(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 113646835 A
(43)申请公布日 2021.11.12

(21)申请号 202080024957.3

(22)申请日 2020.04.06

(30)优先权数据
62/830,306 2019.04.05 US

(85)PCT国际申请进入国家阶段日
2021.09.27

(86)PCT国际申请的申请数据
PCT/US2020/026937 2020.04.06

(87)PCT国际申请的公布数据
WO2020/206455 EN 2020.10.08

(71)申请人 谷歌有限责任公司
地址 美国加利福尼亚州

(72)发明人 劳伦特·艾谢弗 哈根·索尔陶 伊扎克·沙弗兰

(74)专利代理机构 中原信达知识产权代理有限责任公司 11219
代理人 李佳 周亚荣

(51)Int.Cl.
G10L 17/18(2006.01)
G10L 15/16(2006.01)
G06N 3/04(2006.01)

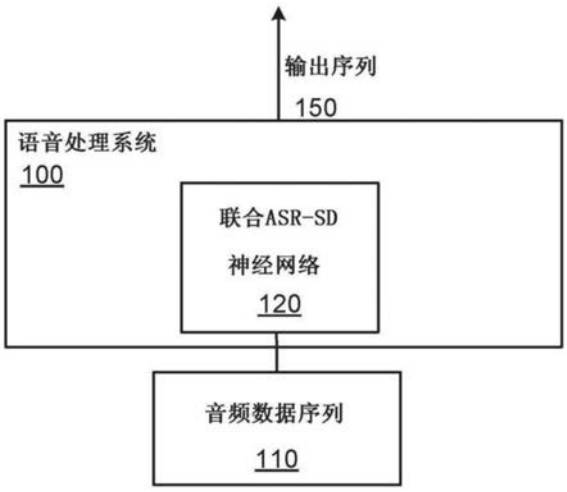
权利要求书2页 说明书9页 附图4页

(54)发明名称

联合自动语音识别和说话人二值化

(57)摘要

用于使用神经网络处理音频数据的方法、系统和装置，包括在计算机存储介质上编码的计算机程序。



CN 113646835 A

1.一种计算机实现的方法,包括:

获得表征音频片段的音频片段序列,所述音频片段序列包括多个音频帧;

使用联合自动语音识别-说话人二值化ASR-SD神经网络将所述音频片段序列映射到输出序列,所述输出序列包括多个时间步中的每一个的相应输出符号,其中,对于所述时间步中的每一个,所述输出序列中的所述时间步的输出符号是从输出符号集合中被选择的,所述输出符号集合包括(i)多个文本符号、(ii)多个说话人标签符号,每个说话人标签符号标识来自可能的说话人集合中的不同说话人、以及(iii)空白符号;以及

从所述输出序列中确定所述音频片段数据的转录,所述音频片段数据的转录(i)标识在所述音频片段中说出的单词,以及(ii)针对说出的单词中的每一个标识来自说出所述单词的可能的说话人集合中的所述说话人。

2.根据权利要求1所述的方法,其中,所述联合ASR-SD神经网络包括转录神经网络,并且其中,映射所述音频片段序列包括:

使用所述转录神经网络来处理所述音频片段序列,其中,所述转录神经网络被配置为处理所述音频片段数据以生成所述多个时间步中的每一个的相应的编码表示。

3.根据权利要求2所述的方法,其中,所述联合ASR-SD神经网络进一步包括预测神经网络,并且其中,映射所述音频片段序列包括,对于每个时间步:

标识所述时间步的当前输出符号,以及

使用所述预测神经网络来处理所述时间步的所述当前输出符号,其中,所述预测神经网络被配置为处理所述当前输出符号,以生成以已经被包括在所述输出序列中的任何更早的时间步处的任何非空白输出符号为条件的所述时间步的预测表示。

4.根据权利要求3所述的方法,其中,所述联合ASR-SD神经网络包括联合神经网络和softmax输出层,并且其中,映射所述音频片段序列包括,对于每个时间步:

处理所述时间步的编码表示和所述时间步的预测表示,以生成所述输出符号集合中的每个输出符号的相应的logit;和

使用所述softmax输出层来处理所述输出符号的logit,以生成所述输出符号集合中的所述输出符号上的概率分布。

5.根据权利要求4所述的方法,其中,映射所述音频片段序列包括,对于每个时间步:

使用所述概率分布从所述输出符号集合中选择输出符号。

6.根据任一前述权利要求所述的方法,其中,所述文本符号表示音素、语素或字符。

7.根据任一前述权利要求所述的方法,其中,从所述输出序列中确定所述音频片段数据的转录、所述音频片段数据的转录(i)标识在所述音频片段中说出的单词以及(ii)针对说出的单词中的每一个标识来自说出所述单词的可能的说话人集合中的所述说话人包括:

标识由所述输出序列中的所述文本符号表示的单词;和

对于每个所标识的单词,将所述单词标识为由紧接在表示所述输出序列中的所述单词的所述文本符号之后的所述说话人标签符号所表示的所述说话人说出。

8.根据任一前述权利要求所述的方法,其中,所述可能的说话人集合是对话中的可能的说话角色的集合,并且其中,每个说话人标签符号标识来自所述多个可能的说话角色的不同说话角色。

9.一种系统,包括一个或多个计算机和存储指令的一个或多个存储设备,所述指令当

由所述一个或多个计算机执行时使所述一个或多个计算机执行任何一项前述权利要求所述的相应方法的操作。

10.一种利用指令编码的计算机存储介质,所述指令当由一个或多个计算机执行时使所述一个或多个计算机执行任一前述权利要求所述的操作。

联合自动语音识别和说话人二值化

[0001] 相关申请的交叉引用

[0002] 本申请要求2019年4月5日提交的美国专利申请No.62/830,306的优先权,其全部内容通过引用并入本文。

背景技术

[0003] 本说明书涉及执行语音识别和说话人二值化(speaker diarization)的神经网络。

[0004] 神经网络是机器学习模型,其采用一层或多层非线性单元来预测接收到的输入的输出。一些神经网络除输出层外还包括一个或多个隐藏层。每个隐藏层的输出被用作网络中下一层,即,下一隐藏层或输出层的输入。网络的每一层根据相应的参数集合的当前值从接收的输入生成输出。

[0005] 一些神经网络是递归神经网络。递归神经网络是一种接收输入序列并从输入序列生成输出序列的神经网络。特别地,递归神经网络可以在当前时间步计算输出时使用来自先前时间步的部分或全部网络内部状态。递归神经网络的示例是长短期(LSTM)神经网络,其包括一个或多个LSTM存储块。每个LSTM存储块能够包括一个或多个单元,每个单元包括输入门、忘记门和输出门,允许该单元存储该单元的先前状态,例如,用于生成当前激活或要被提供给LSTM神经网络的其他组件。

发明内容

[0006] 本说明书描述一种作为计算机程序实现在一个或多个位置中的一个或多个计算机上的系统,该系统生成音频数据的转录。特别地,由该系统生成的转录标识给定音频分段中说出的单词,并且对于说出的单词中的每一个,标识说出了该单词的说话人。对说话人的标识能够是从可能的说话人角色的集合中标识说话人在对话中的说话人角色或从可能的唯一说话人的集合中标识唯一说话人。

[0007] 本说明书中描述的主题的特定实施例能够被实现以便实现以下优点中的一个或多个。

[0008] 需要既识别音频分段中说出的单词又标识单词的说话人的常规系统组合分离的ASR和SD系统的输出,该分离的ASR和SD系统各自对声学数据(即,音频帧)进行操作并且被独立地训练。在推理时,即,在分别训练每个系统之后组合两个这种系统出于若干原因而产生不佳的输出。特别地,可能难以跨时间准确地对准ASR和SD系统的输出,由于SD系统不被约束遵守由ASR系统生成的输出中的单词边界(即,因为SD系统也对仅声学数据进行操作)。另一方面,描述的系统生成既转录音频中的单词又标识说出的单词中的每一个的说话人的输出序列。在这样做时,SD输出遵守单词边界,因为神经网络通过训练学习不在说出的单词中间输出说话人身份标签。另外,描述的系统能够在任何给定时间步生成以来自输入音频数据的声学线索和来自已经识别的语音的语言线索两者为条件的SD输出。通过并入这些另外的语言线索并且通过将神经网络配置为在生成SD输出时固有地遵守单词边界,该系统能

够生成高质量的SD输出，即，比独立于ASR过程操作的常规系统更高质量的SD输出。另外，组合声学线索和语言线索(说出的单词)的现有尝试尚未成功地改善二值化。然而，描述的技术有效地组合这些线索以生成高质量的说话人二值化结果。描述的系统大幅度地简化在生产中为该模型服务的工程开销，因为后处理(例如，先前系统中的语音识别输出与二值化输出之间的对准)被消除。最后，描述的技术很适合于生成包括标点符号和大写的丰富转录。

[0009] 本说明书主题的一个或多个实施例的细节在附图和以下描述中被阐述。本主题的其他特征、方面和优点将从描述、附图和权利要求中变得显然。

附图说明

[0010] 图1示出示例语音处理系统。

[0011] 图2示出联合ASR-SD神经网络的示例架构。

[0012] 图3是用于确定输入音频片段序列的转录的示例过程的流程图。

[0013] 图4示出使用联合ASR-SD神经网络生成的示例转录。

[0014] 图5是示出所描述的系统相对于基线系统的性能的示意图。

[0015] 不同附图中相同的附图标记和命名指示相同的元件。

具体实施方式

[0016] 图1示出示例语音处理系统100。语音处理系统100是作为计算机程序实现在一个或多个位置中的一个或多个计算机上的系统的示例，其中下述系统、组件和技术能够被实现。

[0017] 该系统100生成音频数据的转录。特别地，由系统100生成的转录标识给定音频分段中说出的单词，并且对于说出的单词中的每一个，标识说出了该单词的说话人，即，标识在对话中说出了该单词的说话人的角色或者唯一地标识个体说话人。

[0018] 更具体地，系统100通过使用神经网络120将音频数据110的输入序列110转录即映射到输出符号的输出序列150来执行联合自动语音识别(ASR)和说话人二值化(SD)。该神经网络120在本说明书中被称为“联合ASR-SD神经网络”。

[0019] 系统100被称为执行“联合”ASR和SD是因为使用神经网络120生成的单个输出序列150既定义音频数据的ASR输出，即，在音频分段中说出了哪些单词，又定义音频数据的SD输出，即，哪个说话人说出了每个单词。

[0020] 更具体地，音频数据110的输入序列是音频帧的序列，例如，原始音频的log-mel滤波器组能量或其他表示，并且输出序列150中的每个输出符号各自从包括文本符号和说话人标签符号两者的输出符号集合被选择。

[0021] 文本符号是表示自然语言中的某个文本单位的符号，例如，音素、字素、语素、字符、词块或某种自然语言中的单词。可选地，文本符号也能够包括其他书写单位，例如，标点符号。

[0022] 输出符号集合中的说话人标签符号(也称为“讲话者身份标签”)各自从可能的说话人的集合中标识不同的说话人。

[0023] 在一些情况下，每个说话人标签符号从说话人在对话中能够具有的可能的角色的集合中标识不同的角色。例如，输出符号集合可以包括标识患者正在说话的患者说话人标

签符号和标识医生或其他医疗专业人员正在说话的医生说话人标签符号。作为另一示例，输出符号集合可以包括标识客户正在说话的客户说话人标签符号和标识客户服务代表正在说话的代表说话人标签符号。

[0024] 在一些其他情况下，每个说话人标签符号从可能的个体说话人的集合中标识不同的唯一个体说话人。例如，可能的个体说话人的集合能够包括John Smith、Jane Doe和John Doe。

[0025] 输出符号集合通常也包括空白输出符号，当在给定时间步被选择为输出时，该空白输出符号指示系统在给定时间步没有发出说话人标签符号或文本符号。

[0026] 因此，系统100通过在多个时间步中的每一个时间步生成相应的输出符号来生成输出序列150。通过允许神经网络120在每个时间步从包括定义在音频输入中说出了什么单词的文本符号以及定义谁被标识为说出每个单词的说话人标签符号的符号集合中选择，系统100将联合ASR-SD神经网络120配置成执行联合ASR和SD，即，代替对同一输入独立地执行ASR和SD并且然后合并两个过程的结果。

[0027] 如本说明书中使用的，术语“嵌入”和“表示”指数值的有序集合，例如，浮点或其他数值的向量或矩阵，该数值的有序集合表示输入，例如，其表示输入文本符号或者表示文本符号的跨度。

[0028] 一旦系统100已经生成了输出序列，系统100能够提供输出序列150作为音频数据序列110的输出，即，通过将输出序列存储在一个或多个存储器中或者提供标识输出序列中的输出的数据以用于向用户呈现，或者能够从输出序列150生成音频数据序列110的转录并且提供该转录作为系统对于音频数据序列110的输出。

[0029] 转录从输出序列150中的文本符号中标识在音频数据序列110中说出的单词，并且从说话人标签符号中标识哪个说话人说出了每个单词。在下面参考图4描述转录的示例以及转录如何从输出序列被生成。

[0030] 在一些实施方式中，在推理时，系统100使用神经网络120来执行束搜索以生成最终输出序列150。

[0031] 特别地，在波搜索解码中，系统100维持一定数量的评分最高的部分序列的“束”，并且在每个输出时间步，将束中的每个序列扩展一个输出符号(即，通过将每个可能的输出符号添加到每个部分序列)。换句话说，对于给定时间步并且对于束中的每个部分输出序列，系统100使用下述技术来确定部分输出序列的分数分布。系统100然后从所有部分输出序列中选择将具有最高总分数的一定数量的扩展输出序列作为要被维持用于下一个时间步的部分序列。每个部分序列的总分数能够是，例如，根据生成用于在对应的时间步的部分序列的分数分布的部分序列中的输出符号的对数似然性。

[0032] 图2示出联合ASR-SD神经网络120的示例架构。

[0033] 如图2的示例所示，神经网络120包括转录神经网络210、预测神经网络220、联合神经网络230以及softmax输出层240。

[0034] 转录神经网络210被配置成处理音频分段数据以为输出序列中的每个时间步生成相应的编码表示h_t^{enc}。

[0035] 例如，转录神经网络210能够是深度循环神经网络，例如，包括单向或双向长短期记忆(LSTM)神经网络层或其他类型的循环层的堆叠的循环神经网络。在一些情况下，为了

证明与在音频分段数据中存在音频帧比可能存在更少的输出时间步的事实,转录神经网络210能够包括散布在循环层的堆叠当中的一个或多个时间延迟神经网络(TDNN)层。TDNN层用来降低音频分段数据的时间分辨率。

[0036] 预测神经网络220是被配置成在每个时间步处理该时间步的当前输出符号y_{u-1}以生成该时间步的预测表示h_u^{pred}的神经网络,其以已经被包括在输出序列中的任何较早的时间步的任何非空白输出符号为条件。

[0037] 在任何给定时间步的当前输出符号y_{u-1}通常是输出序列中最近发出的非空白输出符号,即,在相对于给定时间步在输出符号为空白输出符号的时间步之后最近的时间步的输出符号已经被忽略。当在输出序列中的任何较早的时间步,例如,在输出序列中的第一时间步,非空白输出符号尚未被包括时,系统能够使用固定占位符输入作为当前输出符号。

[0038] 例如,预测神经网络220能够包括将每个非空白输出符号(和占位符输出)映射到相应的嵌入的嵌入层,其由一个或多个单向LSTM或其他递归层跟随。在一些情况下,最后递归层直接生成预测表示,而在其他情况下,最后递归层由生成预测表示的全连接层跟随。

[0039] 联合神经网络230是被配置成在每个时间步处理(i)在该时间步的音频帧的编码表示和(ii)该时间步的预测表示以生成logit集合$l_{t,u}$的神经网络,该logit集合包括输出符号集合中的输出符号中的每一个的相应的logit。如上所述,输出符号集合包括文本符号和说话人标签符号两者。

[0040] 例如,联合神经网络230能够是将(i)编码表示和(ii)预测表示的级联映射到logit的单个全连接层或将(i)编码表示和(ii)预测表示的级联映射到logit的多层感知器(MLP)。

[0041] softmax输出层240被配置成接收输出符号中的每一个的相应的logit$l_{t,u}$并且生成输出符号集合中的输出符号上的概率分布P(y|t,u),即,包括每个文本符号、每个说话人标签符号和空白符号的相应的概率的概率分布。

[0042] 因此,当神经网络120具有图2中描述的架构时,为了使用神经网络220将音频分段序列映射到输出序列,系统在每个时间步执行以下操作:

[0043] (1)使用预测神经网络220处理时间步的当前输出符号以生成时间步的预测表示,其以已经被包括在输出序列中的任何较早的时间步的任何非空白输出符号为条件,

[0044] (2)使用联合神经网络230来处理(i)时间步的编码表示和(ii)时间步的预测表示以生成输出符号集合中的输出符号中的每一个的相应的logit,

[0045] (3)使用softmax输出层240来处理输出符号中的每一个的相应的logit以生成输出符号集合中的输出符号上的概率分布,以及

[0046] (4)使用概率分布来选择在时间步的输出符号,例如,通过根据概率分布来采样或者贪婪地选择具有最高概率的符号。

[0047] 为了生成时间步的编码表示,系统能够在第一时间步之前使用转录神经网络210来预处理音频序列以生成所有时间步的编码表示,或者执行使用转录神经网络210在每个时间步生成时间步的编码表示所必需的所需的附加处理。

[0048] 在一些情况下,例如,当如上所述执行束搜索时,不是执行步骤(4),而是系统自时间步起对于束中的k个候选输出序列中的每一个执行步骤(1)-(3)并且然后使用候选输出序列的概率分布来更新波束,例如,通过生成各自将候选输出序列中的相应的候选输出序

列扩展一个符号的扩展的候选输出序列的候选集合并且然后对于下一个时间步维持具有最高总分数的k个扩展的候选输出序列。

[0049] 为使神经网络120有效地被用于生成输出序列，系统在训练数据上训练神经网络120，该训练数据包括训练输入音频分段序列，并且对于每个训练输入音频分段序列，包括对应的输出目标。每个训练输入序列的对应的输出目标是包括文本符号和说话人标签符号的输出序列。更具体地，对于在训练输入音频分段中说出的每个单词，对应的输出目标包括标识说出了该单词的说话人标签符号作为在与该单词相对应的文本符号之后的下一个说话人标签符号。

[0050] 为了在训练数据上训练神经网络120，系统能够使用称为前向-后向算法的算法来优化目标函数，该目标函数测量给定的对应的输入音频分段情况下指配给由神经网络120的地面实况输出序列的条件概率(即，通过对在如果空白输出符号被移除则会产生对应的音频分段的可能的对准进行边缘化)。用于使用前后-后向算法来训练具有图2中描述的架构的神经网络的示例技术在K.C.Sim、A.Narayanan、T.Bagby、T.N.Sainath和M.Bacchiani的“Improving the efficiency of forward-backward algorithm using batched computation in tensorflow,”in IEEE Automatic Speech Recognition and Understanding Workshop(ASRU),2017(使用张量流中的批量计算提高前向后向算法的效率”，IEEE自动语音识别和理解研讨会(ASRU)，2017)和T.Bagby and K.Rao,的“Efficient implementation of recurrent neural network transducer in tensorflow,”in IEEE Spoken Language Technology Workshop(SLT).IEEE,2018(“在tensorflow中高效实现循环神经网络传感器”，IEEE口语技术研讨会(SLT)，IEEE,2018)中被描述，其全部内容特此通过引用整体地并入本文。

[0051] 图3是用于处理音频分段序列的示例过程300的流程图。为了方便，过程300将被描述为由位于一个或多个位置中的一个或多个计算机的系统执行。例如，适当地编程的音频处理系统，例如图1的系统100，能够执行过程300。

[0052] 系统获得表征音频分段的音频分段序列(步骤302)。音频分段可以是整个对话或较大对话的例如十、十五或三十秒的固定长度的部分。

[0053] 更具体地，音频分段序列包括多个音频帧。例如，每个音频帧能够是d维log-mel滤波器组能量，其中d是固定常数，例如五十、八十或一百，或者是音频分段的对应部分的不同声学特征表示。

[0054] 系统使用联合ASR-SD神经网络将音频分段序列映射到包括多个时间步中的每一个时间步的相应的输出符号的输出序列(步骤304)。

[0055] 如上所述，对于时间步中的每一个时间步，输出序列中的时间步的输出符号从包括以下的输出符号集合被选择：(i)多个文本符号、(ii)多个说话人标签符号，以及(iii)空白符号。

[0056] 此外，如上面另外描述的，系统能够通过维护候选序列的束并且然后选择该束中评分最高的候选序列或者通过从由联合ASR-SD神经网络在时间步生成的概率分布采样或者贪婪地选择输出符号来维护并更新在每个时间步的单个候选序列来生成输出序列。

[0057] 系统然后从输出序列确定(i)标识在音频分段中讲出的单词并且(ii)对于说出的单词中的每一个标识来自说出了该单词的可能的说话人的集合中的说话人的音频分段数

据的转录(步骤306)。例如,当词汇表中的文本符号是语素时,系统能够通过移除所有空白输出并且然后适当地接合输出序列中的相邻语素,即通过联合利用指示语素位于单词中间的标签标记的语素来标识转录中说出的单词。系统然后能够通过将每个单词标识为由通过紧接在输出序列中表示该单词的文本符号之后的说话人标签符号所表示的说话人说出而标识每个单词的说话人。

[0058] 图4示出使用联合ASR-SD神经网络生成的示例转录400。

[0059] 在图4的示例中,说话人标签符号标识说话人在对话中的角色,即,代替唯一地标识个体说话人。

[0060] 因此,为了生成示例转录400,系统生成了包括与多个单词“hello dr smith”(“你好史密斯博士”)相对应的文本符号、其由与“患者”角色相对应的说话人标签<spk:pt>的输出序列。因此,输出序列的该部分指示多个单词“hello dr smith”由在生成了转录400的对话中具有患者角色的说话人说出。

[0061] 更一般地,在输出序列中包括说话人标签指示与在说话人标签之前(即,从前一个说话人标签开始,或者,如果不存在前一个说话人标签,则从输出序列的开头开始)的文本标签相对应的单词被预测为已经由通过说话人标签标识的说话人说出。换句话说,对于通过输出系统中的文本符号集合表示的每个单词,系统将该单词标识为由通过紧接在输出序列中表示该单词的文本符号之后的说话人标签符号所表示的说话人说出。

[0062] 类似地,跟随说话人标签<spk:pt>之后,输出序列包括了与多个单词“hello mr jones what brings you here today”(“琼斯先生你好,今天是什么让你来这里”)相对应的文本符号并且然后包括与“医生”角色相对应的说话人标签<spk:dr>。系统因此在转录400中将多个单词“hello mr jones what brings you here today”标识为由具有医生角色的说话人说出。

[0063] 输出序列然后包括了与系统标识为由具有患者角色的说话人说出的单词“I am struggling again with my back pain”(“我再次因我的背痛而挣扎”)相对应的文本符号,因为这些文本符号后面是说话人标签<spk:pt>。

[0064] 图5是示出所描述的系统相对于基线系统的性能的示意图500。

[0065] 特别地,图500示出使用所描述的系统处理的对话的单词二值化错误率(WDER)的分布和由基线系统处理的对话的WDER的分布。

[0066] 基线系统是使用高质量的ASR神经网络以为对话生成文本符号并且单独地使用高质量的SD系统以标识对话的各部分的说话人标签的系统。基线系统然后使用复杂的技术以用于确定说话人何时在对话期间改变并且对准ASR系统和SD系统的输出。

[0067] 然而,如能够从图5看到的,描述的系统一致地生成与基线系统比具有更低的,即,更好的WDER的转录。更具体地,图500中示出的分布反映了使用描述的系统产生WDER的很大的改善,即从15.8%下降到2.2%,关于基线约86%的相对改善。WDER的这种增益在单词错误率(WER)的约0.6%退化情况下以小ASR性能成本为代价。因此,如能够从图5看到的,描述的系统以相对于高质量的ASR系统最小或无ASR性能退化来显著地执行系统的SD性能。

[0068] 本说明书连同系统和计算机程序组件一起使用术语“被配置”与系统和计算机程序组件相结合。对于要被配置成执行特定操作或动作的一个或多个计算机的系统意味着系统已经在其上安装了在操作中使该系统执行这些操作或动作的软件、固件、硬件或软件、固

件、硬件的组合。对于要被配置成执行特定操作或动作的一个或多个计算机程序意味着该一个或多个程序包括指令，该指令当由数据处理装置执行时，使该装置执行操作或动作。

[0069] 本说明书中描述的主题和功能操作的实施例能够以数字电子电路、以有形地体现的计算机软件或固件、以包括本说明书中公开的结构及其结构等价物的计算机硬件或者以它们中的一个或多个的组合而被实现。本说明书中描述的主题的实施例能够被实现为一个或多个计算机程序，即，在有形非暂时性存储介质上编码以用于由数据处理装置执行或者控制数据处理装置的操作的计算机程序指令的一个或多个模块。计算机存储介质能够是机器可读存储设备、机器可读存储基板、随机或串行访问存储设备或它们中的一个或多个的组合。可替代地或另外，程序指令能够被编码在人工生成的传播信号上，该传播信号例如是机器生成的电、光或电磁信号，其被生成以对信息进行编码以向适合的接收器装置传输以用于由数据处理装置执行。

[0070] 术语“数据处理装置”指数据处理硬件并且涵盖用于处理数据的所有种类的装置、设备和机器，作为示例包括可编程处理器、计算机或多个处理器或计算机。装置还能够是或者进一步包括专用逻辑电路，例如，FPGA(现场可编程门阵列)或ASIC(专用集成电路)。装置除了包括硬件之外还能够可选地包括为计算机程序创建执行环境的代码，例如，构成处理器固件、协议栈、数据库管理系统、操作系统或它们中的一个或多个的组合的代码。

[0071] 也可以被称为或者描述为程序、软件、软件应用、app、模块、软件模块、脚本或代码的计算机程序能够以包括编译或解释语言或声明或过程语言的任何形式的编程语言被编写；并且它能够以任何形式被部署，包括作为独立程序或者作为模块、组件、子例行程序或适合于在计算环境中使用的其它单元。程序可以但是不需要对应于文件系统中的文件。程序能够被存储在保持其它程序或数据的文件的一部分中，例如存储在标记语言文档中的一个或多个脚本；在专用于所述所讨论的程序的单个文件中或者在多个协调文件中，例如存储代码的一个或多个模块、子程序或部分的文件。计算机程序能够被部署为在一个计算机上或者在位于一个站点处或者分布在多个站点上并通过数据通信网络互连的多个计算机上被执行。

[0072] 在本说明书中，术语“数据库”被广泛用于指任何数据的合集：数据不需要以任何特定方式被结构化，或者根本不需要被结构化，并且其能够被存储在一个或多个位置中的存储设备中。因此，例如，索引数据库能够包括多个数据合集，每个数据合集可以被不同地组织和访问。

[0073] 类似地，在本说明书中术语“引擎”被广泛地用于指被编程为执行一个或多个具体功能的基于软件的系统、子系统或过程。通常，引擎将被实现为安装在一个或多个位置中的一个或多个计算机上的一个或多个软件模块或组件。在一些情况下，一个或多个计算机将被专用于特定引擎；在其它情况下，多个引擎能够在同一计算机或多个计算机上被安装并运行。

[0074] 本说明书中描述的过程和逻辑流程能够由执行一个或多个计算机程序的一个或多个可编程计算机执行以通过对输入数据进行操作并生成输出来执行功能。过程和逻辑流程还能够由例如是FPGA或ASIC的专用逻辑电路执行，或者通过专用逻辑电路和一个或多个编程计算机的组合而被执行。

[0075] 适合于执行计算机程序的计算机能够基于通用微处理器或专用微处理器或两者，

或任何其它种类的中央处理器。通常，中央处理单元将从只读存储器或随机存取存储器或两者接收指令和数据。计算机的必要元件是用于执行或者实行指令的中央处理单元以及用于存储指令和数据的一个或多个存储设备。中央处理单元和存储器能够由专用逻辑电路补充或者被并入在专用逻辑电路中。通常，计算机还将包括用于存储数据的一个或多个大容量存储设备，例如磁盘、磁光盘或光盘，或者操作上被耦合为从该一个或多个大容量存储设备接收数据或者将数据传送到该一个或多个大容量存储设备，或者两者。然而，计算机不需要具有这样的设备。此外，计算机能够被嵌入在另一设备中，该另一设备例如是移动电话、个人数字助理(PDA)、移动音频或视频播放器、游戏控制器、全球定位系统(GPS)接收器或便携式存储设备，例如通用串行总线(USB)闪存驱动器等。

[0076] 适合于存储计算机程序指令和数据的计算机可读介质包括所有形式的非易失性存储器、介质和存储设备，作为示例包括半导体存储设备，例如EPROM、EEPROM和闪速存储器设备；磁盘，例如内部硬盘或可移动盘；磁光盘；以及CD ROM和DVD-ROM盘。

[0077] 为了提供与用户的交互，本说明书中描述的主题的实施例能够在计算机上被实现，该计算机具有用于向用户显示信息的显示设备，例如，CRT(阴极射线管)或LCD(液晶显示器)监视器，以及用户能够通过其向计算机提供输入的键盘和定点指向设备，例如是鼠标或轨迹球。其它种类的设备能够被用于提供与用户的交互；例如，提供给用户的反馈能够是任何形式的感觉反馈，例如，视觉反馈、听觉反馈或触觉反馈；并且能够以任何形式接收来自用户的输入，包括声学、语音或触觉输入。另外，计算机能够通过向由用户使用的设备发送文档并从由用户使用的设备接收文档来与用户交互；例如，通过响应于从web浏览器接收到请求而向用户的设备上的web浏览器发送网页。另外，计算机能够通过向个人设备，例如，正在运行消息传送应用的智能电话，发送文本消息或其它形式的消息并且发过来从用户接收响应消息来与用户交互。

[0078] 用于实现机器学习模型的数据处理装置还能够包括，例如，用于处理机器学习训练或生产，即，推理、工作负载，的公共和计算密集部分的专用硬件加速器单元。

[0079] 机器学习模型能够使用机器学习框架被实现和部署例如，TensorFlow框架、Microsoft Cognitive Toolkit框架、Apache Singa框架或Apache MXNet框架。

[0080] 本说明书中描述的主题的实施例能够被实现在计算系统中，该计算系统包括后端组件，例如，作为数据服务器，或者包括中间件组件，例如应用服务器，或者包括前端组件，例如，具有用户通过其能够与本说明书中描述的主题的实施方式交互的图形用户界面、web浏览器或app的客户端计算机，或者包括一个或多个这种后端、中间件或前端组件的任何组合。系统的组件能够通过例如通信网络的任何形式或介质的数字数据通信而被互连。通信网络的示例包括局域网(LAN)和广域网(WAN)，例如互联网。

[0081] 计算系统能够包括客户端和服务器。客户端和服务器一般地通常彼此远离并通常通过通信网络来交互。客户端和服务器的关系借助于在相应的计算机上运行并且彼此具有客户端-服务器关系的计算机程序而产生。在一些实施例中，服务器向用户设备发送例如HTML页面的数据，例如，处于向与作为客户端的设备交互的用户显示数据并从该用户接收用户输入的目的。在用户设备处生成的数据，例如，用户交互的结果，能够在服务器处从设备被接收，。

[0082] 虽然本说明书包含许多具体实施方式细节，但是这些不应该被解释为对任何发明

的或可能要求保护的范围的限制，而是相反地被解释为对可以特定于特定发明的特定实施例的特征的描述。在本说明书中的单独的实施例的上下文中描述的某些特征也能够在单个实施例中组合地被实现。相反地，在单个实施例的上下文中描述的各种特征也能够单独地或者按照任何适合的子组合在多个实施例中被实现。此外，尽管特征可能在上文被描述按照某些组合起作用并且甚至最初被如此要求保护，但是来自要求保护的组合的一个或多个特征能够在一些情况下被从该组合中除去，并且要求保护的组合可以被引导为子组合或子组合的变形。

[0083] 类似地，虽然操作按照特定顺序在附图中被描绘并在权利要求书中被记载，但是这不应该被理解为需要这种操作以所示的特定顺序或者以先后顺序被执行，或者需要所有图示的操作被执行以实现预期的结果。在某些情况下，多任务处理和并行处理可以是有益的。此外，上述实施例中的各种系统模块和组件的分离不应该被理解为在所有实施例中需要这种分离，并且应该理解的是，描述的程序组件和系统一般地能够被一起集成在单个软件产品中或者封装到多个软件产品中。

[0084] 已经描述了主题的特定实施例。其它实施例在所附权利要求的范围内。例如，权利要求中记载的动作能够以不同的顺序被执行并仍然实现预期的结果。作为一个示例，附图中描绘的过程不一定需要示出的特定顺序或依次的顺序以实现预期的结果。在一些情况下，多任务处理和并行处理可以是有益的。

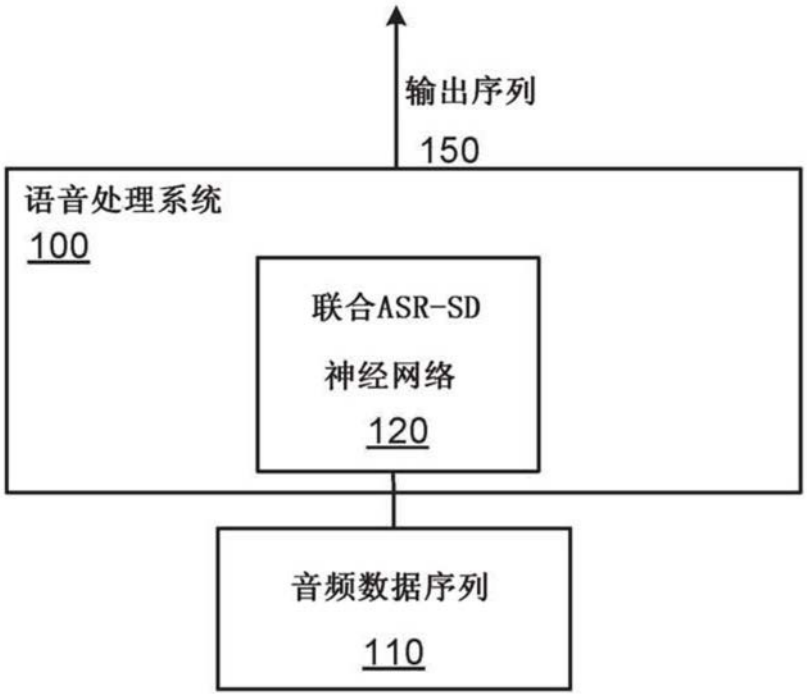

输出序列
150
语音处理系统
100
联合ASR-SD
神经网络
120
音频数据序列
110

图1

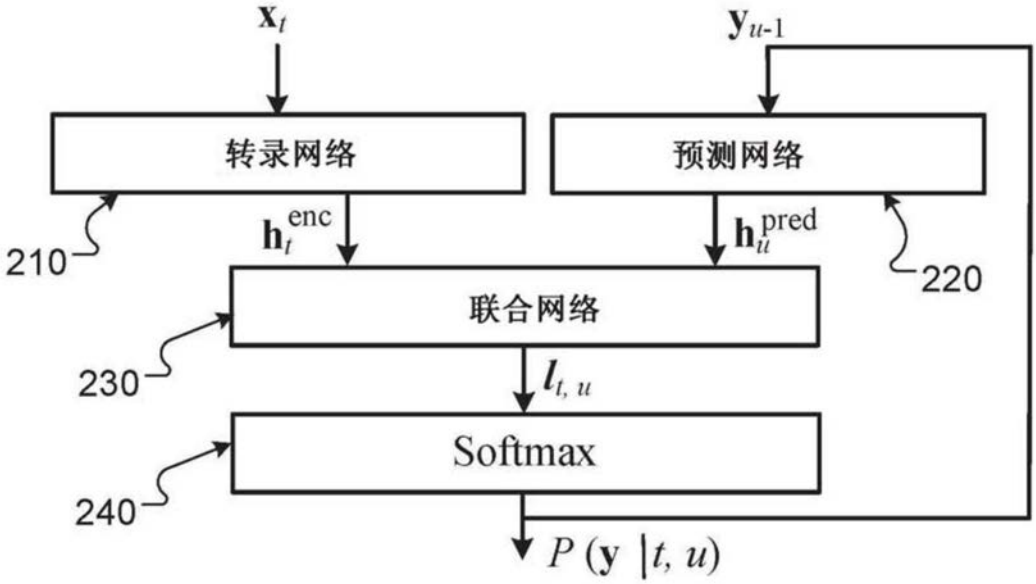

$\mathbf{x}_t$
$\mathbf{y}_{u-1}$
转录网络
预测网络
$\mathbf{h}_t^{enc}$
$\mathbf{h}_u^{pred}$
210
220
联合网络
230
$l_{t,u}$
Softmax
240
$P(\mathbf{y}|t,u)$

图2

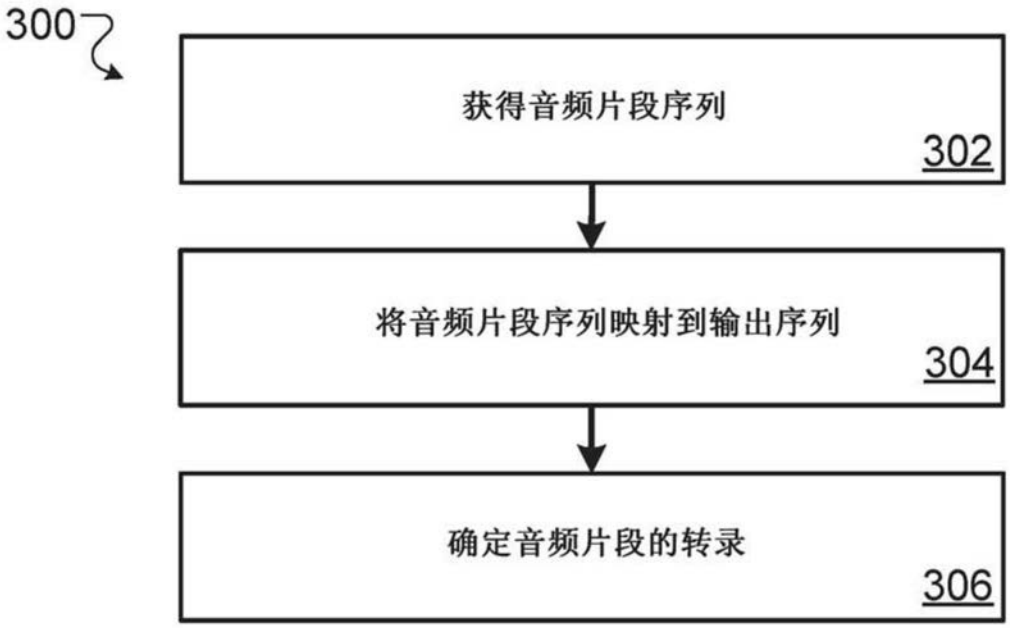


图3

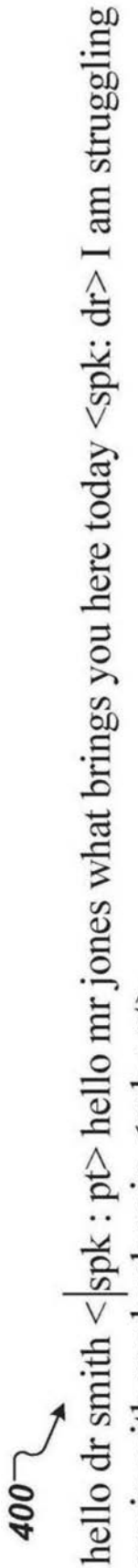

400
hello dr smith <|spk : pt> hello mr jones what brings you here today <spk: dr> I am struggling again with my back pain <spk : pt>

图4

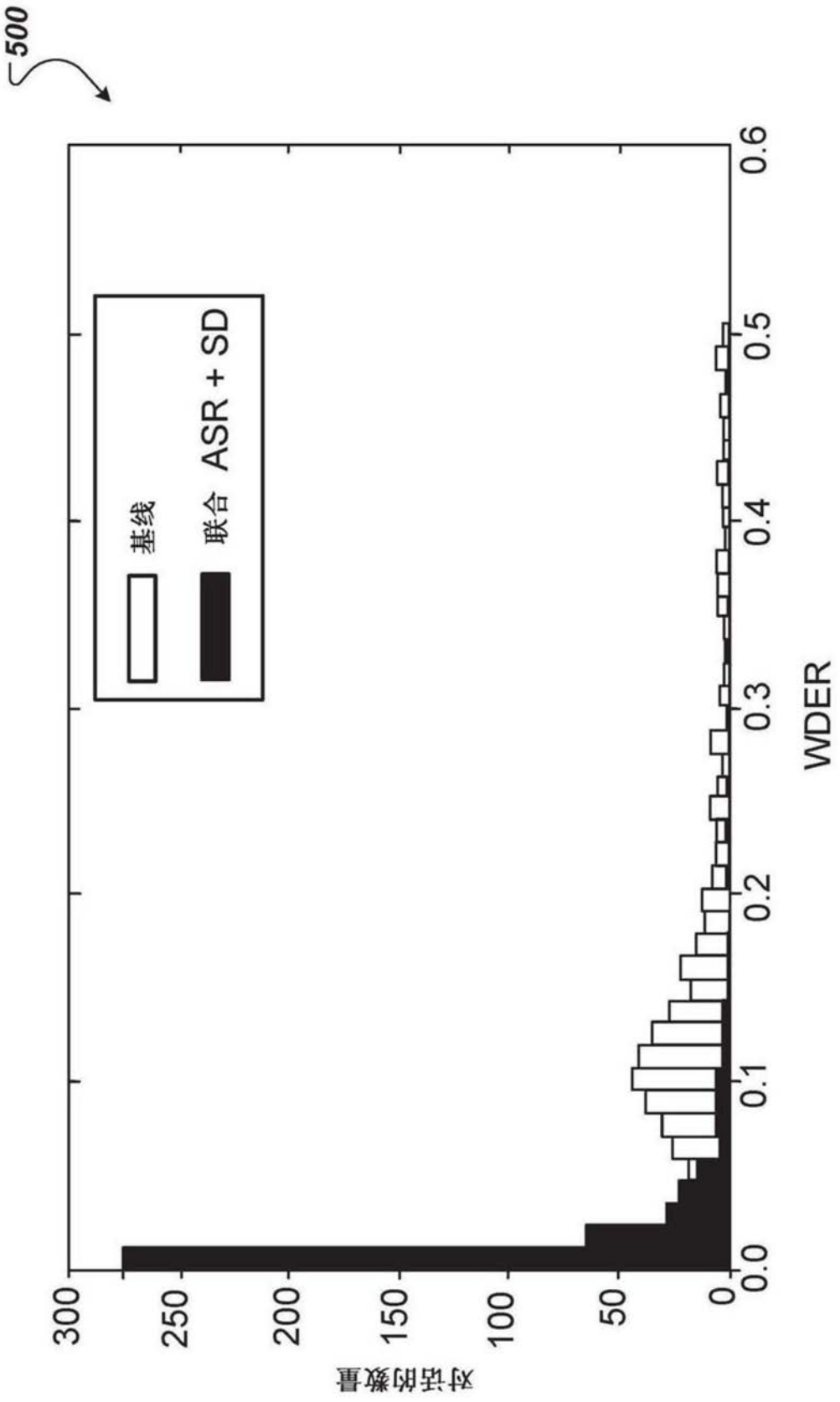

500
基线
联合 ASR + SD
对话的数量
300
250
200
150
100
50
0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
WDER

图5