(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 107221320 A
(43)申请公布日 2017.09.29

(21)申请号 201710359207.1

(22)申请日 2017.05.19

(71)申请人 百度在线网络技术(北京)有限公司
地址 100085 北京市海淀区上地十街10号百度大厦

(72)发明人 蒋兵 马啸空 李超 李先刚

(74)专利代理机构 北京鸿德海业知识产权代理事务所(普通合伙) 11412
代理人 袁媛

(51)Int.Cl.
G10L 15/02(2006.01)
G10L 15/06(2013.01)
G10L 15/16(2006.01)
G10L 25/30(2013.01)

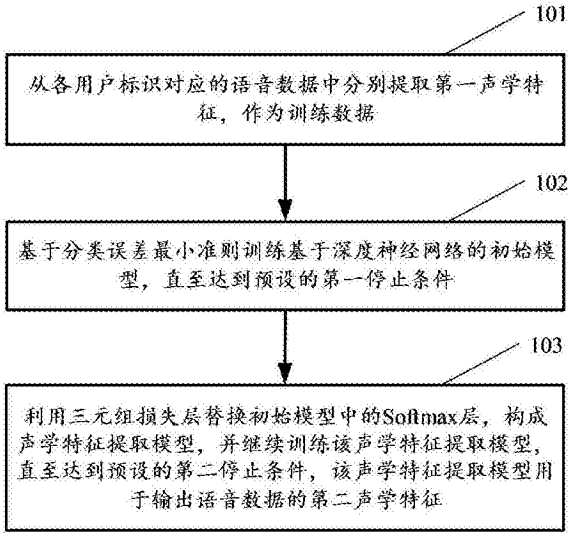
权利要求书3页 说明书13页 附图5页

(54)发明名称

训练声学特征提取模型的方法、装置、设备和计算机存储介质

(57)摘要

本发明提供了一种训练声学特征提取模型的方法、装置、设备和计算机存储介质，其中方法包括：将从各用户标识对应的语音数据中分别提取的第一声学特征，作为训练数据；基于分类误差最小准则训练基于深度神经网络的初始模型，直至达到预设的第一停止条件；利用三元组损失层替换所述初始模型中的Softmax层，构成声学特征提取模型，并继续训练所述声学特征提取模型，直至达到预设的第二停止条件，所述声学特征提取模型用于输出语音数据的第二声学特征；其中所述三元组损失层用于最大化相同用户的第二声学特征之间的相似度且最小化不同用户的第二声学特征之间的相似度。



CN 107221320 A

1.一种训练声学特征提取模型的方法，其特征在于，该方法包括：

将从各用户标识对应的语音数据中分别提取的第一声学特征，作为训练数据；

基于分类误差最小准则训练基于深度神经网络的初始模型，直至达到预设的第一停止条件；

利用三元组损失层替换所述初始模型中的Softmax层，构成声学特征提取模型，并继续训练所述声学特征提取模型，直至达到预设的第二停止条件，所述声学特征提取模型用于输出语音数据的第二声学特征；

其中所述三元组损失层用于最大化相同用户的第二声学特征之间的相似度且最小化不同用户的第二声学特征之间的相似度。

2.根据权利要求1所述的方法，其特征在于，所述第一声学特征包括：Fbank64声学特征。

3.根据权利要求1所述的方法，其特征在于，所述基于深度神经网络的初始模型包括：深度神经网络、池化层和Softmax层；

所述深度神经网络用于对各语音数据的第一声学特征进行学习，输出帧级别的第二声学特征；

所述池化层用于对帧级别的第二声学特征进行句子平均化处理，得到句子级别的第二声学特征；

Softmax层用于将句子级别的第二声学特征映射至各用户标识对应的分类上。

4.根据权利要求1所述的方法，其特征在于，所述分类误差最小准则包括交叉熵准则。

5.根据权利要求1所述的方法，其特征在于，所述深度神经网络包括：卷积神经网络CNN、残差卷积神经网络ResCNN或者门控递归单元GRU。

6.根据权利要求1所述的方法，其特征在于，所述第一停止条件包括以下至少之一：

迭代值趋于收敛；

所述初始模型在测试集上的性能趋于收敛。

7.根据权利要求1所述的方法，其特征在于，所述三元损失层具体用于利用所述各语音数据的第二声学特征计算三元组损失，利用所述三元组损失对所述深度神经网络进行调参，以最小化所述三元组损失；

其中，所述三元组损失体现不同用户的第二声学特征之间的相似度与相同用户的第二声学特征之间的相似度的差值状况。

8.根据权利要求1所述的方法，其特征在于，所述第二停止条件包括以下至少一种：

迭代值趋于收敛；

所述声学特征提取模型在测试集上的性能趋于收敛。

9.一种提取声学特征的方法，其特征在于，该方法包括：

提取待处理语音数据的第一声学特征；

将所述第一声学特征输入声学特征提取模型，得到待处理语音数据的第二声学特征；

其中所述声学特征提取模型是采用如权利要求1至8任一权利要求所述方法预先训练得到的。

10.一种训练声学特征提取模型的装置，其特征在于，该装置包括：

训练数据获取单元，用于将从各用户标识对应的语音数据中分别提取的第一声学特

征，作为训练数据；

初始模型训练单元，用于基于分类误差最小准则训练基于深度神经网络的初始模型，直至达到预设的第一停止条件；

提取模型训练单元，用于利用三元组损失层替换所述初始模型中的Softmax层，构成声学特征提取模型，并继续训练所述声学特征提取模型，直至达到预设的第二停止条件，所述声学特征提取模型用于输出语音数据的第二声学特征；

其中所述三元组损失层用于最大化相同用户的第二声学特征之间的相似度且最小化不同用户的第二声学特征之间的相似度。

11.根据权利要求10所述的装置，其特征在于，所述第一声学特征包括：FBank64声学特征。

12.根据权利要求10所述的装置，其特征在于，所述基于深度神经网络的初始模型包括：深度神经网络、池化层和Softmax层；

所述深度神经网络用于对各语音数据的第一声学特征进行学习，输出帧级别的第二声学特征；

所述池化层用于对帧级别的第二声学特征进行句子平均化处理，得到句子级别的第二声学特征；

Softmax层用于将句子级别的第二声学特征映射至各用户标识对应的分类上。

13.根据权利要求12所述的装置，其特征在于，所述分类误差最小准则包括：交叉熵准则。

14.根据权利要求10所述的装置，其特征在于，所述深度神经网络包括：卷积神经网络CNN、残差卷积神经网络ResCNN或者门控递归单元GRU。

15.根据权利要求10所述的装置，其特征在于，所述第一停止条件包括以下至少之一：

迭代值趋于收敛；

所述初始模型在测试集上的性能趋于收敛。

16.根据权利要求10所述的装置，其特征在于，所述三元损失层具体用于利用所述各语音数据的第二声学特征计算三元组损失，利用所述三元组损失对所述深度神经网络进行调参，以最小化所述三元组损失；

其中，所述三元组损失体现不同用户的第二声学特征之间的相似度与相同用户的第二声学特征之间的相似度的差值状况。

17.根据权利要求10所述的装置，其特征在于，所述第二停止条件包括以下至少一种：

迭代值趋于收敛；

所述声学特征提取模型在测试集上的性能趋于收敛。

18.一种提取声学特征的装置，其特征在于，该装置包括：

预处理单元，用于提取待处理语音数据的第一声学特征；

特征提取单元，用于将所述第一声学特征输入声学特征提取模型，得到待处理语音数据的第二声学特征；

其中所述声学特征提取模型是由权利要求10至17任一权利要求所述装置预先训练得到的。

19.一种设备，其特征在于，所述设备包括：

一个或多个处理器；

存储装置，用于存储一个或多个程序，

当所述一个或多个程序被所述一个或多个处理器执行，使得所述一个或多个处理器实现如权利要求1-9中任一所述的方法。

20.一种包含计算机可执行指令的存储介质，所述计算机可执行指令在由计算机处理器执行时用于执行如权利要求1-9中任一所述的方法。

训练声学特征提取模型的方法、装置、设备和计算机存储介质

【技术领域】

[0001] 本发明涉及计算机应用技术领域，特别涉及一种训练声学特征提取模型的方法、装置、设备和计算机存储介质。

【背景技术】

[0002] 随着人工智能的不断发展，语音交互已经成为最自然的交互方式之一得到日益推广，声音识别技术也越来越得到人们的重视。在声音识别技术中，声学特征的提取是核心技术，其可以用于用户识别、验证或分类等。

[0003] 在现有声学特征提取模型中，越来越多的采用了深度神经网络，在使用深度神经网络进行模型训练过程中，采用的大多是Softmax回归方式。但这种Softmax回归方式在特征空间很大时，性能较差且准确性也较低。

【发明内容】

[0004] 本发明提供了一种训练声学特征提取模型的方法、装置、设备和计算机存储介质，以便于提高模型训练的性能和准确性。

[0005] 具体技术方案如下：

[0006] 本发明提供了一种训练声学特征提取模型的方法，该方法包括：

[0007] 将从各用户标识对应的语音数据中分别提取的第一声学特征，作为训练数据；

[0008] 基于分类误差最小准则训练基于深度神经网络的初始模型，直至达到预设的第一停止条件；

[0009] 利用三元组损失层替换所述初始模型中的Softmax层，构成声学特征提取模型，并继续训练所述声学特征提取模型，直至达到预设的第二停止条件，所述声学特征提取模型用于输出语音数据的第二声学特征；

[0010] 其中所述三元组损失层用于最大化相同用户的第二声学特征之间的相似度且最小化不同用户的第二声学特征之间的相似度。

[0011] 根据本发明一优选实施方式，所述第一声学特征包括：Fbank64声学特征。

[0012] 根据本发明一优选实施方式，所述基于深度神经网络的初始模型包括：深度神经网络、池化层和Softmax层；

[0013] 所述深度神经网络用于对各语音数据的第一声学特征进行学习，输出帧级别的第二声学特征；

[0014] 所述池化层用于对帧级别的第二声学特征进行句子平均化处理，得到句子级别的第二声学特征；

[0015] Softmax层用于将句子级别的第二声学特征映射至各用户标识对应的分类上。

[0016] 根据本发明一优选实施方式，所述分类误差最小准则包括交叉熵准则。

[0017] 根据本发明一优选实施方式，所述深度神经网络包括：卷积神经网络CNN、残差卷积神经网络ResCNN或者门控递归单元GRU。

[0018] 根据本发明一优选实施方式，所述第一停止条件包括以下至少之一：

[0019] 迭代值趋于收敛；

[0020] 所述初始模型在测试集上的性能趋于收敛。

[0021] 根据本发明一优选实施方式，所述三元损失层具体用于利用所述各语音数据的第二声学特征计算三元组损失，利用所述三元组损失对所述深度神经网络进行调参，以最小化所述三元组损失；

[0022] 其中，所述三元组损失体现不同用户的第二声学特征之间的相似度与相同用户的第二声学特征之间的相似度的差值状况。

[0023] 根据本发明一优选实施方式，所述第二停止条件包括以下至少一种：

[0024] 迭代值趋于收敛；

[0025] 所述声学特征提取模型在测试集上的性能趋于收敛。

[0026] 本发明还提供了一种提取声学特征的方法，其特征在于，该方法包括：

[0027] 提取待处理语音数据的第一声学特征；

[0028] 将所述第一声学特征输入声学特征提取模型，得到待处理语音数据的第二声学特征；

[0029] 其中所述声学特征提取模型是采用上述训练声学特征提取模型的方法预先训练得到的。

[0030] 本发明还提供了一种训练声学特征提取模型的装置，该装置包括：

[0031] 训练数据获取单元，用于将从各用户标识对应的语音数据中分别提取的第一声学特征，作为训练数据；

[0032] 初始模型训练单元，用于基于分类误差最小准则训练基于深度神经网络的初始模型，直至达到预设的第一停止条件；

[0033] 提取模型训练单元，用于利用三元组损失层替换所述初始模型中的Softmax层，构成声学特征提取模型，并继续训练所述声学特征提取模型，直至达到预设的第二停止条件，所述声学特征提取模型用于输出语音数据的第二声学特征；

[0034] 其中所述三元组损失层用于最大化相同用户的第二声学特征之间的相似度且最小化不同用户的第二声学特征之间的相似度。

[0035] 根据本发明一优选实施方式，所述第一声学特征包括：FBank64声学特征。

[0036] 根据本发明一优选实施方式，所述基于深度神经网络的初始模型包括：深度神经网络、池化层和Softmax层；

[0037] 所述深度神经网络用于对各语音数据的第一声学特征进行学习，输出帧级别的第二声学特征；

[0038] 所述池化层用于对帧级别的第二声学特征进行句子平均化处理，得到句子级别的第二声学特征；

[0039] Softmax层用于将句子级别的第二声学特征映射至各用户标识对应的分类上。

[0040] 根据本发明一优选实施方式，所述分类误差最小准则包括：交叉熵准则。

[0041] 根据本发明一优选实施方式，所述深度神经网络包括：卷积神经网络CNN、残差卷积神经网络ResCNN或者门控递归单元GRU。

[0042] 根据本发明一优选实施方式，所述第一停止条件包括以下至少之一：

[0043] 迭代值趋于收敛；

[0044] 所述初始模型在测试集上的性能趋于收敛。

[0045] 根据本发明一优选实施方式，所述三元损失层具体用于利用所述各语音数据的第二声学特征计算三元组损失，利用所述三元组损失对所述深度神经网络进行调参，以最小化所述三元组损失；

[0046] 其中，所述三元组损失体现不同用户的第二声学特征之间的相似度与相同用户的第二声学特征之间的相似度的差值状况。

[0047] 根据本发明一优选实施方式，所述第二停止条件包括以下至少一种：

[0048] 迭代值趋于收敛；

[0049] 所述声学特征提取模型在测试集上的性能趋于收敛。

[0050] 本发明还提供了一种提取声学特征的装置，该装置包括：

[0051] 预处理单元，用于提取待处理语音数据的第一声学特征；

[0052] 特征提取单元，用于将所述第一声学特征输入声学特征提取模型，得到待处理语音数据的第二声学特征；

[0053] 其中所述声学特征提取模型是由上述训练声学特征提取模型的装置预先训练得到的。

[0054] 本发明还提供了一种设备，所述设备包括：

[0055] 一个或多个处理器；

[0056] 存储装置，用于存储一个或多个程序，

[0057] 当所述一个或多个程序被所述一个或多个处理器执行，使得所述一个或多个处理器实现上述方法。

[0058] 本发明还提供了一种包含计算机可执行指令的存储介质，所述计算机可执行指令在由计算机处理器执行时用于执行上述方法。

[0059] 由以上技术方案可以看出，在本发明中首先基于分类误差最小准则训练基于深度神经网络的初始模型，直至达到预设第一停止条件后，利用三元组损失层替换初始模型中的Softmax层，继续训练声学特征提取模型，直至达到预设的第二停止条件。这种预训练+调整的模型训练方式，相比较单纯采用Softmax回归的训练方式，在较大特征空间上能够提高模型训练的性能和准确性。

【附图说明】

[0060] 图1为本发明实施例提供的训练声学特征提取模型的方法流程图；

[0061] 图2为本发明实施例提供的基于深度神经网络的初始模型的结构图；

[0062] 图3为本发明实施例提供的堆叠残差块的示意图；

[0063] 图4为本发明实施例提供的声学特征提取模型的结构图；

[0064] 图5为本发明实施例提供的提取声学特征的方法流程图；

[0065] 图6为本发明实施例提供的训练声学特征提取模型的装置结构图；

[0066] 图7为本发明实施例提供的提取声学特征的装置结构图；

[0067] 图8为实现本发明实施方式的示例性计算机系统/服务器的框图。

【具体实施方式】

[0068] 为了使本发明的目的、技术方案和优点更加清楚，下面结合附图和具体实施例对本发明进行详细描述。

[0069] 在本发明实施例中使用的术语是仅仅出于描述特定实施例的目的，而非旨在限制本发明。在本发明实施例和所附权利要求书中所使用的单数形式的“一种”、“所述”和“该”也旨在包括多数形式，除非上下文清楚地表示其他含义。

[0070] 应当理解，本文中使用的术语“和/或”仅仅是一种描述关联对象的关联关系，表示可以存在三种关系，例如，A和/或B，可以表示：单独存在A，同时存在A和B，单独存在B这三种情况。另外，本文中字符“/”，一般表示前后关联对象是一种“或”的关系。

[0071] 取决于语境，如在此所使用的词语“如果”可以被解释成为“在……时”或“当……时”或“响应于确定”或“响应于检测”。类似地，取决于语境，短语“如果确定”或“如果检测(陈述的条件或事件)”可以被解释成为“当确定时”或“响应于确定”或“当检测(陈述的条件或事件)时”或“响应于检测(陈述的条件或事件)”。

[0072] 本发明的核心思想为：将整个声学特征提取模型的训练分为预训练(pre-training)阶段和调整(fine-tuning)阶段。在预训练阶段中基于分类误差最小准则训练基于深度神经网络的初始模型，在调整阶段利用三元组损失层替换Softmax层，继续进行模型训练，得到最终的声学特征提取模型。下面结合实施例对该方法进行详细描述。

[0073] 图1为本发明实施例提供的训练声学特征提取模型的方法流程图，如图1所示，该方法可以包括以下步骤：

[0074] 在101中，将从各用户标识对应的语音数据中分别提取的第一声学特征，作为训练数据。

[0075] 在本发明中进行声学特征提取模型的训练时，需要首先对训练数据中的语音数据进行低层声学特征的提取，即进行预处理。其中该低层声学特征相对于高层声学特征粒度更粗，包含信息量也更粗；相反，经过声学特征提取模型处理后得到的高层声学特征相对于低层声学特征粒度更细，包含信息量也更细致，更适于建立声纹模型，以进行用户声纹的建立。在本发明实施例中，为了对这两种声学特征进行区分，将对语音数据进行预处理后得到的低层声学特征称为第一声学特征；将经过声学特征提取模型对低层声学特征进行处理后，得到的高层声学特征称为第二声学特征。

[0076] 本步骤中，可以预先采集已知用户的语音数据，在选择训练数据时可以对这些语音数据有一些质量要求，例如选取清晰度较好的语音数据，再例如删除长度过长或过短的语音数据，等等。

[0077] 对于采集到的语音数据首先进行预处理，从中提取各语音数据的第一声学特征。如前面所述的，该第一声学特征是低层的声学特征。在本发明实施例中可以采用FBank(Mel-scale Filter Bank，梅尔标度滤波器组)特征作为第一声学特征。例如，以25ms为一帧、10ms为步长提取语音数据的FBank特征。但本发明并不限于FBank特征，还可以采用其他特征作为第一声学特征。

[0078] 这样，就可以得到各用户标识对应的第一声学特征，从而构成训练数据。其中本发明并不限定用户标识的具体类型，可以是任意类型的标识，只要能够区分用户即可。在训练

数据中可以包含同一用户对应的不同语音数据的第一声学特征，不同用户对应的语音数据的第一声学特征，等等。训练数据中各第一声学特征均具有对应的用户标识作为标签。

[0079] 在102中，基于分类误差最小准则训练基于深度神经网络的初始模型，直至达到预设的第一停止条件。

[0080] 本步骤中涉及的基于深度神经网络的初始模型如图2中所示，主要包括以下三层：深度神经网络、池化层和Softmax层。

[0081] 其中，深度神经网络用于对各语音数据的第一声学特征进行学习，输出帧级别的第二声学特征。

[0082] 深度神经网络可以采用CNN、GRU(Gated Recurrent Unit，门控递归单元)等，当然也可以采用其他诸如RNN、LSTM等其他类型的深度神经网络。由于CNN相比较RNN、LSTM等而言，能够更加有效地减小频谱变化以及将频谱相关性在声学特征中进行体现，因此在本发明实施例中优选CNN这种类型的深度神经网络。

[0083] 然而，尽管深度神经网络具有很好地学习能力，但更难进行训练，在一定深度情况下准确性反而下滑。为了解决该问题，本发明可以基于CNN使用但不限于ResNet(Residual Net，残差网络)型CNN，或者采用GRU。

[0084] 首先对ResNET型CNN进行介绍。

[0085] ResNet可以用于简化CNN的训练。ResNet包括若干ResBlock(堆叠残差块)，各ResBlock包括低层输出和高层输入间的直接连接。如图3中所示，各ResBlock可以定义为：

[0086] $h=F(x,W_i)+x$

[0087] 其中，x和h分别表示ResBlock的输入和输出。F表示堆叠的非线性层的映射函数。

[0088] 如图3所示，ResBlock可以包括两个卷积层和两个激活层。其中，两个卷积层可以包括诸如3×3的过滤器和1×1的stride(步幅)。每个ResBlock包括相同的结构，并且跳转连接是对x的相同映射。若通道的数量增加，则可以使用一个卷积层(例如具有5×5的过滤器和2×2的stride)。因此，频率维度始终在卷积层中保持恒定。经过研究发现，语音识别在时间维度上对stride并不敏感。在本发明实施例中，可以使用如下ReLU函数作为所有激活层的非线性处理：

[0089] $\sigma(x)=\min\{\max\{x,0\},20\}$

[0090] 下面对GRU进行介绍。

[0091] GRU相比较LSTM而言，训练速度更快且发散程度更小。本发明实施例中深度神经网络层可以采用多个GRU构成。例如，每个GRU可以包括一个5×5过滤器和2×2stride的卷积层，能够减少时域和频域的维度，从而允许GRU的计算速度更快。紧接着卷积层的是三个具有1024个单元的前向GRU层，在时间维度上进行循环。在GRU中也可以采用诸如ReLU进行激活。

[0092] 深度神经网络的上层是池化层(Pooling)，池化层用于对深度神经网络输出的帧级别的第二声学特征进行句子平均化处理，得到句子级别的第二声学特征。

[0093] 池化层的输出h'可以采用如下公式：

[0094] $$h'=\frac{1}{T}\sum_{t=0}^{T-1}x'(t)$$

[0095] 其中，T为句子包含的帧数目，x'(t)为池化层的输入。

[0096] 经过池化层的处理，使得本发明实施例提供的声学特征提取模型能够处理不同时长的语句，解决了文本无关的情况。

[0097] 另外，在池化层中还可以包括仿射层(图2中未示出)，用于将句子级别的第二声学特征投射到预设的维度，例如投射到512维度。

[0098] 本实施例中基于深度神经网络的初始模型实际上是一个多分类模型，其输出层为Softmax层，主要作用是将句子级别的第二声学特征映射至各用户标识对应的分类上。

[0099] 在本发明中实际上预训练的过程来初始化深度神经网络的参数，使得初始化深度神经网络能够最小化Softmax层的分类误差。然后再使用三元组损失来继续进行模型训练，调整深度神经网络的参数。其中，分类误差最小准则可以包括交叉熵准则。由于Softmax层和交叉熵准则为现有技术中已有的概念，在此不做详述。

[0100] 由于Softmax层对应的是多分类，因此通常训练样本存在多少个用户，Softmax层就对应有多少个节点。因为训练样本通常采用上万个用户作为发声者来采集样本，因此，Softmax层就存在上万个节点。

[0101] 上述的第一停止条件可以体现在收敛速度或性能达到一定程度上，例如迭代值趋于收敛，或者初始模型在测试集上的性能趋于收敛。举个例子，在初始模型的预训练过程中，会进行分类误差的迭代，从而实现深度神经网络的参数调整。但随着训练的进行，该分类误差会趋于收敛，例如连续若干次的迭代值的变化率小于预设阈值，则可以认为分类误差趋于收敛。再举个例子，在训练过程中，可以将初始模型在测试集上进行测试，测试集可以是由不同于训练数据的已知用户的语音数据构成，利用初始模型在测试集上进行测试，提取第二声学特征，并利用提取的第二声学特征进行用户识别等测试，判断EER或ACC等是否逐渐趋于收敛。

[0102] 在103中，利用三元组损失层替换初始模型中的Softmax层，构成声学特征提取模型，并继续训练该声学特征提取模型，直至达到预设的第二停止条件，该声学特征提取模型用于输出语音数据的第二声学特征。

[0103] 经过上述替换后，声学特征提取模型的结构可以如图4所示，深度神经网络和池化层仍保持训练后的初始模型的情况不变，将Softmax层替换为了三元组损失层。

[0104] 三元组损失层用于最大化相同用户的第二声学特征之间的相似度且最小化不同用户的第二声学特征之间的相似度。具体地，三元组损失层可以利用各语音数据的第二声学特征计算三元组损失，利用三元组损失对深度神经网络进行调参，以最小化三元组损失。其中三元组损失体现不同用户的第二声学特征之间的相似度与相同用户的第二声学特征之间的相似度的差值状况。

[0105] 优选地，在池化层和三元组损失层之间还可以包括长度标准化层，长度标准化层用于将仿射层输出的句子级别的第二声学特征的长度进行规整，使模为1。

[0106] 本发明实施例中，三元损失层采用三元损失对深度神经网络层进行反馈训练，以最小化相同用户的第二声学特征之间的相似度且最大化不同用户的第二声学特征之间的相似度。

[0107] 三元损失层可以采用三个样本作为输入：锚样本，包括一个用户的句子级别的第二声学特征；正样本，包括与锚样本同一用户的另一句子级别的第二声学特征；负样本，包括与锚样本不同用户的句子级别的第二声学特征。将上述样本构成一个三元组。

[0108] 三元损失层对深度神经网络层进行反馈，以使得锚样本和正样本之间的余弦相似度(在本发明实施例中样本之间的相似度采用余弦相似度体现，但不排除其他相似度计算方式)大于锚样本和负样本之间的余弦相似度。形式上，

[0109] $s_i^{ap} - \alpha > s_i^{an}$

[0110] 其中，s_i^{ap}为三元组i中锚样本a和正样本p之间的余弦相似度。s_i^{an}为三元组i中锚样本a和正样本n之间的余弦相似度。训练目标是找到这些相似度中的最小边缘α。即计算三元组损失，该三元组损失体现不同用户的第二声学特征之间的相似度与相同用户的第二声学特征之间的相似度的差值状况。例如该三元组损失的计算函数L可以为：

[0111] $L = \sum_{i=0}^{N} [s_i^{an} - s_i^{ap} + \alpha]_+$

[0112] 其中，N为三元组的数目，操作符$[x]_+ = \max(x, 0)$。

[0113] 计算出的三元组损失反馈给深度神经网络层，以不断调整深度神经网络层的参数，从而逐渐训练深度神经网络，最终最小化利用提取的第二声学特征计算的三元组损失。达到第二停止条件后，此次训练过程结束，得到声学特征提取模型。

[0114] 第二停止条件与第一停止条件类似的，可以包括迭代值趋于收敛，或者声学特征提取模型在测试集上的性能趋于收敛。

[0115] 在采用上述训练方法训练得到声学特征提取模型后，就可以利用声学特征提取模型进行声学特征的提取。提取过程可以如图5中所示。包括以下步骤：

[0116] 在501中，提取待处理语音数据的第一声学特征。

[0117] 本步骤是对待处理语音数据的预处理，即从中提取第一声学特征，该第一声学特征是低层的声学特征。此处提取的第一声学特征的类型和方式与图1所示实施例中步骤101中提取第一声学特征的类型和方式一致。在此不再赘述。

[0118] 在502中，将提取出的第一声学特征输入声学特征提取模型，得到待处理语音数据的第二声学特征。

[0119] 对于预先训练得到的声学特征提取模型，由于其从训练数据中已经完成从第一声学特征到第二声学特征的自学习，因此将步骤501中提取出的待处理语音数据的第一声学特征输入声学特征提取模型，声学特征提取模型就能够输出待处理语音数据的第二声学特征。该第二声学特征可以为句子级别的高层声学特征。

[0120] 在得到待处理语音数据的第二声学特征后，可以利用第二声学特征进行后续应用的处理，在503a中，利用待处理语音数据的第二声学特征，注册该待处理语音数据所对应用户标识的声纹模型，或者在503b中，将待处理语音数据的第二声学特征与已注册的各用户标识的声纹模型进行匹配，确定待处理语音数据对应的用户标识。

[0121] 在503a中，若待处理语音数据对应的用户标识已知，则可以利用提取的第二声学特征注册该用户标识对应的声纹模型。在注册声纹模型时，可以将提取的第二声学特征进行处理后，作为声纹信息存储于声纹模型库中。可以利用用户标识对应的一个或多个第二声学特征来进行声学模型的注册，具体注册过程本发明不做具体限制。

[0122] 在503b中，若待处理语音数据对应的用户标识未知，则可以利用提取的第二声学特征与声纹模型库中各已注册的声纹模型进行匹配，例如通过计算提取的第二声学特征与声纹模型库中各声纹模型之间相似度的方式进行匹配。若匹配到某个声纹模型，则可以确

定该待处理语音数据对应该匹配到的声纹模型对应的用户标识。

[0123] 上述503a和503b是本发明实施例提供的两种在提取语音数据的第二声学特征后，对其的应用方式，当然除了这两种应用方式之外，还可以进行其他应用，本发明不做一一穷举。

[0124] 上述方法可以应用于语音识别系统中，执行主体可以为对应装置，该装置可以是位于用户设备的应用，或者还可以为位于用户设备的应用中的插件或软件开发工具包(Software Development Kit，SDK)等功能单元。其中，用户设备可以包括但不限于诸如：智能移动终端、智能家居设备、网络设备、可穿戴式设备、智能医疗设备、PC(个人计算机)等。其中智能移动设备可以包括诸如手机、平板电脑、笔记本电脑、PDA(个人数字助理)、互联网汽车等。智能家居设备可以包括智能家电设备，诸如智能电视、智能空调、智能热水器、智能冰箱、智能空气净化器等等，智能家居设备还可以包括智能门锁、智能电灯、智能摄像头等。网络设备可以包括诸如交换机、无线AP、服务器等。可穿戴式设备可以包括诸如智能手表、智能眼镜、智能手环、虚拟现实设备、增强现实设备、混合现实设备(即可以支持虚拟现实和增强现实的设备)等等。智能医疗设备可以包括诸如智能体温计、智能血压仪、智能血糖仪等等。

[0125] 图6为本发明实施例提供的训练声学特征提取模型的装置结构图，如图6所示，该装置可以包括：训练数据获取单元01、初始模型训练单元02和提取模型训练单元03。其中各组成单元的主要功能如下：

[0126] 训练数据获取单元01负责将从各用户标识对应的语音数据中分别提取的第一声学特征，作为训练数据。

[0127] 可以预先采集已知用户的语音数据，在选择训练数据时可以对这些语音数据有一些质量要求，例如选取清晰度较好的语音数据，再例如删除长度过长或过短的语音数据，等等。

[0128] 对于采集到的语音数据首先进行预处理，从中提取各语音数据的第一声学特征。如前面所述的，该第一声学特征是低层的声学特征。在本发明实施例中可以采用FBank特征作为第一声学特征。例如，以25ms为一帧、10ms为步长提取语音数据的FBank特征。但本发明并不限于FBank特征，还可以采用其他特征作为第一声学特征。

[0129] 初始模型训练单元02负责基于分类误差最小准则训练基于深度神经网络的初始模型，直至达到预设的第一停止条件。

[0130] 基于深度神经网络的初始模型可以如图2中所示，主要包括以下三层：深度神经网络、池化层和Softmax层。

[0131] 深度神经网络可以采用CNN、GRU等，当然也可以采用其他诸如RNN、LSTM等其他类型的深度神经网络。由于CNN相比较RNN、LSTM等而言，能够更加有效地减小频谱变化以及将频谱相关性在声学特征中进行体现，因此在本发明实施例中优选CNN这种类型的深度神经网络。

[0132] 然而，尽管深度神经网络具有很好地学习能力，但更难进行训练，在一定深度情况下准确性反而下滑。为了解决该问题，本发明可以基于CNN使用但不限于ResNet型CNN，或者采用GRU。

[0133] 深度神经网络的上层是池化层(Pooling)，池化层用于对深度神经网络输出的帧

级别的第二声学特征进行句子平均化处理，得到句子级别的第二声学特征。

[0134] 经过池化层的处理，使得本发明实施例提供的声学特征提取模型能够处理不同时长的语句，解决了文本无关的情况。

[0135] 另外，在池化层中还可以包括仿射层(图2中未示出)，用于将句子级别的第二声学特征投射到预设的维度，例如投射到512维度。

[0136] 本实施例中基于深度神经网络的初始模型实际上是一个多分类模型，其输出层为Softmax层，主要作用是将句子级别的第二声学特征映射至各用户标识对应的分类上。

[0137] 在本发明中实际上预训练的过程来初始化深度神经网络的参数，使得初始化深度神经网络能够最小化Softmax层的分类误差。然后再使用三元组损失来继续进行模型训练，调整深度神经网络的参数。其中，分类误差最小准则可以包括交叉熵准则。由于Softmax层和交叉熵准则为现有技术中已有的概念，在此不做详述。

[0138] 由于Softmax层对应的是多分类，因此通常训练样本存在多少个用户，Softmax层就对应有多少个节点。因为训练样本通常采用上万个用户作为发声者来采集样本，因此，Softmax层就存在上万个节点。

[0139] 上述的第一停止条件可以体现在收敛速度或性能达到一定程度上，例如可以包括但不限于以下至少之一：

[0140] 迭代值趋于收敛，或者初始模型在测试集上的性能趋于收敛。举个例子，在初始模型的预训练过程中，会进行分类误差的迭代，从而实现深度神经网络的参数调整。但随着训练的进行，该分类误差会趋于收敛，例如连续若干次的迭代值的变化率小于预设阈值，则可以认为分类误差趋于收敛。再举个例子，在训练过程中，可以将初始模型在测试集上进行测试，测试集可以是由不同于训练数据的已知用户的语音数据构成，利用初始模型在测试集上进行测试，提取第二声学特征，并利用提取的第二声学特征进行用户识别等测试，判断EER或ACC等是否趋于收敛。

[0141] 提取模型训练单元03负责利用三元组损失层替换初始模型中的Softmax层，构成声学特征提取模型，并继续训练声学特征提取模型，直至达到预设的第二停止条件，声学特征提取模型用于输出语音数据的第二声学特征；其中三元组损失层用于最大化相同用户的第二声学特征之间的相似度且最小化不同用户的第二声学特征之间的相似度。

[0142] 经过上述替换后，声学特征提取模型的结构可以如图4所示，深度神经网络和池化层仍保持训练后的初始模型的情况不变，将Softmax层替换为了三元组损失层。

[0143] 三元组损失层用于最大化相同用户的第二声学特征之间的相似度且最小化不同用户的第二声学特征之间的相似度。具体地，三元组损失层可以利用各语音数据的第二声学特征计算三元组损失，利用三元组损失对深度神经网络进行调参，以最小化三元组损失。其中三元组损失体现不同用户的第二声学特征之间的相似度与相同用户的第二声学特征之间的相似度的差值状况。

[0144] 优选地，在池化层和三元组损失层之间还可以包括长度标准化层，长度标准化层用于将仿射层输出的句子级别的第二声学特征的长度进行规整，使模为1。

[0145] 本发明实施例中，三元损失层采用三元损失对深度神经网络层进行反馈训练，以最大化相同用户的第二声学特征之间的相似度且最小化不同用户的第二声学特征之间的相似度。

[0146] 三元损失层可以采用三个样本作为输入:锚样本,包括一个用户的句子级别的第二声学特征;正样本,包括与锚样本同一用户的另一句子级别的第二声学特征;负样本,包括与锚样本不同用户的句子级别的第二声学特征。将上述样本构成一个三元组。

[0147] 三元损失层对深度神经网络层进行反馈,以使得锚样本和正样本之间的余弦相似度(在本发明实施例中样本之间的相似度采用余弦相似度体现,但不排除其他相似度计算方式)大于锚样本和负样本之间的余弦相似度。即计算三元组损失,该三元组损失体现不同用户的第二声学特征之间的相似度与相同用户的第二声学特征之间的相似度的差值状况。

[0148] 计算出的三元组损失反馈给深度神经网络层,以不断调整深度神经网络层的参数,从而逐渐训练深度神经网络,最终最小化利用提取的第二声学特征计算的三元组损失。达到第二停止条件后,此次训练过程结束,得到声学特征提取模型。

[0149] 第二停止条件与第一停止条件类似的,可以以包括迭代值趋于收敛,或者声学特征提取模型在测试集上的性能趋于收敛。

[0150] 图7为本发明实施例提供的提取声学特征的装置结构图,如图7所示,该装置可以包括:预处理单元11和特征提取单元12。其中各组成单元的主要功能如下:

[0151] 预处理单元11负责提取待处理语音数据的第一声学特征。该第一声学特征的类型和提取方式与图6中数据获取单元01获取训练数据时所采用的第一声学特征的类型和提取方式一致。例如,第一声学特征可以采用FBank特征。

[0152] 特征提取单元12负责将第一声学特征输入声学特征提取模型,得到待处理语音数据的第二声学特征。

[0153] 图8示出了适于用来实现本发明实施方式的示例性计算机系统/服务器012的框图。图8显示的计算机系统/服务器012仅仅是一个示例,不应对本发明实施例的功能和使用范围带来任何限制。

[0154] 如图8所示,计算机系统/服务器012以通用计算设备的形式表现。计算机系统/服务器012的组件可以包括但不限于:一个或者多个处理器或者处理单元016,系统存储器028,连接不同系统组件(包括系统存储器028和处理单元016)的总线018。

[0155] 总线018表示几类总线结构中的一种或多种,包括存储器总线或者存储器控制器,外围总线,图形加速端口,处理器或者使用多种总线结构中的任意总线结构的局域总线。举例来说,这些体系结构包括但不限于工业标准体系结构(ISA)总线,微通道体系结构(MAC)总线,增强型ISA总线、视频电子标准协会(VESA)局域总线以及外围组件互连(PCI)总线。

[0156] 计算机系统/服务器012典型地包括多种计算机系统可读介质。这些介质可以是任何能够被计算机系统/服务器012访问的可用介质,包括易失性和非易失性介质,可移动的和不可移动的介质。

[0157] 系统存储器028可以包括易失性存储器形式的计算机系统可读介质,例如随机存取存储器(RAM)030和/或高速缓存存储器032。计算机系统/服务器012可以进一步包括其它可移动/不可移动的、易失性/非易失性计算机系统存储介质。仅作为举例,存储系统034可以用于读写不可移动的、非易失性磁介质(图8未显示,通常称为“硬盘驱动器”)。尽管图8中未示出,可以提供用于对可移动非易失性磁盘(例如“软盘”)读写的磁盘驱动器,以及对可移动非易失性光盘(例如CD-ROM,DVD-ROM或者其它光介质)读写的光盘驱动器。在这些情况下,每个驱动器可以通过一个或者多个数据介质接口与总线018相连。存储器028可以包括

至少一个程序产品，该程序产品具有一组(例如至少一个)程序模块，这些程序模块被配置以执行本发明各实施例的功能。

[0158] 具有一组(至少一个)程序模块042的程序/实用工具040，可以存储在例如存储器028中，这样的程序模块042包括——但不限于——操作系统、一个或者多个应用程序、其它程序模块以及程序数据，这些示例中的每一个或某种组合中可能包括网络环境的实现。程序模块042通常执行本发明所描述的实施例中的功能和/或方法。

[0159] 计算机系统/服务器012也可以与一个或多个外部设备014(例如键盘、指向设备、显示器024等)通信，在本发明中，计算机系统/服务器012与外部雷达设备进行通信，还可与一个或者多个使得用户能与该计算机系统/服务器012交互的设备通信，和/或与使得该计算机系统/服务器012能与一个或多个其它计算设备进行通信的任何设备(例如网卡，调制解调器等等)通信。这种通信可以通过输入/输出(I/O)接口022进行。并且，计算机系统/服务器012还可以通过网络适配器020与一个或者多个网络(例如局域网(LAN)，广域网(WAN)和/或公共网络，例如因特网)通信。如图所示，网络适配器020通过总线018与计算机系统/服务器012的其它模块通信。应当明白，尽管图8中未示出，可以结合计算机系统/服务器012使用其它硬件和/或软件模块，包括但不限于：微代码、设备驱动器、冗余处理单元、外部磁盘驱动阵列、RAID系统、磁带驱动器以及数据备份存储系统等。

[0160] 处理单元016通过运行存储在系统存储器028中的程序，从而执行各种功能应用以及数据处理，例如实现一种训练声学特征提取模型的方法，可以包括：

[0161] 将从各用户标识对应的语音数据中分别提取的第一声学特征，作为训练数据；

[0162] 基于分类误差最小准则训练基于深度神经网络的初始模型，直至达到预设的第一停止条件；

[0163] 利用三元组损失层替换所述初始模型中的Softmax层，构成声学特征提取模型，并继续训练所述声学特征提取模型，直至达到预设的第二停止条件，所述声学特征提取模型用于输出语音数据的第二声学特征；

[0164] 其中所述三元组损失层用于最小化相同用户的第二声学特征之间的相似度且最大化不同用户的第二声学特征之间的相似度。

[0165] 再例如，实现一种提取声学特征的方法，可以包括：

[0166] 提取待处理语音数据的第一声学特征；

[0167] 将所述第一声学特征输入声学特征提取模型，得到待处理语音数据的第二声学特征。

[0168] 上述的计算机程序可以设置于计算机存储介质中，即该计算机存储介质被编码有计算机程序，该程序在被一个或多个计算机执行时，使得一个或多个计算机执行本发明上述实施例中所示的方法流程和/或装置操作。例如，被上述一个或多个处理器执行的方法流程，可以包括：

[0169] 将从各用户标识对应的语音数据中分别提取的第一声学特征，作为训练数据；

[0170] 基于分类误差最小准则训练基于深度神经网络的初始模型，直至达到预设的第一停止条件；

[0171] 利用三元组损失层替换所述初始模型中的Softmax层，构成声学特征提取模型，并继续训练所述声学特征提取模型，直至达到预设的第二停止条件，所述声学特征提取模型

用于输出语音数据的第二声学特征；

[0172] 其中所述三元组损失层用于最小化相同用户的第二声学特征之间的相似度且最大化不同用户的第二声学特征之间的相似度。

[0173] 再例如，被上述一个或多个处理器执行的方法流程，可以包括：

[0174] 提取待处理语音数据的第一声学特征；

[0175] 将所述第一声学特征输入声学特征提取模型，得到待处理语音数据的第二声学特征。

[0176] 随着时间、技术的发展，介质含义越来越广泛，计算机程序的传播途径不再受限于有形介质，还可以直接从网络下载等。可以采用一个或多个计算机可读的介质的任意组合。计算机可读介质可以是计算机可读信号介质或者计算机可读存储介质。计算机可读存储介质例如可以是——但不限于——电、磁、光、电磁、红外线、或半导体的系统、装置或器件，或者任意以上的组合。计算机可读存储介质的更具体的例子(非穷举的列表)包括：具有一个或多个导线的电连接、便携式计算机磁盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦式可编程只读存储器(EPROM或闪存)、光纤、便携式紧凑磁盘只读存储器(CD-ROM)、光存储器件、磁存储器件、或者上述的任意合适的组合。在本文件中，计算机可读存储介质可以是任何包含或存储程序的有形介质，该程序可以被指令执行系统、装置或者器件使用或者与其结合使用。

[0177] 计算机可读的信号介质可以包括在基带中或者作为载波一部分传播的数据信号，其中承载了计算机可读的程序代码。这种传播的数据信号可以采用多种形式，包括——但不限于——电磁信号、光信号或上述的任意合适的组合。计算机可读的信号介质还可以是计算机可读存储介质以外的任何计算机可读介质，该计算机可读介质可以发送、传播或者传输用于由指令执行系统、装置或者器件使用或者与其结合使用的程序。

[0178] 计算机可读介质上包含的程序代码可以用任何适当的介质传输，包括——但不限于——无线、电线、光缆、RF等等，或者上述的任意合适的组合。

[0179] 可以以一种或多种程序设计语言或其组合来编写用于执行本发明操作的计算机程序代码，所述程序设计语言包括面向对象的程序设计语言—诸如Java、Smalltalk、C++，还包括常规的过程式程序设计语言—诸如“C”语言或类似的程序设计语言。程序代码可以完全地在用户计算机上执行、部分地在用户计算机上执行、作为一个独立的软件包执行、部分在用户计算机上部分在远程计算机上执行、或者完全在远程计算机或服务器上执行。在涉及远程计算机的情形中，远程计算机可以通过任意种类的网络——包括局域网(LAN)或广域网(WAN)连接到用户计算机，或者，可以连接到外部计算机(例如利用因特网服务提供商来通过因特网连接)。

[0180] 由以上描述可以看出，本发明提供的方法、装置、设备和计算机存储设备可以具备以下优点：

[0181] 1)在本发明中首先基于分类误差最小准则训练基于深度神经网络的初始模型，直至达到预设第一停止条件后，利用三元组损失层替换初始模型中的Softmax层，继续训练声学特征提取模型，直至达到预设的第二停止条件。这种pre-training+fine-tuning的模型训练方式，相比较单纯采用Softmax回归的训练方式，在较大特征空间上能够提高模型训练的性能和准确性。

[0182] 2)这种pre-training+fine-tuning的模型训练方式，相比较直接利用三元组损失训练声学特征提取模型的方式，除了具备更高的训练效率之外，经过试验证明，具有更高的验证集准确率(ACC)和等错误率(EER)。

[0183] 3)本发明中优选ResCNN或GRU类型的深度神经网络，从而在采用较高层级深度的神经网络情况下，也能够保证特征提取的准确性，且提高深度神经网络的训练速度。

[0184] 4)本发明在训练声学特征提取模型的过程中，对深度神经网络的输出进行池化和句子标准化处理，使得该模型除了能够对文本相关的语音数据进行特征提取之外，也能够对文本无关的语音数据进行很好地特征提取。

[0185] 5)经过试验后发现，本发明能够更好的处理大规模的语音数据并且能够很好地适应不同语言的处理。

[0186] 在本发明所提供的几个实施例中，应该理解到，所揭露的方法和装置，可以通过其它的方式实现。例如，以上所描述的装置实施例仅仅是示意性的，例如，所述单元的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式。

[0187] 所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

[0188] 另外，在本发明各个实施例中的各功能单元可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中。上述集成的单元既可以采用硬件的形式实现，也可以采用硬件加软件功能单元的形式实现。

[0189] 以上所述仅为本发明的较佳实施例而已，并不用以限制本发明，凡在本发明的精神和原则之内，所做的任何修改、等同替换、改进等，均应包含在本发明保护的范围之内。

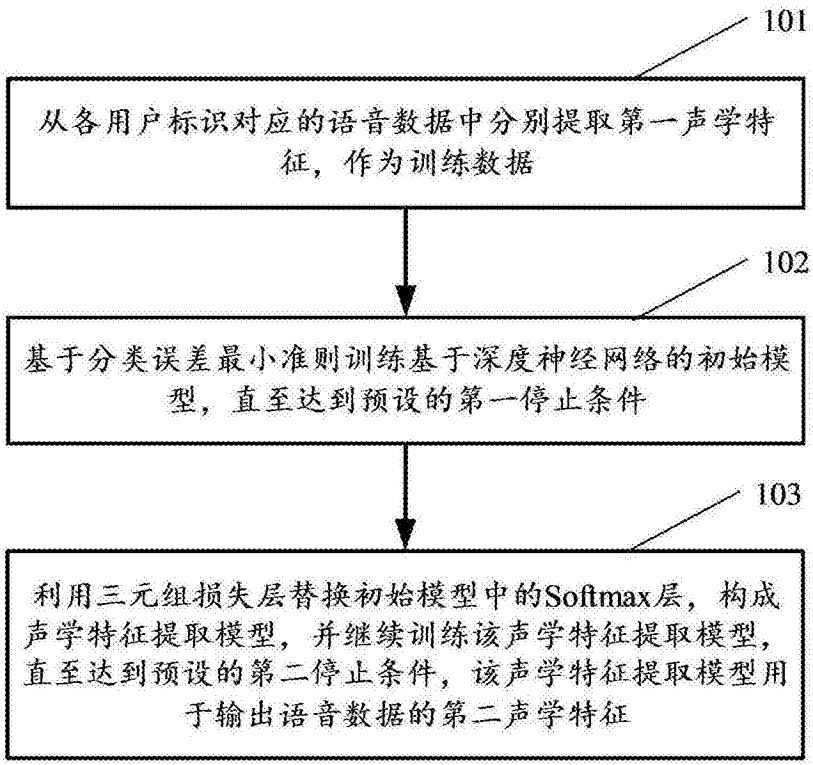

101
从各用户标识对应的语音数据中分别提取第一声学特征，作为训练数据
102
基于分类误差最小准则训练基于深度神经网络的初始模型，直至达到预设的第一停止条件
103
利用三元组损失层替换初始模型中的Softmax层，构成声学特征提取模型，并继续训练该声学特征提取模型，直至达到预设的第二停止条件，该声学特征提取模型用于输出语音数据的第二声学特征

图1


深度神经网络层
池化层
Softmax层

图2

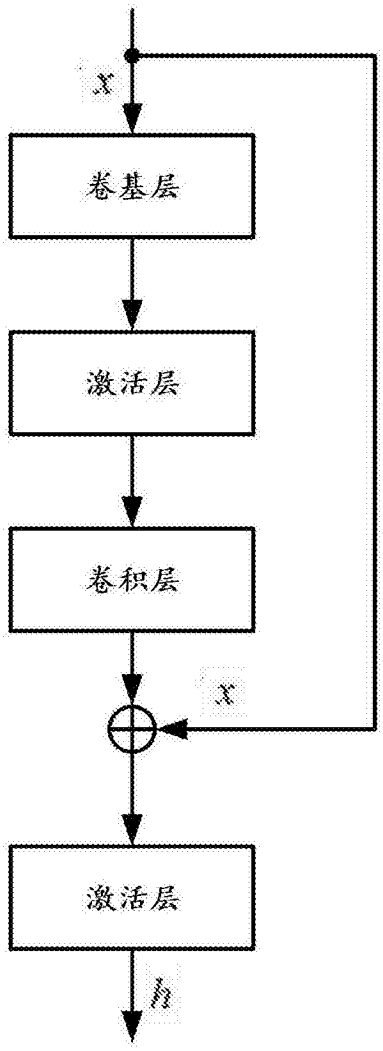


图3

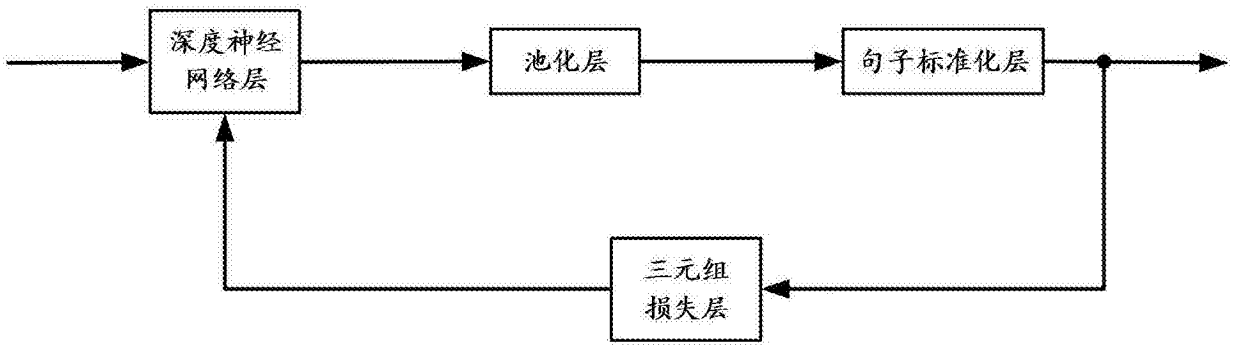


图4

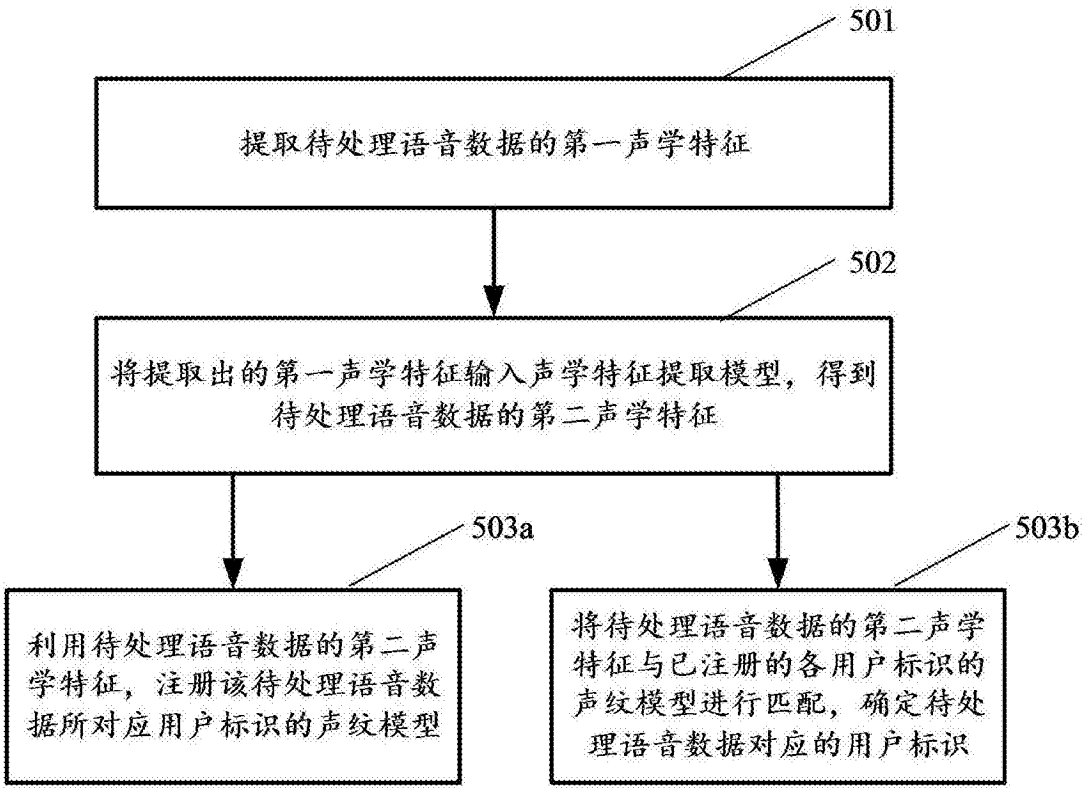

501
提取待处理语音数据的第一声学特征
502
将提取出的第一声学特征输入声学特征提取模型，得到待处理语音数据的第二声学特征
503a
利用待处理语音数据的第二声学特征，注册该待处理语音数据所对应用户标识的声纹模型
503b
将待处理语音数据的第二声学特征与已注册的各用户标识的声纹模型进行匹配，确定待处理语音数据对应的用户标识

图5

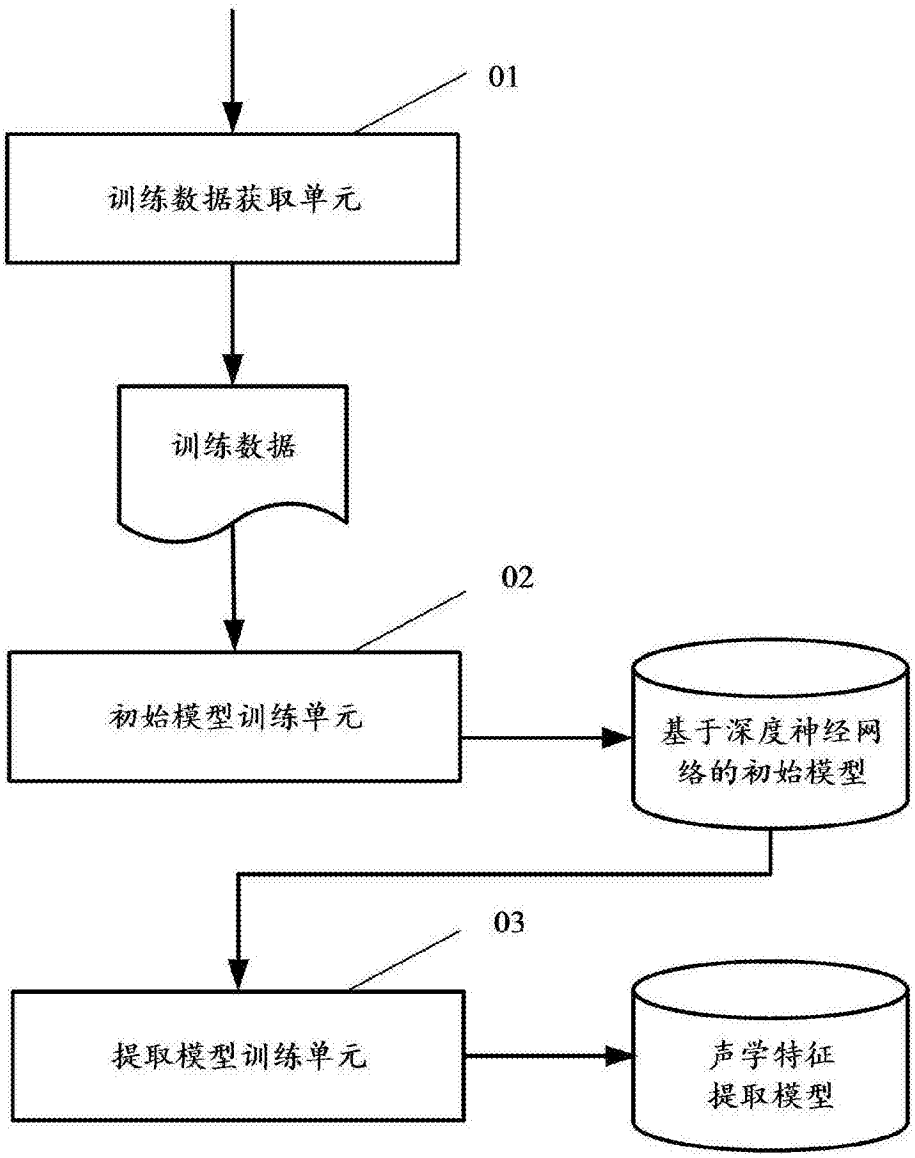

01
训练数据获取单元
训练数据
02
初始模型训练单元
基于深度神经网络的初始模型
03
提取模型训练单元
声学特征提取模型

图6

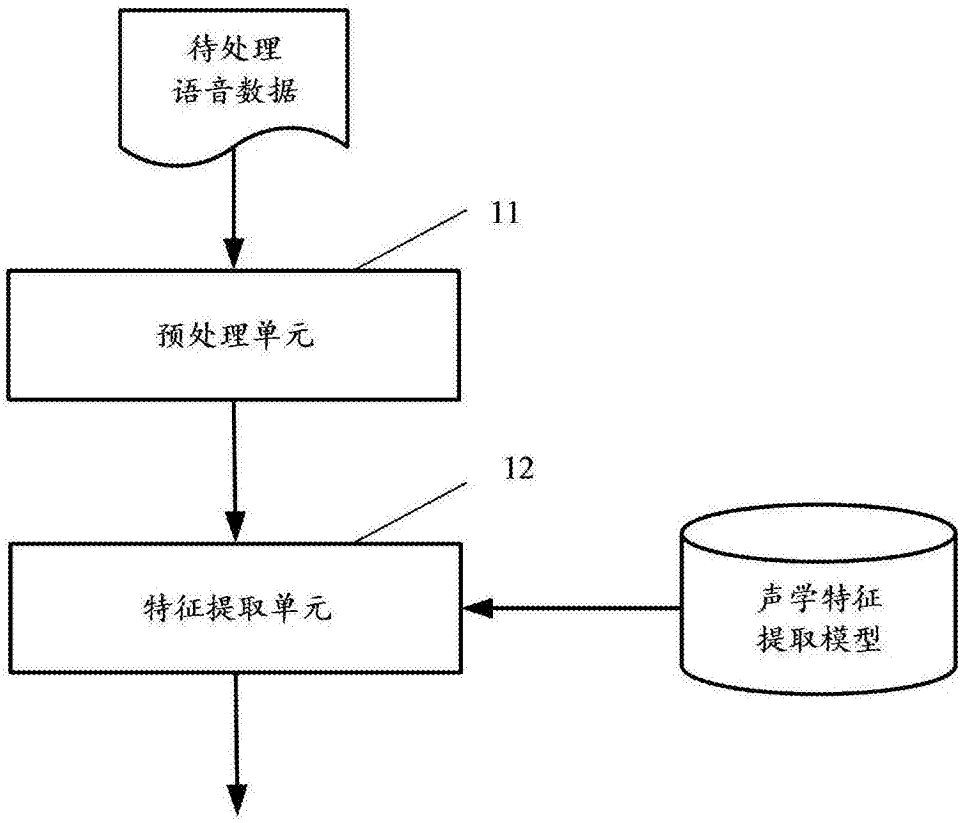

待处理
语音数据
11
预处理单元
12
特征提取单元
声学特征
提取模型

图7

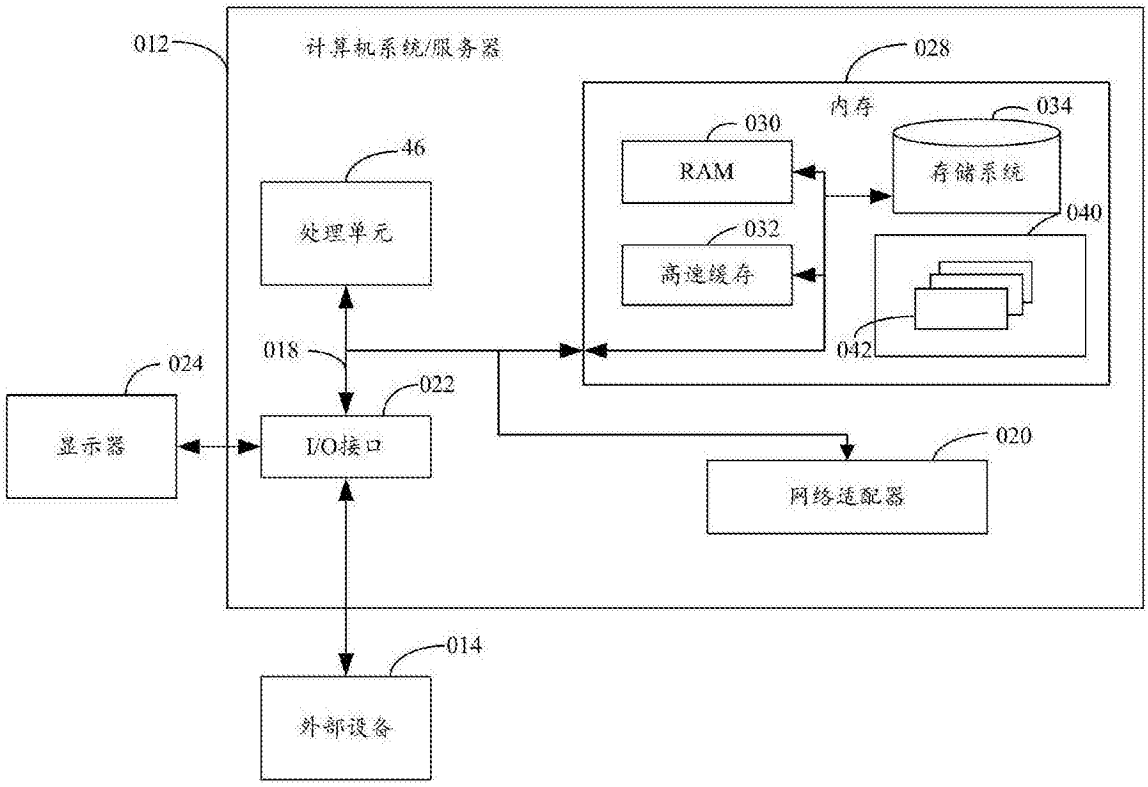

012
计算机系统/服务器
028
内存
030
RAM
034
存储系统
040
032
高速缓存
042
46
处理单元
018
024
显示器
022
I/O接口
020
网络适配器
014
外部设备

图8