(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 113936647 B
(45)授权公告日 2022.04.01

(21)申请号 202111548060.3

(22)申请日 2021.12.17

(65)同一申请的已公布的文献号
申请公布号 CN 113936647 A

(43)申请公布日 2022.01.14

(73)专利权人 中国科学院自动化研究所
地址 100190 北京市海淀区中关村东路95号

(72)发明人 陶建华 田正坤 易江燕

(74)专利代理机构 北京华夏泰和知识产权代理有限公司 11662
代理人 李永叶

(51)Int.Cl.
G10L 15/06(2013.01)
G10L 15/22(2006.01)
G10L 15/26(2006.01)
G10L 19/16(2013.01)
G10L 25/03(2013.01)
G10L 25/24(2013.01)

(56)对比文件
CN 112185352 A,2021.01.05
CN 102968989 A,2013.03.13
US 2021233510 A1,2021.07.29
CN 112599122 A,2021.04.02
姚煜等.基于双向长短时记忆-联结时序分类和加权有限状态转换器的端到端中文语音识别系统.《计算机应用》.2018,(第09期),

审查员 王玥

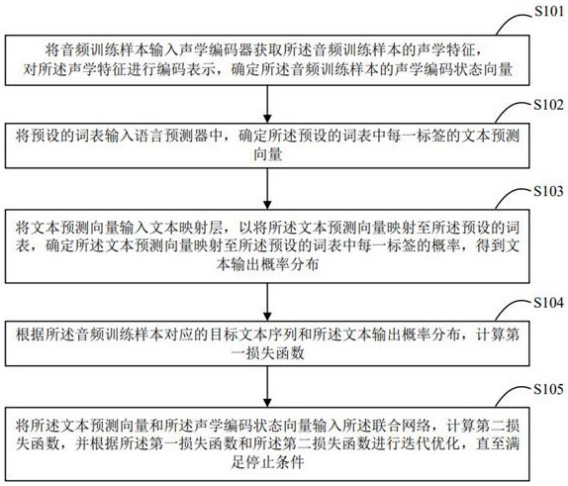
权利要求书2页 说明书8页 附图4页

(54)发明名称

语音识别模型的训练方法、语音识别方法和系统

(57)摘要

本发明实施例公开了一种语音识别模型的训练方法、语音识别方法和系统，涉及语音识别技术领域。该实施例包括：将音频训练样本输入声学编码器，对音频训练样本进行编码表示，确定声学编码状态向量；将预设的词表输入语言预测器中，确定文本预测向量；将文本预测向量线输入文本映射层，得到文本输出概率分布；根据音频训练样本对应的目标文本序列和文本输出概率分布，计算第一损失函数；将文本预测向量和声学编码状态向量输入联合网络，计算第二损失函数，根据第一损失函数和第二损失函数进行迭代优化，直至满足停止条件。本实施例对语音识别模型的训练、预测过程进行了调整，提高了该语义识别模型的建模能力，从而提高了该语音识别模型的准确率。



CN 113936647 B

1.一种语音识别模型的训练方法,其特征在于,所述语音识别模型包括声学编码器、语言预测器、文本映射层和联合网络,所述方法包括:

将音频训练样本输入所述声学编码器,以对所述音频训练样本进行编码表示,确定所述音频训练样本的声学编码状态向量;

将预设的词表输入所述语言预测器中,确定所述预设的词表中每一标签的文本预测向量;

将所述文本预测向量线输入所述文本映射层,确定所述文本预测向量映射至所述预设的词表中每一标签的概率,得到文本输出概率分布;

根据所述音频训练样本对应的目标文本序列和所述文本输出概率分布,计算第一损失函数;

将所述文本预测向量和所述声学编码状态向量输入所述联合网络,计算第二损失函数,并根据所述第一损失函数和所述第二损失函数进行迭代优化,直至满足停止条件,

其中,据所述第一损失函数和所述第二损失函数进行迭代优化包括:

根据所述第一损失函数和所述第二损失函数,确定第三损失函数;

根据所述第三损失函数进行迭代优化。

2.根据权利要求1所述的方法,其特征在于,所述第一损失函数为交叉熵损失函数,所述第二损失函数为Transducer损失函数。

3.根据权利要求2所述的方法,其特征在于,所述方法还包括根据下式确定第三损失函数:

$$L = (1-\alpha) L_{Transducer} + \alpha L_{Text}$$

其中,L表示第三损失函数,L_{Text}表示第一损失函数,$L_{Transducer}$表示第二损失函数,α表示预设的权重。

4.根据权利要求1所述的方法,其特征在于,将音频训练样本输入所述声学编码器,以对所述音频训练样本进行编码表示包括:

将音频训练样本输入所述声学编码器获取所述音频训练样本的声学特征,并对所述音频训练样本的声学特征进行编码表示。

5.一种语音识别方法,其特征在于,所述方法应用于权利要求1-4任一项所训练得到的语音识别模型,所述语音识别模型包括:声学编码器、语言预测器、文本映射层和联合网络;所述方法包括:

将待识别音频输入所述声学编码器进行编码表示,确定所述待识别音频的声学编码状态向量;

将预设的词表输入所述语言预测器,确定所述预设的词表中每一标签的文本预测向量;

将所述文本预测向量输入所述文本映射层,以将所述文本预测向量映射至所述预设的词表,确定所述文本预测向量映射至所述预设的词表中每一标签的第一概率;

将所述文本预测向量和所述声学编码状态向量输入所述联合网络,确定所述待识别音频映射至所述预设的词表中每一标签的第二概率;

根据所述第一概率和所述第二概率,确定所述待识别音频对应的文本内容。

6.根据权利要求5所述的方法，其特征在于，根据所述第一概率和所述第二概率，确定所述待识别音频对应的文本内容包括：

计算所述第一概率和所述第二概率的加权和；

将最大的所述加权和作为所述待识别音频对应的文本内容。

7.根据权利要求6所述的方法，其特征在于，所述方法还包括根据下式确定待识别音频对应的文本内容：

$$\text{Token}=\underset{i}{argmax}\{P_{transducer}(Token_i|A_t,T_u)+\beta P_{text}(Token_i|Token_{0,1,2,\ldots,u})\}$$

其中，Token表示待识别音频对应的文本内容，$P_{text}(Token_i|Token_{0,1,2,\ldots,u})$表示第一概率，$P_{transducer}(Token_i|A_t,T_u)$表示第二概率，$\beta$表示文本融合权重，$A_t$表示t时刻的声学编码状态向量，

T_u表示预设的词表中第u个标签。

8.根据权利要求5所述的方法，其特征在于，将待识别音频输入所述声学编码器进行编码表示包括：

将待识别音频输入所述声学编码器获取所述待识别音频的声学特征，并对所述音频训练样本的声学特征进行编码表示。

9.一种语音识别系统，其特征在于，所述语音识别系统包括声学编码器、语言预测器、文本映射层和联合网络；

其中，所述声学编码器用于对待识别音频进行编码表示，确定所述待识别音频的声学编码状态向量；

所述语言预测器用于确定预设的词表中每一标签的文本预测向量；

所述文本映射层用于将所述文本预测向量映射至所述预设的词表，确定所述文本预测向量映射至所述预设的词表中每一标签的第一概率；

所述联合网络用于根据所述文本预测向量和所述声学编码状态向量确定所述待识别音频映射至所述预设的词表中每一标签的第二概率，并根据所述第一概率和所述第二概率，确定所述待识别音频对应的文本内容。

10.一种电子设备，其特征在于，包括处理器、通信接口、存储器和通信总线，其中，处理器、通信接口和存储器通过通信总线完成相互间的通信；

所述存储器用于存放至少一可执行指令，所述可执行指令使得所述处理器执行权利要求1-4或权利要求5-8中任一项所述的方法。

11.一种计算机可读存储介质，其上存储有计算机程序，其特征在于，所述计算机程序被处理器执行时实现权利要求1-4或5-8中任一项所述的方法。

语音识别模型的训练方法、语音识别方法和系统

技术领域

[0001] 本申请涉及语音识别技术领域，尤其涉及一种语音识别模型的训练方法、语音识别方法和系统。

背景技术

[0002] 基于Transducer的语音识别模型在国内外获得了广泛的应用，其典型特点是能够直接适配流式语音识别任务。其虽然引入了语言预测器，但是其语言建模能力不足，经研究发现，语言预测器在真实推理中并没有起到类似语言模型的作用，而更多的承担了消除重复标签的功能，其建模语言之间依赖关系的能力还有进一步提升的空间。

发明内容

[0003] 为了解决上述技术问题或者至少部分地解决上述技术问题，本发明实施例提供一种语音识别模型的训练方法、语音识别方法、语音识别系统、电子设备和计算机可读存储介质。

[0004] 第一方面，本发明实施例提供了一种语音识别模型的训练方法，所述语音识别模型包括声学编码器、语言预测器、文本映射层和联合网络，所述方法包括：

[0005] 将音频训练样本输入所述声学编码器，以对所述音频训练样本进行编码表示，确定所述音频训练样本的声学编码状态向量；

[0006] 将预设的词表输入所述语言预测器中，确定所述预设的词表中每一标签的文本预测向量；

[0007] 将所述文本预测向量输入所述文本映射层，确定所述文本预测向量映射至所述预设的词表中每一标签的概率，得到文本输出概率分布；

[0008] 根据所述音频训练样本对应的目标文本序列和所述文本输出概率分布，计算第一损失函数；

[0009] 将所述文本预测向量和所述声学编码状态向量输入所述联合网络，计算第二损失函数，并根据所述第一损失函数和所述第二损失函数进行迭代优化，直至满足停止条件。

[0010] 在可选的实施例中，据所述第一损失函数和所述第二损失函数进行迭代优化包括：根据所述第一损失函数和所述第二损失函数，确定第三损失函数；根据所述第三损失函数进行迭代优化。

[0011] 在可选的实施例中，所述第一损失函数为交叉熵损失函数，所述第二损失函数为Transducer损失函数。

[0012] 在可选的实施例中，所述方法还包括根据下式确定第三损失函数：

[0013] $$L = (1-\alpha) L_{Transducer} + \alpha L_{Text}$$

[0014] 其中，L表示第三损失函数，L_{Text}表示第一损失函数，$L_{Transducer}$表示第二损失函数，α表示预设的权重。

[0015] 在可选的实施例中，将音频训练样本输入所述声学编码器，以对所述音频训练样

本进行编码表示包括:将音频训练样本输入所述声学编码器获取所述音频训练样本的声学特征,并对所述音频训练样本的声学特征进行编码表示。

[0016] 第二方面,本发明实施例提供了一种语音识别方法,所述方法应用于上述实施例所训练得到的语音识别模型,所述语音识别模型包括:声学编码器、语言预测器、文本映射层和联合网络;所述方法包括:

[0017] 将待识别音频输入所述声学编码器进行编码表示,确定所述待识别音频的声学编码状态向量;

[0018] 将预设的词表输入所述语言预测器,确定所述预设的词表中每一标签的文本预测向量;

[0019] 将所述文本预测向量输入所述文本映射层,以将所述文本预测向量映射至所述预设的词表,确定所述文本预测向量映射至所述预设的词表中每一标签的第一概率;

[0020] 将所述文本预测向量和所述声学编码状态向量输入所述联合网络,确定所述待识别音频映射至所述预设的词表中每一标签的第二概率;

[0021] 根据所述第一概率和所述第二概率,确定所述待识别音频对应的文本内容。

[0022] 在可选的实施例中,根据所述第一概率和所述第二概率,确定所述待识别音频对应的文本内容包括:计算所述第一概率和所述第二概率的加权和;将最大的所述加权和作为所述待识别音频对应的文本内容。

[0023] 在可选的实施例中,所述方法还包括根据下式确定待识别音频对应的文本内容:

[0024] $\text{Token}=arg\max_i\{P_{transducer}(Token_i|A_t,T_u)+\beta P_{text}(Token_i|Token_{0,1,2,\ldots,u})\}$ 其中,Token表示待识别音频对应的文本内容,$P_{text}(Token_i|Token_{0,1,2,\ldots,u})$ 表示第一概率,$P_{transducer}(Token_i|A_t,T_u)$ 表示第二概率,β 表示文本融合权重,A_t 表示t时刻的声学编码状态向量,T_u 表示预设的词表中第u个标签。

[0025] 在可选的实施例中,将待识别音频输入所述声学编码器进行编码表示包括:将待识别音频输入所述声学编码器获取所述待识别音频的声学特征,并对所述音频训练样本的声学特征进行编码表示。

[0026] 第三方面,本发明实施例还提供了一种语音识别系统,所述语音识别系统包括声学编码器、语言预测器、文本映射层和联合网络;

[0027] 其中,所述声学编码器用于对待识别音频进行编码表示,确定所述待识别音频的声学编码状态向量;

[0028] 所述语言预测器用于确定预设的词表中每一标签的文本预测向量;

[0029] 所述文本映射层用于将所述文本预测向量映射至所述预设的词表,确定所述文本预测向量映射至所述预设的词表中每一标签的第一概率;

[0030] 所述联合网络用于根据所述文本预测向量和所述声学编码状态向量确定所述待识别音频映射至所述预设的词表中每一标签的第二概率;并根据所述第一概率和所述第二概率,确定所述待识别音频对应的文本内容。

[0031] 第四方面,本发明实施例还提供了一种电子设备,包括处理器、通信接口、存储器和通信总线,其中,处理器、通信接口和存储器通过通信总线完成相互间的通信;所述存储

器用于存放至少一可执行指令，所述可执行指令使得所述处理器执行本发明实施例的语音识别模型训练方法或语音识别方法。

[0032] 第五方面，本发明实施例还提供了一种计算机可读存储介质，其上存储有计算机程序，所述计算机程序被处理器执行时实现本发明实施例的语音识别模型训练方法或语音识别方法。

[0033] 上述实施例中的一个或多个技术方案至少具有如下优点的部分或全部：

[0034] 本发明实施例的语音识别模型加入了文本映射层，并对语音识别模型的训练过程和预测过程进行了调整，提高了该语义识别模型的建模能力，从而提高了该语音识别模型的准确率。

附图说明

[0035] 此处的附图被并入说明书中并构成本说明书的一部分，示出了符合本发明的实施例，并与说明书一起用于解释本发明的原理。

[0036] 为了更清楚地说明本发明实施例或现有技术中的技术方案，下面将对实施例或相关技术描述中所需要使用的附图作简单地介绍，显而易见地，对于本领域普通技术人员而言，在不付出创造性劳动性的前提下，还可以根据这些附图获得其他的附图。

[0037] 图1示意性地示出了本发明实施例的语音识别模型训练方法的主要步骤的流程图；

[0038] 图2示意性地示出了本发明实施例的语音识别模型训练方法得到的语音识别模型的结构图；

[0039] 图3示意性地示出了本发明实施例的语音识别方法的主要步骤的流程图；

[0040] 图4示意性地示出了适用于本发明实施例的语音识别模型训练方法或语音识别方法的系统架构；

[0041] 图5示意性示出了本发明实施例提供的电子设备的结构框图。

具体实施方式

[0042] 为使本发明实施例的目的、技术方案和优点更加清楚，下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本发明的一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有做出创造性劳动的前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0043] 基于Transducer语音识别模型在国内外获得了广泛的应用。该模型通常包含三部分，分别是声学编码器，语言预测器和联合网络。声学编码器负责将输入声学特征编码为声学编码状态向量，语言预测器输入为预设的词表(该预测的词表包括空格标签或者先前预测得到的文本标签)，输出为当前时刻预测的文本预测状态向量，联合网络输入为当前时刻声学编码器输出的声学编码状态向量和语言预测器输出的文本预测状态向量，输出得到词表中所有标签的概率。该模型虽然引入了语言预测器，但是其语言建模能力不足，经研究发现，语言预测器在真实推理中并没有起到类似语言模型的作用，而更多的承担了消除重复标签的作用。针对该技术问题，常见的方法一般是在解码时添加辅助语言模型，这种方法虽然能提升语音识别系统的准确率，但是对于模型本身语言建模能力的提升没有帮助。为了

解决该技术问题，本发明实施例提供了一种语音识别模型的训练方法以及语音识别方法，该方法对Transducer语音识别模型的训练和解码过程进行了调整，以通过提高其语言建模能力来提升模型语音识别的准确率。

[0044] 为方便理解本发明实施例的语音识别模型的训练方法，下面结合附图对其进行说明。

[0045] 图1示意性地示出了本发明实施例的语音识别模型的训练方法的主要步骤的示意图。图2示意性地示出了本发明实施例所训练得到的语音识别模型的结构示意图。如图2所示，该语音识别模型200包括：声学编码器201、语言预测器202、文本映射层203和联合网络204。

[0046] 如图1所示，该语音识别模型的训练方法包括：

[0047] 步骤S101：将音频训练样本输入所述声学编码器获取所述音频训练样本的声学特征，对所述音频训练样本的声学特征进行编码表示，确定所述音频训练样本的声学编码状态向量；

[0048] 步骤S102：将预设的词表输入所述语言预测器中，确定所述预设的词表中每一标签的文本预测向量；

[0049] 步骤S103：将所述文本预测向量输入所述文本映射层，以将所述文本预测向量映射至所述预设的词表，确定所述文本预测向量映射至所述预设的词表中每一标签的概率，得到文本输出概率分布；

[0050] 步骤S104：根据所述音频训练样本对应的目标文本序列和所述文本输出概率分布，计算第一损失函数；

[0051] 步骤S105：将所述文本预测向量和所述声学编码状态向量输入所述联合网络，计算第二损失函数，并根据所述第一损失函数和所述第二损失函数进行迭代优化，直至满足停止条件。

[0052] 对于步骤S101，在本实施例中，声学特征例如可以是MFCC特征，也可以是FBank特征等。MFCC(Mel-Frequency Cepstral Coefficients，梅尔倒谱系数)和FBank(Filterbank，滤波器组特征)特征是语音识别常用的一种特征。在获得音频训练样本的声学特征之后，对音频训练样本的声学特征进行编码表示，获得音频训练样本的声学编码状态向量。结合图2，可以将音频训练样本输入声学编码器201中，获得该音频训练样本的声学特征，并对该音频训练样本的声学特征进行编码表示，确定该音频训练样本的声学编码状态向量。作为示例，该声学编码器201可以采用基于循环神经网络、卷积神经网络或者是Transformer模型以及这些模型的变体结构或者组合结构构成。

[0053] 结合图2，对于步骤S102-S104，将预设的词表(该词表中包括空格标签和非空格标签，非空格标签包括统计后的常用词语)输入语言预测202中，计算得到文本预测向量。在得到文本预测向量之后，将该文本预测向量输入文本映射层203，该文本映射层203仅包括一层线性映射，将输入的文本预测向量映射至上述预设的词表，并计算得到文本预测向量映射至词表中每一标签的概率，得到文本输出概率分布。然后，根据音频训练样本对应的目标文本序列以及该文本输出概率分布，计算第一损失函数。具体的，该过程包括：确定音频训练样本对应的目标文本序列在上述预设的词表中的索引，根据该索引，确定与该目标文本序列对应的第一概率。对于第一损失函数，作为示例该第一损失函数可以是交叉熵损失函

数。

[0054] 对于步骤S105,将上述文本预测向量和上述声学编码状态向量输入联合网络204,计算得到音频训练样本映射至上述词表中每一标签的第二概率,并基于该第二概率计算第二损失函数。作为示例,该第二损失函数可以是Transducer损失函数。其中,Transducer损失函数是一种用于基于Transducer的语音识别模型计算的负对数损失函数,其利用基于动态规划思路的前后向算法对所有可行的音频-输出标签对齐路径进行概率求和,并对概率和的负对数形式进行优化。在得到第二损失函数之后,对第一损失函数和第二损失函数进行加权求和,并进行联合优化迭代,直至达到停止条件如模型收敛,至此得到语音识别模型。其中,对第一损失函数和第二损失函数进行加权求和,并进行联合优化迭代包括根据所述第一损失函数和所述第二损失函数,确定第三损失函数;根据所述第三损失函数进行迭代优化。

[0055] 更具体的,可以根据下式确定第三损失函数:

[0056] $$L = (1-\alpha) L_{Transducer} + \alpha L_{Text}$$

[0057] 其中,L表示第三损失函数,L_{Text}表示第一损失函数,$L_{Transducer}$表示第二损失函数,α表示预设的权重。

[0058] 本发明实施例的语音识别模型训练方法,对语音识别模型的训练过程进行了调整,提高了该语义识别模型的建模能力,从而提高了该语音识别模型的准确率。

[0059] 图3示意性地示出了本发明实施例的语音识别方法的主要步骤的示意图。该语音识别方法可以应用于图1所示的实施例训练得到的语音识别模型。

[0060] 如图3所示,该方法包括:

[0061] 步骤S301:将待识别音频输入所述声学编码器获取所述待识别音频的声学特征,并对所述音频训练样本的声学特征进行编码表示;

[0062] 步骤S302:将预设的词表输入所述语言预测器,确定所述预设的词表中每一标签的文本预测向量;

[0063] 步骤S303:将所述文本预测向量输入所述文本映射层,以将所述文本预测向量映射至所述预设的词表,确定所述文本预测向量映射至所述预设的词表中每一标签的第一概率;

[0064] 步骤S304:将所述文本预测向量和所述声学编码状态向量输入所述联合网络,确定所述待识别音频映射至所述预设的词表中每一标签的第二概率;

[0065] 步骤S305:根据所述第一概率和所述第二概率,确定所述待识别音频对应的文本内容。

[0066] 在本实施例中,将待识别的音频输入声学编码器中,获取该待识别音频的声学特征,例如可以是MFCC特征,也可以是FBank特征,并对该待识别音频的声学特征进行编码表示,得到声学编码状态向量A_t,其中,t表示t时刻。然后将预设的词表中的空格标签或非空格标签输入语言预测器中,计算得到每一标签的文本预测向量T_u,u表示第u个标签。将计算得到的文本预测向量输入文本映射层,将该文本预测向量映射至所述预设的词表,确定文本预测向量映射至所述预设的词表中每一标签的第一概率,从而得到文本输出概率分

布，其中，映射到词表中第u+1个标签 $Token_{u+1}$ 的第一概率为 $P_{text}(Token_i|Token_{0,1,2,\ldots,u})$。然后，将文本预测向量和声学编码状态向量输入至联合网络，计算得到待识别音频映射至预设的词表中每一标签的第二概率 $P_{transducer}(Token_i|A_t, T_u)$。最后，根据上述第一概率和上述第二概率，确定待识别音频对应的文本内容。若联合网络根据第一概率和第二概率的加权和，预测得到空格标签(即预测的待识别音频为空格标签)，则保持联合网络输入的文本预测向量不变，更新下一个声学编码状态向量，如果预测得非空格标签，则保持声学编码状态向量不变，更新文本预测向量。重复上述步骤，直至语音识别模型在基于最后一个声学编码状态向量预测得到空格标签或者提前达到其他停止条件。在可选的实施例中，根据第一概率和第二概率，确定待识别音频对应的文本内容的步骤包括：计算所述第一概率和所述第二概率的加权和；将最大的所述加权和作为所述待识别音频对应的文本内容。更具体的，该步骤根据下式确定第一概率和第二概率的加权和，以及确定待识别音频对应的文本内容：$\text{Token}=\underset{i}{argmax}\{P_{transducer}(Token_i|A_t, T_u)+\beta P_{text}(Token_i|Token_{0,1,2,\ldots,u})\}$

[0067] 其中，Token表示待识别音频对应的文本内容，$P_{text}(Token_i|Token_{0,1,2,\ldots,u})$ 表示第一概率，$P_{transducer}(Token_i|A_t, T_u)$ 表示第二概率，β 表示文本融合权重，A_t 表示t时刻的声学编码状态向量，T_u 表示预设的词表中第u个标签。在可选的实施例中，β 的典型值为0.1。

[0068] 本发明实施例的语音识别过程对语音识别模型的预测过程进行了调整，提高了该语义识别模型的建模能力，从而提高了该语音识别模型的准确率。

[0069] 图4示意性地示出了适用于本发明实施例的语音识别模型的训练方法和语音识别方法的系统架构。

[0070] 如图4所示，适用于本发明实施例的语音识别模型的训练方法和语音识别方法的系统架构400包括：终端设备401、402、403，网络404和服务器405。网络404用以在终端设备401、402、403和服务器405之间提供通信链路的介质。网络404可以包括各种连接类型，例如有线、无线通信链路或者光纤电缆等。

[0071] 终端设备401、402、403通过网络404与服务器405交互，以接收或发送消息等。终端设备401、402、403上可以安装有各种通讯客户端应用。终端设备401、402、403可以是具有数据采集功能例如音频采集功能的电子设备。

[0072] 服务器405可以是提供各种服务的服务器。服务器可以对接收到的请求或消息进行分析和处理，并将数据处理后得到的结果反馈给终端设备。

[0073] 需要说明的是，本发明实施例所提供的语音识别模型的训练方法和语音识别方法一般可以由服务器405执行。本发明实施例所提供的语音识别模型的训练方法和语音识别方法也可以由不同于服务器405且能够与终端设备401、402、403和/或服务器405通信的服务器或服务器集群执行。

[0074] 应该理解的是，图4中的终端设备、网络和服务器的数目仅仅是示意性的。根据实现需要，可以具有任意数目的终端设备、网络和服务器。

[0075] 图5示意性示出了本发明一实施例的电子设备的示意图。如图5所示，本发明实施例提供的电子设备500包括处理器501、通信接口502、存储器503和通信总线504，其中，处理器501、通信接口502和存储器503通过通信总线504完成相互间的通信；存储器503，用于存放至少一可执行指令；处理器501，用于执行存储器上所存放的可执行指令时，实现如上所述的语音识别模型的训练方法和语音识别方法。

[0076] 具体而言，当实现上述语音识别模型的训练方法时，上述可执行指令使得上述处理器执行以下步骤：将音频训练样本输入所述声学编码器，以对所述音频训练样本进行编码表示，确定所述音频训练样本的声学编码状态向量；将预设的词表输入所述语言预测器中，确定所述预设的词表中每一标签的文本预测向量；将所述文本预测向量线输入所述文本映射层，确定所述文本预测向量映射至所述预设的词表中每一标签的概率，得到文本输出概率分布；根据所述音频训练样本对应的目标文本序列和所述文本输出概率分布，计算第一损失函数；将所述文本预测向量和所述声学编码状态向量输入所述联合网络，计算第二损失函数，并根据所述第一损失函数和所述第二损失函数进行迭代优化，直至满足停止条件。

[0077] 当实现上述语音识别方法时，上述可执行指令使得上述处理器执行以下步骤：将待识别音频输入所述声学编码器进行编码表示，确定所述待识别音频的声学编码状态向量；将预设的词表输入所述语言预测器，确定所述预设的词表中每一标签的文本预测向量；将所述文本预测向量输入所述文本映射层，以将所述文本预测向量映射至所述预设的词表，确定所述文本预测向量映射至所述预设的词表中每一标签的第一概率；将所述文本预测向量和所述声学编码状态向量输入所述联合网络，确定所述待识别音频映射至所述预设的词表中每一标签的第二概率；根据所述第一概率和所述第二概率，确定所述待识别音频对应的文本内容。

[0078] 上述存储器503可以是诸如闪存、EEPROM(电可擦除可编程只读存储器)、EPROM、硬盘或者ROM之类的电子存储器。存储器503具有用于执行上述方法中的任何方法步骤的程序代码的存储空间。例如，用于程序代码的存储空间可以包括分别用于实现上面的方法中的各个步骤的各个程序代码。这些程序代码可以从一个或者多个计算机程序产品中读出或者写入到这一个或者多个计算机程序产品中。这些计算机程序产品包括诸如硬盘，光盘(CD)、存储卡或者软盘之类的程序代码载体。这样的计算机程序产品通常为便携式或者固定存储单元。该存储单元可以具有与上述电子设备中的存储器503类似布置的存储段或者存储空间等。程序代码可以例如以适当形式进行压缩。通常，存储单元包括用于执行根据本发明的实施例的方法步骤的程序，即可以由例如诸如501之类的处理器读取的代码，这些代码当由电子设备运行时，导致该电子设备执行上面所描述的方法中的各个步骤。

[0079] 本发明实施例还提供了一种计算机可读存储介质。上述计算机可读存储介质上存储有计算机程序，上述计算机程序被处理器执行时实现如上所述的语音识别模型的训练方法和语音识别方法。

[0080] 该计算机可读存储介质可以是上述实施例中描述的设备/装置中所包含的；也可以是单独存在，而未装配入该设备/装置中。上述计算机可读存储介质承载有一个或者多个程序，当上述一个或者多个程序被执行时，实现根据本发明实施例的方法。

[0081] 根据本发明的实施例，计算机可读存储介质可以是非易失性的计算机可读存储介

质，例如可以包括但不限于：便携式计算机磁盘、硬盘、随机访问存储器(RAM)、只读存储器(ROM)、可擦式可编程只读存储器(EPROM或闪存)、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。在本发明中，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。

[0082] 本发明的实施例提供的上述各个技术方案可以全部或部分步骤以硬件实现，或者以在一个或者多个处理器上运行的软件模块实现，或者以它们的组合实现。本领域的技术人员应当理解，可以在实践中使用微处理器或者数字信号处理器(DSP)来实现根据本发明的实施例的电子设备中的一些或者全部部件的一些或者全部功能。本发明的实施例还可以实现为用于执行这里所描述的方法的一部分或者全部的设备或者装置程序(例如，计算机程序和计算机程序产品)。实现本发明的实施例的程序可以存储在计算机可读介质上，或者可以具有一个或者多个信号的形式。这样的信号可以从因特网网站上下载得到，或者在载体信号上提供，或者以任何其他形式提供。

[0083] 需要说明的是，在本文中，诸如“第一”和“第二”等之类的关系术语仅仅用来将一个实体或者步骤与另一个实体或步骤区分开来，而不一定要求或者暗示这些实体或步骤之间存在任何这种实际的关系或者顺序。而且，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、物品或者设备不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、物品或者设备所固有的要素。在没有更多限制的情况下，由语句“包括一……”限定的要素，并不排除在包括所述要素的过程、方法、物品或者设备中还存在另外的相同要素。

[0084] 以上所述仅是本发明的具体实施方式，使本领域技术人员能够理解或实现本发明。对这些实施例的多种修改对本领域的技术人员来说将是显而易见的，本文中所定义的一般原理可以在不脱离本发明的精神或范围的情况下，在其它实施例中实现。因此，本发明将不会被限制于本文所示的这些实施例，而是要符合与本文所申请的原理和新颖特点相一致的最宽的范围。

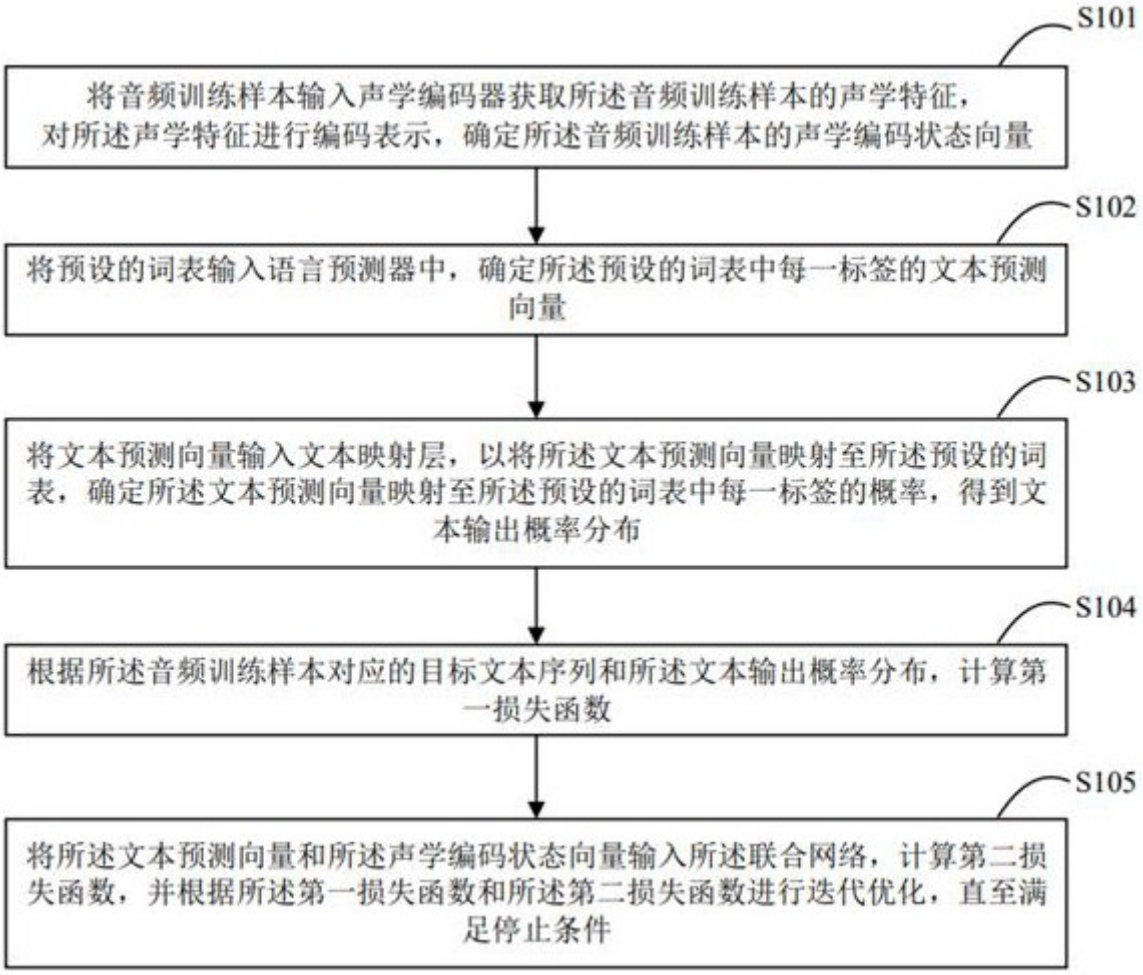

S101
将音频训练样本输入声学编码器获取所述音频训练样本的声学特征，对所述声学特征进行编码表示，确定所述音频训练样本的声学编码状态向量
S102
将预设的词表输入语言预测器中，确定所述预设的词表中每一标签的文本预测向量
S103
将文本预测向量输入文本映射层，以将所述文本预测向量映射至所述预设的词表，确定所述文本预测向量映射至所述预设的词表中每一标签的概率，得到文本输出概率分布
S104
根据所述音频训练样本对应的目标文本序列和所述文本输出概率分布，计算第一损失函数
S105
将所述文本预测向量和所述声学编码状态向量输入所述联合网络，计算第二损失函数，并根据所述第一损失函数和所述第二损失函数进行迭代优化，直至满足停止条件

图1

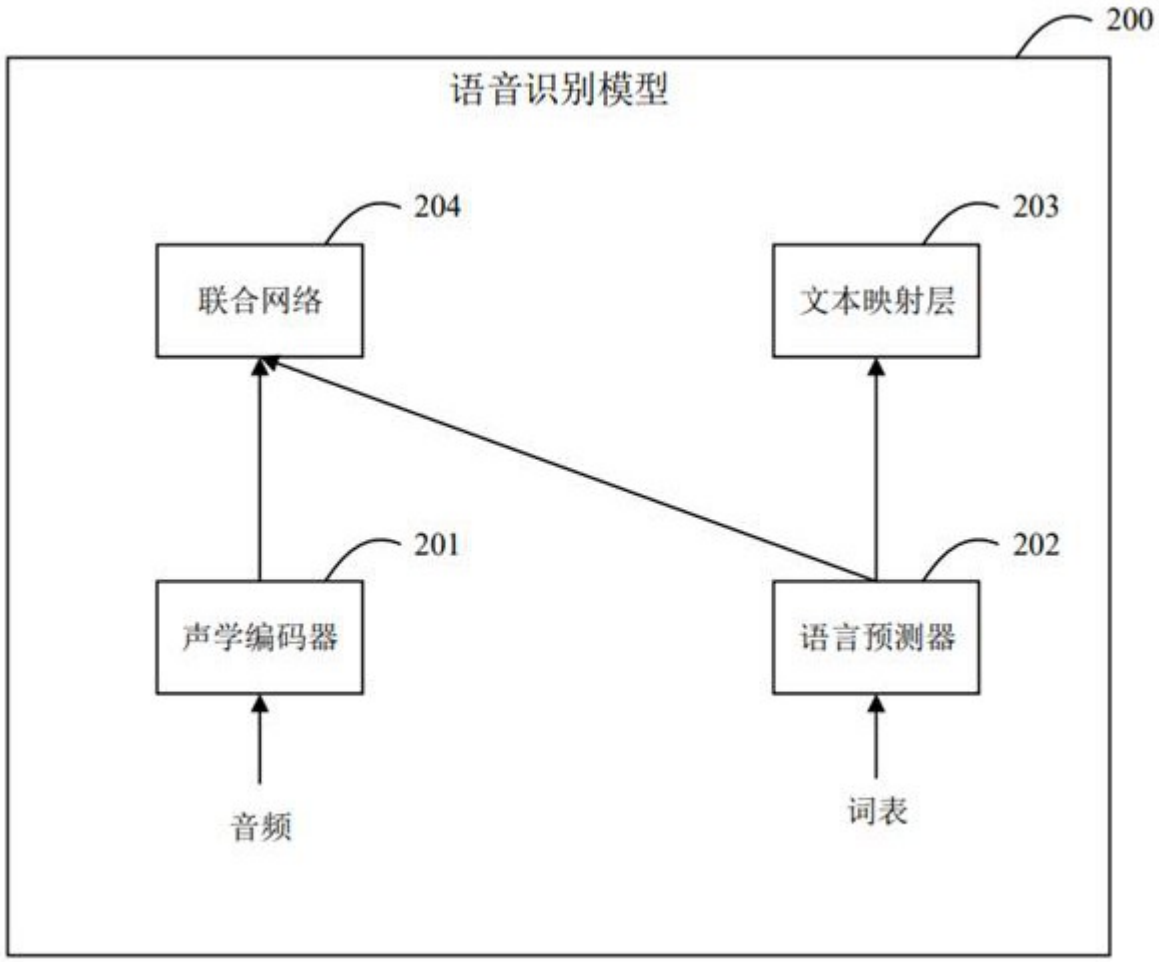


图2

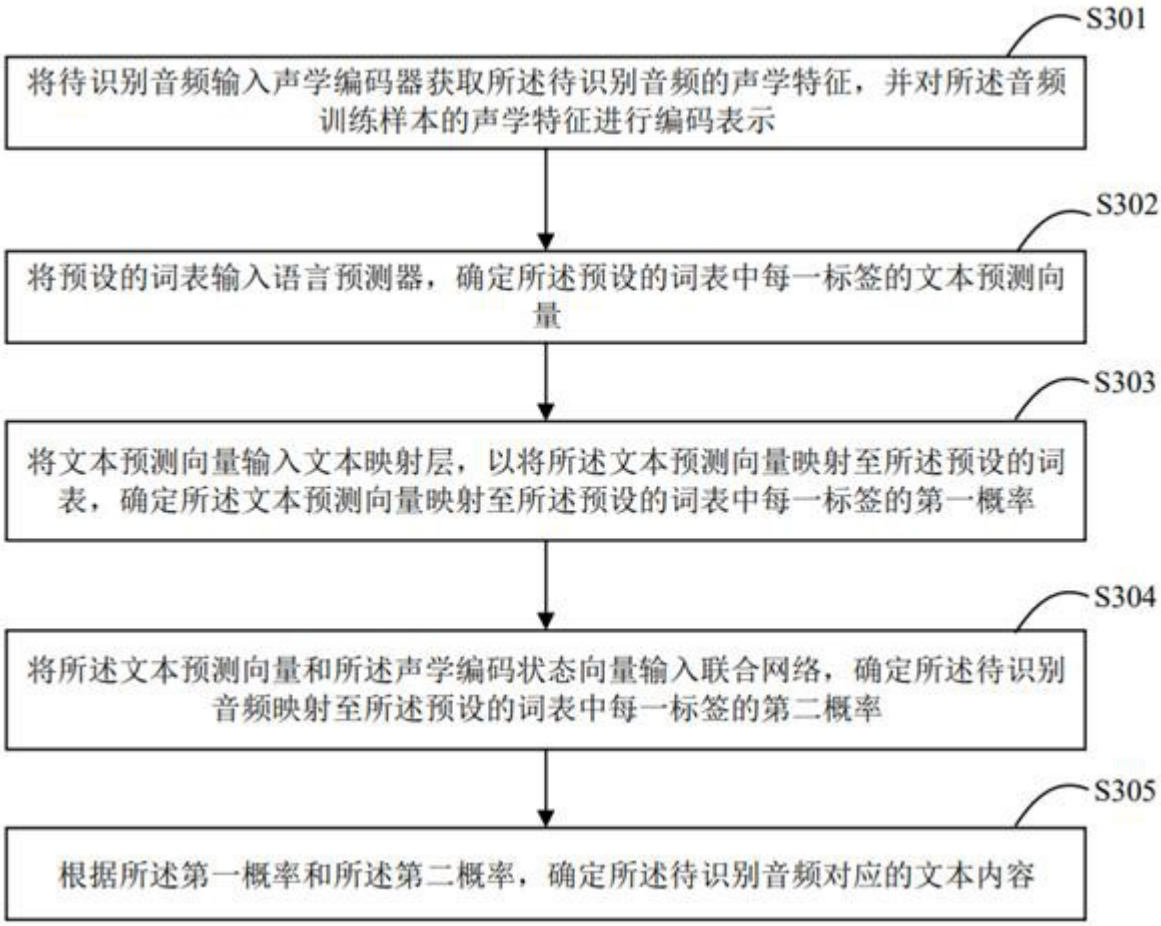

S301
将待识别音频输入声学编码器获取所述待识别音频的声学特征，并对所述音频训练样本的声学特征进行编码表示
S302
将预设的词表输入语言预测器，确定所述预设的词表中每一标签的文本预测向量
S303
将文本预测向量输入文本映射层，以将所述文本预测向量映射至所述预设的词表，确定所述文本预测向量映射至所述预设的词表中每一标签的第一概率
S304
将所述文本预测向量和所述声学编码状态向量输入联合网络，确定所述待识别音频映射至所述预设的词表中每一标签的第二概率
S305
根据所述第一概率和所述第二概率，确定所述待识别音频对应的文本内容

图3

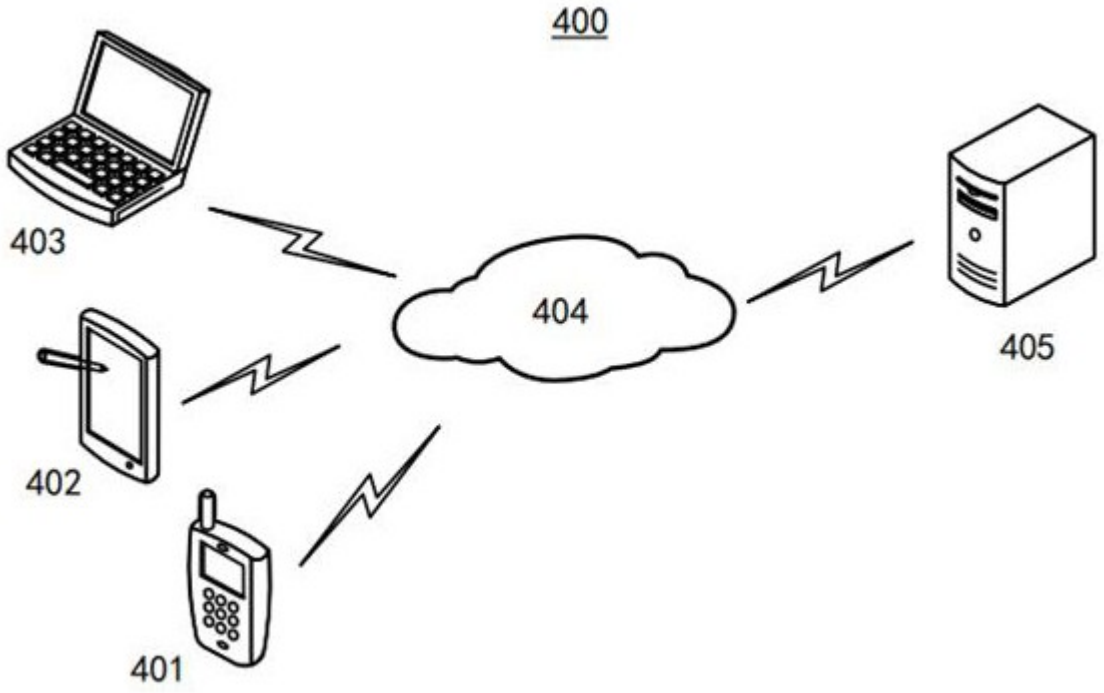


图4

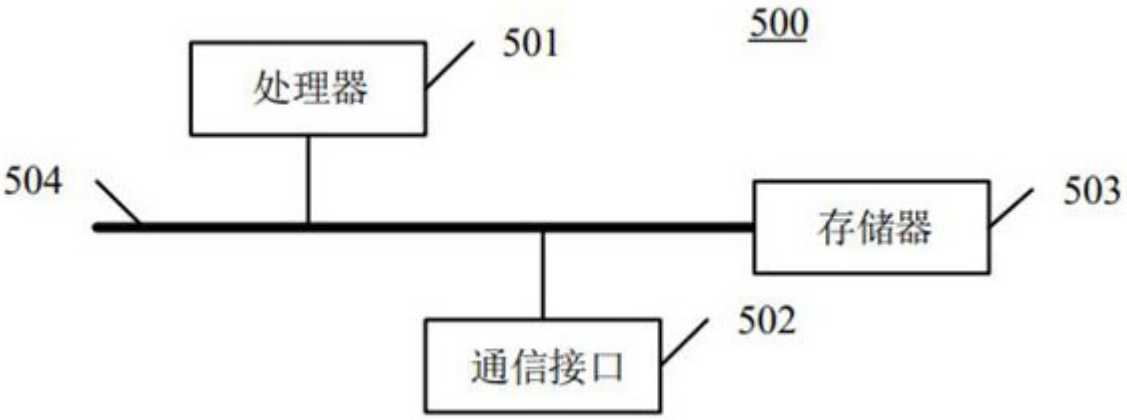


图5