**语音前端信号处理产品设计规格模板**

(仅供内部使用）

**深圳市普渡科技有限公司**

版权所有 侵权必究

1、简介

1.1 范围

麦阵语音前端信号处理，代号aispeech，属产品线机器猫，葫芦以及未来会使用语音技术的产品。

1.2 目的

1. cutdown 方案(淘宝价格讯飞的软硬件pcba是600元)。
2. 可控可扩展(声纹，多模态)。
3. 提供标准接口的PCBA软硬件方案(直接插入usb接口使用)。
4. 可优化某些场景下的信号处理（比如餐厅，机器本体发出噪声）。

1.3 标识和图例

通过图例说明功能、子系统、接口的标识规则和图例。

2、概述

语音前端处理是指在特征提取之前，先对原始语音进行处理，部分消除噪声和不同说话人带来的影响，使处理后的信号更能反映语音的本质特征。最常用的前端处理有端点检测和语音增强。端点检测是指在语音信号中将语音和非语音信号时段区分开来，准确地确定出语音信号的起始点。经过端点检测后，后续处理就可以只对语音信号进行，这对提高模型的精确度和识别正确率有重要作用。语音增强的主要任务就是消除环境噪声对语音的影响。目前通用的方法是采用维纳滤波，该方法在噪声较大的情况下效果好于其它滤波器。语音前端处理最终服务于高准确率的语音识别。

2.1 V1.0.0版本描述

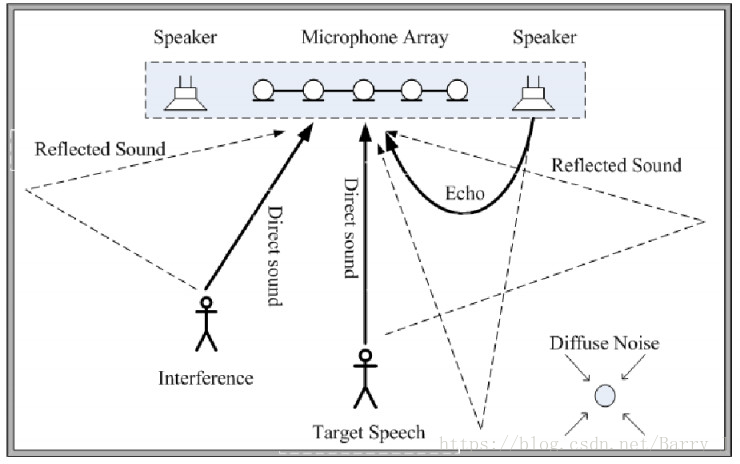
1. **实现端点检测：**可以检测声音的开始位置和结束位置，并且提供设置阈值，检测时长。
2. **实现降噪算法：**可以有效降低餐厅，机器本身发出来的噪音。
3. **实现波束成型：**可以定位某一方向声音增强。
4. **角度定位：**可以实现15度的方向定位。
5. **离线识别：**在2GB内存，1.5GHZ的CPU实现离线识别。
6. **唤醒功能：**可以唤醒设备
7. **回声消除：**消除机器本体扬声器发出的声音。

2.2 V版本各R版本描述

描述同属一个V版本的各R版本的关系及主要特征和区别。注明本规格书属于哪个版本。

2.3 业务简述

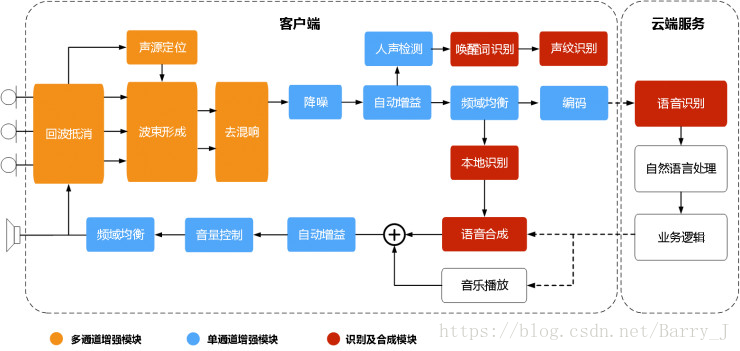
通过前端处理模块在特征提取之前，先对原始语音进行处理，部分消除噪声和不同说话人带来的影响，抑制各种干扰，使待识别的语音更干净更能反映语音的本质特征。语音前端处理算法在语音通信和语音修复中也有着广泛的应用。



语音处前端处理主要会处理这几个方面接收到的声信号在时间序列可分为三部分：直达声、早期反射声(经过一次、两次的反射，能量较大、时延较短的反射声) 、混响声(经过多次反射以后到达的数目众多、能量较小、密集的反射声群），前期反射声是指未达到稳定状态时的反射声，也就是在直达声之后混响开始衰减之前的这段时间差内的反射声，以下是一些声音的来源：

1. **Echo：**远端扬声器播放的声音回传给麦克。
2. **Diffuse Noise：**无向噪声的干扰。
3. **Reflected Sound：**声音通过墙壁反射，造成混响干扰。
4. **Interference：**其他方向的干扰源。
5. **Target Speech：**目标方向声音。

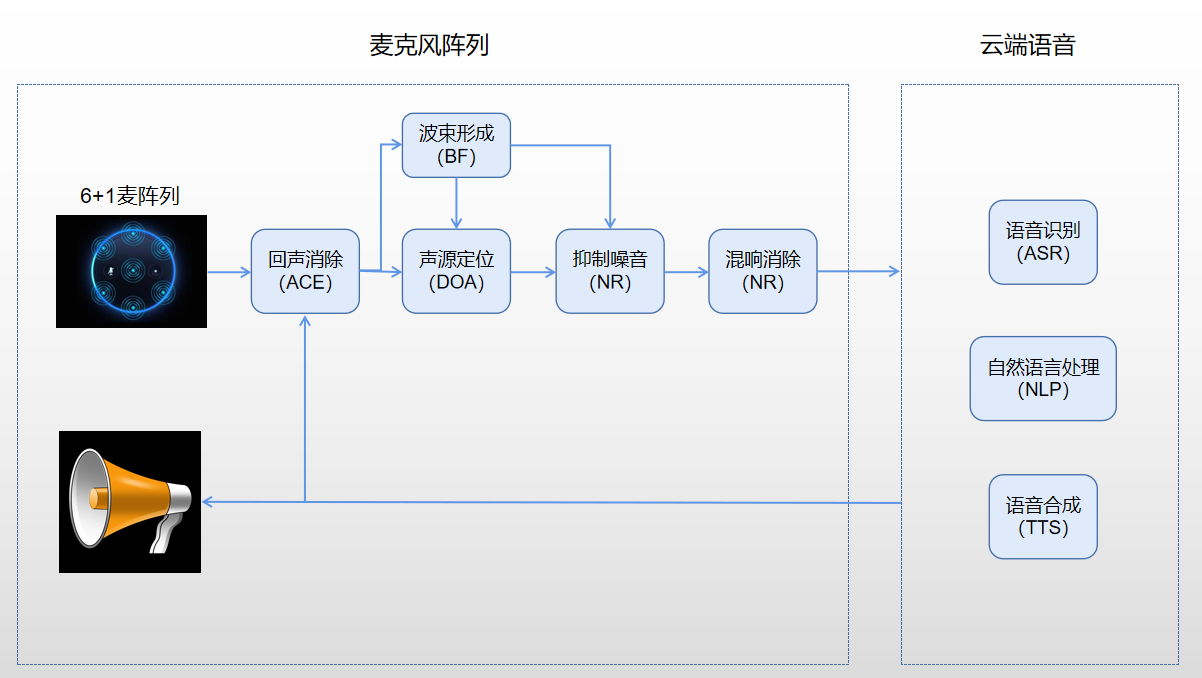
由于噪声来源众多，在不同的应用场合其特性各不相同，增加了语音增强算法的复杂性。而且语音增强不仅仅是一个数字信号处理技术方面的问题，还涉及人的听觉感知和语音学，是一门很复杂的技术。要想理想地设计出一种算法来消除所有的噪声是不现实的，只能针对不同的噪声情况，采取不同的语音增强算法。



**语音前端处理模块跟语音交互系统的关系：**橙色部分表示多通道处理模块，蓝色部分表示单通道处理模块，红色部分表示后端识别合成等模块。麦克风阵列采集的语音首先利用参考源对各通道的信号进行回波消除，然后确定声源的方向信息，进而通过波束形成算法来增强目标方向的声音，再通过混响消除方法抑制混响；需要强调的是可以先进行多通道混响消除，然后再进行波束形成，也可以先进行波束形成再进行单通道混响消除。经过上述处理后的单路语音进行后置滤波消除残留的音乐噪声，然后通过自动增益算法调节各个频带的能量后最为前端处理的输出，将输出的音频传递给后端进行识别和理解。

3、系统总体描述

3.1 系统上下文



麦克风阵列是由一定数目的声学传感器（麦克风）按照一定规则排列的多麦克风系统，对声场的空间特性进行采样并滤波的系统。

麦克风阵列除了看到的麦克风数量以外，还有一系列的前端算法，两者结合的系统才是完整的麦克风阵列。而麦克风阵列也只是完成了物理世界的音频信号处理，想要完成语音识别，还是需要云端的ASR模型，两个系统配合在一起才能得到最好的识别效果。语音前端算法包含：

**（1）声源定位**

人有两个耳朵，可以通过声音判断发声的方向，机器人也同样可以做到。这个功能就是声源定位，通过声音感知人所在的方向，从而实现对目标声源方向的跟踪。这也为后续的波束形成做技术铺垫。

比如机器人场景，我们在机器人左边叫它，机器人听到声音后就会把头转向左边，我们在机器人背后叫它，机器人听到声音后就会转过去，这就是声源定位最典型的应用。通常声源定位会应用在语音唤醒阶段，能够检测一个大致的方向。

常用到的技术是TDOA（Time Difference Of Arrival，到达时间差），简单理解就是通过计算信号到达麦克风之间的时间差，从而计算出声源的位置坐标，需要毫秒级的响应和计算。

**（2）抑制噪音/增强人声**

在语音识别中，语音信息中往往夹杂着噪音，常见的有环境噪音和人声干扰，通常不会掩盖正常的语音，只是影响声音的清晰度。麦克风阵列主要通过波束形成技术，来抑制噪音，增强人声。可以理解为只识别某个角度的声音（一般角度可以进行调节），其他角度的声音都会受到抑制，从而实现抑制噪音的目的。反过来也可以增强角度内的人声，就是增强人声。

比如家庭场景，如果我们开着电视和空调在和音箱说话，音箱会以唤醒的它的角度为拾音区域，抑制非该角度的噪音（电视声音和空调噪音）。一般我们根据使用场景去设置拾音角度，使用距离越远，角度越小，常见为60°-120°之间。

抑制噪音能够满足日常家庭的使用场景，但对于强噪音环境的抑制效果并不理想，典型的就是鸡尾酒效音。

**（3） 回声消除**

如果不做特殊处理的话，机器人会识别自己发出来的声音，很有可能就会变成无休止的自问自答，或者拾音错误。回声消除就是为了解决这个问题，消除掉机器自己发出的声音。

比如家庭场景下，你的音箱正在放音乐周杰伦的新歌，但是你想要查一下天气，这个时候你就会说“小X小X，今天天气怎么样”。回声消除的目的就是要去掉其中音乐信息而保留你的声音。

其实回声消除可能不太好理解，有时也被称作为“自识别”，即自己识别自己发出的声音。

**（4）混响消除**

在某些场景，发音会有回音，人能听到的是17米左右距离返回的回音。但是机器的感知要比人敏感的多，如果不做处理，就会出现一句话叠加识别的情况。混响常指声波在室内传播时，被墙壁、天花板、地板等障碍物形成反射声，并和直达声形成叠加的现象。

比如在演播厅，我们能够感受到较为明显的回音，机器同样能够识别到这些回音。混响消除就是消除之后带来的回音，只识别第一遍的内容。

解决了这些问题，基本上就可以在日常环境下进行一个正常拾音，从而保证整个语音识别的正常。

5、系统性能列表

5.1 环境条件

1. 麦克风阵列：瑞星微3399开发板，六路麦克风，两路回采
2. 背景噪声：安静环境<40db，办公室环境60db,商场餐厅环境75db
3. 声源：60-65db，说话人正常音量，在麦克风处测得声源60db
4. 声源和麦克风距离：1m

5.2 VAD 端点检测技术

要求信噪比<10db,vad效果>95%；说话者音量>55db

5.3 回声消除

消除自己麦克风播放声音，六麦，回声消除性能最低能消掉约-25dB，在背景噪声特别强的情况（75db），能消除-30db到-40db

5.4 去混响+降噪

噪声抑制量，15-20db，背景噪声75db,降噪后背景噪声55-60db,信噪比从0db变化成15db

5.5 唤醒词验收

一般性指标参考

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 测试环境 | 唤醒音量 | 噪声音量 | 距离 | 唤醒率 |
| 安静环境 | 65db | <40db | 1m | 98% |
| 安静环境 | 65db | <40db | 3m | 97% |
| 办公室环境 | 65db | 60-65db | 1m | 96% |
| 办公室环境 | 65db | 60-65db | 3m | 92% |
| 商场餐厅环境 | 65db | 75db | 1m | 82%(讯飞提供) |
| 商场餐厅环境 | 65db | 75db | 3m | 74%(讯飞提供) |
| 误唤醒<1/24h | | | | |

普渡验收方案说明：唤醒方案测试，对第三方验收，办公室场景，播放音乐，mic收音处背景音70-80dB，背景音乐选择新闻广播、脱口秀、访谈节目、影视剧等，选择成年两男两女，1m,2m,3m,唤醒词各200遍，音量65db，响应时间：说唤醒词之后到唤醒响应时间<500ms，测试状态：1.单纯测试，2.满负荷状态下测试,3399的cpu占比

误唤醒测试：办公室场景下，24h,保存唤醒音频，前后个10s,查看唤醒次数结果；

5.6 离线语音识别

最多指令词识别两千条，识别准确率：95%以上，说话结束之后响应时间：<500ms，芯片3399，4小核，2大核，cpu占比<25%，背景噪声65-75db,距离1m范围，单人说话，音量65db

5.7 在线语音识别

背景噪声65-75db,距离1m,识别准确率：>95%, 握手阶段连接服务器时间200ms,发送到返回首字时间1000ms,最后延迟时间500ms

目前讯飞识别字准确率指标

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 测试场景 | 声源到达MIC处音量 | 噪声到达MIC处音量 | 声源与麦克风距离 | 6mic字准率（定制） |
| 安静 | 65db | <40db | 1m | 96% |
| 安静 | 65db | <40db | 3m | 93% |
| 外噪（非人声） | 65db | 65db | 1m | 85% |
| 外噪（非人声） | 65db | 65db | 3m | 82% |

5.8 音频压缩

拾音压缩上传，云端接受解压，参考连接

<https://github.com/dreamflyforever/speex_package>

6、资料交付件清单

时间：预计2023年3月份

22年的计划：

唤醒wakeup & 回声消除AEC：可以定制唤醒词语，快捷唤醒

2月15号份调研采用端到端的方案，输入语音，输出为各唤醒的概率，一个模型解决。

3月15号实现基于神经网络的kws，CNN建模

4月份30号，调优提供标准API接口以及单元测试程序，唤醒率>95%, 误唤醒率< 24h/c次，唤醒速度<500ms

本地asr：

5月15号调研本地asr的几种方式的对本地cpu性能和内存使用大小。

6月15号实现第一版本，可以识别20条

7月15号调优提高性能，降低功耗

8月30号，提供API接口，单元测试程序。

降噪&去混响&角度定位

9月15号调研降噪的最前沿和效果最佳的技术方案

10月15号实现降噪第一版本

11月15号，调优提高性能，降低功耗

12月30号提供API和单元测试例子

2023年1月份-2023年3月份

调优debug，整合各个模块，并且配合云端识别做调整。

**人力资源: 6人，分别3个有1到3年经验的，3个刚毕业(实习生)。**