1、一种非母语语音识别系统，其特征在于，所述系统包括：

一非母语语音接口，用于采集非母语语音数据，并将该非母语语音数据送入非母语状态强制对齐模块和母语状态解码模块；

一母语模型模块，用于向母语状态解码模块和非母语状态修正模型解码模块提供母语声学模型；

一非母语模型模块，用于向非母语状态强制对齐模块和非母语状态修正模型解码模块提供非母语声学模型；

一母语状态解码模块，用于根据标准母语声学模型对非母语语音数据进行解码，得到母语语音状态级分段信息，即母语语音状态解码信息，并将母语语音状态级分段信息送入母语与非母语状态相似度矩阵计算模块；

一非母语状态强制对齐模块，用于根据非母语声学模型将非母语语音数据进行强制对齐，得到非母语语音状态级分段信息，即非母语语音状态参考信息，并将非母语语音状态级分段信息送入母语与非母语状态相似度矩阵计算模块；

一母语与非母语状态相似度矩阵计算模块，用于将母语语音状态级分段信息和非母语语音状态级分段信息在时间上对齐，当母语语音状态与非母语语音状态的重合时间大于预先设定的门限值时，认为这两个状态出现一次“同现”，统计出所有的“同现”情况，并计算得到非母语语音状态对应于母语语音状态的相似度矩阵，并将该相似度矩阵信息送入母语与非母语状态映射表计算模块；

一母语与非母语状态映射表计算模块，用于根据相似度矩阵计算得到状态映射表；及

一非母语状态修正模型解码模块，用于在语音识别的解码过程中，用状态映射表中找到的母语声学模型状态修正相应的非母语声学模型状态，得到修正后的非母语声学模型；最终利用该修正后的非母语声学模型完成非母语语音识别。

2、根据权利要求1所述的非母语语音识别系统，其特征在于，所述非母语语音状态对应于母语语音状态的相似度矩阵通过以下公式得到：

$$A_{i,j}=\frac{\mathrm{count}(t_j|s_i)}{\sum_{n=1}^{N}\mathrm{count}(t_n|s_i)}$$

其中，M，N分别为中文和英文的状态数量；$A_{S,T}(M,N)$为相似度矩阵；$A_{i,j}$为矩阵的第i^{th}行j^{th}列元素；t_j为非母语语音状态，s_i为母语语音状态；$A_{i,j}\in A_{S,T}(M,N)$，$i=1\cdots\cdots M$，$j=1\cdots\cdots N$；$\mathrm{count}(t_j|s_i)$是母语语音状态s_i和非母语语音状态t_j之间的“同现”出现次数。

Petitioner Exhibit 1004, Page 1002 of 1518

3、根据权利要求1所述的非母语语音识别系统，其特征在于，所述状态映射表通过以下方式得到：

如果$A_{i,j}$矩阵中n^{th}行j^{th}列元素是j^{th}列的第n大元素，则状态n^{th}是状态j^{th}的第n候选修正状态；每个非母语语言的状态都能在矩阵中找到母语语言的n个候选修正状态，其中，n<M。

4、根据权利要求1所述的非母语语音识别系统，其特征在于，所述非母语状态修正模型解码模块在语音识别解码过程中，对于观察值o_t，用n个候选状态修正以后的非母语状态j^{th}的观察概率$p_{j^{th}}(o_t)$变为：

$$p_{j^{th}}(o_t)=\alpha p_{j^{th}}(o_t)_{tar}+(1-\alpha)\sum_{i=1}^{n}a_{lj}p_{l^{th}}(o_t)_{sou}$$

其中，α表示非母语状态的修正权重，a_{lj}是相应第l^{th}个候选状态，n是候选数；$p_{j^{th}}(o_t)_{tar}$和$p_{l^{th}}(o_t)_{sou}$分别表示非母语状态j^{th}和母语状态l^{th}在观察值o_t下的原始观
察概率。

5、一种非母语语音识别方法，包括如下步骤：

(1)非母语语音接口采集一定数量的非母语语音数据，用于得到模型状态映射表；

(2)非母语状态强制对齐模块用非母语声学模型对非母语语音数据进行强制对齐，得到非母语语音状态级分段信息，即非母语语音状态参考信息；

(3)母语状态解码模块用标准母语声学模型对非母语语音数据进行解码，得到母语语音状态级分段信息，即母语语音状态解码信息；

(4)母语与非母语状态相似度矩阵计算模块将得到的非母语语音状态级分段信息和母语语音状态级分段信息在时间上对齐，当某两个状态的重合时间大于预先设定的门限值时，这两个状态出现一次“同现”；

(5)母语与非母语状态相似度矩阵计算模块统计出所有的“同现”情况并计算得到非母语语音状态对应于母语语音状态的相似度矩阵；

(6)母语与非母语状态映射表计算模块根据该相似度矩阵得到状态映射表；

(7)非母语状态修正模型解码模块根据得到的状态映射表，在语音识别的解码过程中，用映射表中找到的母语声学模型状态修正相应的非母语声学模型状态，得到修正后的非母语声学模型。

6、根据权利要求5所述的非母语语音识别方法，其特征在于，所述非母语语音状态对应于母语语音状态的相似度矩阵通过以下公式得到：

Petitioner Exhibit 1004, Page 1003 of 1518

$$A_{i,j}=\frac{\mathrm{count}(t_j|s_i)}{\sum_{n=1}^{M}\mathrm{count}(t_n|s_i)}$$

其中，M，N分别为中文和英文的状态数量；$A_{S,T}(M,N)$为相似度矩阵；$A_{i,j}$为矩阵的第i^{th}行j^{th}列元素；t_j为非母语语音状态，s_i为母语语音状态；$A_{i,j}\in A_{S,T}(M,N)$，$i=1\cdots\cdots M$，$j=1\cdots\cdots N$；$\mathrm{count}(t_j|s_i)$是母语语音状态s_i和非母语语音状态t_j之间的“同现”出现次数。

7、根据权利要求5所述的非母语语音识别方法，其特征在于，所述状态映射表通过以下方式得到：

如果$A_{i,j}$矩阵中n^{th}行j^{th}列元素是j^{th}列的第n大元素，则状态n^{th}是状态j^{th}的第n候选修正状态；每个非母语语言的状态都能在矩阵中找到母语语言的n个候选修正状态，其中，$n<M$。

8、根据权利要求5所述的非母语语音识别方法，其特征在于，所述非母语状态修正模型解码模块在语音识别解码过程中，对于观察值o_t，用n个候选状态修正以后的非母语状态j^{th}的观察概率$p_{j^{th}}(o_t)$变为：

$$p_{j^{th}}(o_t)=\alpha p_{j^{th}}(o_t)_{tar}+(1-\alpha)\sum_{l=1}^{n}a_{lj}p_{l^{th}}(o_t)_{sou}$$

其中，α表示非母语状态的修正权重，a_{lj}是相应第l^{th}个候选状态，n是候选数；$p_{j^{th}}(o_t)_{tar}$和$p_{l^{th}}(o_t)_{sou}$分别表示非母语状态j^{th}和母语状态l^{th}在观察值o_t下的原始观

察概率。

Petitioner Exhibit 1004, Page 1004 of 1518

一种非母语语音识别系统及方法

技术领域

本发明涉及语音识别系统及方法，特别涉及基于混合模型状态修正的非母语语音识别系统及方法。

背景技术

随着现代社会信息的全球化，非母语语音识别成为语音识别技术领域研究热点之一。对非母语的语音识别相对母语的语音识别而言，识别性能将会大幅度降低，特别是对于有着严重口音的说话人。如何在保证语音识别系统在仅有少量非母语训练数据的前提下，对带有不同程度口音的非母语语音识别性能的提高是非母语语音识别研究的重点。

文献(Bohn，O.-S.，Flege，J.E.，"The production of new and similar vowels by adult German learners of English."[M]Stud.Second Lang.Acquis.14，131-158，1992.)指出，非母语说话人在进行目标语言发音时，可能会用说话人自身的母语发音进行替代，或者产生出有说话人母语发音特点和目标语言发音特点相结合的发音。这个结论启发我们，说话人自身的母语语言训练数据，可能对说话人非母语语音识别有帮助，特别是针对那些发音口音较重的语音数据。

当前，说话人自适应技术（如MAP、MLLR）已广泛的应用于处理非母语语音识别中(Z.Wang，T.Schultz，A.Waibel，"Comparison of acoustic model adaptation techniques on non-native speech"[C]，Proc.ICASSP 2003.)。这类方法主要是通过对少量非母语语音数据进行自适应，使得母语发音声学模型能够在一定程度上逼近非母语的发音特点。在这些方法中，自适应数据和测试数据的相似性是决定识别性能好坏的关键因素。尽管自适应技术在提高非母语语音识别性能上有贡献，但通过自适应以后的模型在非母语语音识别中的性能相比母语语音模型在母语语音识别的性能，仍然较低。文献(J.Humphries，P.Woodland，and D.Pearce."Using accent-specific pronunciation modeling for robust speech recognition."[C]In Proc.ICSLP'96，pages 2324-2327，Philadelphia，PA，October 1996.)研究了自适应算法的局限性，指出导致非母语语音识别率低的主要原因来自于说话人母语语系中无法涵盖的非母语语言发音。如何使非母语语音声学模型较好的模拟出这类语言发音是非母语语音识别研究工作的重点。

Petitioner Exhibit 1004, Page 1005 of 1518

发明内容

为了克服上述现有技术的不足，本发明提供一种基于状态修正的非母语语音识别系统及方法，该系统及方法通过不同模型间的状态映射，利用说话人母语的声学模型在状态级别上对非母语声学模型进行修正，从而得到更为符合非母语发音特点的模型。

为了实现上述目的，本发明提出的一种非母语语音识别系统，包括：

一非母语语音接口，用于采集非母语语音数据，并将该非母语语音数据送入非母语状态强制对齐模块和母语状态解码模块。

一母语模型模块，用于向母语状态解码模块和非母语状态修正模型解码模块提供母语声学模型。

一非母语模型模块，用于向非母语状态强制对齐模块和非母语状态修正模型解码模块提供非母语声学模型。

一母语状态解码模块，用于根据标准母语声学模型对非母语语音数据进行解码，得到母语语音状态级分段信息，即母语语音状态解码信息，并将母语语音状态级分段信息送入母语与非母语状态相似度矩阵计算模块。

一非母语状态强制对齐模块，用于根据非母语声学模型将非母语语音数据进行强制对齐，得到非母语语音状态级分段信息，即非母语语音状态参考信息，并将非母语语音状态级分段信息送入母语与非母语状态相似度矩阵计算模块。

一母语与非母语状态相似度矩阵计算模块，用于将母语语音状态级分段信息和非母语语音状态级分段信息在时间上对齐，当母语语音状态与非母语语音状态的重合时间大于预先设定的门限值时，认为这两个状态出现一次“同现”，统计出所有的“同现”情况，并计算得到非母语语音状态对应于母语语音状态的相似度矩阵，并将该相似度矩阵信息送入母语与非母语状态映射表计算模块。

一母语与非母语状态映射表计算模块，用于根据相似度矩阵计算得到状态映射表。及

一非母语状态修正模型解码模块，用于在语音识别的解码过程中，用状态映射表中找到的母语声学模型状态修正相应的非母语声学模型状态，得到修正后的非母语声学模型；最终利用该修正后的非母语声学模型完成非母语语音识别。

其中，所述非母语语音状态对应于母语语音状态的相似度矩阵通过以下公式得到：

Petitioner Exhibit 1004, Page 1006 of 1518

$$A_{i,j}=\frac{count(t_j|s_i)}{\sum_{m=1}^{N} count(t_m|s_i)}$$

其中，M，N分别为中文和英文的状态数量；$A_{S,T}(M,N)$为相似度矩阵；$A_{i,j}$为矩阵的第i^{th}行j^{th}列元素；t_j为非母语语音状态，s_i为母语语音状态；$A_{i,j}\in A_{S,T}(M,N)$，$i=1\cdots\cdots M$，$j=1\cdots\cdots N$；$count(t_j|s_i)$是母语语音状态s_i和非母语语音状态t_j之间的“同现”出现次数。

其中，所述状态映射表通过以下方式得到：

如果$A_{i,j}$矩阵中n^{th}行j^{th}列元素是j^{th}列的第n大元素，则状态n^{th}是状态j^{th}的第n候选修正状态；例如：如果$A_{i,j}$矩阵中i^{th}行j^{th}列元素是矩阵中j^{th}列的最大元素，这意味着来自相应语言的状态i^{th}和状态j^{th}最为相似，状态i^{th}是状态j^{th}的第一候选修正状态。如果k^{th}行j^{th}列元素是j^{th}列的第二大元素，那么状态k^{th}是状态j^{th}的第二候选修正状态。如此类推，每个非母语语言的状态都能在矩阵中找到母语语言的n个候选修正状态(n<M)。

其中，所述非母语状态修正模型解码模块在语音识别解码过程中，对于观察值o_t，用n个候选状态修正以后的非母语状态j^{th}的观察概率$p_{j^{th}}(o_t)$变为：

$$p_{j^{th}}(o_t)=\alpha p_{j^{th}}(o_t)_{tar}+(1-\alpha)\sum_{l=1}^{n} a_{lj}p_{l^{th}}(o_t)_{sou}$$

其中，α表示非母语状态的修正权重，a_{lj}是相应第l^{th}个候选状态，n是候选数；$p_{j^{th}}(o_t)_{tar}$和$p_{l^{th}}(o_t)_{sou}$分别表示非母语状态j^{th}和母语状态l^{th}在观察值o_t下的原始观

察概率。

本发明提供的一种非母语语音识别方法，包括如下步骤：

(1)非母语语音接口采集一定数量的非母语语音数据，用于得到模型状态映射表。

(2)非母语状态强制对齐模块用非母语声学模型对非母语语音数据进行强制对齐，得到非母语语音状态级分段信息，即非母语语音状态参考信息。

(3)母语状态解码模块用标准母语声学模型对非母语语音数据进行解码，得到母语语音状态级分段信息，即母语语音状态解码信息。

(4)母语与非母语状态相似度矩阵计算模块将得到的非母语语音状态级分段信息和母语语音状态级分段信息在时间上对齐，当某两个状态的重合时间大于预先设定的门限值时，这两个状态出现一次“同现”。

(5)母语与非母语状态相似度矩阵计算模块统计出所有的“同现”情况并计算

Petitioner Exhibit 1004, Page 1007 of 1518

得到非母语语音状态对应于母语语音状态的相似度矩阵。

(6)母语与非母语状态映射表计算模块根据该相似度矩阵得到状态映射表。

(7)非母语状态修正模型解码模块根据得到的状态映射表，在语音识别的解码过程中，用映射表中找到的母语声学模型状态修正相应的非母语声学模型状态，得到修正后的非母语声学模型。

本发明的优点在于：

(1)本发明的非母语语音识别系统及方法使非母语声学模型更适应带有母语口音的非母语语音特点。

(2)本发明的非母语语音识别系统及方法，采用不同模型间的状态映射，利用说话人母语的声学模型在状态级别上对非母语声学模型进行修正，解码器根据得到的状态映射表实现非母语语音识别。混合模型状态修正的非母语语音识别系统对非母语语音的识别率与未采用该方法修正过的识别系统相比较；并且在不增加任何非母语语音训练数据，仅依靠标准母语训练数据的前提下，对带有母语口音的非母语语音的识别情况相对未采用该方法修正过的识别系统有了明显提高；同时系统识别语音的速度并未明显下降。

(3)本发明的非母语语音识别系统及方法实现了利用母语声学模型对非母语声学模型的修正，提高系统对非母语语音的识别性能。相比语音自适应技术，基于混合模型状态的修正方法不需要增加更多的非母语训练数据。由于用于修正的母语声学模型的规模可以控制在很小的范围，修正后的模型在计算速度上也不会有明显增加。我们进行了大量实网数据的测试，结果显示在识别性能方面，由于该系统在状态级别上的映射信息是基于解码器输出的状态序列情况进行统计得到的，这种准则较为直接，基本能真实的反映出两种语言状态间的相似程度。与未使用该方法的系统相比，基于混合模型状态修正的非母语语音识别系统对非母语语音的识别错误率相对下降可达到5-10%。

附图说明

图1本发明的基于混合模型状态修正算法的非母语语音识别系统及方法的基本原理框图；

图2本发明的非母语语音识别系统的流程框图；

图3本发明的具体实施例母语语音状态与非母语语音状态之间的“同现”说明图。

Petitioner Exhibit 1004, Page 1008 of 1518

具体实施方式

下面结合附图对本发明的具体实施方式做进一步详细描述：

图1是基于混合模型状态修正算法的非母语语音识别系统的基本原理框图，它描述了混合模型状态修正算法的核心组成成分，主要由几个部分构成：非母语语音、母语/非母语声学模型、强制对齐、解码、状态映射表生成以及非母语状态修正模型的解码。图2是基于混合模型状态修正算法的非母语语音识别系统的具体实施流程框图。

本发明涉及的混合模型状态修正算法的非母语语音识别的核心技术在于状态映射表的获取（如图2所示的模块1到7）。混合模型状态修正算法是一种新型的基于相似度矩阵的状态映射算法，通过对同现次数的统计，得到两种语言状态间的对应关系，利用该关系确定用于修正的状态对。

本发明的系统包括：一非母语语音接口，用于采集非母语语音数据，并将该非母语语音数据送入非母语状态强制对齐模块和母语状态解码模块。

一母语模型模块，用于向母语状态解码模块和非母语状态修正模型解码模块提供母语声学模型。

一非母语模型模块，用于向非母语状态强制对齐模块和非母语状态修正模型解码模块提供非母语声学模型。

一母语状态解码模块，用于根据标准母语声学模型对非母语语音数据进行解码，得到母语语音状态级分段信息，即母语语音状态解码信息，并将母语语音状态级分段信息送入母语与非母语状态相似度矩阵计算模块。

一非母语状态强制对齐模块，用于根据非母语声学模型将非母语语音数据进行强制对齐，得到非母语语音状态级分段信息，即非母语语音状态参考信息，并将非母语语音状态级分段信息送入母语与非母语状态相似度矩阵计算模块。

一母语与非母语状态相似度矩阵计算模块，用于将母语语音状态级分段信息和非母语语音状态级分段信息在时间上对齐，当母语语音状态与非母语语音状态的重合时间大于预先设定的门限值时，认为这两个状态出现一次“同现”，统计出所有的“同现”情况，并计算得到非母语语音状态对应于母语语音状态的相似度矩阵，并将该相似度矩阵信息送入母语与非母语状态映射表计算模块。

一母语与非母语状态映射表计算模块，用于根据相似度矩阵计算得到状态映射表。及

一非母语状态修正模型解码模块，用于在语音识别的解码过程中，用状态映射表中找到的母语声学模型状态修正相应的非母语声学模型状态，得到修正后的非母

Petitioner Exhibit 1004, Page 1009 of 1518

语声学模型；最终利用该修正后的非母语声学模型完成非母语语音识别。

本发明的具体计算流程如下：(如图2)

第一步，非母语状态参考答案：选择一定量的非母语语音数据(非母语语音接口1)，这些数据用于生成母语对非母语的相似度矩阵。用非母语声学模型(非母语模型模块2)对这些数据进行强制对齐(Forced-alignment)，得到非母语状态序列，同时记录下每个状态的时间信息(非母语状态强制对齐模块4)。

第二步，母语状态识别结果：用母语声学模型(母语模型模块3)对第一步中的非母语语音数据进行解码(Decode)，得到母语状态序列，同时记录下每个状态的时间信息(母语状态解码模块5)。

第三步，“同现”准则：通过第一步和第二步，在同一批非母语语音数据上，分别得到了非母语和母语的状态序列以及相应的时间段信息，根据这两个状态序列在时间轴上的位置关系，可以定义两个状态的“同现”现象。在计算“同现”前，首先定义一个“同现”矩阵，该矩阵为(母语语言状态个数*非母语语言状态个数)的矩阵，每个元素位置上记录了对应行和列的状态之间的“同现”次数。在本发明所述方法的实验中，定义当两个状态在时间段上重叠的时间占到非母语语言状态持续时间一半时，算作一次“同现”。当非母语语言状态i^{th}和母语语言状态j^{th}出现一次“同现”时，在对应同现矩阵的i^{th}行j^{th}列的位置上加1记录。如图3所示，母语语音状态“aa_native”和非母语语音状态“ae_nonnative”之间出现的一次“同现”。

第四步，状态相似度矩阵计算(母语与非母语状态相似度矩阵计算模块6)：若设定M，N分别为母语语言和非母语语言状态个数，$A_{S\cdot T}(M, N)$为从同现矩阵派生出来的相似度矩阵，矩阵元素记录了母语语言和非母语语言状态之间的相似度情况。$A_{i\cdot j}$为矩阵i^{th}行j^{th}列的元素，设定t_j为非母语语言状态，s_i为母语语言状态，两者之间的相似度计算为：

$$A_{ij}=\frac{count(t_j|s_i)}{\sum_{n=1}^{N} count(t_n|s_i)} \quad ---(1)$$

其中，$A_{i\cdot j}\in A_{S\cdot T}(M, N)$，i=1……M，j=1……N。

第五步，得到相似度矩阵后，根据该矩阵得到状态映射表(母语与非母语状态映射表计算模块7)。如果$A_{i\cdot j}$矩阵中i^{th}行j^{th}列元素是矩阵中j^{th}列的最大元素，这意味着来自相应语言的状态i^{th}和状态j^{th}最为相似，状态i^{th}是状态j^{th}的第一候选修正状态。如果k^{th}行j^{th}列元素是j^{th}列的第二大元素，那么状态k^{th}是状态j^{th}的第二候选修正状态。如此类推，每个非母语语言的状态都能在矩阵中找到母语语言的n个候

Petitioner Exhibit 1004, Page 1010 of 1518

选修正状态(n＜M)。

到此，我们基于混合模型状态修正算法，得到了母语语言和非母语语言状态间的映射信息。

第六步，接下来根据得到的状态间的映射关系，在解码过程中用母语声学模型对非母语声学模型进行修正(非母语状态修正模型解码模块8)。根据第五步得到的状态映射表，在语音识别的解码过程中，对于观察值o_t，用n个候选状态修正以后的非母语状态j^{th}的观察概率$p_{j^{th}}(o_t)$变为：

$$p_{j^{th}}(o_t)=\alpha p_{j^{th}}(o_t)_{tar}+(1-\alpha)\sum_{l=1}^{n}a_{lj}p_{l^{th}}(o_t)_{sou} \quad ---(2)$$

这里α表示非母语状态的修正权重，a_{lj}是相应第1^{th}个候选状态，n是候选数。$p_{j^{th}}(o_t)_{tar}$和$p_{l^{th}}(o_t)_{sou}$分别表示非母语状态j^{th}和母语状态1^{th}在观察值o_t下的原始观

察概率。

本发明的特点：(1)使非母语声学模型更适应带有母语口音的非母语语音特点(2)提出一种新型基于混合模型状态修正的非母语语音识别方法，(3)在仅依靠标准母语声学模型的前提下，提高系统对非母语语音的识别性能。

在进行基于给定语法的带有中文口音的英文实网数据测试中显示，基于混合模型状态修正算法的非母语语音识别系统在带有中文口音的英文测试集上的错误率相对未使用该方法的非母语语音识别系统下降5％-10％(相对值)；计算速度相对于未使用该方法的非母语语音识别系统降低20％-25％(相对值)。

Petitioner Exhibit 1004, Page 1011 of 1518

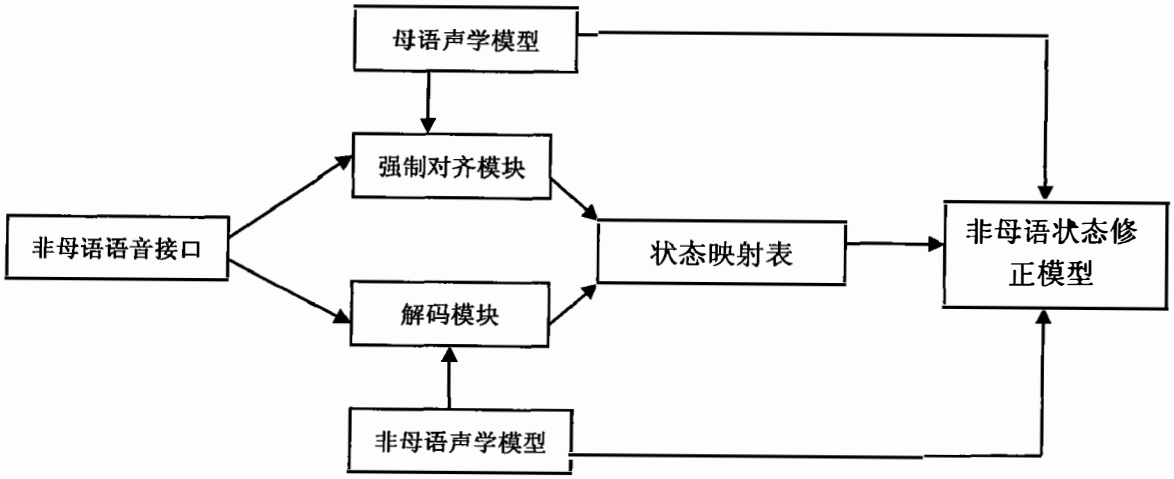


图1

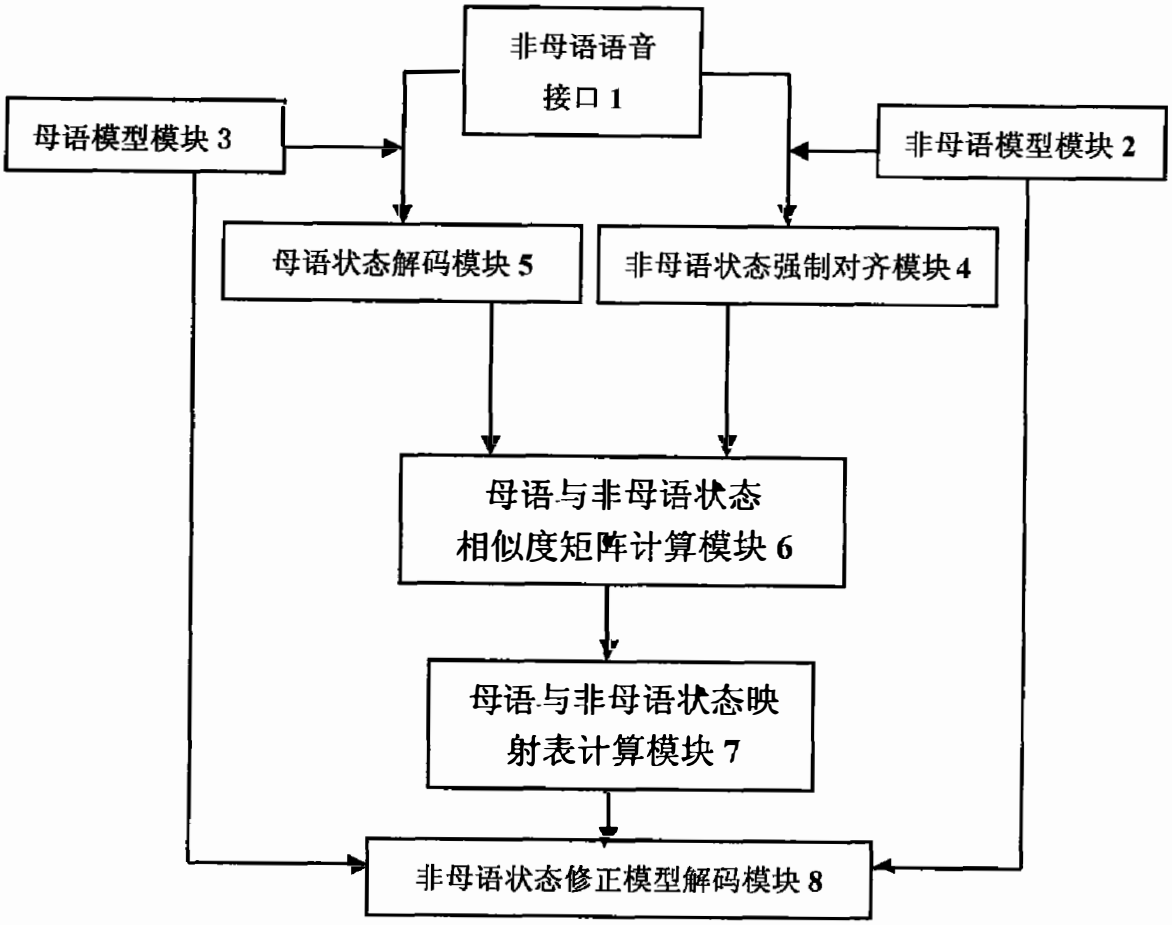


图2

Petitioner Exhibit 1004, Page 1012 of 1518

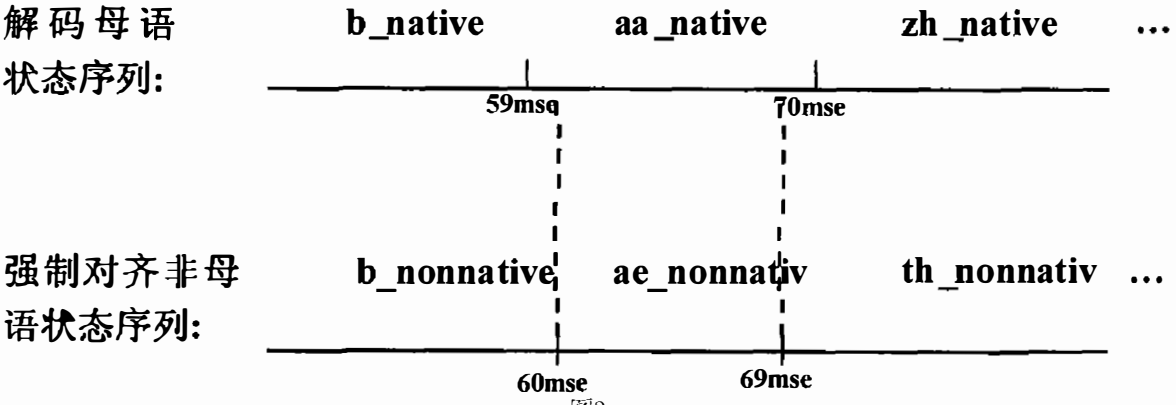


图3

Petitioner Exhibit 1004, Page 1013 of 1518

Doc Code: TRAN.LET
Document Description: Transmittal Letter

USED IN LIEU OF PTO/SB/21 (07-09)

TRANSMITTAL FORM

(to be used for all correspondence after initial filing)

Application Number	13/795,933-Conf. #2321
Filing Date	March 12, 2013
First Named Inventor	William F. Ganong, III
Art Unit	2644
Examiner Name	H. C. Ho
Attorney Docket Number	N0484.71090US01

Total Number of Pages in This Submission

ENCLOSURES (*Check all that apply*)

- [ ] Fee Transmittal Form
  - [ ] Fee Attached
- [ ] Amendment/Reply
  - [ ] After Final
  - [ ] Affidavits/declaration(s)
- [ ] Extension of Time Request
- [ ] Express Abandonment Request
- [x] Information Disclosure Statement
- [ ] Certified Copy of Priority Document(s)
- [ ] Reply to Missing Parts/ Incomplete Application
  - [ ] Reply to Missing Parts under 37 CFR 1.52 or 1.53
- [ ] Drawing(s)
- [ ] Licensing-related Papers
- [ ] Petition
- [ ] Petition to Convert to a Provisional Application
- [ ] Power of Attorney, Revocation Change of Correspondence Address
- [ ] Terminal Disclaimer
- [ ] Request for Refund
- [ ] CD, Number of CD(s) _____
  - [ ] Landscape Table on CD
- [ ] After Allowance Communication to TC
- [ ] Appeal Communication to Board of Appeals and Interferences
- [ ] Appeal Communication to TC **(Appeal Notice, Brief, Reply Brief)**
- [ ] Proprietary Information
- [ ] Status Letter
- [x] Other Enclosure(s) (please Identify below):

  Form PTO - 1449

  Copies of Cited References

Remarks

SIGNATURE OF APPLICANT, ATTORNEY, OR AGENT

Firm Name	WOLF, GREENFIELD & SACKS, P.C.		
Signature	/Andrew J. Tibbetts/		
Printed name	Andrew J. Tibbetts		
Date	May 22, 2019	Reg. No.	65,139

Certificate of Electronic Filing under 37 C.F.R. § 1.8

I hereby certify that this paper (along with any paper referred to as being attached or enclosed) is being transmitted via the Office's electronic filing system in accordance with 37 C.F.R. § 1.6(a)(4).

Dated: May 22, 2019 Electronic Signature for MacAulay S. Rush: /MacAulay S. Rush/

7243441.1

Publication number: CN106098059 A
Publication country: CHINA
Publication type: UNEXAMINED APPLIC. OPEN TO PUBLIC INSPECTION
Publication date: 20161109
Application number: CN201610462976
Application date: 20160623
Priority: CN201610462976201606 23CN201614629 76;
Assignee: ?????? ; SHANGHAI JIAO TONG UNIV ;
Assignee[std]: UNIV SHANGHAI JIAOTONG ;
Inventor: ??? ; ??? ; ??? ; ?? ; ??? ;
Inventor[std]: CHANG XUANKAI ; YU KAI ; QIAN YANMIN ; CHEN ZHEHUAI ; ZHUANG YIMENG ;
International class[8]: G10L15/02; G10L15/06; G10L15/22;
Cooperative Patent class: G10L15/02 ; G10L15/063 ; G10L15/22 ; G10L2015/025 ; G10L2015/0631 ; G10L2015/223 ;
Family members: CN106098059 A

Title: CUSTOMIZABLE VOICE WAKE-UP METHOD AND SYSTEM ????????????

Title: 可定制语音唤醒方法及系统

Abstract:

The invention discloses a customizable voice wake-up method and a system. Through using a long and short memory network and a connection time sequence classification model, phoneme information of the voice information is modeled, the model is trained, the model after training is adopted for testing, and a possible phoneme sequence most similar to a customized wake-up word is searched in a generated Lattice network structure and serves as a judgment basis. The feature of CTC model output posteriori probability sparsity is used for high-efficiency searching, and the technique of calculating the confidence of the wake-up word is completed. On one hand, high wake-up performance can be obtained, that is, high accuracy and low error wake-up can be obtained; and on the other hand, the calculation resources of the application system are relatively little consumed.

?????????????,??,???????????????? Lattice???????????????????????????????????? CTC??????????????????,??????????????????????????????,?????,????,?????????????????????

Abstract:

一种可定制语音唤醒方法及系统，通过使用基于长短时记忆网络和连接时序分类模型对语音信息的音素信息进行建模并对模型进行训练，采用训练后模型进行测试并在生成的Lattice网络结构上搜索与定制的唤醒

A customizable voice wake methods and systems, by using short-term memory network and connection time series classification model for the modeling of phoneme information voice messages and training

词最相似的可能音素序列来作为判断依据。本发明利用CTC模型输出后验概率稀疏的特点进行高效搜索，从而完成对唤醒词置信度计算的技术。本发明一方面可以得到较高的唤醒性，即高准确率，低误唤醒，另一方面对应用系统的计算资源消耗相对较少。

model, training model is used to test and build on a Lattice network structure search and customized wake the most similar word of possible phoneme sequences as a judge. By CTC model outputs of the invention a posteriori probability characteristics for efficient search of sparse, thus completing the wake the word confidence measure technology. The invention can obtain a higher sexual arousal, high accuracy and low false alarm wake up, on the application of the system of consumption of computing resources is relatively small.

Claims:

1.一种可定制语音唤醒方法，其特征在于，包括以下步骤：

步骤1)使用基于长短时记忆网络和连接时序分类模型对语音信息的音素信息进行建 模；

步骤2)对模型进行训练：将事先收集并标注好的音频数据首先要经过传统的信号处理 方法做预处理并提取可供模型训练的声学特征，模型将特征数据作为输入，将标注的音素 信息作为输出，在海量数据下通过深度学习的方法完成模型各参数量的训练，得到可使用 的长短时记忆网络和连接时序分类模型；

步骤3)采用训练后模型进行测试：对测试语音做同样的预处理和特征提取，并输入模 型，模型将输出每一帧所有建模单元，即音素可能出现的后验概率；

步骤4)唤醒词搜索：通过在生成的Lattice网络结构上搜索与定制的唤醒词最相似的 可能音素序列来作为判断依据。

2.根据权利要求1所述的可定制语音唤醒方法，其特征是，所述的最相似的可能音素序 列H_{max}，通过判断音素序列在语音中的出现概率，即出现概率越高且越与唤醒词音素序列相 似的序列越可能是唤醒词的方式得到，具体为：

$$H_{max} = \underset{H}{\operatorname{argmax}} P(H|T) = \underset{H}{\operatorname{argmax}} \frac{P(T|H)P(H)}{P(T)}$$
$$= \underset{H}{\operatorname{argmax}} P(T|H)P(H)$$

其中：P(T)为目标音素序列T，T＝ $\{t_1,t_2,\cdots,t_n\}$的观察概率，t_i表示T中的第i个音素，i为1～n；

目标音素序列T位于CTC音素网格结构中的所有音素序列的概率为：

$P(T|L_H) \propto P(L_H|T)P(T) \approx P(H_{max}|T)P(T)$，其中：$L_H$表示CTC音素网格结构中的所有音素 序列，而可能音素序列H_{max}即为T已知时L_H中的最高概率；

P(H)为音素序列H，H＝$\{n_{ij1},n_{(i+1)j2},\cdots,n_{(i+m-1)jm}\}$的观察概率，$n_{ij}$为lattice网格 结构中第i栏第j列的音素，音素序列H也可以表示为H＝$\{h_1,h_2,\cdots,h_m\}$，其中：$h_k = n_{(i+k-1)jk}$；

1. a customizable voice awakens, which is characterized, comprising the following steps:

Step 1) use short-term memory network and connection time series classification model for the modeling of phoneme information voice messages;

Steps 2) on model for training: will prior collection and mark good of audio data first to after traditional of signal processing method do pretreatment and extraction can for model training of acoustic features, model will features data as entered, will mark of phonemes information as output, in mass data Xia through depth learning of method completed model the parameter number of training, get can using of length Shi memory network and connection timing classification model;

Step 3) training model to test: to test voice doing the same preprocessing and feature extracting and enter the model, the model output each frame of all model units, that phoneme posterior probabilities that may occur;

Step 4) wake the word search: Lattice network structure generated by searches and customized wake the most similar word of possible

P(H)通过unigram假设，即累积乘音素序列中每个音素的后验概率获得，具体为：

$$P(H)=\prod_{i=1}^{m}P(h_i|H_{1:i-1})\approx\prod_{i=1}^{m}P(h_i)$$;

P(T/H)为目标音素序列T和音素序列H之间的相似程 度，

$$P(T|H)\triangleq\prod_{i=1}^{MED(T,H)}P(op_i|R=T,E=H)$$,

即目标音素序列和假设序列之间的每个编辑 操作的概率之积作为衡量音素序列相似性的指标，MED(T,H)表示目标音素序列T和音素序 列H的最少编辑操作次数，P(op$_i$|R＝T，E＝H)表示当参考音素序列R为T时猜测音素序列E为 H时，序列E和R之间的第i次编辑操作op$_i$的概率。

3.根据权利要求2所述的可定制语音唤醒方法，其特征是，所述的编辑操作是指：插入、 删除、替换操作，其概率，即P(insert(e$_i$))、P(delete(r$_i$))和P(r$_i$/e$_i$)直接由先验知识得 出，r$_i$和e$_i$分别取自参考音素序列R和猜测音素序列E。

4.根据权利要求2所述的可定制语音唤醒方法，其特征是，所述的步骤1包括：

1.1)确定模型结构：根据应用场景设备的计算能力预估模型复杂度，首先设定参数量 的上限，然后在此限定下指定网络结构；

1.2)初始化长短时记忆网络：使用随机初始化的参数或model transferring方式初始 化。

5.根据权利要求4所述的可定制语音唤醒方法，其特征是，所述的初始化长短时记忆网 络，使用cross entropy准则预训练一个标准的声学长短时记忆网络，然后通过复制参数的 方式初始化模型。

6.根据权利要求2所述的可定制语音唤醒方法，其特征是，所述的步骤2包括：

2.1)对训练数据提取fbank，即Filter-bank声学特征；

2.2)使用随机梯度下降方式完成模型的训练，训练参数需根据模型结构以及训练数据 大小设定。

7.根据权利要求2所述的可定制语音唤醒方法，其特征是，所述的步骤3包括：

3.1)对测试数据提取fbank声学特征，特征提取过程要求与训练数据特征提取 一致；

phoneme sequences as a judge.

2. According to claim 1 can be customized voice-wake method, which is characterized, the closest possible phoneme sequences described h $_{Max}$, by determining the probability in a sequence of phonemes in speech, the higher the probability of occurrence and more and wake the word phoneme sequences more similar sequences may be wake the word manner, in particular:

[IMAGE]

: P (t) is the target phoneme sequence T,T= {t $_1$, t $_2$, ���, t $_n$ } the observed probability, t $_i$ t I in phonemes, I 1~n;

T target phoneme sequences in the CTC phoneme lattice structure probabilities of all sequences of phonemes are:

P(T| L$_H$)?P(L$_H$| T)P(T)�P(H$_{Max}$| T) p (t), of which: l $_h$ all phoneme sequences in the CTC phoneme lattice structure, possible phoneme sequences h $_{Max}$ is known that t l $_h$, the highest probability;

P (h) for phonemes sequence H,H={n $_{ij1}$, n $_{(i 1) j2}$, ��, n $_{(i m-1) JM}$ } of observation probability, n $_{IJ}$ for lattice network grid structure in the subsection I bar subsection j column of phonemes, phonemes sequence h also can said for H={h $_1$, h $_2$, ��, H $_m$ } where: h $_k$ = n $_{(i k-1) JK}$;

P (h) through unigram assumptions, cumulative phoneme sequences, each phoneme posterior probabilities obtained, in particular to:

3.2)将提取的特征逐帧输入训练得到的模型，直接计算出每一帧的后验概率。

8.根据权利要求2所述的可定制语音唤醒方法，其特征是，所述的步骤4包括：

4.1)对每一句测试语音生成lattice网络结构，扫描每一帧空白的后验概率，当其后验 低于预设值时判定该帧为一个尖峰，将一句话中所有尖峰后，将时间上连续的尖峰合并为 一个尖峰，对于每个尖峰，选择出这一帧上后验概率超过预设值的音素以组成lattice网络 结构中的一列；在构造出所有列后，两两连接lattice中相邻两列中的每个节点音素，得到 所需lattice网络结构；

4.2)在生成的lattice网络结构上执行搜索与定制，找到与目标音素序列最相似的音 素序列；

4.3)计算找到的音素序列的观察概率与相似程度的乘积，并与设置的阈值做比较，当 大于阈值，则判断测试语音包含唤醒词，否则不唤醒。

9.一种实现权利要求1～8中任一所述方法的可定制语音唤醒系统，其特征在于，包括： 声学特征提取模块、记忆网络模块、分类模型模块、唤醒词搜索模块、决策模块和阈值估计 模块，其中：声学特征提取模块与记忆网络模块相连并输出待测语音的声学特征信息，记忆 网络模块与分类模型模块相连并传输音素后验信息，分类模型模块与唤醒词搜索模块相连 并传输最大相似音素序列，唤醒词搜索模块与决策模块相连并根据收到的唤醒词输出待测 语音可能解，决策模块根据来自阈值估计模块的判断阈值以及待测语音可能解得出判断结 果；

所述的唤醒词搜索模块采用基于连接时序分类模型的Lattice网络结构，其中包含了 所有可能的语音识别结果及其概率。

P (T/H) for the target phoneme sequence similarity between the t and h phoneme sequences,

The target phoneme sequences and hypothetical sequence of product of the probabilities of each edit operation as an indicator of the phoneme sequence similarity, MED (T,H) representing the target phoneme sequences t and h phoneme sequences of a minimum number of edits, p (op $_l$ | R=T,E=H) represents the phoneme sequences reference r to t guess phoneme sequences e h, series e and r l between editing operation op $_l$ probability.

3. According to right requirements 2 by said of can custom voice wake method, its features is, by said of edit operation is refers to: Insert, and delete, and replaced operation, its probability, that p (Insert (e $_l$)), and p (delete (r $_l$)) and p (r $_l$/e $_l$) directly by prior knowledge obtained, r $_l$ And e $_l$ taken from reference sequence of phonemes respectively r and guess the phoneme sequence e.

4. According to claim 2 customizable voice awakens, which is, as described in step 1, including:

1.1) determines the structure of the model: a predictive model of the computing power of the device depending on the application scenario complexity and quantity ceiling set by first, and then under this limit specifies the network structure;

1.2) memory to initialize the length: using initialization parameters or model of transferring method initialize.

5. According to claim 4 customizable voice awakens, which is characterized, the length of initialization of memory, using cross entropy criterion when training the length of a standard acoustic memory and initialized by copying parameters model.

6. According to claim 2 customizable voice awakens, which is, as described in step 2, including:

2.1) the training data extraction fbank, namely Filter LFAT acoustic characteristics of the Bank;

2.2) model is accomplished using stochastic gradient descent training training parameter based on model structure and the training data set size.

7. According to claim 2 customizable voice awakens, which is, as described in step 3, including:

3.1) to extract test data fbank acoustic features, the feature extraction process requires consistent with the training data extraction;

3.2) extraction characteristics of frame-by-frame to enter training to get the model directly compute the posterior probability of each frame.

8. According to claim 2 customizable voice awakens, which is, as described in step 4 include:

4.1) on each sentence test voice generated lattice network structure, scan each frame blank of Hou inspection probability, dang subsequently inspection below preset value Shi judge the frame for a pointed peak, will word in the all

pointed peak Hou, will time Shang continuous of pointed peak merged for a pointed peak, for each pointed peak, select out this a frame Shang Hou inspection probability over preset value of sound are renowned for their composition lattice network structure in the of a column; in structure out all column Hou, 22 connecting two adjacent columns of each node in the lattice phonemes, given the lattice network structure;

4.2) generated by performing a search on a lattice network structure and customization, the target phoneme sequences are found the most similar sequences of phonemes;

4.3) calculation found observing the product of the probability and the degree of similarity of the phoneme sequence, and compare with the set threshold value, when greater than a threshold value, then test with wake Word speech, you do not wake up.

9. a achieved right requirements 1~8 in the any a by said method of can custom voice wake system, its features is, including: acoustic features extraction module, and memory network module, and classification model module, and wake word search module, and decision module and threshold value estimated module, which: acoustic features extraction module and memory network module connected and output stay measuring voice of acoustic features information, memory network module and classification model module connected and transmission phonemes Hou inspection information, Classification model of module and wake up word search module and transfer the most similar

phoneme sequences, wake up word search module connected with decision-making module and select the wake the word output sound to be measured according to possible solutions, decision-making module based on threshold estimation module to determine thresholds and sound judgment findings of possible solutions to be tested;

The wake of Word search module based on connection time series classification model of Lattice structure, which contains all possible speech recognition results and probabilities.

Description:

可定制语音唤醒方法及系统

技术领域

本发明涉及的是一种计算机辅助控制领域的技术，具体是一种基于长短时记忆网 络

(LSTM)和连接时序分类模型(CTC)的可定制语音唤醒方法及系统。

背景技术

近年来，随着信息技术的发展，包括语音在内多媒体信息技术越来越成为研究的 关注热点。语言唤醒技术是语音识别中的一个重要领域，并被广泛应用于语音命令控制系 统中。一个可定制的语音唤醒(Voice Wake-up)系统的任务是从一段连续的语音中自动发 现并定位一些事先指定的命令词(唤醒词)。可定制的特性体现在唤醒词检测模型不依赖于 用户指定的唤醒词，从而实现无需修改模型即可方便更改用户唤醒词。相反，非可定制的唤 醒技术与指定的唤醒词有关，唤醒词固定，无法轻易变更唤醒词。语音唤醒与连续语音识别 关系密切，但语音唤醒技术并不要求将完整的语音句子全部识别出来，而是仅关心用户指 定的关键信息。因此，唤醒技术降低了对识别系统的要求。与传统的文本态文档比较，语音 数据作为一种对声音的编码形式，使得语音信息成为一种对于计算机来说更加难以直接检 索和提取有效信息的数据形态。此外由于多种潜在的因素(如背景噪声、说话人口音等)，也 使得开发出一套有效的语音唤醒系统变得更加复杂和困难。主要的语音唤醒技术包括早期 的动态时间规整法，如今的基于隐马尔科夫模型的方法，以及基于深度学习的方法。

发明内容

本发明针对现有技术中唤醒词无法定制且依赖预设语言模型的不足，提出一种可 定制语音唤醒方法及系统，利用CTC模型输出后验概率稀疏的特点进行高效搜索，从而完成 对唤醒词置信度计算的技术。本发明一方面可以

Customizable voice wake method and system

technology the invention relates to the field of computer-aided control technology, which is based on the short-term memory network

(LSTM) time series classification model and connection (CTC) and customizable voice wake methods and systems. background technology in recent years, with the development of information technologies, multimedia and information technology, including voice more and more become the focus of study. Language wake on LAN technology is an important area of speech recognition, and is widely used in voice command-and-control systems. A customizable wake voice (Voice Wake LFAT up) system is a continuous voice commands that automatically find and locate some pre-specified in words (wake up). Customizable characteristics reflected in the wake of the word wake model

得到较高的唤醒性能(准确率、召回率)，高 准确率，低误唤醒，另一方面对应用系统的计算资源消耗相对较少。

本发明是通过以下技术方案实现的：

本发明涉及一种可定制语音唤醒方法，包括以下步骤：

步骤1)使用基于长短时记忆网络和连接时序分类模型对语音信息的音素信息进 行建模；

步骤2)对模型进行训练：将事先收集并标注好的音频数据首先要经过传统的信号 处理方法做预处理并提取可供模型训练的声学特征，模型将特征数据作为输入，将标注的 音素信息作为输出，在海量数据下通过深度学习的方法完成模型各参数量的训练，得到可 使用的长短时记忆网络和连接时序分类模型；

步骤3)采用训练后模型进行测试：对测试语音做同样的预处理和特征提取，并输 入模型，模型将输出每一帧所有建模单元，即音素可能出现的后验概率；

步骤4)唤醒词搜索：通过在生成的Lattice网络结构上搜索与定制的唤醒词最相 似的可能音素序列来作为判断依据。

所述的最相似的可能音素序列H_{max}，通过判断音素序列在语音中的出现概率，即出 现概率越高且越与唤醒词音素序列相似的序列越可能是唤醒词的方式得到，具体为：

$$H_{max} = \underset{H}{\operatorname{argmax}} P(H|T) = \underset{H}{\operatorname{argmax}} \frac{P(T|H)P(H)}{P(T)} = \underset{H}{\operatorname{argmax}} P(T|H)P(H)$$

，

其中：P(T)为目标音素序列 T，$T=\{t_1,t_2,\cdots,t_n\}$的观察概率，t_i表示T中的第i个音素，i为1～n；

目标音素序列T位于CTC音素网格结构中的所有音素序列的概率为：

$P(T|L_H) \propto P(L_H|T)P(T) \approx P(H_{max}|T)P(T)$，其中：$L_H$表示CTC音素网格结构中的所有 音素序列，而可能音素序列H_{max}即为T已知时L_H中的最高概率。

P(H)为音素序列H，$H=\{n_{ij1},n_{(i+1)j2},\cdots,n_{(i+m-1)jm}\}$的观察概率，$n_{ij}$为lattice网格结构中第i栏第j列的音素，音素序列H也可以表示为$H=\{h_1,h_2,\cdots,h_m\}$，其中：$h_k = n_{(i+k-1)jk}$；

P(H)通过unigram假设，即累积乘音素序列中每个音素的后验概率获得，具体为：

$$P(H) = \prod_{i=1}^{m} P(h_i|H_{1:i-1}) \approx \prod_{i=1}^{m} P(h_i)$$

；

does not rely on the user to specify words, which do not need to modify the model to easily change the user wake up Word. In contrast, non-customizable wake wake technology specified words related to wake the word fixed wake cannot easily change the word. Voice awakens and closely related to the continuous speech recognition, voice wake on LAN technology does not require complete voice sentences all identified, but only concerned with user-specified key information. Therefore, the wake on LAN technology reduces the requirements for recognition system. Compared with the conventional text mode documents, voice data as a form of coding of sound makes voice messages became more difficult to direct a computer to retrieve and extract data information form. In addition various potential factors (such as background noise, the speaker's accent), and develop an effective voice to wake up systems become more complex and difficult. Main voice of wake on LAN technology, including an early method of dynamic time warping, now based on hidden Markov model methods, as well as the deep learning method. contents of invention wake up word in the existing technology of the invention cannot be customized and depends on the preset language model is insufficient, presents a customizable voice wake method and system using CTC model outputs a posteriori sparse efficient search features to complete technology to wake the word confidence measure. The invention can obtain a higher wake performance (precision and recall), high accuracy and low false alarm wake up, on the application of the system of consumption of

P(T/H)为目标音素序列T和音素序列H之间的相 似程度，

$$P(T|H) \triangleq \prod_{i=1}^{MED(T,H)} P(op_i|R=T,E=H)$$

，

即目标音素序列和假设序列之间的每个 编辑操作的概率之积作为衡量音素序列相似性的指标，MED(T,H)表示目标音素序列T和音 素序列H的最少编辑操作次数，$P(op_i|R=T, E=H)$表示当参考音素序列R为T时猜测音素序列E为H时，序列E和R之间的第i次编辑操作op_i的概率。

所述的编辑操作是指：插入、删除、替换操作，其概率，即P(insert(e_i))、P(delete (r_i))和P(r_i/e_i)直接由先验知识得出，r_i和e_i分别取自参考音素序列R和猜测音素序列E。

本发明涉及一种实现上述方法的可定制语音唤醒系统，包括：声学特征提取模块、 记忆网络模块、分类模型模块、唤醒词搜索模块、决策模块和阈值估计模块，其中：声学特征 提取模块与记忆网络模块相连并输出待测语音的声学特征信息，记忆网络模块与分类模型 模块相连并传输音素后验信息，分类模型模块与唤醒词搜索模块相连并传输最大相似音素 序列，唤醒词搜索模块与决策模块相连并根据收到的唤醒词输出待测语音可能解，决策模 块根据来自阈值估计模块的判断阈值以及待测语音可能解得出判断结果。

所述的唤醒词搜索模块采用基于连接时序分类(Connectionist Temporal Classification,CTC)模型的Lattice网络结构，其中包含了所有可能的语音识别结果及其 概率。

技术效果

本发明相比较传统方法主要有以下不同：

系统构建各环节	传统方法	本发明方法	优点比较
声学建模 Acoustic Modeling	隐马尔科夫模型	连接时序分类模型	更准确的声学建模
置信度打分 Confidence scoring	声学解码	基于编辑距离的动态搜索	模型搜索计算量和次数减少

附图说明

图1为本发明系统结构示意图；

图2为神经网络结构示意图；

图中给出了不同手机对应的后验概率，下半部分的网格中的实线表示潜在路径， 虚线表示所有有效连接。

具体实施方式

本实施例包括：声学特征提取模块、记忆网络模块、分类模型模块、唤醒词搜索模 块、决策模块和阈值估计模块，其中：声学特征提取模块与记忆

computing resources is relatively small. the invention is realized through the following technology: the invention relates to a customizable voice wake method, comprising the following steps: step 1) use short-term memory network and connection time series classification model for the modeling of phoneme information voice messages; steps 2) on model for training: will prior collection and mark good of audio data first to after traditional of signal processing method do pretreatment and extraction can for model training of acoustic features, model will features data as entered, will mark of phonemes information as output, in mass data Xia through depth learning of method completed model the parameter number of training, get can using of length Shi memory network and connection timing classification model; step 3) training model to test: to test voice doing the same preprocessing and feature extracting and enter the model, the model output each frame of all model units, that phoneme posterior probabilities that may occur; step 4) wake the word search: Lattice network structure generated by searches and customized wake the most similar word of possible phoneme sequences as a judge. the closest possible phoneme sequences described h_{Max}, by determining the probability in a sequence of phonemes in speech, the higher the probability of occurrence and more and wake the word phoneme sequences more similar sequences may be wake the word manner, in particular:

[IMAGE]

: P (t) is the target phoneme sequence T,T={t_1, t_2, ���, t

网络模块相连并输出待测语 音的声学特征信息，记忆网络模块与分类模型模块相连并传输音素后验信息，分类模型模 块与唤醒词搜索模块相连并传输最大相似音素序列，唤醒词搜索模块与决策模块相连并根 据收到的唤醒词输出待测语音可能解，决策模块根据来自阈值估计模块的判断阈值以及待 测语音可能解得出判断结果。

本实施例涉及上述系统的语音唤醒方法，具体包括以下步骤：

步骤1)使用基于长短时记忆网络和连接时序分类模型对语音信息的音素信息进 行建模，具体步骤包括：

1.1)确定模型结构：根据应用场景设备的计算能力预估模型复杂度，由于计算复 杂度与模型的参数量正相关，因此首先要设定参数量的上限，例如不超过5.5M大小。然后在 此限定下指定网络结构，例如使用3隐层网络，每次256个节点，投影为96个节点。

1.2)初始化长短时记忆网络：使用随机初始化的参数，也可以通过model transferring方式初始化，由于随机初始化会造成之后CTC训练困难的问题，这里推荐使用 cross entropy准则预训练一个标准的声学长短时记忆网络，然后通过复制参数的方式初 始化模型。

步骤2)对模型进行训练：将事先收集并标注好的音频数据首先要经过传统的信号 处理方法做预处理并提取可供模型训练的声学特征，模型将特征数据作为输入，将标注的 音素信息作为输出，在海量数据下通过深度学习的方法完成模型各参数量的训练，得到可 使用的长短时记忆网络和连接时序分类模型，具体步骤包括：

2.1)对训练数据提取fbank，即Filter-bank声学特征。

2.2)使用随机梯度下降方式完成模型的训练，训练参数需根据模型结构以及训练 数据大小设定，例如对于上述例子中的模型，可采用0.00001的学习率，动量值为0.9，batch 大小256等等。

步骤3)采用训练后模型进行测试：对测试语音做同样的预处理和特征提取，并输 入模型，模型将输出每一帧所有建模单元可能出现的后验概率，具体步骤包括：

3.1)对测试数据提取fbank声学特征，特征提取过程要求与训练数据特征提取一 致。

3.2)将提取的特征逐帧输入训练得到的模型，直接计算出每一帧的后验概率。

步骤4)唤醒词搜索：通过在生成的lattice网络结构上搜索与定制的唤醒词最相 似的可能音素序列来作为判断依据，具体步骤包括：

4.1)对每一句测试语音生成lattice网络结构，扫描每一帧"空白"的后验概率，当 其后验低于预设值，例如0.8，那么认为这帧为一个尖峰。在找出一句话的所有尖峰后，将时 间上连续的尖峰合并为一个尖峰，对于每个尖峰，选择出这一帧上后验概率较大的音素，例 如后验概率大于0.005的音素，组成lattice网络结构中的一列；在构造出所有列后，两两连 接lattice中相邻两列中的每个节点音素，得到所需lattice网络结构。

4.2)根据前面提到的搜索算法公式，在生成的lattice网络结构上执行搜索算法， 找到与目标音素序列最相似的音素序列。

$_n$ } the observed probability, t_l t l in phonemes, l 1~n; t target phoneme sequences in the CTC phoneme lattice structure probabilities of all sequences of phonemes are:

P(T| L_H)?P(L_H| T)P(T)$\hat{}$P(H_{max}| T) p (t), of which: l_h all phoneme sequences in the CTC phoneme lattice structure, possible phoneme sequences h $_{Max}$ is known that t l_h , the highest probability. p (h) for phonemes sequence H, H={n_{ij1} , $n_{(i\ 1)\ j2}$, $\hat{}$ $\hat{}$ $\hat{}$, $n_{(i\ m\text{-}1)\ JM}$ } of observation probability, n_{IJ} for lattice network grid structure in the subsection l bar subsection j column of phonemes, phonemes sequence h also can said for H={h_1 , h_2 , $\hat{}$ $\hat{}$ $\hat{}$, H_m } where: $h_k = n_{(i\ k\text{-}1)\ JK}$; p (h) through unigram assumptions, cumulative phoneme sequences, each phoneme posterior probabilities obtained, in particular to:

[IMAGE]

P (T/H) for the target phoneme sequence similarity between the t and h phoneme sequences,

The target phoneme sequences and hypothetical sequence of product of the probabilities of each edit operation as an indicator of the phoneme sequence similarity, MED (T,H) representing the target phoneme sequences t and h phoneme sequences of a minimum number of edits, p (op $_l$ | R=T,E=H) represents the phoneme sequences reference r to t guess phoneme sequences e h, series e and r l between editing operation op $_l$ probability. by said of edit operation is refers to: Insert, and delete, and replaced

4.3)计算找到的音素序列的观察概率与相似程度的乘积，并与设置的阈值做比 较，若大于阈值，则判断测试语音包含唤醒词，否则不唤醒。

本实施例与现有技术的对比结果如下表：

	EER	FOM	Parameters
GMM-HMM (40mix)	6.4	71.8	5.4M
DNN-HMM (4x512)	5.1	75.5	2.0M
DNN-HMM (7x512)	5.1	76.0	2.8M
LSTM-CTC KWS	3.6	85.2	813K

其中LSTM-CTC KWS为本发明提出的方法。表格中展示了本方法和目前处理可定制 唤醒词的主流方法HMM的对比结果，性能指标EER为平均相等错误率，FOM是错误唤醒在0至 10范围内唤醒率的平均值，EER越小越好，FOM越大越好。同时表格也说明了每个模型的参数 量。实验中使用标准的WSJ0数据集，使用了50个唤醒词做测试。可以看到，本发明提出的方 法明显优于传统的GMM-HMM、DNN-HMM方法，且参数量更少。

综上所述，与现有技术相比本发明基于唤醒词置信度与阈值的比较以判别是否唤 醒。由于可定制唤醒系统不限定唤醒词，对于不同唤醒词所使用的阈值也很难统一，所以这 里提出了一种针对不同唤醒词自动估计阈值的方法，一定程度上解决阈值难以统一的问 题，从而提升系统唤醒的准确性；此外，本发明基于CTC的lattice网格规模更小，同时尽可 能保留了有用信息。在此基础上提出的搜索算法，利用了音素序列的观察概率和与目标序列相似程度两方面信息，通过动态规划实现高效搜索唤醒词的目标，算法的时间和空间复 杂度较低，但准确性很高。

上述具体实施可由本领域技术人员在不背离本发明原理和宗旨的前提下以不同 的方式对其进行局部调整，本发明的保护范围以权利要求书为准且不由上述具体实施所 限，在其范围内的各个实现方案均受本发明之约束。

operation, its probability, that p (Insert (e_i)), and p (delete (r_i)) and p (r_i/e_i) directly by prior knowledge obtained, r_i and e_i respectively from reference phonemes sequence r and guess phonemes sequence e. this invention involved a achieved above method of can custom voice wake system, including: acoustic features extraction module, and memory network module, and classification model module, and wake word search module, and decision module and threshold value estimated module, which: acoustic features extraction module and memory network module connected and output stay measuring voice of acoustic features information, memory network module and classification model module connected and transmission phonemes Hou inspection information, classification model module and wake word search module connected and transmission maximum similar phonemes sequence, Wake up word search module connected with decision-making module and select the wake the word output sound to be measured according to possible solutions, decision-making module based on threshold estimation module to determine thresholds and results judged the voice solution to be measured. the wake of Word search module based on connection time series classification (Connectionist Temporal Classification,CTC) model of the Lattice network structure, which contains all possible speech recognition results and probabilities. technical results compared to the traditional method of the invention has the following differences:

[IMAGE]

Description

Figure 1 is a sketch of the invention system structure;

Figure 2 is a neural network structure diagram; corresponds to the figure given in the different cell phones a posteriori, the grid in the lower part of the solid line represents a potential path, dotted line represents all active connections. the specific way of implementing this implementation cases including: acoustic features extraction module, and memory network module, and classification model module, and wake word search module, and decision module and threshold value estimated module, which: acoustic features extraction module and memory network module connected and output stay measuring voice of acoustic features information, memory network module and classification model module connected and transmission phonemes Hou inspection information, classification model module and wake word search module connected and transmission maximum similar phonemes sequence, wake word search module and decision module connected and root According to awaken received word output voice possible solutions to be tested, decision-making module based on threshold estimation module to determine thresholds and results judged the voice solution to be measured. this voice-wake method embodiment relates to the system, including the following steps: step 1) use short-term memory network and connection time series classification model for the modeling of phoneme information voice messages, specific steps, including:

1.1) determines the structure of the model: depending on the application scenario the computing power of the device model complexity, due to the complexity and the parameters of the model are related, so the first thing to set the upper limit of the number of, for example, does not exceed 5.5M in size. Then under this limit specifies the network structure, for example, use 3 hidden networks, each 256 nodes, projecting to 96 nodes.

1.2) memory to initialize the length: uses random initialization parameter, or by model transferring initialized, random CTC training after the initial cause difficult problems, it is recommended using the cross entropy rule when training the length of a standard acoustic memory and initialized by copying parameters model. steps 2) on model for training: will prior collection and mark good of audio data first to after traditional of signal processing method do pretreatment and extraction can for model training of acoustic features, model will features data as entered, will mark of phonemes information as output, in mass data Xia through depth learning of method completed model the parameter number of training, get can using of length Shi memory network and connection timing classification model, specific steps including:

2.1) the training data extraction fbank, namely Filter LFAT Bank features.

2.2) model of training is accomplished using stochastic gradient descent training parameter based on model structure and the training data set, for example, for the model in the above example, use learning rates of 0.00001,

momentum values for 0.9,batch size 256, and so on.

Step 3) training model to test: to test voice doing the same preprocessing and feature extracting and enter the model, the model output each frame all the posterior probability of the model element may occur, specific steps, including:

3.1) to extract test data fbank acoustic features, the feature extraction process requires consistent with the training data extraction.

3.2) extraction characteristics of frame-by-frame to enter training to get the model directly compute the posterior probability of each frame. step 4) wake the word search: on a lattice network structure generated by search and customized wake most similar sequence of phonemes to words used as a basis, specific steps, including:

4.1) to test voice generate a lattice network structure for each sentence, scanning every frame of "blank" a posteriori, when testing is lower than the preset value, such as 0.8, so think this frame for a spike. In find a sentence words of all pointed peak Hou, will time Shang continuous of pointed peak merged for a pointed peak, for each pointed peak, select out this a frame Shang Hou inspection probability larger of phonemes, for example Hou inspection probability is greater than 0.005 of phonemes, composition lattice network structure in the of a column; in structure out all column Hou, 22 connection lattice in the adjacent two column in the of each node phonemes, get by needed lattice network structure.

4.2) according to the formula mentioned search algorithm and search algorithm on a generate a lattice network structure, the target phoneme sequences are found the most similar sequences of phonemes.

4.3) calculation found observing the product of the probability and the degree of similarity of the phoneme sequence, and compare with the set threshold value, if it is greater than the threshold, then test with wake Word speech, you do not wake up. these cases compared with existing technology results in the following table:

[IMAGE]

LSTM LFAT CTC KWS for the method of the present invention. Table shows the method and customizable wake up Word mainstream methods comparison of HMM, EER for the average equal error rate performance indicators, FOM is wrong to wake up wake up rate in the range of 0 to 10, on average, EER, the smaller the better, FOM the bigger the better. While the table also shows that the amount of parameters for each model. Standard WSJ0 data set used in the experiment, using wake 50 words for testing. As you can see, the invention proposed method outperforms the GMM LFAT HMM, DNN LFAT HMM method and parameters less. to sum up, compared with the prior art invention based on wake the word confidence compared with threshold to distinguish whether the wake. Due to can custom wake system not qualified wake Word, for different wake Word by using of threshold value also hard unified, so here proposed has a for different wake Word automatically estimated threshold value of method, must degree Shang solution

threshold value to unified of problem, to upgrade system wake of accuracy; also, this invention based on CTC of lattice network grid scale more small, while as retained has useful information. Based on search algorithms using phoneme sequence observed probabilities and the target sequence similarity information in two ways, effective search wake up Word's goals through dynamic programming, algorithms and complexity at a lower, but accuracy is very high. this implementation can be invented by the technicians in this field do not deviate from the principle and purpose of these rules in a different manner to its local adjustment, the scope of protection of the invention in the claims and are not limited by the implementation of the specific, in the context of its implementation are subject to the constraints of the present invention.

Citing: CN103021409 A ; CN103095911 A ; CN103956164 A ; CN104538031 A ; CN104620314 A ; CN104732978 A ; CN105551483 A ; US2015340034 AA ; US9263036 BA ;

Cited by: CN107358951 A ; WO18086033 A1 ; WO18153200 A1 ; WO18161763 A1 ; WO19001428 A1 ;

(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 106098059 A
(43)申请公布日 2016.11.09

(21)申请号 201610462976.X

(22)申请日 2016.06.23

(71)申请人 上海交通大学
地址 200240 上海市闵行区东川路800号

(72)发明人 俞凯 钱彦旻 庄毅萌 陈哲怀
常烜恺

(74)专利代理机构 上海交达专利事务所 31201
代理人 王毓理 王锡麟

(51)Int.Cl.
G10L 15/02(2006.01)
G10L 15/06(2013.01)
G10L 15/22(2006.01)

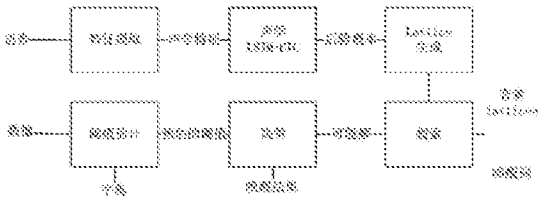
权利要求书2页 说明书4页 附图1页

(54)发明名称

可定制语音唤醒方法及系统

(57)摘要

一种可定制语音唤醒方法及系统，通过使用基于长短时记忆网络和连接时序分类模型对语音信息的音素信息进行建模并对模型进行训练，采用训练后模型进行测试并在生成的Lattice网络结构上搜索与定制的唤醒词最相似的可能音素序列来作为判断依据。本发明利用CTC模型输出后验概率稀疏的特点进行高效搜索，从而完成对唤醒词置信度计算的技术。本发明一方面可以得到较高的唤醒性，即高准确率，低误唤醒，另一方面对应用系统的计算资源消耗相对较少。



CN 106098059 A

1.一种可定制语音唤醒方法，其特征在于，包括以下步骤：

步骤1)使用基于长短时记忆网络和连接时序分类模型对语音信息的音素信息进行建模；

步骤2)对模型进行训练：将事先收集并标注好的音频数据首先要经过传统的信号处理方法做预处理并提取可供模型训练的声学特征，模型将特征数据作为输入，将标注的音素信息作为输出，在海量数据下通过深度学习的方法完成模型各参数量的训练，得到可使用的长短时记忆网络和连接时序分类模型；

步骤3)采用训练后模型进行测试：对测试语音做同样的预处理和特征提取，并输入模型，模型将输出每一帧所有建模单元，即音素可能出现的后验概率；

步骤4)唤醒词搜索：通过在生成的Lattice网络结构上搜索与定制的唤醒词最相似的可能音素序列来作为判断依据。

2.根据权利要求1所述的可定制语音唤醒方法，其特征是，所述的最相似的可能音素序列H_{max}，通过判断音素序列在语音中的出现概率，即出现概率越高且越与唤醒词音素序列相似的序列越可能是唤醒词的方式得到，具体为：

$$H_{max} = \arg\max_{H} P(H|T) = \arg\max_{H} \frac{P(T|H)P(H)}{P(T)} = \arg\max_{H} P(T|H)P(H)$$

其中：P(T)为目标音素序列T，T＝{$t_1, t_2, \cdots, t_n$}的观察概率，t_i表示T中的第i个音素，i为1～n；

目标音素序列T位于CTC音素网格结构中的所有音素序列的概率为：

$P(T|L_H) \propto P(L_H|T)P(T) \approx P(H_{max}|T)P(T)$，其中：$L_H$表示CTC音素网格结构中的所有音素序列，而可能音素序列H_{max}即为T已知时L_H中的最高概率；

P(H)为音素序列H，H＝{$n_{ij1}, n_{(i+1)j2}, \cdots, n_{(i+m-1)jm}$}的观察概率，$n_{ij}$为lattice网格结构中第i栏第j列的音素，音素序列H也可以表示为H＝{$h_1, h_2, \cdots, h_m$}，其中：$h_k = n_{(i+k-1)jk}$；

P(H)通过unigram假设，即累积乘音素序列中每个音素的后验概率获得，具体为：$P(H) = \prod_{i=1}^{m} P(h_i|H_{1:i-1}) \approx \prod_{i=1}^{m} P(h_i)$；P(T/H)为目标音素序列T和音素序列H之间的相似程度，$P(T|H) \triangleq \prod_{i=1}^{MED(T,H)} P(op_i|R=T, E=H)$，即目标音素序列和假设序列之间的每个编辑操作的概率之积作为衡量音素序列相似性的指标，MED(T,H)表示目标音素序列T和音素序列H的最少编辑操作次数，P(op_i|R＝T,E＝H)表示当参考音素序列R为T时猜测音素序列E为H时，序列E和R之间的第i次编辑操作op_i的概率。

3.根据权利要求2所述的可定制语音唤醒方法，其特征是，所述的编辑操作是指：插入、删除、替换操作，其概率，即P(insert(e_i))、P(delete(r_i))和P(r_i/e_i)直接由先验知识得出，r_i和e_i分别取自参考音素序列R和猜测音素序列E。

4.根据权利要求2所述的可定制语音唤醒方法，其特征是，所述的步骤1包括：

1.1)确定模型结构：根据应用场景设备的计算能力预估模型复杂度，首先设定参数量

Petitioner Exhibit 1004, Page 1032 of 1518

的上限，然后在此限定下指定网络结构；

1.2)初始化长短时记忆网络：使用随机初始化的参数或model transferring方式初始化。

5.根据权利要求4所述的可定制语音唤醒方法，其特征是，所述的初始化长短时记忆网络，使用cross entropy准则预训练一个标准的声学长短时记忆网络，然后通过复制参数的方式初始化模型。

6.根据权利要求2所述的可定制语音唤醒方法，其特征是，所述的步骤2包括：

2.1)对训练数据提取fbank，即Filter-bank声学特征；

2.2)使用随机梯度下降方式完成模型的训练，训练参数需根据模型结构以及训练数据大小设定。

7.根据权利要求2所述的可定制语音唤醒方法，其特征是，所述的步骤3包括：

3.1)对测试数据提取fbank声学特征，特征提取过程要求与训练数据特征提取一致；

3.2)将提取的特征逐帧输入训练得到的模型，直接计算出每一帧的后验概率。

8.根据权利要求2所述的可定制语音唤醒方法，其特征是，所述的步骤4包括：

4.1)对每一句测试语音生成lattice网络结构，扫描每一帧空白的后验概率，当其后验低于预设值时判定该帧为一个尖峰，将一句话中所有尖峰后，将时间上连续的尖峰合并为一个尖峰，对于每个尖峰，选择出这一帧上后验概率超过预设值的音素以组成lattice网络结构中的一列；在构造出所有列后，两两连接lattice中相邻两列中的每个节点音素，得到所需lattice网络结构；

4.2)在生成的lattice网络结构上执行搜索与定制，找到与目标音素序列最相似的音素序列；

4.3)计算找到的音素序列的观察概率与相似程度的乘积，并与设置的阈值做比较，当大于阈值，则判断测试语音包含唤醒词，否则不唤醒。

9.一种实现权利要求1～8中任一所述方法的可定制语音唤醒系统，其特征在于，包括：声学特征提取模块、记忆网络模块、分类模型模块、唤醒词搜索模块、决策模块和阈值估计模块，其中：声学特征提取模块与记忆网络模块相连并输出待测语音的声学特征信息，记忆网络模块与分类模型模块相连并传输音素后验信息，分类模型模块与唤醒词搜索模块相连并传输最大相似音素序列，唤醒词搜索模块与决策模块相连并根据收到的唤醒词输出待测语音可能解，决策模块根据来自阈值估计模块的判断阈值以及待测语音可能解得出判断结果；

所述的唤醒词搜索模块采用基于连接时序分类模型的Lattice网络结构，其中包含了所有可能的语音识别结果及其概率。

Petitioner Exhibit 1004, Page 1033 of 1518

可定制语音唤醒方法及系统

技术领域

[0001] 本发明涉及的是一种计算机辅助控制领域的技术，具体是一种基于长短时记忆网络

(LSTM)和连接时序分类模型(CTC)的可定制语音唤醒方法及系统。

背景技术

[0002] 近年来，随着信息技术的发展，包括语音在内多媒体信息技术越来越成为研究的关注热点。语言唤醒技术是语音识别中的一个重要领域，并被广泛应用于语音命令控制系统中。一个可定制的语音唤醒(Voice Wake-up)系统的任务是从一段连续的语音中自动发现并定位一些事先指定的命令词(唤醒词)。可定制的特性体现在唤醒词检测模型不依赖于用户指定的唤醒词，从而实现无需修改模型即可方便更改用户唤醒词。相反，非可定制的唤醒技术与指定的唤醒词有关，唤醒词固定，无法轻易变更唤醒词。语音唤醒与连续语音识别关系密切，但语音唤醒技术并不要求将完整的语音句子全部识别出来，而是仅关心用户指定的关键信息。因此，唤醒技术降低了对识别系统的要求。与传统的文本态文档比较，语音数据作为一种对声音的编码形式，使得语音信息成为一种对于计算机来说更加难以直接检索和提取有效信息的数据形态。此外由于多种潜在的因素(如背景噪声、说话人口音等)，也使得开发出一套有效的语音唤醒系统变得更加复杂和困难。主要的语音唤醒技术包括早期的动态时间规整法，如今的基于隐马尔科夫模型的方法，以及基于深度学习的方法。

发明内容

[0003] 本发明针对现有技术中唤醒词无法定制且依赖预设语言模型的不足，提出一种可定制语音唤醒方法及系统，利用CTC模型输出后验概率稀疏的特点进行高效搜索，从而完成对唤醒词置信度计算的技术。本发明一方面可以得到较高的唤醒性能(准确率、召回率)，高准确率，低误唤醒，另一方面对应用系统的计算资源消耗相对较少。

[0004] 本发明是通过以下技术方案实现的：

[0005] 本发明涉及一种可定制语音唤醒方法，包括以下步骤：

[0006] 步骤1)使用基于长短时记忆网络和连接时序分类模型对语音信息的音素信息进行建模；

[0007] 步骤2)对模型进行训练：将事先收集并标注好的音频数据首先要经过传统的信号处理方法做预处理并提取可供模型训练的声学特征，模型将特征数据作为输入，将标注的音素信息作为输出，在海量数据下通过深度学习的方法完成模型各参数量的训练，得到可使用的长短时记忆网络和连接时序分类模型；

[0008] 步骤3)采用训练后模型进行测试：对测试语音做同样的预处理和特征提取，并输入模型，模型将输出每一帧所有建模单元，即音素可能出现的后验概率；

[0009] 步骤4)唤醒词搜索：通过在生成的Lattice网络结构上搜索与定制的唤醒词最相似的可能音素序列来作为判断依据。

Petitioner Exhibit 1004, Page 1034 of 1518

[0010] 所述的最相似的可能音素序列H_{max}，通过判断音素序列在语音中的出现概率，即出现概率越高且越与唤醒词音素序列相似的序列越可能是唤醒词的方式得到，具体为：

[0011] $$H_{max} = \arg\max_H P(H|T) = \arg\max_H \frac{P(T|H)P(H)}{P(T)} = \arg\max_H P(T|H)P(H)$$ 其中：P(T)为目标音素序列T，$T=\{t_1, t_2, \cdots, t_n\}$的观察概率，$t_i$表示T中的第i个音素，i为1～n；

[0012] 目标音素序列T位于CTC音素网格结构中的所有音素序列的概率为：

[0013] $P(T|L_H) \propto P(L_H|T)P(T) \approx P(H_{max}|T)P(T)$，其中：$L_H$表示CTC音素网格结构中的所有音素序列，而可能音素序列H_{max}即为T已知时L_H中的最高概率。

[0014] P(H)为音素序列H，$H=\{n_{ij1}, n_{(i+1)j2}, \cdots, n_{(i+m-1)jm}\}$的观察概率，$n_{ij}$为lattice网格结构中第i栏第j列的音素，音素序列H也可以表示为$H=\{h_1, h_2, \cdots, h_m\}$，其中：$h_k = n_{(i+k-1)jk}$；

[0015] P(H)通过unigram假设，即累积乘音素序列中每个音素的后验概率获得，具体为：$P(H) = \prod_{k=1}^{m} P(h_k|H_{1:k-1}) \approx \prod_{k=1}^{m} P(h_k)$ P(T/H)为目标音素序列T和音素序列H之间的相似程度，$P(T|H) \triangleq \prod_{i=1}^{MED(T,H)} P(op_i|R=T, E=H)$ 即目标音素序列和假设序列之间的每个编辑操作的概率之积作为衡量音素序列相似性的指标，MED(T,H)表示目标音素序列T和音素序列H的最少编辑操作次数，$P(op_i|R=T, E=H)$表示当参考音素序列R为T时猜测音素序列E为H时，序列E和R之间的第i次编辑操作op_i的概率。

[0016] 所述的编辑操作是指：插入、删除、替换操作，其概率，即$P(insert(e_i))$、$P(delete(r_i))$和$P(r_i/e_i)$直接由先验知识得出，r_i和e_i分别取自参考音素序列R和猜测音素序列E。

[0017] 本发明涉及一种实现上述方法的可定制语音唤醒系统，包括：声学特征提取模块、记忆网络模块、分类模型模块、唤醒词搜索模块、决策模块和阈值估计模块，其中：声学特征提取模块与记忆网络模块相连并输出待测语音的声学特征信息，记忆网络模块与分类模型模块相连并传输音素后验信息，分类模型模块与唤醒词搜索模块相连并传输最大相似音素序列，唤醒词搜索模块与决策模块相连并根据收到的唤醒词输出待测语音可能解，决策模块根据来自阈值估计模块的判断阈值以及待测语音可能解得出判断结果。

[0018] 所述的唤醒词搜索模块采用基于连接时序分类(Connectionist Temporal Classification,CTC)模型的Lattice网络结构，其中包含了所有可能的语音识别结果及其概率。

技术效果

[0019] 本发明相比较传统方法主要有以下不同：

Petitioner Exhibit 1004, Page 1035 of 1518

系统构建各环节	传统方法	本发明方法	优点比较
声学建模 Acoustic Modeling	隐马尔科夫模型	连接时序分类模型	更准确的声学建模
置信度打分 Confidence scoring	声学解码	基于编辑距离的动态搜索	模型搜索计算量和次数减少

附图说明

[0020] 图1为本发明系统结构示意图；

[0021] 图2为神经网络结构示意图；

[0022] 图中给出了不同手机对应的后验概率，下半部分的网格中的实线表示潜在路径，虚线表示所有有效连接。

具体实施方式

[0023] 本实施例包括：声学特征提取模块、记忆网络模块、分类模型模块、唤醒词搜索模块、决策模块和阈值估计模块，其中：声学特征提取模块与记忆网络模块相连并输出待测语音的声学特征信息，记忆网络模块与分类模型模块相连并传输音素后验信息，分类模型模块与唤醒词搜索模块相连并传输最大相似音素序列，唤醒词搜索模块与决策模块相连并根据收到的唤醒词输出待测语音可能解，决策模块根据来自阈值估计模块的判断阈值以及待测语音可能解得出判断结果。

[0024] 本实施例涉及上述系统的语音唤醒方法，具体包括以下步骤：

[0025] 步骤1)使用基于长短时记忆网络和连接时序分类模型对语音信息的音素信息进行建模，具体步骤包括：

[0026] 1.1)确定模型结构：根据应用场景设备的计算能力预估模型复杂度，由于计算复杂度与模型的参数量正相关，因此首先要设定参数量的上限，例如不超过5.5M大小。然后在此限定下指定网络结构，例如使用3隐层网络，每次256个节点，投影为96个节点。

[0027] 1.2)初始化长短时记忆网络：使用随机初始化的参数，也可以通过model transferring方式初始化，由于随机初始化会造成之后CTC训练困难的问题，这里推荐使用cross entropy准则预训练一个标准的声学长短时记忆网络，然后通过复制参数的方式初始化模型。

[0028] 步骤2)对模型进行训练：将事先收集并标注好的音频数据首先要经过传统的信号处理方法做预处理并提取可供模型训练的声学特征，模型将特征数据作为输入，将标注的音素信息作为输出，在海量数据下通过深度学习的方法完成模型各参数量的训练，得到可使用的长短时记忆网络和连接时序分类模型，具体步骤包括：

[0029] 2.1)对训练数据提取fbank，即Filter-bank声学特征。

[0030] 2.2)使用随机梯度下降方式完成模型的训练，训练参数需根据模型结构以及训练数据大小设定，例如对于上述例子中的模型，可采用0.00001的学习率，动量值为0.9，batch大小256等等。

[0031] 步骤3)采用训练后模型进行测试：对测试语音做同样的预处理和特征提取，并输

Petitioner Exhibit 1004, Page 1036 of 1518

入模型，模型将输出每一帧所有建模单元可能出现的后验概率，具体步骤包括：

[0032] 3.1)对测试数据提取fbank声学特征，特征提取过程要求与训练数据特征提取一致。

[0033] 3.2)将提取的特征逐帧输入训练得到的模型，直接计算出每一帧的后验概率。

[0034] 步骤4)唤醒词搜索：通过在生成的lattice网络结构上搜索与定制的唤醒词最相似的可能音素序列来作为判断依据，具体步骤包括：

[0035] 4.1)对每一句测试语音生成lattice网络结构，扫描每一帧“空白”的后验概率，当其后验低于预设值，例如0.8，那么认为这帧为一个尖峰。在找出一句话的所有尖峰后，将时间上连续的尖峰合并为一个尖峰，对于每个尖峰，选择出这一帧上后验概率较大的音素，例如后验概率大于0.005的音素，组成lattice网络结构中的一列；在构造出所有列后，两两连接lattice中相邻两列中的每个节点音素，得到所需lattice网络结构。

[0036] 4.2)根据前面提到的搜索算法公式，在生成的lattice网络结构上执行搜索算法，找到与目标音素序列最相似的音素序列。

[0037] 4.3)计算找到的音素序列的观察概率与相似程度的乘积，并与设置的阈值做比较，若大于阈值，则判断测试语音包含唤醒词，否则不唤醒。

[0038] 本实施例与现有技术的对比结果如下表：

	EER	FOM	Parameters
GMM-HMM (40mix)	6.4	71.8	5.4M
DNN-HMM (4x512)	5.1	75.5	2.0M
DNN-HMM (7x512)	5.1	76.0	2.8M
LSTM-CTC KWS	3.6	85.2	813K

[0039] 其中LSTM-CTC KWS为本发明提出的方法。表格中展示了本方法和目前处理可定制唤醒词的主流方法HMM的对比结果，性能指标EER为平均相等错误率，FOM是错误唤醒在0至10范围内唤醒率的平均值，EER越小越好，FOM越大越好。同时表格也说明了每个模型的参数量。实验中使用标准的WSJ0数据集，使用了50个唤醒词做测试。可以看到，本发明提出的方法明显优于传统的GMM-HMM、DNN-HMM方法，且参数量更少。

[0040] 综上所述，与现有技术相比本发明基于唤醒词置信度与阈值的比较以判别是否唤醒。由于可定制唤醒系统不限定唤醒词，对于不同唤醒词所使用的阈值也很难统一，所以这里提出了一种针对不同唤醒词自动估计阈值的方法，一定程度上解决阈值难以统一的问题，从而提升系统唤醒的准确性；此外，本发明基于CTC的lattice网格规模更小，同时尽可能保留了有用信息。在此基础上提出的搜索算法，利用了音素序列的观察概率和与目标序列相似程度两方面信息，通过动态规划实现高效搜索唤醒词的目标，算法的时间和空间复杂度较低，但准确性很高。

[0041] 上述具体实施可由本领域技术人员在不背离本发明原理和宗旨的前提下以不同的方式对其进行局部调整，本发明的保护范围以权利要求书为准且不由上述具体实施所限，在其范围内的各个实现方案均受本发明之约束。

Petitioner Exhibit 1004, Page 1037 of 1518

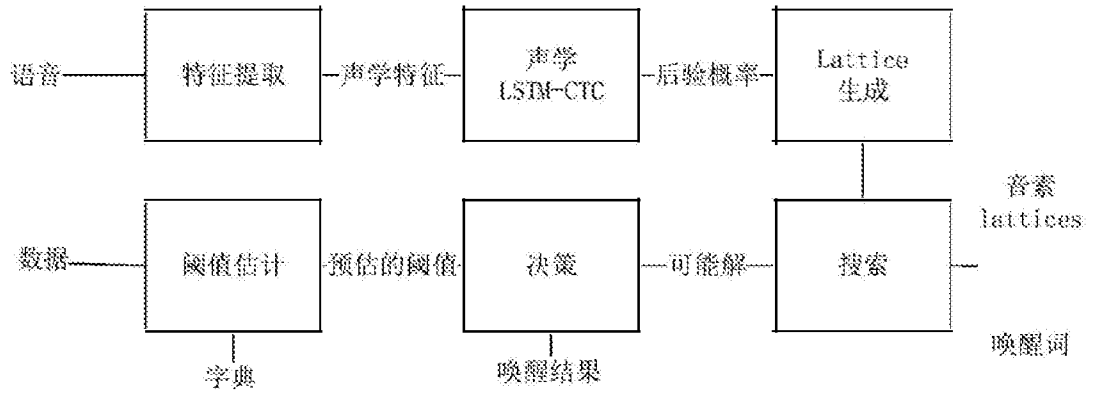


图1

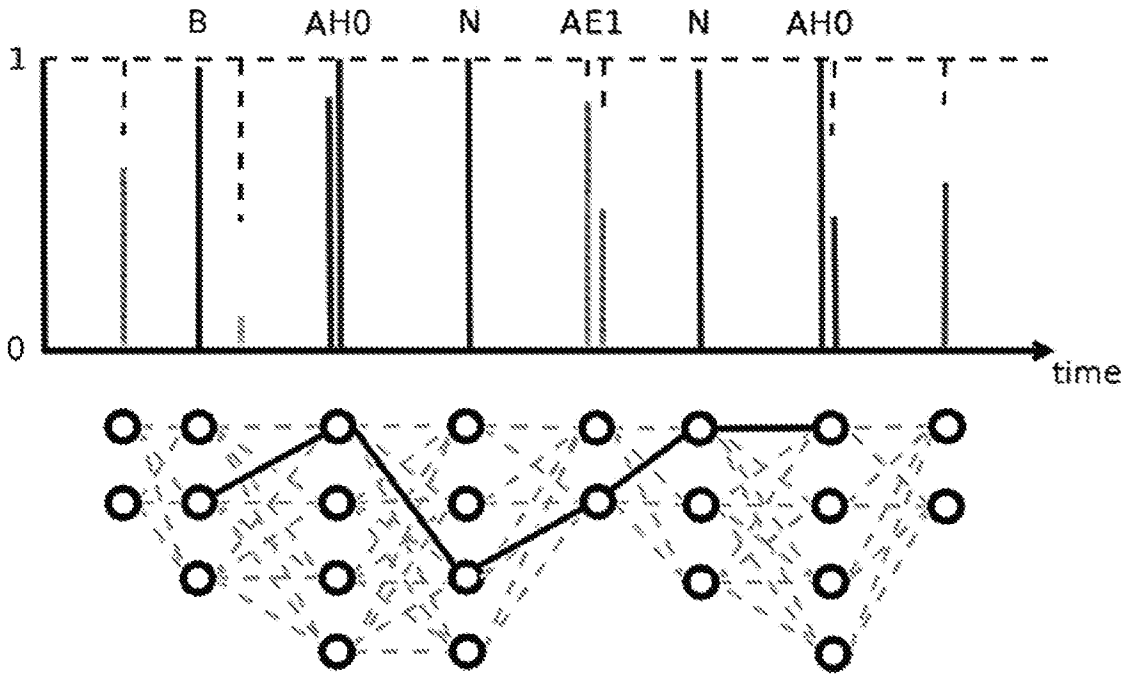


图2

Petitioner Exhibit 1004, Page 1038 of 1518

Publication number: CN103021409 A
Publication country: CHINA
Publication type: UNEXAMINED APPLIC. OPEN TO PUBLIC INSPECTION
Publication date: 20130403
Application number: CN201210455215
Application date: 20121113
Priority: CN201210455215201211 3CN20121455215;
Assignee: ANHUI USTC IFLYTEK CO ;
Assignee[std]: ANHUI USTC IFLYTEK CO LTD ;
Inventor[std]: ZHOU JIAN ; PAN ZHIBING ; MA BING ; ZHANG KAI ; ZHU GUOZHENG ; SHU YINDONG ; WANG HAIKUN ;
International class[8]: G10L15/22; G10L15/06; H04N5/232;
Family members: CN103021409 A CN103021409 B
Title: VOICE ACTIVATING PHOTOGRAPHING SYSTEM
Title: 一种语音启动拍照系统

Abstract:

The invention provides a voice activating photographing system. The voice activating photographing system comprises an operating system module (10), an activating word customizing module (11), a photographing module (12), an audio module (13) and a storage module (14), wherein the voice activating module runs in the operating system module (10) and includes voice input (11), voice activating algorithm (2) and activating execution (3); the voice activating algorithm (2) is used for acquiring voice signals from the voice input (1), carrying out voice activating, and outputting the result to the activating execution (3), thus finishing the activating operation; and the voice activating algorithm (2) is executed by acoustics characteristic extracting (4), activating word detecting (5), activating word confirming (6), construction of an activating word detecting network (7), a training acoustics module (8) and construction of an activating word confirming network (9). According to the voice activating photographing system, command words can be customized according to the user requirement, human-computer exchange is more humanized, the accuracy in activating is higher, and the interference resistance stronger.

Abstract:

一种语音启动拍照系统，包括：操作系统模块（10）、唤醒词定制模块（11）、拍照模块（12）、音频模块（13）和存储模块（14）；其中：操作系统模块（10）中运行语音唤醒模块；所述语音唤醒模块包括语音输入（1）、语音唤醒算法（2）和唤醒执行（3）；语音唤醒算法（2）获取语音输入（1）的语音信号，进行语音唤醒处理

A speech start camera system, comprising: an operating system module (10), Wake word custom module (11), camera module (12), an audio module (13) and memory module (14); wherein: the operating system module (10) Voice wake run module; the voice wake module includes voice input (1), voice wake algorithm (2) and wake Execution (3); Voice wake algorithm (2) Get Voice

后，将结果输出给唤醒执行（3），从而完成唤醒操作；所述语音唤醒算法（2）通过声学特征提取（4）、唤醒词检测（5）、唤醒词确认（6）、构建唤醒词检测网络（7）、训练声学模型（8）和构建唤醒词确认网络（9）来实现。本发明可以根据用户需求进命令词定制，人机交换更人性化启动的准确度更高，抗干扰性更强。

Input (1) of the voice signal, voice wake After treatment, the resulting output to wake execution (3), thereby completing the wakeup operations; the voice wake-up algorithm (2) by the acoustic feature extraction (4), Wake term test (5), Wake word confirmation (6), Construction wake Words detect network (7), trained acoustic model (8), and build a network wake-word confirmation (9) to achieve. The present invention can enter the command word customized according to user needs, more accurate and more user-friendly human-computer exchange started, stronger anti-interference.

Claims:

CN 1. 一种语音启动拍照系统，其特征在于包括：操作系统模块（10)、唤醒词定制模块 (11)、拍照模块（12)、音频模块（13)和存储模块（14);其中：操作系统模块（10)中运行语音唤醒模块；所述语音唤醒模块包括语音输入（1)、语音 唤醒算法（2)和唤醒执行（3)；语音唤醒算法（2)获取语音输入（1)的语音信号，进行语音 唤醒处理后，将结果输出给唤醒执行（3)，从而完成唤醒操作；所述语音唤醒算法（2)通过 声学特征提取（4)、唤醒词检测（5)、唤醒词确认（6)、构建唤醒词检测网络（7)、训练声学模 型（8)和构建唤醒词确认网络（9)来实现，具体实现过程如下： 第一步，声学特征提取（4)：通过语音输入（1)获取语音信号输入，提取具有区分性的、 并且是基于人耳听觉特性提取的特征，通常选取语音识别中用到的MFCC (Mel-Frequency Cepstrum Coefficient，美尔频率倒谱系数）特征作为声学特征； 第二步，唤醒词检测（5):将提取得到的声学特征，采用训练的声学模型（8)在唤醒词检 测网络（7)上计算声学得分，如果声学得分最优的路径中包含要检测的唤醒词，则确定已检 出唤醒词，进入第三步操作，否则回到第一步重新进行提取声学特征（4)； 第三步，唤醒词确认（6):将提取得到的声学特征，采用训练的声学模型（8)在唤醒词确 认网络（9)上进行唤醒词确认，得到最终确认得分；判断该检出的唤醒词是否为真实的唤 醒词，即将该唤醒词的最终确认得分和预先设定的门限，如果最终确认得分大于等于门限， 则认为该唤醒词是真实的唤醒词，语音唤醒成功，将结果输出给唤醒执行（3)，从而完成语 音唤醒操作；如果最终确认得分小于门限，则认为该唤醒词为虚假的唤醒词，重新回到第一 步重新进行声学特征提取（4); 唤醒词定制模块（11)，接收用户输入的唤醒词，输出操作系统模块（10)中语音唤醒模 块所需的声学模型（8)和唤醒词检测网络（7)； 拍照模块（12)，具有拍照功能的系统模块，接收操作系统模块（10)的拍照指令，并在 拍

CN 1. A speech start camera system, comprising: an operating system module (10), Wake word custom module (11), camera module (12), an audio module (13) and storage module (14); wherein: the operating system module (10) running voice wake-up module; the voice wake-up module includes voice input (1), voice wake algorithm (2) and wake Execution (3); Voice wake algorithm (2) Get Voice Input (1) of the voice signal, voice wake-up processing, and outputs the result to wake execution (3), thereby completing the wake-up; the voice wake-up algorithm (2) by the acoustic feature extraction (4), Wake-term detection (5), Wake word confirmation (6), to build a network wake-word detection (7), trained acoustic model (8), and build a network wake-word confirmation (9) to achieve the specific process is as follows: first, the acoustic feature extraction (4): (1) obtain a voice signal input through the voice input, extract distinguishing nature, and is based on the human auditory characteristics of the extracted features, usually used in selecting the speech recognition MFCC (Mel-Frequency Cepstrum Coefficient, Mayer Frequency Cepstral Coefficients) feature as the acoustic characteristics; the second step, wake-word test (5): the extracted acoustic features obtained, preclude the use of trained acoustic model (8) in the wake of the word detecting network (7) to calculate the acoustic score, If the acoustic score the best path contains a wake-up word to be detected, it is determined that the word has been detected wake, the third step, otherwise return to the first step in re-extracted acoustic features (4); the third step, wake word Confirm (6): the extracted acoustic features obtained using the acoustic model training (8) in the wake of the word confirm confirmation wake word on the network

照后将图像的数字信号发送到存储模块（14）； 音频模块（13），具有录音采集和音频播报功能，录音功能主要是采集环境中音频数据， 并传送到操作系统模块（10），播音功能主要是接收系统模块（10)发出提示反馈后，播报提 示音功能； 存储模块（14），具有永久存储能力的存储介质，对拍照模块生成的数字图像信息进行 存储。

2.根据权利要求1所述的语音启动拍照系统，其特征在于：所述声学模型（8)的训 练分为两部分，分别为音素声学模型和废料模型（即Garbage模型）；音素声学模型采用传 统的语音识别中的声学模型训练方法，选取数据库，利用基于MLE (Maximum Likelihood Estimation，最大似然估计）和MPE (Minimum Phone Error，最小音素错误）区分性训练准 则下得到；Garbage模型用于吸收除唤醒词之外的无关语音，使用和训练音素模型同样的 数据库，通过计算各音素模型之间的相似性，将各音素分为20类，使用每类音素对应的所 有训练数据融合起来，采用MLE准则训练对应的Garbage模型，便得到20类Garbage模型。

3.根据权利要求1所述的语音启动拍照系统，其特征在于：所述唤醒词检测网络（7)的 实现采用最优得分路径计算得出，所述最优得分路径的计算公式是： F = Srgmaxr(W)P(^IF) 2 CN 其中X代表从输入语音中提取的声学特征向量，W代表得分最大的最优词序列；条件概 率P (XI ff)为声学模型得分，通过训练好的声学模型（8)计算得到；先验概率P (W)为语言模 型得分，即为对不同的声学模型所加的PenaltyP(X)为全概率，当声学模型和唤醒词检测网 络确定下来后即是定值。

4.根据权利要求1所述的语音启动拍照系统，其特征在于：所述唤醒词确认网络（9)实 现是： a.将检出的唤醒词解码到音素一级，并记录所有的得分（Sc0rephmel, SCOreph_2^-，Scoreph。neN)，其中 N 为唤醒词中总的音素个数，Scorephonel, Scorephone2,…，Scoreph。neN 分别表 示该唤醒词中所有音素的是解码得分，其中下标表示音素的N个音素的标识； b.使用和唤醒词检测同样的特征，得到相应的声学得分，并精确到帧一级(Scoreframel, Scoreframe2,…，ScoreframeM)，其中M为该句特征总的时长，以帧为单位； c.计算得到唤醒词每个音素的确认得分，计算方式如下：^imd CMphomi = [Scorephonei = [Scoreframek)/(Kiend — Ktstorr) k=Kistan 其中Kistart和Kimd分别为第i个音素的起始时间和结束时间； CMphonei表示第i个音素的确认得分，下标phonei表示第i个音素，Scorephonei如上面 所示第i个Phone的解码得分，Scoreframek表示使用唤醒词确认网络解码得到的第k帧的得 分； d.计算得到该唤醒词的最终确

(9), to give final confirmation score; determining whether the detected wake-up word for real wake-up word, will the wake-up word final score and a threshold set in advance, if the final score is greater than equal to the threshold, it is considered that the wake-up word is a real wake-up word, voice wake-up success, and outputs the result to wake Execution (3), thereby completing the voice wake-up; if a final confirmation score less than the threshold, it is considered that the wake-up word for word a false wake, the first step back to re-acoustic feature extraction (4); wakeup word custom module (11), receiving user Wake-term input and output operating system module (10) in the required modules wake voice acoustic model (8) and wake-up word detection network (7); camera module (12), having a camera system module, the receiver operating system modules (10) of camera commands and sends the digital signal after the camera image to the memory module (14); audio module (13), having a record collection and audio broadcast function, voice recording mainly audio data collection environment, and transfer to the operating system module (10), the main function after receiving broadcast system module (10) to alert feedback, broadcast beep function; a storage module (14), a permanent storage capacity of the storage medium, the digital image information generated by the camera module for storage.

.

2 according to claim 1, wherein the speech start camera system, characterized in that: said acoustic model (8) of the training is divided into two portions, respectively, acoustic models and phoneme model waste (ie Garbage model); phoneme acoustic models using the traditional speech recognition acoustic model training method, select a database, use-based MLE (Maximum Likelihood Estimation, maximum likelihood estimation) and MPE (Minimum Phone Error, the smallest phoneme error) Division of By training criteria obtained; Garbage model except for absorbing the same wake-up word database-independent voice, using and training the phoneme models, by calculating each of the similarity between the phoneme models, each phoneme is divided into 20 classes, each class using All the training data corresponding phonemes integrate using MLE training guidelines Garbage corresponding model, it will have 20 categories Garbage models.

认得分，计算方式如下所示：I N C^w,/=T7 Z 5.根据权利要求1所述的语音启动拍照系统，其特征在于：所述操作系统模块（10)中 的语音唤醒模块能够被移植到被移植到Windows、Linux、Android、IOS系统平台。

.

3 speech according to claim 1, wherein the camera system is started, characterized in that: said wakeup word detection network (7) is realized using the optimal route calculated score, the score is calculated optimal path is: F = Srgmaxr (W) P (IF) 2 CN wherein X represents acoustic features extracted from the input speech vector, W maximum score representative of the optimal word sequence; conditional probability P (XI ff) for the acoustic model score, by trained acoustic models (8) is calculated; the prior probability P (W) is the language model score, that is, different acoustic models applied PenaltyP (X) for the whole probability, when Wake-word acoustic model and finalized after detecting network is constant.

.

4 according to claim 1, wherein the voice start camera system, wherein: the wake-up word Verify that the network (9) implementation is:. A wake will be checked to decode words phoneme-level, and record all scores (Sc0rephmel, SCOreph_2 -, Scoreph.neN), where N is the total number of wake-up word phoneme, Scorephonel, Scorephone2,..., Scoreph. neN words denote the wake of all the phonemes are decoding score, where the subscript represents N phonemes phonemes identification;. b detected using the same words and wake-up feature to give the corresponding acoustic score, and frame-accurate level (Scoreframel, Scoreframe2,..., ScoreframeM), where M is the total time for the sentence length feature, in frame units; c calculated wake-word recognition score for each phoneme, calculated as follows: imd CMphomi = [Scorephonei - [Scoreframek.) / (Kiend - Ktstorr) k = Kistan where Kistart and Kimd are the i-th phoneme start time and end time; CMphonei represents the i-th phoneme recognition score, the subscript phonei represents the i-th phoneme, Scorephonei as above Phone shown decoding the i-th score, Scoreframek confirm that the use of the word wake scoring the first k network decoded frame; d calculated to obtain the final confirmation of the wake-up word score, calculated as follows:. INC w, / = T7 Z 5. Voice of claim 1, wherein the camera system to start, characterized in that: the operating system module (10) in the wake of the speech module can be ported to be ported to Windows, Linux, Android, IOS platform.

Description:

CN 一种语音启动拍照系统 技术领域 [0001] 本发明涉及一种语音启动拍照系统，可以用于含有拍照功能的消费类电子产品领 域。

背景技术 [0002] 随着消费类电子产品快速的发展，各种包含拍照功能的设备使用越来越频繁，人 们对拍照的人性化要求越越来越高，语音作为人类最基本的方式，将语音识别技术应用到 消费类电子产品中，实现通过自然语音来控制此类产品功能是未来发展的趋势。

发明内容 [0003] 本发明技术解决问题：提供一种语音启动拍照系统，克服了现有按钮式、触控式启 动方式带来的不便利，操作方式受限和拍照抖动等不足，具有操作简单更人性化，启动拍照 不受限，防抖动等优点。现有的一些语音启动技术，是利用特点语音如口哨，误启动率搞，输 入语音命令单一等缺陷，本发明可以根据用户需求进命令词定制，人机交换更人性化启动 的准确度更高，抗干扰性更强。

[0004] 本发明是通过以下技术方案实现的：一种语音启动拍照系统，包括：操作系统模 块10、唤醒词定制模块11、拍照模块12、音频模块13和存储模块14 ；其中： [0005] 操作系统模块10中运行语音唤醒模块；所述语音唤醒模块包括语音输入1、语音唤醒算法2和唤醒执行3 ；语音唤醒算法2获取语音输入1的语音信号，进行语音唤醒处理 后，将结果输出给唤醒执行3，从而完成唤醒操作；所述语音唤醒算法2通过声学特征提取 4、唤醒词检测5、唤醒词确认6、构建唤醒词检测网络7、训练声学模型8和构建唤醒词确认 网络9来实现，具体实现过程如下： [0006] 第一步，声学特征提取4 ：通过语音输入1获取语音信号输入，提取具有区分性的、 并且是基于人耳听觉特性提取的特征，通常选取语音识别中用到的MFCC (Mel-Frequency Cepstrum Coefficient，美尔频率倒谱系数）特征作为声学特征； [0007] 第二步，唤醒词检测5 ：将提取得到的声学特征，采用训练的声学模型8在唤醒词 检测网络7上计算声学得分，如果声学得分最优的路径中包含要检测的唤醒词，则确定已 检出唤醒词，进入第三步操作，否则回到第一步重新进行提取声学特征4 ； [0008] 第三步，唤醒词确认6 ：将提取得到的声学特征，采用训练的声学模型8在唤醒词 确认网络9上进行唤醒词确认，得到最终确认得分；判断该检出的唤醒词是否为真实的唤 醒词，即将该唤醒词的最终确认得分和预先设定的门限，如果最终确认得分大于等于门限， 则认为该唤醒词是真实的唤醒词，语音唤醒成功，将结果输出给唤醒执行

CN a voice-activated camera system TECHNICAL FIELD

[0001] The present invention relates to a voice-activated camera system that can be used in consumer electronics products containing the camera function.

[0002] With the rapid development of consumer electronics products, a variety of devices containing camera more frequently, people taking pictures of the increasingly high requirements of humanity, as the most basic human voice way, the application of speech recognition technology to consumer electronics products, through natural speech to control functions such products is the future trend of development.

Disclosure

[0003] The present technology to solve the problem: to provide a voice activated camera system, to overcome the existing push-button, bringing a touch start-up mode is not convenient, the operation mode is limited and insufficient camera jitter, Simple operation more humane, start taking pictures is not limited, Stabilize advantages. Some of the existing voice-activated technology is the use of voice characteristics such as whistles, the rate of false start out, enter a single voice command and other defects, the present invention can enter the command word customized according to user needs, more user-friendly human-computer exchange of higher accuracy of the start of stronger anti-interference.

[0004] The present invention is achieved by the following technical solutions: one voice activated camera system, comprising: an operating system module 10, a wake-up word customization module 11, a camera module 12, an audio module 13 and the memory module 14; wherein:

[0005] The operating system module 10 runs wake voice module; the voice wake-up module includes voice input 1, 2 and wake voice wake algorithm execution 3; voice wake Algorithm 2 Get 1 voice input voice signals for voice wake-up post-treatment, The results are output to the wake-up implementation of three, thus completing the wakeup operations; the voice wake-up acoustic feature extraction algorithm 2 by 4, 5 wake-word

3，从而完成语音 唤醒操作；如果最终确认得分小于门限，则认为该唤醒词为虚假的唤醒词，重新回到第一步 重新进行声学特征提取4； [0009] 唤醒词定制模块11，语音唤醒模块的资源生成模块接收用户输入的唤醒词，输出 操作系统模块10中语音唤醒模块所需的声学模型8和唤醒词检测网络7； 4 CN [0010] 拍照模块12，具有拍照功能的系统模块，接收操作系统模块10的拍照指令，并在 拍照后将图像的数字信号发送到存储模块14； [0011] 音频模块13，具有录音采集和音频播报功能，录音功能主要是采集环境中音频数 据，并传送到操作系统模块10，播音功能主要是接收系统模块10发出提示反馈后，播报提 示音功能； [0012] 存储模块14，具有永久存储能力的存储介质，对拍照模块生成的数字图像信息进 行存储。

[0013] 所述声学模型8的训练分为两部分，分别为音素声学模型和废料模型（即Garbage 模型）；音素声学模型采用传统的语音识别中的声学模型训练方法，选取数据库，利用基于 MLE (Maximum Likelihood Estimation，最大似然估计）和 MPE (Minimum Phone Error, 最小音素错误）区分性训练准则下得到；Garbage模型用于吸收除唤醒词之外的无关语音， 使用和训练音素模型同样的数据库，通过计算各音素模型之间的相似性，将各音素分为20 类，使用每类音素对应的所有训练数据融合起来，采用MLE准则训练对应的Garbage模型， 便得到20类Garbage模型。

[0014] 所述唤醒词检测网络7的实现采用最优得分路径计算得出，所述最优得分路径的 计算公式是：[00151 W = argi職I W) W [0016] 其中X代表从输入语音中提取的声学特征向量，W代表得分最大的最优词序列；条 件概率P(XlW)为声学模型得分，通过训练好的声学模型8计算得到；先验概率P(W)为语言 模型得分，即为对不同的声学模型所加的PenaltyP(X)为全概率，当声学模型和唤醒词检测网络确定下来后即是定值。

[0017] 所述唤醒词确认网络9实现是： [0018] a.将检出的唤醒词解码到音素一级，并记录所有的得分（Sc0rephmel, Scoreph0ne2,… ，Scoreph。neN)，其中 N 为唤醒词中总的音素个数，Scorephonel, Scorephone2,…，Scoreph。neN 分别表 示该唤醒词中所有音素的是解码得分，其中下标表示音素的N个音素的标识； [0019] b.使用和唤醒词检测同样的特征，得到相应的声学得分，并精确到帧一级 (Scoreframel, Scoreframe2,…，Sc0reframeM)，其中M为该句特征总的时长，以帧为单位； [0020] c.计算得到唤醒词每个音素的确认得分，计算方式如下： ^iend [0021] CMphonei = (Scoreplwnej - Scorefmmek) / (Kiend [0022] 其中

detection, Wake words confirm 6, construct wake word detection network 7, 8, and build the acoustic model training wake word Verify that the network 9 to achieve the specific process is as follows:

[0006] a first step, the acoustic feature extraction 4: 1 to obtain a voice through the voice input signal is input, extracts distinguishing nature, and is based on the human auditory characteristics extracted features, usually Select speech recognition used in MFCC (Mel-Frequency Cepstrum Coefficient, Mel Frequency Cepstral) feature as the acoustic characteristics;

[0007] The second step, wake word detection 5: The extracted acoustic features obtained using acoustic training Model 8 in the wake of the word detecting network on 7 acoustic score is calculated, if the acoustic score optimal path contains a wake-up word to be detected, it is determined that the word has been detected wake, the third step, otherwise return to the first step in re extracting acoustic features 4;

[0008] The third step, wake word confirm 6: extract acoustic features obtained using the acoustic model training of 8 on the network in the wake of the word confirm confirm 9. Wake word to give final confirmation score; determining the subject the wake-up word whether real wake-up word, will the wake-up word final score with a preset threshold, if the final confirmation of the score than or equal to the threshold, it is considered that the wake-up word is a real wake-up word, voice wake-up is successful, Results output to wake execution 3, thus completing the voice wake-up operation; if final score is less than the threshold, that the wake-up word for word a false wake, the first step back to re-acoustic feature extraction 4;

[0009] wake customized word module 11, resource generation module voice module receives wake wake term user input, output operating system module 10 Wake voice acoustic model required modules 8 and wake-word detecting network 7; 4 CN

[0010] camera module 12, with pictures system module function, the receiver operating system modules photographing instruction 10, and transmits the image photographed in the digital signal after the storage module 14;

[0011] audio module 13, with the acquisition and recording audio broadcast function, the main

Kistot和Kimd分别为第i个音素的起始时间和结束时间； [0023] CMphonei表示第i个音素的确认得分，下标phonei表示第i个音素，Scorephonei如上 面所示第i个Phone的解码得分，Scoreframek表示使用唤醒词确认网络解码得到的第k帧的 得分； [0024] d.计算得到该唤醒词的最终确认得分，计算方式如下所示： 1 N . [0025] CMword=-YjCMphonm iv j=i [0026] 所述操作系统模块中的语音唤醒模块能够被移植到被移植到Windows、Linux、 2 CN Android、IOS系统平台。

[0027] 本发明与现有技术相比的优点在于： [0028] (1)本发明通过用户特定的语音唤醒词作为触发源，即使是在嘈杂的环境下，也无 需用户利用双手操作，仅通过语音命令快速实现唤醒功能，进行下一步交互操作。

[0029] (2)本发明带来的价值是，使用本系统后可通过说约定的语音唤醒词开启语音唤 醒功能，无需事先暂停音频播放，同时通过实际测试验证，正确识别唤醒率可以达到90%以 上。

[0030] (3)本发明实现，成本低，代码移植方便，具有很好的推广应用价值。

[0031] (4)本发明中的语音唤醒功能均通过软件算法实现，并且算法可以很方便的被移 植到 Windows、Linux、Android、IOS 等系统平台。

附图说明 [0032] 图1为本发明的结构示意图； [0033] 图2为本发明中的语音唤醒模块实现框图； [0034] 图3为本发明的构建唤醒词检测网络示意图； [0035] 图4为本发明的构建唤醒词确认网络示意图。

具体实施方式 [0036] 如图1所示，本发明的语音启动拍照系统包括：操作系统模块10、唤醒词定制模块 11、拍照模块12、音频模块13和存储模块14、操作系统模块10中运行语音唤醒模块。

[0037] 如图2所示，语音唤醒模块包括语音输入1、语音唤醒算法2和唤醒执行3。语音 唤醒算法2实现主要由声学特征提取4、唤醒词检测5、唤醒词确认6、构建唤醒词检测网络 7、训练声学模型8和构建唤醒词确认网络9完成，具体实现过程是： [0038] (1)训练声学模型8 ：声学模型的训练分为两部分，分别为音素声学模型和废料模 型（即Garbage模型)。音素声学模型采用传统的语音识别中的声学模型训练方法，选取合适 的数据库，利用基于MLE(Maximum Likelihood Estimation,最大似然估计)和MPE(Minimum Phone Error，最小音素错误）区分性训练准则下的得到。Garbage模型用于

function is to collect environmental record audio data, and transferred to the operating system module 10, the main function is to receive broadcast system to alert the feedback module 10, the broadcast beep function;

[0012] The storage module 14, with a permanent storage capacity of the storage medium, for photographing module generated storing the digital image information.

[0013] The acoustic model training 8 is divided into two parts, namely phoneme acoustic model and scrap models (ie Garbage model); phoneme acoustic models using the traditional speech recognition acoustic model training method, select the database, use based on the following MLE (Maximum Likelihood Estimation, maximum likelihood estimation) and MPE (Minimum Phone Error, the smallest phoneme error) discriminative training guidelines are; Garbage model-independent voice except for absorbing wake of the word, the use and training of phoneme models The same database, each phoneme model by calculating the similarity between, each phoneme is divided into 20 categories, each category using the corresponding phoneme integrate all training data, using the MLE criteria Garbage corresponding training model, it will have 20 categories Garbage Model.

[0014] wake up the word detection networks 7 uses the best path score is calculated, the formula for calculating the optimal score path is: [00151 W = argi grade IW) W [0016] wherein X represents extracted from the input speech acoustic feature vector, W on behalf of the score's largest optimal word sequence; conditional probability P (XlW) for the acoustic model score 8 by trained acoustic model is calculated; the prior probability P (W) is the language score model, is different acoustic models applied PenaltyP (X) for the whole probability, when the word acoustic model and wake finalized after detecting network is constant.

[0017] 9 Wake words the acknowledgment network implementations are:. [0018] a detection of the wake-up word decoding to phoneme level, and record all scores (Sc0rephmel, Scoreph0ne2,..., Scoreph.neN), where N is the total number of wake-up word phoneme, Scorephonel, Scorephone2,..., Scoreph. neN words denote the wake of all the phonemes are decoding score, where the subscript represents N