(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 109584896 A
(43)申请公布日 2019.04.05

(21)申请号 201811293499.4

(22)申请日 2018.11.01

(71)申请人 苏州奇梦者网络科技有限公司
地址 215024 江苏省苏州市工业园区若水路388号纳米技术国家大学科技园E栋1604

(72)发明人 肖佳林 王欢良 唐浩元 王佳珺 吴洪宇 马殿昌 李志

(74)专利代理机构 苏州国诚专利代理有限公司 32293
代理人 王丽

(51)Int.Cl.
G10L 21/0208(2013.01)
G10L 21/0224(2013.01)
G10L 21/0232(2013.01)
G10L 21/028(2013.01)
G10L 25/24(2013.01)
G10L 15/06(2013.01)
G10L 15/08(2006.01)
G10L 15/22(2006.01)

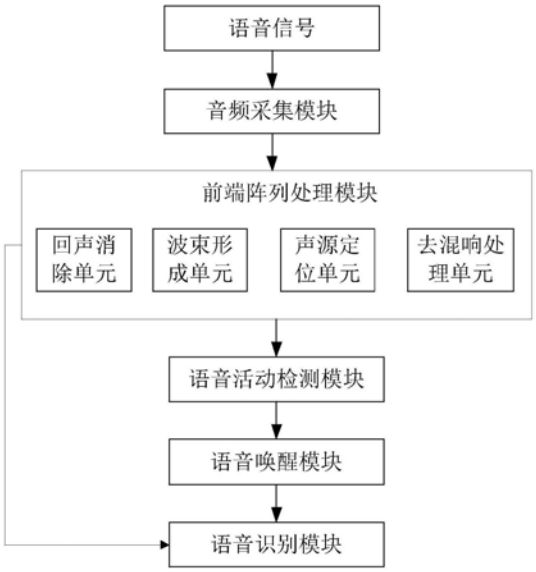
权利要求书2页 说明书9页 附图1页

(54)发明名称

一种语音芯片及电子设备

(57)摘要

本发明涉及一种语音芯片及应用的电子设备，包括：音频采集模块，用于采集语音信号；前端阵列处理模块，与音频采集模块连接，用于对语音信号进行处理；语音活动检测模块，与前端阵列处理模块连接，用于确定前端阵列处理模块处理后的语音信号的语音端点，其中，语音端点包括语音信号的开始端点和结束端点，每个语音端点对应语音信号中的至少一帧；语音唤醒模块，与语音活动检测模块连接，用于基于语音端点确定语音信号包括预设的唤醒语音时，唤醒电子设备；语音识别模块，与前端阵列处理模块连接，用于在电子设备被唤醒后，识别前端阵列处理模块处理后的语音信号中的交互指令，并使电子设备执行交互指令。本发明可提高语音识别的准确性。



CN 109584896 A

1.一种语音芯片,应用于电子设备,其特征在于,包括:

音频采集模块,用于采集语音信号;

前端阵列处理模块,与所述音频采集模块连接,用于对所述语音信号进行处理;

语音活动检测模块,与所述前端阵列处理模块连接,用于确定所述前端阵列处理模块处理后的语音信号的语音端点,其中,所述语音端点包括所述语音信号的开始端点和结束端点,每个所述语音端点对应语音信号中的至少一帧;

语音唤醒模块,与所述语音活动检测模块连接,用于基于所述语音端点确定所述语音信号包括预设的唤醒语音时,唤醒所述电子设备;

语音识别模块,与所述前端阵列处理模块连接,用于在所述电子设备被唤醒后,识别所述前端阵列处理模块处理后的语音信号中的交互指令,并使电子设备执行所述交互指令。

2.根据权利要求1所述的语音芯片,其特征在于,所述前端阵列处理模块包括:

回声消除单元,用于对所述语音信号进行回去消除处理;

波束形成单元,用于对所述语音信号进行波束形成处理;

声源定位单元,用于对所述语音信号进行声源定位;

去混响处理单元,用于以所述语音信号去除混响。

3.根据权利要求2所述的语音芯片,其特征在于,所述回声消除单元具体用于:

对所述语音信号进行陷波滤除直流分量,并做预加重处理,形成第一输入信号;

采用前级滤波对所述第一输入信号进行滤波处理,将得到的误差信号方差存储给Sff;

采用后级滤波对所述第一输入信号进行滤波处理,将得到的误差信号方差给See;

基于所述Sff和See输出最终滤波后的信号。

4.根据权利要求2所述的语音芯片,其特征在于,所述波束形成单元具体用于:

对所述语音信号进行傅里叶变换,并计算所述语音信号的语音信号协方差;

对所述语音信号协方差进行特征值分解,确定最大特征值;

确定所述最大特征值对应的特征向量;

基于所述语音信号协方差和所述特征向量计算最终的增强信号;

对所述增强信号进行傅里叶变换后,并将转换后的信号转到时域。

5.根据权利要求2所述的语音芯片,其特征在于,所述声源定位单元具体用于:

计算所述语音信号到达阵列中不同麦克风之间由于所述语音信号传输距离不同而引起的时间差;

将所述时间差乘以声速得到距离差;

根据几何关系和所述距离差计算一系列双曲面,通过所述双曲面的交点得到声源位置。

6.根据权利要求2所述的语音芯片,其特征在于,所述去混响处理单元具体用于:

对所述语音信号进行短时傅里叶变换;

基于变换后的信号计算去除混响后的频域信号;

对所述频域信号进行逆短时傅里叶变换,并把变换后的语音信号转到时域。

7.根据权利要求1所述的语音芯片,其特征在于,所述语音活动检测模块具体用于:

对所述语音信号进行预处理;

将经过预处理的信号逐帧使用FilterBank算法来提取FBank特征;

将每一帧的FBank特征输入深度神经网络模型，由深度神经网络模型输出每一帧在音素集中每一个音素的输出概率；

将每一帧中对应所有非噪声、非静音的音素对应的输出概率进行求和；

如果所述和值大于预设阈值，则判断对应帧为语音端点；

当最后一帧被判断后，则对之前的判决结果进行平滑滤波处理，得到最终的语音端点判决结果。

8.根据权利要求1所述的语音芯片，其特征在于，所述语音唤醒模块具体用于：

对所述语音信号进行语音特征提取，得到对应的语音特征向量；

将所述语音特征向量输入到DNN模型，得到该语音特征向量对应的语音信号是关键词或者是非关键词的后验概率；

对所述后验概率进行平滑得到对应的置信度；

如果所述置信度大于预设值，则判断所述语音特征向量对应的语音信号包含关键词；

如果所述关键词按设定顺序出现则判断唤醒所述电子设备。

9.根据权利要求1所述的语音芯片，其特征在于，所述语音识别模块具体用于：

提取所述语音信号中的语音特征向量；

对所述语音特征向量进行语音解码，得到最优的输出词序列；

基于所述输出词序列输出对应的文本；

基于所述文本确定所述文本表达的交互指令，并使电子设备执行所述交互指令。

10.一种电子设备，其特征在于，所述电子设备包含权利要求1-9中任一项所述的语音芯片。

一种语音芯片及电子设备

技术领域

[0001] 本发明涉及语音处理领域,特别是涉及一种语音芯片及电子设备。

背景技术

[0002] 电子设备所具备的智能语音功能可以实现完美的人机语音交互,其最终目 标是让电子设备听懂人类的语言,并执行相应的功能。目前电子设备中所应用 的芯片不具备智能功能,且价格昂贵,功耗高。对于目前支持语音交互功能的 芯片,其支持的语音识别基于深度学习的神经网络算法,但深度学习神经网络 算法计算量较大,导致目前支持语音交互功能的芯片计算,功耗大,速度慢。为了满足计算量,支持语音交互功能的芯片会对深度学习神经网络算法会进行 简化,导致识别性能降低,进一步导致用户在与电子设备进行语音交互式时体 验感极差。

发明内容

[0003] 基于此,有必要针对目前语音识别度低的问题,提供一种语音芯片及电子 设备。

[0004] 一种语音芯片,应用于电子设备,,包括:

[0005] 音频采集模块,用于采集语音信号;

[0006] 前端阵列处理模块,与所述音频采集模块连接,用于对所述语音信号进行 处理;

[0007] 语音活动检测模块,与所述前端阵列处理模块连接,用于确定所述前端阵 列处理模块处理后的语音信号的语音端点,其中,所述语音端点包括所述语音 信号的开始端点和结束端点,每个所述语音端点对应语音信号中的至少一帧;

[0008] 语音唤醒模块,与所述语音活动检测模块连接,用于基于所述语音端点确 定所述语音信号包括预设的唤醒语音时,唤醒所述电子设备;

[0009] 语音识别模块,与所述前端阵列处理模块连接,用于在所述电子设备被唤 醒后,识别所述前端阵列处理模块处理后的语音信号中的交互指令,并使电子 设备执行所述交互指令。

[0010] 优选的,所述前端阵列处理模块包括:

[0011] 回声消除单元,用于对所述语音信号进行回去消除处理;

[0012] 波束形成单元,用于对所述语音信号进行波束形成处理;

[0013] 声源定位单元,用于对所述语音信号进行声源定位;

[0014] 去混响处理单元,用于以所述语音信号去除混响。

[0015] 优选的,所述回声消除单元具体用于:

[0016] 对所述语音信号进行陷波滤除直流分量,并做预加重处理,形成第一输入 信号;

[0017] 采用前级滤波对所述第一输入信号进行滤波处理,将得到的误差信号方差 存储给Sff;

[0018] 采用后级滤波对所述第一输入信号进行滤波处理,将得到的误差信号方差 给See;

[0019] 基于所述Sff和See输出最终滤波后的信号。

[0020] 优选的，所述波束形成单元具体用于：

[0021] 对所述语音信号进行傅里叶变换，并计算所述语音信号的语音信号协方差；

[0022] 对所述语音信号协方差进行特征值分解，确定最大特征值；

[0023] 确定所述最大特征值对应的特征向量；

[0024] 基于所述语音信号协方差和所述特征向量计算最终的增强信号；

[0025] 对所述增强信号进行傅里叶变换后，并将转换后的信号转到时域。

[0026] 优选的，所述声源定位单元具体用于：

[0027] 计算所述语音信号到达阵列中不同麦克风之间由于所述语音信号传输距离 不同而引起的时间差；

[0028] 将所述时间差乘以声速得到距离差；

[0029] 根据几何关系和所述距离差计算一系列双曲面，通过所述双曲面的交点得 到声源位置。

[0030] 优选的，所述去混响处理单元具体用于：

[0031] 对所述语音信号进行短时傅里叶变换；

[0032] 基于变换后的信号计算去除混响后的频域信号；

[0033] 对所述频域信号进行逆短时傅里叶变换，并把变换后的语音信号转到时域。

[0034] 优选的，所述语音活动检测模块具体用于：

[0035] 对所述语音信号进行预处理；

[0036] 将经过预处理的信号逐帧使用FilterBank算法来提取FBank特征；

[0037] 将每一帧的FBank特征输入深度神经网络模型，由深度神经网络模型输出 每一帧在音素集中每一个音素的输出概率；

[0038] 将每一帧中对应所有非噪声、非静音的音素对应的输出概率进行求和；

[0039] 如果所述和值大于预设阈值，则判断对应帧为语音端点；

[0040] 当最后一帧被判断后，则对之前的判决结果进行平滑滤波处理，得到最终 的语音端点判决结果。

[0041] 优选的，所述语音唤醒模块具体用于：

[0042] 对所述语音信号进行语音特征提取，得到对应的语音特征向量；

[0043] 将所述语音特征向量输入到DNN模型，得到该语音特征向量对应的语音信 号是关键词或者是非关键词的后验概率；

[0044] 对所述后验概率进行平滑得到对应的置信度；

[0045] 如果所述置信度大于预设值，则判断所述语音特征向量对应的语音信号包 含关键词；

[0046] 如果所述关键词按设定顺序出现则判断唤醒所述电子设备。

[0047] 优选的，所述语音识别模块具体用于：

[0048] 提取所述语音信号中的语音特征向量；

[0049] 对所述语音特征向量进行语音解码，得到最优的输出词序列；

[0050] 基于所述输出词序列输出对应的文本；

[0051] 基于所述文本确定所述文本表达的交互指令，并使电子设备执行所述交互 指令。

[0052] 一种电子设备,所述电子设备包含以上所述的语音芯片。

[0053] 以上所述语音芯片包括音频采集模块,前端阵列处理模块,语音活动检测(Voice activity detection,VAD)模块(VAD模块),语音唤醒模块,语音识别 模块。音频采集模块首先采集到语音信号,然后将语音信号传给前端阵列处理 模块,进行处理。VAD模块与前端阵列处理模块相连接,VAD模块会对经过处 理的语音信号检测用户语音的开始与结束。语音唤醒模块对用户语音与设定的 唤醒词进行相似度比较,如果匹配可将设备从休眠状态中唤醒。唤醒成功后通 过前端阵列算法处理,定位说话人方位,进行定向语音增强,将增强后的语音 送入语音识别模块,电子设备对于识别的指令做出相应的动作,即可实现人机语音交互。由此,本发明在唤醒电子设备后,可以将定位以及定向语音增强后 的语音信号送入语音识别模块,从而提升语音识别的准确性。

附图说明

[0054] 图1为一实施例的语音芯片的结构图。

具体实施方式

[0055] 为了使本发明的目的、技术方案及优点更加清楚明白,以下结合附图及实 施例,对本发明进行进一步详细说明。应当理解,此处所描述的具体实施例仅 仅用以解释本发明,并不用于限定本发明。

[0056] 图1为一实施例的语音芯片的结构图,如图1所示,该语音芯片应用于电 子设备,包括:

[0057] 音频采集模块,用于采集语音信号;

[0058] 前端阵列处理模块,与所述音频采集模块连接,用于对所述语音信号进行 处理;

[0059] 语音活动检测模块,与所述前端阵列处理模块连接,用于确定所述前端阵 列处理模块处理后的语音信号的语音端点,其中,所述语音端点包括所述语音 信号的开始端点和结束端点,每个所述语音端点对应语音信号中的至少一帧;

[0060] 语音唤醒模块,与所述语音活动检测模块连接,用于基于所述语音端点确 定所述语音信号包括预设的唤醒语音时,唤醒所述电子设备;

[0061] 语音识别模块,与所述前端阵列处理模块连接,用于在所述电子设备被唤 醒后,识别所述前端阵列处理模块处理后的语音信号中的交互指令,并使电子 设备执行所述交互指令。

[0062] 以上所述语音芯片包括音频采集模块,前端阵列处理模块,语音活动检测 (Voice activity detection,VAD)模块(本实施例语音活动检测模块简称VAD 模块),语音唤醒模块,语音识别模块。音频采集模块首先采集到语音信号,然 后将语音信号传给前端阵列处理模块,进行处理。VAD模块与前端阵列处理模 块相连接,VAD模块会对经过处理的语音信号检测用户语音的开始与结束。语 音唤醒模块对用户语音与设定的唤醒词进行相似度比较,如果匹配可将设备从 休眠状态中唤醒。唤醒成功后通过前端阵列算法处理,定位说话人方位,进行 定向语音增强,将增强后的语音送入语音识别模块,电子设备对于识别的指令 做出相应的动作,即可实现人机语音交互。由此,本发明在唤醒电子设备后, 可以将定位以及定向语音增强后的语音信号送入语音识别模块,从而提升语音 识别的准确性。

[0063] 本实施例中,前端阵列处理模块、VAD模块,语音唤醒模块和语音识别模 块是利用加速器进行的实现,从而提高程序运行速度。此语音芯片可以应用于 各种智能设备来实现人机语音交互。

[0064] 本实施例中,音频采集模块可以是Mic(microphone,麦克风),用于将接 收的语音信号传给前端阵列处理模块。

[0065] 本实施例中,前端阵列处理模块可以进行回声消除、波束形成、声源定位、 去混响等处理操作,且可以加速实现以上处理过程,其主要使用了以下硬件加 速器:fft/ifft加速器,矩阵乘加速器,求逆加速器,求行列式加速器,求特征值 特征向量加速器,simd加速器,数学运算加速器(dma加速器)以及求cholesky 积加速器。其中数学运算加速器的功能主要有:三角函数,对数函数,指数函 数,求和运算,求倒运算,除法运算,开方运算,幂运算,求绝对值,浮点/整 数转换运算等。

[0066] 本实施例以上涉及以下名词,即:fft(Fast FourierTransform,快速傅里叶变换),IFFT(Inverse Fast-Fourier-Transformation,快速傅立叶反变换),SIMD (Single Instruction Multiple Data,单指令多数据流),DMA(Direct Memory Access,直接内存存取),Cholesky分解是把一个对称正定的矩阵表示成一个下 三角矩阵L和其转置的乘积的分解。

[0067] 本发明一实施例中,所述前端阵列处理模块包括:

[0068] 回声消除单元,用于对所述语音信号进行回去消除处理;

[0069] 波束形成单元,用于对所述语音信号进行波束形成处理;

[0070] 声源定位单元,用于对所述语音信号进行声源定位;

[0071] 去混响处理单元,用于以所述语音信号去除混响。

[0072] 本发发明一实施例中,所述回声消除单元具体用于:

[0073] 对所述语音信号进行陷波滤除直流分量,并做预加重处理,形成第一输入 信号;

[0074] 采用前级滤波对所述第一输入信号进行滤波处理,将得到的误差信号方差 存储给Sff;

[0075] 采用后级滤波对所述第一输入信号进行滤波处理,将得到的误差信号方差 给See;

[0076] 基于所述Sff和See输出最终滤波后的信号。

[0077] 本实施例中,回声消除的具体实现是:首先对麦克风阵列接收的语音信号 进行陷波滤除直流分量,并做预加重处理,形成第一输入信号。然后采用前级 滤波对第一输入信号进行滤波处理,其中使用simd加速器和dma加速器来加速 计算,然后将滤波输出信号存在e[]的后半部分,误差信号方差存储给Sff(一种 高频同步旋转坐标系滤波器(Synchronous Frame Filter,SFF),电力控制中用于 提取电流负序分量)。之后计算后级滤波器抽头系数W,其中以归一化最小均方 自适应滤波器(Normalized least mean square,NLMS)为基础,多延迟块频域自 适应滤波器(The multi-delay block frequency-domain adaptive filter,MDF)频域 实现,最终推导出最优步长等于残余回声方差与误差信号方差之比。其中残余 回声方差是通过定义一个泄漏系数并使用simd加速器来计算的,而泄漏系数是 通过递归平均处理方法得到每个频点的自相关、输入信号与误差信号的互相关 最终得到的。之后对第一输入信号用后级滤波再次进行滤波处理,使用到simd 加速

器和dma加速器来加速计算，得到的误差信号方差给See，误差存储于e[] 前半部分。然后结合See与Sff综合判断是否需要更新前级滤波系数或者重置后 级滤波，如果需要则进行自适应滤波，权值更新，并更新误差信号在时域的能 量值，其中使用到simd加速器。最后再用simd加速器来计算最终的滤波输出 out＝input-滤波输出e[]后半部分，并进行去加重处理，到此回声消除完成。

[0078] 需要指出的是，本实施例以上模块的具体实现过程，只是可以使用于本实 施例中的一种最优的选择，本实施例不限于对以上具体的过程、方法进行适当 的变形或者改变，以实现本发明的具体技术方案，这均在本发明保护的范围之 内。

[0079] 本实施例的一实现方式中，所述波束形成单元具体用于：

[0080] 对所述语音信号进行傅里叶变换，并计算所述语音信号的语音信号协方差；

[0081] 对所述语音信号协方差进行特征值分解，确定最大特征值；

[0082] 确定所述最大特征值对应的特征向量；

[0083] 基于所述语音信号协方差和所述特征向量计算最终的增强信号；

[0084] 对所述增强信号进行傅里叶变换后，并将转换后的信号转到时域。

[0085] 本实施例中，波束形成的实现是：首先利用fft加速器来对麦克风阵列接收 到的信号y_t进行短时傅里叶变换(Short-time fourier transform,STFT)。然后用 simd加速器来加速$\varphi_{f,t}^{(v)}$和$R_f(v)$的初始化。之后根据CGMM原理，利用矩阵乘 加速器和dma加速器估计出$\lambda_{f,t}^{(v)}$。然后开始估计噪声协方差$R_n(f)$，带噪语音协 方差$R_{k+n}(f)$和语音信号协方差$R_k(f)$，使用到dma加速器，矩阵乘加速器和simd 加速器来加速计算。之后利用求特征值特征向量加速器来对矩阵$R_k(f)$进行特 征值分解，最大特征值对应的特征向量即为目标语音的方向向量r_f。根据得到 的$R_n(f)$和r_f计算得到权值$w_{f,t}^{(k)}$，其中使用矩阵乘加速器和simd加速器来对计算 进行加速。最后使用simd加速器来计算最终要得到的增强信号$S_{f,t}^{(k)}$，并利用ifft 加速器对其进行逆短时傅里叶变换(Inverse short-time fourier transform,ISTFT)，之后把信号转到时域。至此波束形成结束。

[0086] 需要指出的是，本实施例以上模块的具体实现过程，只是可以使用于本实 施例中的一种最优的选择，本实施例不限于对以上具体的过程、方法进行适当 的变形或者改变，以实现本发明的具体技术方案，这均在本发明保护的范围之 内。

[0087] 本实施例中，所述声源定位单元具体用于：

[0088] 计算所述语音信号到达阵列中不同麦克风之间由于所述语音信号传输距离 不同而引起的时间差；

[0089] 将所述时间差乘以声速得到距离差；

[0090] 根据几何关系和所述距离差计算一系列双曲面，通过所述双曲面的交点得 到声源位置。

[0091] 本实施例中，声源定位的实现是：首先估计语音信号到达阵列中不同麦克 风之间由于信号传输距离不同而引起的时间差(Time delay of arrival,TDOA)，即进行时间延迟估计(Time delay estimation,TDE)。这里使用到广义互相关 (Generalized cross correlation,GCC)法来进行时间延迟估计，先利用fft加速 器对不同传声器接受到的音频

信号进行快速傅里叶变换(Fast fourier transformation,FFT)。然后定义广义互相关函数,先在频域利用加权函数来加 强语音信号直达部分、抑制噪声以及混响信号来突出相应的峰值,其中使用simd 加速器来加速进行。之后利用ifft加速器对加权后的信号进行逆向快速傅里叶变 换(Inverse fast fourier transformation,IFFT)。然后检测广义互相关函数的峰值 来获得TDOA。之后将获得的TDOA乘以声速得到距离差,根据几何关系得到 一系列双曲面,通过双曲面的交点即可得到声源位置。声源定位完成。

[0092] 需要指出的是,本实施例以上模块的具体实现过程,只是可以使用于本实 施例中的一种最优的选择,本实施例不限于对以上具体的过程、方法进行适当 的变形或者改变,以实现本发明的具体技术方案,这均在本发明保护的范围之 内。

[0093] 本实施例中,所述去混响处理单元具体用于:

[0094] 对所述语音信号进行短时傅里叶变换;

[0095] 基于变换后的信号计算去除混响后的频域信号;

[0096] 对所述频域信号进行逆短时傅里叶变换,并把变换后的语音信号转到时域。

[0097] 本实施例的一实现方式中,去混响的实现是:首先利用fft加速器在初始化 时对麦克风阵列接受到的语音信号y_t进行短时傅里叶变换(STFT)。然后使用矩 阵乘加速器来计算$\hat{\Lambda}_i^2$和$\hat{w}_t$。之后使用矩阵乘加速器来加速计算去除混响后的频 域信号$\hat{Y}_t$,并利用ifft加速器对$\hat{Y}_t$进行逆短时傅里叶变换(ISTFT),然后把信号 转到时域,完成去混响。而在更新时同样先使用fft加速器对阵列接受的信号y_t进 行短时傅里叶变换(STFT),用矩阵乘加速器来重新计算$\hat{\Lambda}_i^2$,对每一个频点的 每一帧数据进行更新,并利用前一次更新的$\hat{w}_t$和$\hat{Y}_t$计算得到更新后的$\hat{w}_t$。最后用 矩阵乘加速器来加速计算去除混响后的频域信号$\hat{Y}_t$,并利用ifft加速器对$\hat{Y}_t$进行 逆短时傅里叶变换(ISTFT),之后把信号转到时域,到此去混响完成。

[0098] 需要指出的是,本实施例以上模块的具体实现过程,只是可以使用于本实 施例中的一种最优的选择,本实施例不限于对以上具体的过程、方法进行适当 的变形或者改变,以实现本发明的具体技术方案,这均在本发明保护的范围之 内。

[0099] 本实施例的一实现方式中,所述语音活动检测模块具体用于:

[0100] 对所述语音信号进行预处理;

[0101] 将经过预处理的信号逐帧使用FilterBank算法来提取FBank特征;

[0102] 将每一帧的FBank特征输入深度神经网络模型,由深度神经网络模型输出 每一帧在音素集中每一个音素的输出概率;

[0103] 将每一帧中对应所有非噪声、非静音的音素对应的输出概率进行求和;

[0104] 如果所述和值大于预设阈值,则判断对应帧为语音端点;

[0105] 当最后一帧被判断后,则对之前的判决结果进行平滑滤波处理,得到最终 的语音端点判决结果。

[0106] 需要指出的是,本实施例中的VAD模块、语音唤醒模块和语音识别模块主 要使用了如下硬件加速器:simd加速器,数学运算加速器(dma),fft/ifft加速器, 神经网络加速器(Neural-network process units,NPU)。其中NPU可以灵活的支持 各类神经网络模型,

主要有:深度神经网络(Deep neural network,DNN),循环 递归神经网络(Recurrent neural network,RNN),卷积神经网络(Convolutional neural network,CNN),时延神经网络(Time delay neural network,TDNN)等。

[0107] 在本实施例中VAD模块的实现是:首先对传入的语音信号进行预处理,其 中包括分帧和预滤波等。然后将经过预处理的信号逐帧使用FilterBank算法来提 取FBank特征。之后进行端点判决,即通过一个训练好的对音素进行分类的深 度神经网络(Deep neural network,DNN)模型,输入每一帧的FBank特征,由 该模型输出对应每一帧在音素集中每一个音素的后验概率(也叫输出概率)。然 后将所有非噪声,非静音的音素对应的输出概率进行求和,如果大于设定的阈 值,则认为该帧为语音。当最后一帧信号经过端点判决后,则进行后处理操作, 即对之前的判决结果进行平滑滤波处理,得到最终的语音端点判决结果,至此 语音活动检测完成。其中,VAD模块是利用神经网络加速器(Neural-network process units,NPU)来加速实现的。

[0108] 需要指出的是,本实施例以上模块的具体实现过程,只是可以使用于本实 施例中的一种最优的选择,本实施例不限于对以上具体的过程、方法进行适当 的变形或者改变,以实现本发明的具体技术方案,这均在本发明保护的范围之 内。

[0109] 本实施例的一实现方式中,所述语音唤醒模块具体用于:

[0110] 对所述语音信号进行语音特征提取,得到对应的语音特征向量;

[0111] 将所述语音特征向量输入到DNN模型,得到该语音特征向量对应的语音信 号是关键词或者是非关键词的后验概率;

[0112] 对所述后验概率进行平滑得到对应的置信度;

[0113] 如果所述置信度大于预设值,则判断所述语音特征向量对应的语音信号包 含关键词;

[0114] 如果所述关键词按设定顺序出现则判断唤醒所述电子设备。

[0115] 本实施例的一实现方式中,语音唤醒的实现是:采用端到端的模式,即输 入的是语音信号,输出直接为关键词。首先对输入的语音信号进行语音特征提 取,采用的是MFCC(Mel-frequency cepstral coefficients)算法。其中,在采用 MFCC算法提取语音特征之前先对传入的语音信号做前期处理,包括模数转换, 预加重和分帧加窗。之后进行快速离散傅里叶变换和Mel滤波,最后进行倒谱、能量和差分即可得到MFCC参数向量。之后将得到的语音特征向量输入到DNN 模型(深度神经网络,Deep Neural Networks,DNN),通过训练DNN来预测输 入的语音特征是关键词或者是非关键词的后验概率然后将其输出。之后将得到 的的后验值通过后处理模型,因为后验值是以帧为单位输出所以需要以一定的 窗长来进行平滑,对后验值进行平滑后即可得到关键词的置信度。如果该置信 度大于设定的阈值,则认为关键词出现。而如果关键词按设定顺序出现则认为 唤醒,同时设置一系列参数来限制可能的误唤醒,到此语音唤醒结束。其中, 语音唤醒模块是利用神经网络加速器(NPU)来加速实现的。

[0116] 需要指出的是,本实施例以上模块的具体实现过程,只是可以使用于本实 施例中的一种最优的选择,本实施例不限于对以上具体的过程、方法进行适当 的变形或者改变,以实现本发明的具体技术方案,这均在本发明保护的范围之 内。

[0117] 本实施例的一实现方式中,所述语音识别模块具体用于:

[0118] 提取所述语音信号中的语音特征向量；

[0119] 对所述语音特征向量进行语音解码，得到最优的输出词序列；

[0120] 基于所述输出词序列输出对应的文本；

[0121] 基于所述文本确定所述文本表达的交互指令，并使电子设备执行所述交互 指令。

[0122] 本实施例的一实现方式中，语音识别的一种实现方式是：先采用MFCC算 法提取语音特征向量。然后将提取到的语音特征向量进行语音解码，而语音解 码过程就是通过声学模型，发音字典和语言模型对提取特征后的语音数据进行 文字输出。声学模型是使用TDNN-HMM模型，其中TDNN即为时延深度神经 网络模型，HMM是隐马尔可夫模型(Hidden markov model，HMM)，是根据语 音数据库的特征参数训练出的声学模型参数，然后在识别时对该模型输入提取 到的语音特征向量与声学模型进行匹配，得到识别结果也就是音素信息。其中 使用TDNN模型来拟合概率密度函数，进行HMM的状态建模。HMM模型中 是利用前向算法和后向算法来解决概率计算问题，用Baum-Welch算法解决学习 问题，而且使用的是三音素HMM模型并通过决策树来提高每一类的训练量。发音字典是根据声学模型识别出来的音素信息，找到对应的字或者词，将声学 模型与语言模型联结起来。而语言模型是通过对大量文本信息进行训练得到的，结合语法和语义的知识描述词之间的内在关系，对于发音字典找到的字或者词 得到概率最大的词序列。之后将训练好的声学模型，发音词典，语言模型构建 为一个状态网络。解码是使用Viterbi算法来进行的，即从构建的状态网络中找 到与语音最匹配的路径，得到最优的输出词序列。最终输出文本，就完成了语 法识别过程。其中，语音识别模块是利用神经网络加速器(NPU)来加速进行的。

[0123] 语音识别的另一种实现方式是：采用RNN-CTC来构建声学模型，其中RNN 即为循环神经网络，CTC(Connectionist temporal classification)用作损失函数的 声学模型训练，省去数据对齐和标注，同时对汉语声韵母，音素以及状态等多 种语言结构进行分析建模；此方法通过BP算法(errorBackPropagation)进行 训练，最后语音输出是一段尖峰序列，非语音部分为空白部分；由于输出的尖 峰序列对应多条路径，所以采用前后向算法进行计算简化。发音字典是根据声 学模型识别出来的音素信息，找到对应的字或者词，将声学模型与语言模型联 结起来。而语言模型采用N-gram+LSTM建模而成，其中N-gram是一种统计 语言模型，根据前(n-1)个item预测第n个item，这些item可以是音素，字符， 词等等，是目前最常用的语言模型；LSTM(Long Short Term Memory networks) 是一种特殊的循环神经网络(RNN)，通过元胞状态(Cell State)结构，能够学习到长 期的依赖关系；N-gram+LSTM模型克服了单独的N-gram模型对于长时间依赖 关系失效的问题，此模型通过对大量文本信息进行训练得到，结合语法和语义 的知识描述词之间的内在关系，对于发音字典找到的字或者词得到概率最大的 词序列。之后将训练好的声学模型，发音词典，语言模型构建为一个状态网络。 解码是使用Viterbi算法来进行的，即从构建的状态网络中找到与语音最匹配的 路径，得到最优的输出词序列。最终输出文本，就完成了语法识别过程。本地语音识别是利用神经网络加速器(NPU)来加速进行的。

[0124] 需要指出的是，本实施例以上所述每个模块的具体实现过程，只是可以使 用于本实施例中的一种最优的选择，本实施例不限于对以上具体的过程、方法 进行适当的变形或者改变，以实现本发明的具体技术方案，这均在本发明保护 的范围之内。

[0125] 本实施是由基本的音频采集模块，前端阵列信号处理模块，VAD模块，语 音唤醒模

块，语法识别模块组成的智能语音芯片。在此基础上，前端阵列处理 模块是基于fft/ifft加速器，矩阵乘加速器，求逆加速器，求行列式加速器，求 特征值特征向量加速器，simd加速器，dma加速器，求cholesky积加速器实现 的；VAD模块是基于神经网络加速器实现的；语音唤醒模块是基于神经网络加 速器实现的；本地语音识别模块是基于神经网络加速器实现的。

[0126] 本实施例还提供了一种电子设备，所述电子设备包含权利要求1-9中任一项 所述的语音芯片。

[0127] 以上所述实施例的各技术特征可以进行任意的组合，为使描述简洁，未对 上述实施例中的各个技术特征所有可能的组合都进行描述，然而，只要这些技 术特征的组合不存在矛盾，都应当认为是本说明书记载的范围。

[0128] 以上所述实施例仅表达了本发明的几种实施方式，其描述较为具体和详细， 但并不能因此而理解为对发明专利范围的限制。应当指出的是，对于本领域的 普通技术人员来说，在不脱离本发明构思的前提下，还可以做出若干变形和改 进，这些都属于本发明的保护范围。因此，本发明专利的保护范围应以所附权 利要求为准。

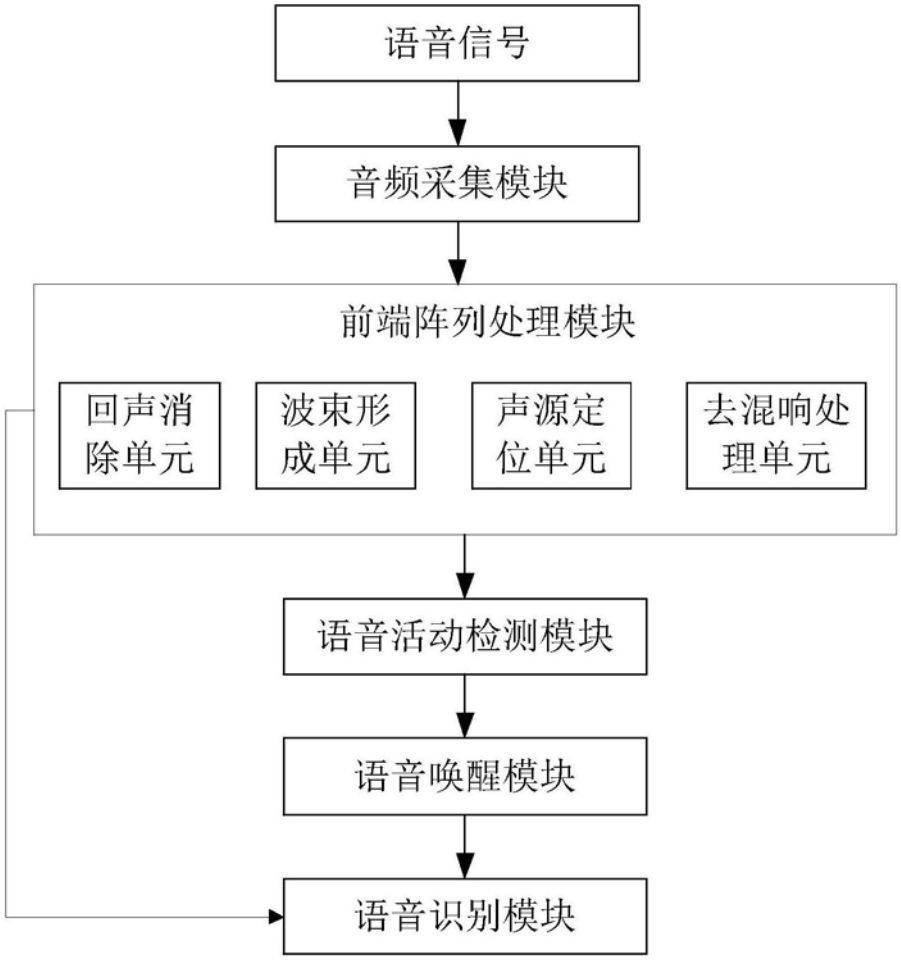

语音信号
音频采集模块
前端阵列处理模块
回声消除单元
波束形成单元
声源定位单元
去混响处理单元
语音活动检测模块
语音唤醒模块
语音识别模块

图1