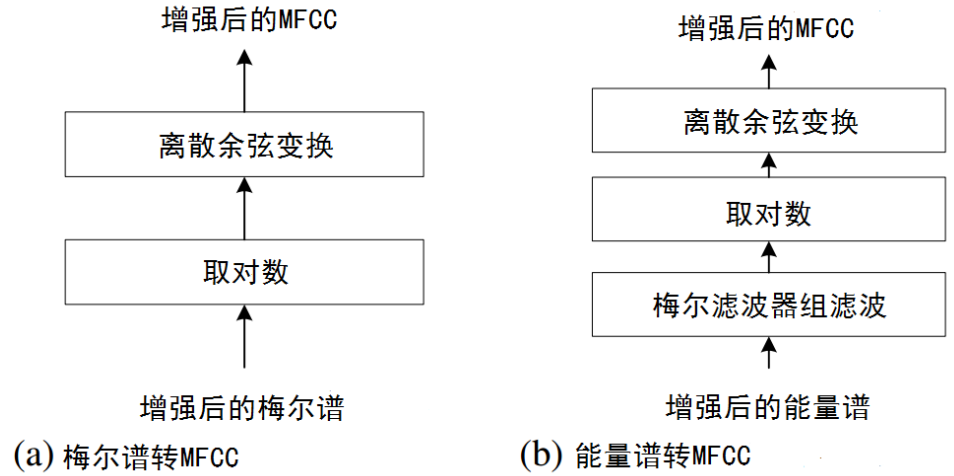

图 4．2 梅尔谱和能量谱的特征转换模块

Figure 4.2 The feature transformation block for (a) Mel-spectrogram and (b) power spectrogram

在研究中，我们一共设计了两种特征域的网络结构，其中每种特征域我们给出了完整版和压缩版两个版本。表4.1和4.3列出了完整的能量谱特征和梅尔谱特征的网络结构，其中卷积神经网络的通道数设置为32，BiLSTM的层结点数为64，记为PowCRN32和MelCRN32。表4.2和4.4为压缩的能量谱特征和梅尔谱特征的网络结构，其中卷积神经网络的通道数设置为16，BiLSTM的层结点数为32，记为PowCRN16和MelCRN16。

表 4．1 PowCRN32 网络结构

Table 4.1 Architecture of the PowCRN32

layer name	input size	hyperparameters	output size
reshape_1	T×241	-	1×T×241
conv2d_1	1×T×241	8,4,32	32×T/4×60
conv2d_2	32×T/4×60	8,4,32	32×T/16×15
reshape_2	32×T/16×15	-	T/16×15×32
BiLSTM	T/16×15×32	64	T/16×64
FC	T/16×64	15×32	T/16×15×32
reshape_3	T/16×15×32	-	64×T/16×15
deconv2d_2	64×T/16×15	8,4,32	64×T/4×60
deconv2d_1	64×T/4×60	9,4,32	32×T×241
conv2d_out	32×T×241	3,1,1	1×T×241
reshape_4	1×T×241	-	T×241

表 4．2 PowCRN16 网络结构

Table 4.2 Architecture of the PowCRN16

layer name	input size	hyperparameters	output size
reshape_1	T×241	-	1×T×241
conv2d_1	1×T×241	8,4,16	16×T/4×60
conv2d_2	16×T/4×60	8,4,16	16×T/16×15
reshape_2	16×T/16×15	-	T/16×15×16
BiLSTM	T/16×15×16	32	T/16×32
FC	T/16×32	15×16	T/16×15×16
reshape_3	T/16×15×16	-	32×T/16×15
deconv2d_2	32×T/16×15	8,4,16	32×T/4×60
deconv2d_1	32×T/4×60	9,4,16	16×T×241
conv2d_out	16×T×241	3,1,1	1×T×241
reshape_4	1×T×241	-	T×241

表 4．3 MelCRN32结构

Table 4.3 Architecture of the MelCRN32

layer name	input size	hyperparameters	output size
reshape_1	T×40	-	1×T×40
conv2d_1	1×T×40	4,2,32	32×T/2×20
conv2d_2	32×T/2×20	4,2,232	64×T/4×10
conv2d_3	64×T/4×10	(3,4),(1,2),4×32	128×T/4×5
reshape_2	128×T/4×5	-	T/4×20×32
BiLSTM	T/4×20×32	64	T/4×64
FC	T/4×64	20×32	T/4×20×32
reshape_3	T/4×20×32	-	256×T/4×5
deconv2d_3	256×T/4×5	(3,4),(1,2),64	128×T/4×10
deconv2d_2	128×T/4×10	4,2,32	64×T/2×20
deconv2d_1	64×T/2×20	4,2,32	32×T×40
conv2d_out	32×T×40	3,1,1	1×T×40
reshape_4	32×T×40	-	T×40

表 4．4 MelCRN16结构

Table 4.4 Architecture of the MelCRN16

layer name	input size	hyperparameters	output size
reshape_1	T×40	-	1×T×40

conv2d_1	1×T×40	4,2,16	16×T/2×20
conv2d_2	16×T/2×20	4,2,32	32×T/4×10
conv2d_3	32×T/4×10	(3,4),(1,2),64	64×T/4×5
reshape_2	64×T/4×5	-	T/4×20×16
BiLSTM	T/4×20×16	32	T/4×32
FC	T/4×32	320	T/4×20×16
reshape_3	T/4×20×16	-	128×T/4×5
deconv2d_3	128×T/4×5	(3,4),(1,2),32	64×T/4×10
deconv2d_2	64×T/4×10	4,2,16	32×T/2×20
deconv2d_1	32×T/2×20	4,2,16	16×T×40
conv2d_out	16×T×40	3,1,1	1×T×40
reshape_4	1×T×40	-	T×40

4.3 实验结果与分析

本节中，将对我们提出的几个模型与训练策略在噪声匹配和噪声不匹配的测试集上给出相关实验结果。实验语料描述可以参考2.4.1小节，本节综合了2.4.2小节和3.5.
3小节的实验结果。我们给出了所有训练模型在噪声匹配测试集上的实验结果，结果如表4.
5所示，评价指标为准确率、AUC和EER（相关概念见2.4.
2小节），其中准确率越高、AUC和EER值越小代表性能越好。cnn-trad-pool2是使用带噪语音训练的基线系统，其他模型为不同训练策略（策略解释见3.3节）和网络结构的模型。例如MelCRN32+
KWD表示使用第一种策略，直接将增强模型放在关键词检测模型的前端。MelCRN32+
Retraining表示固定预训练的增强模型重新训练关键词检测模型。MelCRN32 +
Joint表示第三种训练策略，将预训练的增强模型和关键词检测模型联合在一起，进行联合训练。表4.
6中给出了基于联合训练策略的模型在噪声不匹配的测试集上的实验结果。接下来，我们将围绕表4.5和4.
6给出的实验结果从五个方面对模型性能展开分析。表4.7给出了关键词检测模型和所有增强模型的计算资源需求。

表 4．5 所有模型在噪声匹配测试集上的实验结果

Table 4.5 The result of all models on matched test set

Model	Test accuracy(%)	AUC(%)	EER(%)
cnn-trad-pool2	80.89	1.99	7.28
BiLSTM+KWD	87.64	1.30	6.66
BiLSTM+Retraining	90.18	1.17	5.92
BiLSTM+Joint	91.64	1.01	6.15
PowCRN32+KWD	86.42	1.52	6.67
PowCRN32+Retraining	87.69	1.53	6.63
PowCRN32+Joint	91.07	1.20	6.27
PowCRN16+KWD	86.20	1.61	6.73
PowCRN16+Retraining	87.01	1.67	6.88
PowCRN16+Joint	90.68	1.22	6.50
MelCRN32+KWD	87.59	1.59	6.97
MelCRN32+Retraining	89.17	1.35	6.10
MelCRN32+Joint	93.17	1.19	6.20
MelCRN16+KWD	86.87	1.64	7.00
MelCRN16+Retraining	88.20	1.42	6.49
MelCRN16+Joint	92.56	1.28	6.39

表 4．6 基于联合训练策略模型在噪声不匹配测试集上的实验结果

Table 4.6 The result of all models which based on Joint-training strategy on unmatched test set

Model	Test accuracy(%)
cnn-trad-pool2	68.81
BiLSTM+Joint	73.74
PowCRN32+Joint	75.19
PowCRN16+Joint	72.49
MelCRN32+Joint	78.12
MelCRN16+Joint	75.67

表 4．7 关键词检测模型及所有增强模型的参数量和计算量

Table 4.7 The number of parameters and multiplies used for the KWD model and different enhancement models

Model	Parameters	Multiplies
cnn-trad-pool2	493.7K	95.87M
BiLSTM+Joint	5661.0K	432.7M
PowCRN32+Joint	724.0K	280.1M
PowCRN16+Joint	182.3K	73.0M
MelCRN32+Joint	881.3K	115.1M
MelCRN16+Joint	221.5K	29.2M

1、模型对比：从表4.5和图4.3中，我们可以看到所有的对比模型都优于基线系统cnn-trad-pool2。基于BiLSTM的模型性能很好，但是参数量大和计算复杂度高（见表4.7），不能满足低资源需求的要求。我们所提出的CRNs模型，不仅计算资源需求少，而且有着跟BiLSTM模型一样优秀甚至更好的性能。压缩的模型PowCRN16和MelCRN16对比BiLSTM模型也是有着比较不错的性能。

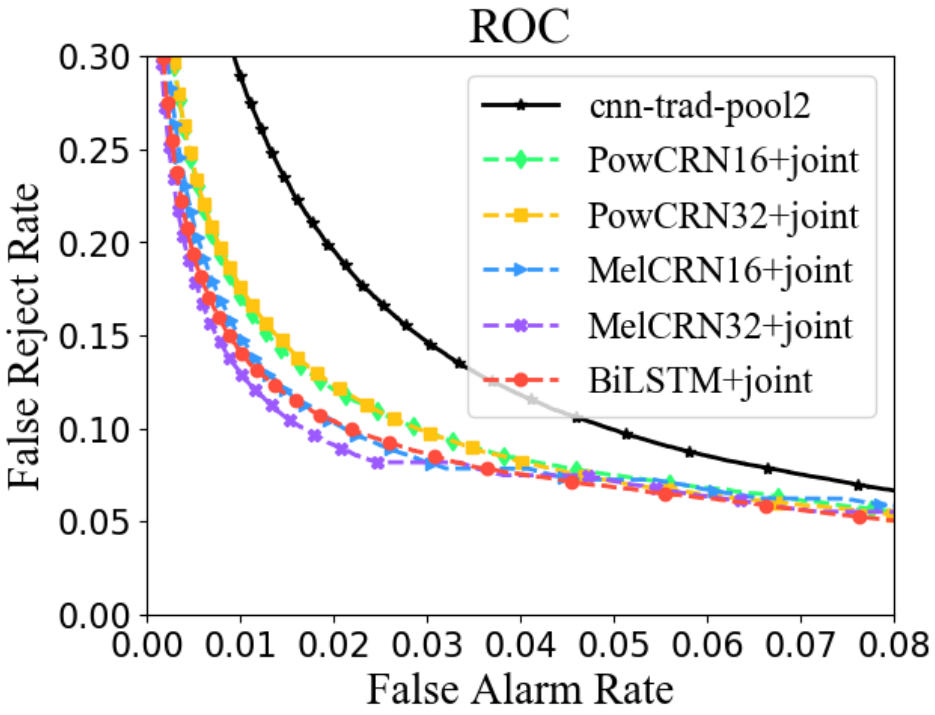

图 4．3 不同增强模型的ROC曲线

Figure 4.3 ROCs from the perspective of different enhancement models

2、训练策略对比：从表4.5和图4.4中，我们可以看到所有的基于多环境训练策略优于纯净语音训练，基于前端增强的模型都优于基于多环境训练策略的模型，基于联合优化的策略优于基于再训练和直接增强的策略，对于CRNs模型而言，联合优化的策略最有助于提升关键词检测系统的噪声鲁棒性。

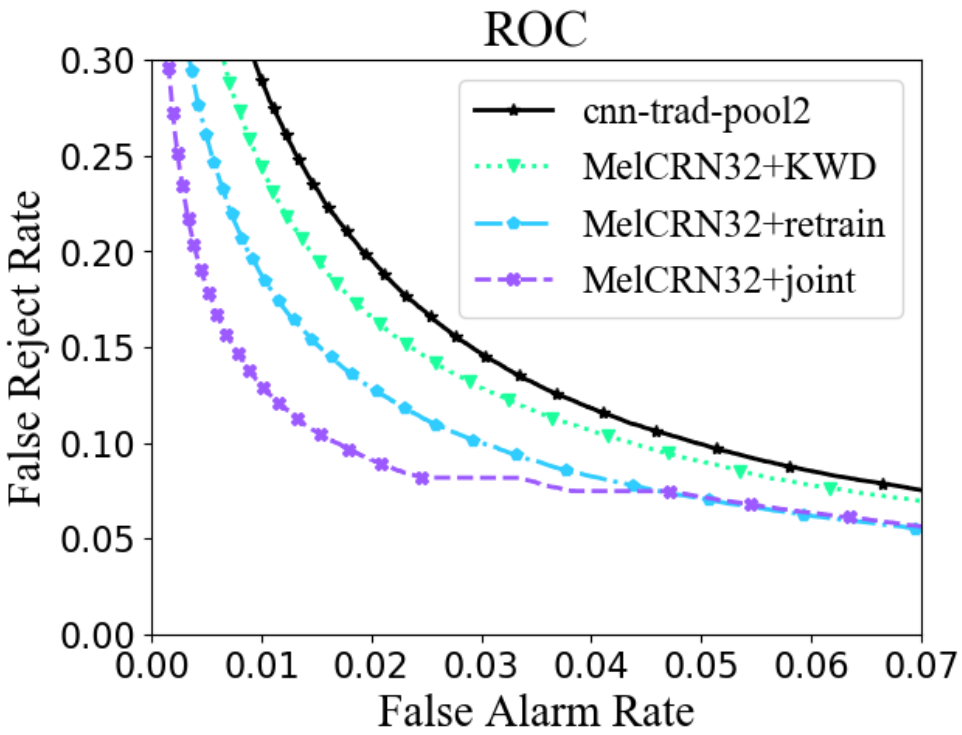

图 4. 4 不同训练策略的ROC曲线

Figure 4.4 ROCs from the perspective of different training strategy

3、梅尔谱和能量谱对比：从表4.5和图4.5中，两种特征域所需的计算资源是差不多的。我们认为梅尔谱的特征更加适合关键词检测系统。由于关键词检测系统的输入更多是静音、背景噪声和非语音，因此虚警率必须尽可能小。在虚警率尽可能小的时候，MelCRN32达到了最好的性能，这一结论也可以从压缩的MelCRN16看出。

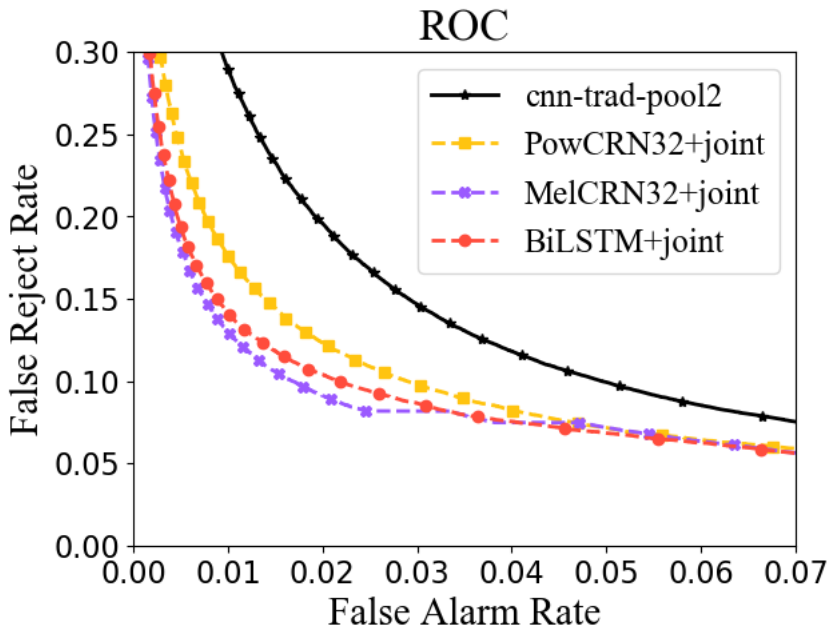

图 4. 5 不同特征域的ROC曲线

Figure 4.5 ROCs from the perspective of different feature domain

发音音标数量敏感程度对比：因为关键词的音标数量不一样，我们评价了增强模型对发音音标数量的敏感度。噪声匹配的测试集分成两类，一类是包含1-2个发音音标的关键词类别，另一种是超出两种发音音标的关键词。图4.5中显示的是AUC相对基线系统的减少量，可以看出，基于梅尔谱的方法相对于其他方法，它对音标数量敏感度较低。

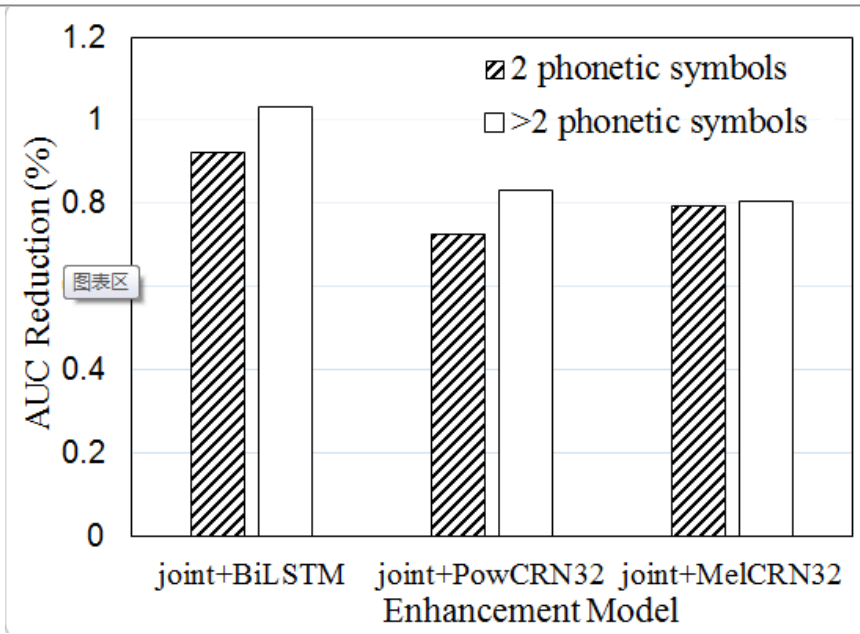

图 4．6 音标数量敏感性对比

Figure 4.6 Sensibility on phonetic symbol length

泛化能力对比：表4.6显示的是基于联合优化策略的方法在噪声不匹配测试集上的结果。从表4.6中可以看出MelCRN32和PowCRN32模型对于训练时从未见过的噪声比基于BiLSTM的模型泛化能力强。基于梅尔谱的模型优于基于能量谱域特征的模型。

4.4 本章小结

本章首先介绍了低资源占用的卷积循环神经网络，并将其应用于语音增强任务。引入了两种增强特征：能量谱和梅尔谱，分别针对两种特征设计了基于CRN的网络。对比了本实验中全部的模型，从模型类别、特征域、音标数量敏感、训练策略以及噪声泛化五个方面进行了分析。实验结果表明，本文提出的低资源占用的CRN在噪声匹配和噪声不匹配的情况下，与基于BiLSTM模型相比
，都有着相近甚至更好的性能，同时仅需要少量的计算资源。本章提出的基于梅尔域的增强特征在关键词检测任务上性能表现优于能量谱特征，并且对关键词的音标数量不敏感。

第五章 总结与展望

5.1 本文工作总结

关键词检测系统作为人机语音交互的接口，其重要程度不言而喻。随着各种移动手持智能设备的兴起，关键词检测系统应用场景愈加广泛，不同场景下背景噪声、无关人声的干扰成为关键词检测系统的挑战，人们对关键词检测系统的鲁棒性要求越来越高。不仅如此，关键词检测系统通常要保持一个“时刻监听”的模式，会给大多数以电池为电源的智能设备带来巨大的能量消耗，因而
关键词需要在尽可能降低关键词计算资源占用的要求下，提高模型的噪声鲁棒能力。因此，本文在设计模型时，以降低模型参数量和计算复杂度为原则。主要工作如下：

1、本文借鉴鲁棒自动语音识别任务，在鲁棒关键词检测任务上引入三种训练策略，其中基于联合优化的训练策略是首次应用在关键词检测领域。实验表明，直接加语音增强前端的训练策略和基于再训练的训练策略性能优于多环境训练策略，而基于联合优化的策略又优于前面这两种策略。对比基于多环境训练策略的基线系统，所有基于联合优化训练策略的关键词检测系统都至少绝对提升10%。

2、长短时记忆神经网络是语音增强领域常用的增强模型结构，本文设计了一种两层的BiLSTM增强模型，其本身具备强大的降噪能力，但由于其网络结构复杂计算量大，不适合关键词检测任务。考虑到卷积神经网络强大的编码、解码能力和循环神经网络的序列表达能力，本文设计了低资源占用的CRN的系列增强模型，不仅在性能上和基于BiLSTM的增强模型相当甚至更优，而且计算资源占用远少于BiLSTM模型。

3、由于深度神经网络强大的特征提取和表达能力，近几年语音增强领域常用的特征表达越来越简单，能量谱的应用越来越流行。对于关键词检测任务而言，能量谱并不适用于鲁棒性关键词检测系统，本文提出了直接在梅尔域进行特征增强，并设计了特征转换模块。最终实验结果表明，基于梅尔谱特征的关键词检测系统性能更优、资源占用更少，
并且对关键词中音标数量不敏感，更适合鲁棒的关键词系统。

最终，本文通过改进训练策略、模型结构、增强特征，得到较好的鲁棒关键词检测方案，即MelCRN32+Joint模型。与基线系统cnn-trad-pool2相比，在噪声匹配的测试集上准确率从80.89%提升到93.17%，在噪声不匹配的测试集上准确率从68.86%提升至78.12%。

5.1 后续工作展望

虽然本文提出的方法在噪声鲁棒性和计算资源占用上都有显著的优势。然而，鲁棒性关键词检测系统还有很大的优化和改进空间。

1、本文使用的关键词检测模型为CNN，本身识别性能一般，2018年最新提出的基于残差网络的识别器拥有更好的识别性能并且计算资源占用更小。在下一步的实验中，将更换识别器，进一步提高识别能力。

2、在本文中仅使用了单个增强特征，实验证明特征组合的方式会有效提高模型的鲁棒性能，在后续工作中，将尝试多种特征组合。

3、在本文的增强实验中，使用的损失函数为均方误差。最新的损失函数研究表明，MSE会弱化低能量部分的重要性，而散度类的损失函数使用比值的形式避免了这个问题。在接下来的工作中将尝试散度类的损失函数。

参考文献

[1] 韩纪庆. 语音信号处理[M]. 清华大学出版社, 2004.
[2] Raphael Tang, Jimmy Lin. Deep residual learning for small-footprint keyword spotting[C]. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2018: 5484-5488.
[3] Peter Warden. Speech Commands: A dataset for limited-vocabulary speech recognition[J]. arXiv preprint at arXiv:1804.03209, 2018.
[4] Prabhavalkar Rohit, Alvarez Raziel, Parada Carolina, etc. Automatic gain control and multi-style training for robust small-footprint keyword spotting with deep neural networks[C]. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015: 4704--4708.
[5] Chen G, Parada C, Heigold G. Small-footprint keyword spotting using deep neural networks[C]. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2014: 4087-4091.
[6] Shan Changhao, Zhang Junbo, Wang Yujun, etc. Attention-based End-to-End Models for Small-Footprint Keyword Spotting[C]. Annual Conference of the International Speech Communication Association(interspeech). 2018: 2037--2041.
[7] 刘加.汉语大词汇量连续语音识别系统研究进展[J].电子学报,2000(01):85-91.
[8] Chan Chun-an, Lee Lin-shan. Unsupervised spoken-term detection with spoken queries using segment-based dynamic time warping[C]. Annual Conference of the International Speech Communication Association(interspeech).2010: 693-696.
[9] Chan C A, Lee L S. Integrating Frame-based and Segment-based Dynamic time Warping for Unsupervised Spoken Term Detection with Spoken Queries[C]. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2011:5652-5655.
[10] Zhang Y, Glass J R. Unsupervised Spoken Keyword Spotting via Segmental DTW on Gaussian Posteriorgrams[C]. IEEE Workshop on Automatic Speech Recognition & Understanding(ASRU). 2009:398-403.
[11] Sun Ming, David Snyder, Yixin Gao, etc. Compressed Time Delay Neural Network for Small-Footprint Keyword Spotting[C]. Annual Conference of the International Speech Communication Association(interspeech). 2017: 3607-3611.
[12] Grangier D, Keshet J, Bengio S, etc. Discriminative Keyword Spotting[J]. In International Workshop on Non-LInear Speech Processing (NOLISP). 2009, 51(4):317–329.
[13] Chiu J, Wang Y, Trmal J, etc. Combination of FST and CN Search in Spoken Term Detection[J]. Annual Conference of the International Speech Communication Association(interspeech). 2014:2784-2788.
[14] Chen G, Yilmaz O, Trmal J, etc. Using Proxies for OOV Keywords in the Keyword Search Task[C]. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). 2013:416-421.
[15]李海洋. 汉语语音关键词检测中置信测度研究[D].哈尔滨工业大学,2014.
[16] Chen G, Parada C, Heigold G. Small-footprint Keyword Spotting Using Deep Neural Networks[C]. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2014:4087-4091.
[17] Sercan Ö. Arık, Markus Kliegl, Rewon Child, etc. Convolutional Recurrent Neural Networks for Small-Footprint Keyword Spotting[C]. Annual Conference of the International Speech Communication Association (interspeech). 2017: 1606-1610.
[18] S. Tamura. An analysis of a noise reduction neural network[C]. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 1989: 2001–2004.
[19] S. Tamura, A. Waibel. Noise reduction using connectionist models[C]. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 1988: 553–556.
[20] F. Xie, D. Van Compernolle. A family of MLP based nonlinear spectral estimators for noise reduction[C]. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 1994: 53–56.
[21] Y. Wang, D. Wang. A structure-preserving training target for supervised speech separation[C]. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2014: 6148–6152.
[22] M. Seltzer, B. Raj, R. Stern. A Bayesian classifier for spectrographic mask estimation for missing feature speech recognition[C]. Speech Communication, 2004 (4):379–393.
[23] N. Roman, D. Wang. Speech segregation based on sound localization[C]. In Proc. International Joint Conference on Neural Networks. 2001: 2861–2866.
[24] J. Tchorz, B. Kollmeier. SNR estimation based on amplitude modulation analysis with applications to noise suppression[J]. IEEE Trans. on Speech and Audio Processing. 2003 (3):184–192.
[25] G. Kim, Y. Lu, Y. Hu, P. Loizou. An algorithm that improves speech intelligibility in noise for normal-hearing listeners[J]. Journal of the Acoustical Society of America. 2009: 1486–1494.
[26] Y. Wang, D. Wang. Towards scaling up classification-based speech separation[J]. IEEE Trans. Audio, Speech, Lang. Process. 2013: 1381–1390.
[27] Y. Wang, A. Narayanan, D. Wang. On training targets for supervised speech separation[J]. IEEE Trans. Audio, Speech, Lang. Process. 2014(22):1849–1858.

[28] Y. Wang, D. Wang. A structure-preserving training target for supervised speech separation[C]. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2014: 6148–6152.
[29] Y. Wang, D. Wang. Cocktail party processing via structured prediction[J]. In Advances in Neural Information Processing Systems. 2012: 224–232.
[30] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, etc. Singing-voice separation from monaural recordings using deep recurrent neural networks[J]. In Proceedings of the International Society for Music Information Retrieval (ISMIR). 2014.
[31] F. J. Weninger, F. Eyben, B. Schuller. Single-channel speech separation with memory-enhanced recurrent neural networks[C]. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2014: 3737–3741.
[32] S. Hochreiter, J. Schmidhuber. Long short-term memory[C]. Neural Computation. 1997 (8):1735–1780.
[33] Prabhavalkar Rohit, Alvarez Raziel, Parada Carolina, etc. Automatic gain control and multi-style training for robust small-footprint keyword spotting with deep neural networks[C]. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015: 4704--4708.
[34] Sainath Tara, Parada Carolina. Convolutional neural networks for small-footprint keyword spotting[C]. Annual Conference of the International Speech Communication Association(interspeech). 2015: 1478-1482.
[35] Shan Changhao, Zhang Junbo, Wang Yujun, etc. Attention-based End-to-End Models for Small-Footprint Keyword Spotting[C]. Annual Conference of the International Speech Communication Association(interspeech). 2018: 2037--2041.
[36] Yu M, Ji X, Gao Y, Chen L, Chen J, Zheng J, Su D, Yu D. Text-Dependent Speech Enhancement for Small-Footprint Robust Keyword Detection[C]. Annual Conference of the International Speech Communication Association(interspeech). 2018:2613-7.
[37] Huang Y, Hughes T, Shabestary TZ, etc. Supervised noise reduction for multichannel keyword spotting[C]. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2018: 5474-5478.
[38] Hermansky H. Perceptual linear predictive (PLP) analysis of speech[J]. the Journal of the Acoustical Society of America. 1990 Apr;87(4):1738-52.
[39] Farhang-Boroujeny B. Filter bank spectrum sensing for cognitive radios. IEEE Transactions on signal processing. 2008 May;56(5):1801-11.
[40] Muda L, KM B, Elamvazuthi I. Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques[J]. Journal of Computing. 2010;2(3):138-143.
[41] Warden P. Speech commands: A dataset for limited-vocabulary speech recognition[J]. arXiv preprint arXiv:1804.03209, 2018.
[42] Erdogan Hakan, Hershey John, Watanabe, etc. Phase-Sensitive and Recognition-Boosted Speech Separation using Deep Recurrent Neural Networks[C]. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2015: 708-712.
[43] C. Taal, R. Hendriks, R. Heusdens, etc. An algorithm for intelligibility prediction of time-frequency weighted noisy speech[J]. IEEE Trans. Audio, Speech, Lang. Process. 2011: 2125–2136.
[44] A. Rix, J. Beerends, M. Hollier, etc. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2001: 749–752.
[45] Y. Wang, D. Wang. Towards scaling up classification-based speech separation[J]. IEEE Trans. Audio, Speech, Lang. Process. 2013: 1381–1390.
[46] Chen J, Wang Y, Yoho SE, etc. Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises[J]. The Journal of the Acoustical Society of America. 2016,139(5):2604-12.
[47] Kolbk M, Tan ZH, Jensen J, etc. Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems[J]. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP). 2017 Jan 1;25(1):153-67.
[48] Chen J, Wang D. Long short-term memory for speaker generalization in supervised speech separation[J]. The Journal of the Acoustical Society of America. 2017 Jun 23;141(6):4705-14.
[49] Naithani G, Barker T, Parascandolo G, etc. Low-latency sound source separation using convolutional recurrent deep neural networks[J]. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. 2017:1-5.

致 谢

两年的硕士时光匆匆而过，马上我就要离开学校踏入社会。在这段时间中，我收获颇多，这些成长使我顺利的从一个学生的角色转变成语音工程师的角色。有幸在这期间进行实习，实习过程中，我更感到内蒙古大学计算机学院语音信号处理组对我的帮助巨大。在专业知识领域、人际交往方面，我都由衷感谢我的老师们和实验室的同学们。红尘有幸，在语音信号处理组跟各位博学睿智的老师、聪慧机敏的同门认识、合作。

首先，感谢我的导师张学良教授。俗话说，师父领进门修行在个人。我一直认同张老师的引导教导法，于关键之处点拨，细节上学生自己摸索，这使我能更加快速地成长为能独当一面的工程师。除此之外，张老师的勤奋与努力也给我留下深刻的印象。这种自上而下的奋进精神使得实验室工作氛围良好，大家都在为了自己的理想而奋斗，找到了满意的工作，形成了良性循环。在此，对张学良教授致以最诚挚的敬意和由衷的感谢，祝福张老师和张老师的家人健康、和顺，平安喜乐。

其次，感谢哈尔滨工业大学语音组的杜志浩同学和实验室的张晖师兄，在我进行具体的科研工作时，两位师兄给了我关键性的指导与支持，帮助我进一步推进工作。希望在未来有机会与两位优秀的学者继续同行，进行合作。

再则，感谢实验室的各位老师和各位同门在日常工作中的帮助与指导。科研工作辛苦，实验室的温馨与良好互动正是这平淡生活中的浪花，使人积极又活跃。

最后，感谢我的父亲和母亲，很开心工作之后可以真正的回报辛苦的父母。

求学之路暂时结束了，但学习远没有停止。生命不止，学习不息，砥砺前行。