文件编号：从容应队-SWC2020-从容应队

受控状态：■受控 □非受控

保密级别：□公司级 □部门级 ■项目级 □普通级

采纳标准：CMMI DEV V1.2



****

素音

**Divoice**

项目开发文档

**Version 1**

2019.12.19

**Written by 从容应队**



**All Rights Reserved**

目录

[1 问题聚焦 1](#_Toc22848191)

[1.1 问题描述 1](#_Toc22848192)

[1.2 问题抽象 1](#_Toc22848193)

[1.3 问题定位 1](#_Toc22848194)

[1.4 问题评估 1](#_Toc22848195)

[1.5 问题分解 1](#_Toc22848196)

[2 相关工作 1](#_Toc22848197)

[3 技术方案 1](#_Toc22848198)

[3.1 技术方向 1](#_Toc22848199)

[3.2 技术选择 1](#_Toc22848200)

[3.3 结果期望 1](#_Toc22848201)

[4 技术实践 1](#_Toc22848202)

[4.1 使用的开发框架及依赖的Library 1](#_Toc22848203)

[4.2 技术实践过程 1](#_Toc22848204)

[5 结果验证 1](#_Toc22848205)

记录更改历史

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **序号** | **更改原因** | **版本** | **作者** | **更改日期** | **备 注** |
| 1 | 创建 | 1.0.0 | 队员A | 2019/12/19 | 创建 |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

# 问题聚焦

## 问题描述

Divoice有三个主要问题待解决：

1. 如何利用用户提供的采样音频优化音频分离效果
2. 对于用户提供的混音素材，如何使用训练好的模型去比对识别、分离
3. 对用户音频编辑内容提供相关素材推荐

## 问题抽象

基于问题一：Divoice需要将用户的采样音频中提取声纹特征，然后将此声纹特征写入用户常用采样音的数据集中。

基于问题二：根据用户选定的音频内含有的声纹标签，在相应的类中的分别进行声纹特征比对并拣选出匹配程度高的音频波段，形成新的音频文件，从而实现音频分离。对于比对度不高的剩余音频波段，再选用树表查找算法进行进一步比对分离，并返回比对度较高的声纹特征类别标签。

基于问题三：基于训练模型，对于不同场景中的常出现的声音特征写成集合，根据识别用户提供音频中的声纹特征，以及用户使用历史，猜测用户音频编辑风格，并对用户提供推荐素材。

## 问题定位

上述三个问题均主要涉及深度学习技术。

问题一主要针对用户的使用体验以及音频分离的效果和准确性。

问题二是本项目的核心，是对用户提供素材的处理及分离。

问题三考虑了用户再使用过程中是否便捷高效，是否具有良好的用户体验。

## 问题评估

## 对于问题一，声纹识别是一个十分基础并且重要的问题，目前声纹识别的技术已经相对成熟，网络上有部分开源项目可供参考。

对于问题二，这是整个项目的核心，也是整个项目的难点，当前没有较好的项目可以借鉴，需要自己从头做起，利用LSTM神经网络对不同数据进行训练分类，再通过比对，使用DNN+feed-forward architecture，分离出所需的音频波段。这部分问题较难，还需要进行充分的研究讨论。

对于问题三，利用深度学习技术，不断利用用户数据进行反复训练，猜测用户喜欢的风格，然后向用户推送素材，此类问题目前已有较为广泛的应用，实现可能性较大。

## 问题分解

问题可分解为声音降噪与声音分离两部分。

对于第一部分，声音降噪主要根据数据点的音量强度，对音频数据进行分割筛选并重新拼接，从而过滤掉大部分噪声。

第二部分较为复杂，声音分离是在声音降噪基础上将其进行再次处理。通过tensorflow环境，建立数据集分析音频中的主音和泛音，搭建深度学习框架达到识别音色并分离的功能。目前网络上已有较为成熟的音频分离学习框架，例如spleeter。可以参考其学习框架并进一步完善，使其能有更强大的分离功能。

# 相关工作

1. ResNet（Residual Neural Network）由微软研究院的Kaiming He等四名华人提出，通过使用ResNet Unit成功训练出了152层的神经网络，并在ILSVRC2015比赛中取得冠军，在top5上的错误率为3.57%，同时参数量比VGGNet低，效果非常突出。ResNet的结构可以极快的加速神经网络的训练，模型的准确率也有比较大的提升。同时ResNet的推广性非常好，甚至可以直接用到InceptionNet网络中。
2. 2017年，北京联合大学北京市信息服务工程重点实验室发表了一篇论文《噪声环境下的FastICA算法研究》，介绍了独立成分分析(ICA)算法的改进算法——FastICA，并引入牛顿下山法来降低初值选择的敏感性，提出差商法来减少计算量来针对FastICA算法的缺陷进行改良，使得收敛更加稳定。
3. 2019年王敏发表《基于深度神经网络和多元损失的说话人识别》提出一种基于深度神经网络和beyond triplet loss相结合的说话人识别方法,模型通过梅尔频率倒谱系数(MFCC)提取MFCC声学特征,对MFCC声学特征提取说话人声纹特征,然后进行多元损失的模型训练。实验结果表明,DNN-BTL算法在说话人识别领域比高斯混合-隐马尔可夫模型(GMM-HMM)具有更好的识别效果。

# 技术方案

## 技术方向

本项目中对音频文件进行处理，需使用深度学习相关知识。

## 技术选择

所用模型为DNN（ResNet，即残差网络）

## 

图3.2.1.DNN加feed-forward architecture结构图

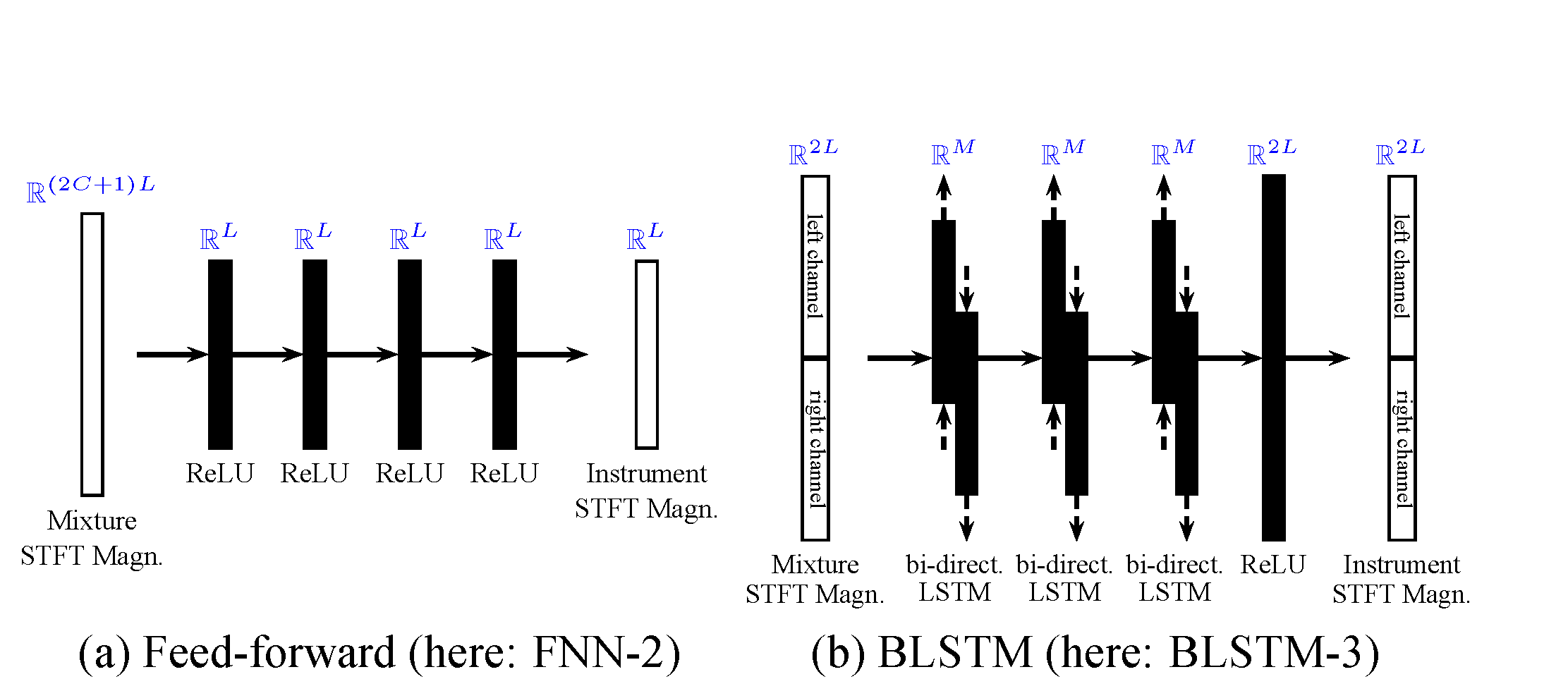


图3.2.2 DNN提取声纹结构示意图

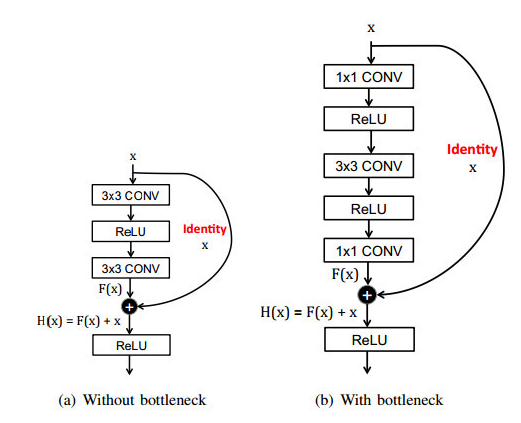


图3.2.3 残差模型

使用改进的基于差商法的FastICA算法（ISFastICA）：

1. 对观测信号进行中心话处理师其均值为0：
2. 对零均值向量进行白化处理，得到白化向量：
3. 随机选择初始化向量
4. 更新w：
5. 标准化w：
6. 若，则算法收敛，估计独立成分，否则返回步骤4)
7. 将收敛后的最优解混矩阵w与观测信号x相乘，得到源信号的估计y

## 结果期望

对于有采样的分析，基于FastICA技术较为成熟，预计分离的成功率在85%以上

对于盲源分析，基于BSS（盲源分离算法）技术一定的局限性，预计的分类成功率在75%以上。

# 技术实践

## 使用的开发框架及依赖的Library

## 技术实践过程

# 结果验证