(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 106469552 B
(45)授权公告日 2021.11.30

(21)申请号 201610685199.5

(22)申请日 2016.08.18

(65)同一申请的已公布的文献号
申请公布号 CN 106469552 A

(43)申请公布日 2017.03.01

(30)优先权数据
10-2015-0117422 2015.08.20 KR

(73)专利权人 三星电子株式会社
地址 韩国京畿道水原市

(72)发明人 崔喜烈 洪锡珒

(74)专利代理机构 北京铭硕知识产权代理有限公司 11286
代理人 王兆赓 张川绪

(51)Int.Cl.
G10L 15/06(2013.01)
G10L 15/08(2006.01)
G10L 15/16(2006.01)
G10L 15/18(2013.01)
G10L 15/183(2013.01)

(56)对比文件
CN 103544955 A,2014.01.29
CN 103544955 A,2014.01.29
GB 0820908 D0,2008.12.24
US 2015/0095026 A1,2015.04.02
US 8965763 B1,2015.02.24
CN 104575497 A,2015.04.29
CN 101158947 A,2008.04.09
JP 5184467 B2,2013.04.17
WO 2008/004666 A1,2008.01.10
US 5606644 A,1997.02.25
CN 103262156 A,2013.08.21
US 2014/0114655 A1,2014.04.24
Jan Chorowski.“Attention-Based Models for Speech Recognition”.《https://arxiv.org/abs/1506.07503》.2015,
M. Lehr 等.“Discriminatively estimated joint acoustic, duration, and language model for speech recognition”.《2010 IEEE International Conference on Acoustics, Speech and Signal Processing》.2010,
周艳萍.“机器人嵌入式语音识别系统设计与开发”.《中国优秀硕士学位论文全文数据库(信息科技辑)》.2013,

审查员 张彩

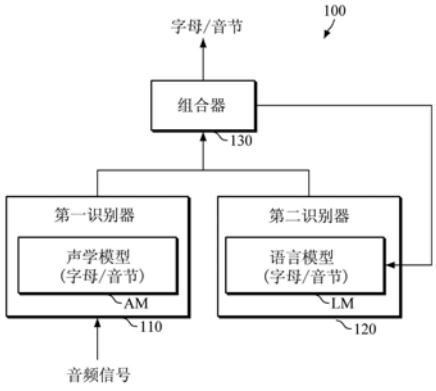
权利要求书3页 说明书15页 附图4页

(54)发明名称

语音识别设备和方法

(57)摘要

一种语音识别设备和方法。所述语音识别设备包括:第一识别器,被配置为通过使用声学模型以第一语言识别单位产生音频信号的第一识别结果;第二识别器,被配置为通过使用语言模型来以第二语言识别单位产生音频信号的第二识别结果;和组合器,被配置为组合第一识别结果和第二识别结果从而以第二语言识别单位产生最终识别结果并且在语言模型中反映最终识别结果。第一语言识别单位可以是与第二语言识别单位相同的语言单位类型。在相同的神经网络中配置第一识别器和第二识别器,并且使用提供到第一识别器的音频训练数据在神经网络中同时地/共同地训练第一识别器和第二识别器。



CN 106469552 B

1.一种语音识别设备,包括:

第一识别器,被配置为通过使用声学模型以第一语言识别单位产生音频信号的第一识别结果;

第二识别器,被配置为通过使用语言模型以第二语言识别单位产生音频信号的第二识别结果;和

组合器,被配置为组合第一识别结果和第二识别结果从而以第二语言识别单位产生音频信号的最终识别结果并且将最终识别结果提供给语言模型作为语言模型的输入。

2.如权利要求1所述的语音识别设备,

其中,第一识别器被配置为通过使用声学模型以第一语言识别单位产生随后的音频信号的第一识别结果;

其中,第二识别器被配置为通过将所述最终识别结果输入到语言模型来使用语言模型以第二语言单位产生随后的音频信号的第二识别结果,以及

其中,组合器还被配置为组合随后的音频信号的第一识别结果和随后的音频信号的第二识别结果作为随后的音频信号的最终识别结果。

3.如权利要求1所述的语音识别设备,其中,声学模型是基于注意机制的模型,并且第一识别结果表示音频信号的不基于针对语言识别单位的连接时间分类(CTC)的语言识别单位的概率,以及

其中,第二识别结果表示音频信号的基于识别的语言识别单位之间的时间连接性的概率。

4.如权利要求1所述的语音识别设备,其中,第一语言识别单位是与第二语言识别单位相同的语言单位类型。

5.如权利要求1所述的语音识别设备,其中,第一识别器被配置为通过使用第一声学模型来以第一语言识别单位产生音频信号的识别结果,并且通过使用被提供按照第一语言识别单位的音频信号的识别结果的第二声学模型来以第二语言识别单位产生音频信号的第一识别结果。

6.如权利要求1所述的语音识别设备,其中,第一识别结果和第二识别结果包括关于第一语言识别单位和第二语言识别单位的各自概率或状态的信息。

7.如权利要求1所述的语音识别设备,其中,组合器将第一识别结果和第二识别结果输入到统一模型中,统一模型的结果是最终识别结果。

8.如权利要求7所述的语音识别设备,其中,在同一神经网络中表示第一识别器、第二识别器和统一模型,神经网络被配置为将神经网络的表示声学模型的输出的节点和神经网络的表示语言模型的输出的节点连接到神经网络的表示统一模型的输入的各节点。

9.如权利要求8所述的语音识别设备,其中,神经网络被配置为将神经网络的表示提供最终识别结果的统一模型的输出的节点连接到神经网络的表示语言模型的输入的节点。

10.如权利要求9所述的语音识别设备,其中,神经网络的表示统一模型的输出的节点的数量取决于神经网络的表示语言模型的输入的节点的数量。

11.如权利要求9所述的语音识别设备,其中,神经网络被配置为已基于包括反向传播学习算法的学习算法在学习处理中被训练。

12.如权利要求7所述的语音识别设备,其中,神经网络被配置为已使用训练数据在学

习处理中被训练，其中，学习处理包括同时训练声学模型、语言模型和统一模型。

13.如权利要求1所述的语音识别设备，其中，第一识别器基于注意机制执行基于神经网络的解码以确定以第一语言识别单位的第一识别结果。

14.一种语音识别方法，包括：

通过使用声学模型以第一语言识别单位产生音频信号的第一识别结果；

通过使用语言模型来以第二语言识别单位产生音频信号的第二识别结果；

组合第一识别结果和第二识别结果从而以第二语言识别单位产生音频信号的最终识别结果；以及

将最终识别结果提供给语言模型作为语言模型的输入。

15.如权利要求14所述的语音识别方法，其中，第一语言识别单位是与第二语言识别单位相同的语言单位类型。

16.如权利要求14所述的语音识别方法，其中，所述产生第一识别结果的步骤包括：通过使用第一声学模型来以第一语言识别单位产生音频信号的识别结果并且通过使用被提供按照第一语言识别单位的音频信号的识别结果的第二声学模型来以第二语言识别单位产生音频信号的第一识别结果。

17.如权利要求14所述的语音识别方法，其中，第一识别结果和第二识别结果包括关于第一语言识别单位和第二语言识别单位的各自概率或状态的信息。

18.如权利要求14所述的语音识别方法，其中，所述产生最终识别结果的步骤包括：将第一识别结果和第二识别结果输入到统一模型中，统一模型的结果是最终识别结果。

19.如权利要求18所述的语音识别方法，其中，在同一神经网络中表示声学模型、语言模型和统一模型，神经网络被配置为将神经网络的表示声学模型的输出的节点和神经网络的表示语言模型的输出的节点连接到神经网络的表示统一模型的输入的各节点。

20.如权利要求19所述的语音识别方法，其中，神经网络被配置为将神经网络的表示提供最终识别结果的统一模型的输出的节点连接到神经网络的表示语言模型的输入的节点。

21.如权利要求19所述的语音识别方法，其中，神经网络的表示统一模型的输出的节点的数量取决于神经网络的表示语言模型的输入的节点的数量。

22.如权利要求19所述的语音识别方法，其中，神经网络被配置为已基于包括反向传播学习算法的学习算法在学习处理中被训练。

23.如权利要求18所述的语音识别方法，其中，神经网络被配置为已使用训练数据在学习处理中被训练，其中，学习处理包括同时训练声学模型、语言模型和统一模型。

24.如权利要求18所述的语音识别方法，还包括：在产生第一识别结果和产生第二识别结果之前，使用训练数据在学习处理中同时训练声学模型、语言模型和统一模型。

25.如权利要求14所述的语音识别方法，其中，产生第一识别结果的步骤包括：基于注意机制执行基于神经网络的解码以确定第一识别结果。

26.一种语音识别设备，包括：

第一识别器，被配置为通过使用第一声学模型来以第一语言识别单位产生音频信号的识别结果并且通过使用被提供按照第一语言识别单位的音频信号的识别结果的第二声学模型来以第二语言识别单位产生第一识别结果；

第二识别器，被配置为通过使用语言模型来以第二语言识别单位产生第二识别结果；

和

组合器，被配置为组合第一识别结果和第二识别结果以产生音频信号的最终识别结果。

27.如权利要求26所述的语音识别设备，其中，组合器将第一识别结果和第二识别结果输入到统一模型中，统一模型的结果是最终识别结果。

28.如权利要求27所述的语音识别设备，其中，在同一神经网络中表示第一识别器、第二识别器和统一模型，神经网络被配置为将神经网络的表示第二声学模型的输出的节点和神经网络的表示语言模型的输出的节点连接到神经网络的表示统一模型的输入的各节点。

29.一种语音识别系统，包括：

第一识别器，被配置为使用第一语言识别单位和声学模型产生音频信号的第一识别结果；

第二识别器，被配置为通过使用语言模型来以第二语言识别单位产生音频信号的第二识别结果；和

组合器，被配置为使用被配置为实现统一模型的神经网络组合第一识别结果和第二识别结果，统一模型以第二语言识别单位产生音频信号的最终识别结果并将最终识别结果提供给语言模型作为语言模型的输入。

30.如权利要求29所述的语音识别系统，其中，声学模型和语言模型是使用独立训练处理预先训练的模型，并且统一模型是使用训练处理训练的模型，所述训练处理使用训练数据与预先训练的声学模型和语言模型并将最终识别结果提供给语言模型作为语言模型的输入以进一步训练语言模型。

31.一种语音识别设备，包括：

语音接收器，被配置为捕获用户的音频并且基于捕获的音频产生音频信号；

包括一个或多个处理器中的第一处理器的语音识别器，被配置为将考虑音频信号的发音的声学模型和考虑音频信号的语言单位的连接性的语言模型的结果提供给统一模型，输出统一模型的结果作为音频信号的最终识别结果，并将最终识别结果提供给语言模型作为语言模型的输入；和

所述一个或多个处理器中的第二处理器，被配置为执行预定操作并且基于输出的最终识别结果执行所述预定操作中的特定操作。

32.如权利要求31所述的语音识别设备，其中，第一处理器和第二处理器是所述一个或多个处理器中的相同的处理器。

33.如权利要求31所述的语音识别设备，其中，在语音识别器中，在同一神经网络中表示声学模型、语言模型和统一模型，神经网络被配置为已通过使用包括反向传播学习算法的学习算法在学习处理中被训练。

34.如权利要求31所述的语音识别设备，其中，第一处理器被配置为执行下面的至少一项操作：控制通过所述语音识别设备的扬声器以听觉方式或通过所述语音识别设备的显示器以文本格式输出最终识别结果，将最终识别结果翻译成另一语言，以及通过第二处理器处理用于控制所述特定操作的执行的命令。

语音识别设备和方法

[0001] 本申请要求于2015年8月20日在韩国知识产权局提交的第10-2015-0117422号韩国专利申请的优先权权益，其全部公开为了所有目的而通过引用包含于此。

技术领域

[0002] 下面的描述涉及语音识别技术。

背景技术

[0003] 电子装置或服务器的语音识别引擎通常包括声学模型、语言模型和解码器。声学模型可以是静态模型，该静态模型基于输入音频信号的发音及其连接性输出输入音频信号的音素和发音的概率。语言模型是静态模型，该静态模型可基于独立地训练或指导的音素、发音、词语、句子等的连接性独立地输出与音素、发音、词语、句子等关联的信息。解码器对声学模型和语言模型的输出进行解码以基于声学模型和语言模型的输出返回输入音频信号的最终识别结果。高斯混合模型(GMM)已通常在过去被用于声学模型，但最近，已通过使用深度神经网络(DNN)声学模型来提高语音识别性能。如上所述，这种语音识别技术使用已被彼此独立地训练的声学模型和语言模型。另外，维特比解码方案已通常被用在声学模型中。

发明内容

[0004] 一个或多个实施例提供一种语音识别设备，所述语音识别设备包括：第一识别器，被配置为通过使用声学模型以第一语言识别单位产生音频信号的第一识别结果；第二识别器，被配置为通过使用语言模型来以第二语言识别单位产生音频信号的第二识别结果；和组合器，被配置为组合第一识别结果和第二识别结果从而以第二语言识别单位产生音频信号的最终识别结果并且在语言模型中反映最终识别结果。

[0005] 第二识别器可被配置为通过使用反映最终识别结果的语言模型来以第二语言单位产生随后的音频信号的第二识别结果，其中组合器还可被配置为组合由声学模型产生的随后的音频信号的第一识别结果和随后的音频信号的第二识别结果作为随后的音频信号的最终识别结果。

[0006] 声学模型可以是基于注意机制的模型，并且第一识别结果可表示不基于针对语言识别单位的连接时间分类的音频信号的语言识别单位的概率，其中第二识别结果可表示基于音频信号的识别的语言识别单位之间的时间连接性的概率。

[0007] 第一语言识别单位可与第二语言识别单位是相同的语言单位类型。

[0008] 第一识别器可被配置为通过使用第一声学模型来以第一语言识别单位产生音频信号的识别结果并且通过使用被提供按照第一语言识别单位的音频信号的识别结果的第二声学模型来以第二语言识别单位产生音频信号的第一识别结果。

[0009] 第一识别结果和第二识别结果可包括关于第一语言识别单位和第二语言识别单位的各自概率或状态的信息。

[0010] 组合器可将第一识别结果和第二识别结果输入到统一模型中，统一模型的结果是最终识别结果。

[0011] 可在同一神经网络中表示第一识别器、第二识别器和统一模型，神经网络被配置为将表示声学模型的输出的神经网络的节点和表示语言模型的输出的神经网络的节点连接到表示统一模型的输入的神经网络的各节点。

[0012] 神经网络可被配置为将表示提供最终识别结果的统一模型的输出的神经网络的节点连接到表示语言模型的输入的神经网络的节点。

[0013] 表示统一模型的输出的神经网络的节点的数量可取决于表示语言模型的输入的神经网络的节点的数量。

[0014] 神经网络被配置为已基于包括反向传播学习算法的学习算法在学习处理中被训练。

[0015] 神经网络被配置为已使用训练数据在学习处理中被训练，其中学习处理包括同时训练声学模型、语言模型和统一模型。

[0016] 第一识别器可基于注意机制执行基于神经网络的解码以确定以第一语言识别单位的第一识别结果。

[0017] 一个或多个实施例提供一种语音识别方法，所述语音识别方法包括：通过使用声学模型以第一语言识别单位产生音频信号的第一识别结果；通过使用语言模型来以第二语言识别单位产生音频信号的第二识别结果；组合第一识别结果和第二识别结果从而以第二语言识别单位产生音频信号的最终识别结果；以及在语言模型中反映最终识别结果。

[0018] 第一语言识别单位可与第二语言识别单位是相同的语言单位类型。

[0019] 第一识别结果的产生可包括通过使用第一声学模型来以第一语言识别单位产生音频信号的识别结果并且通过使用被提供按照第一语言识别单位的音频信号的识别结果的第二声学模型来以第二语言识别单位产生音频信号的第一识别结果。

[0020] 第一识别结果和第二识别结果可包括关于第一语言识别单位和第二语言识别单位的各自概率或状态的信息。

[0021] 最终识别结果的产生可包括将第一识别结果和第二识别结果输入到统一模型中，统一模型的结果是最终识别结果。

[0022] 可在同一神经网络中表示声学模型、语言模型和统一模型，神经网络被配置为将表示声学模型的输出的神经网络的节点和表示语言模型的输出的神经网络的节点连接到表示统一模型的输入的神经网络的各节点。

[0023] 神经网络可被配置为将表示提供最终识别结果的统一模型的输出的神经网络的节点连接到表示语言模型的输入的神经网络的节点。

[0024] 表示统一模型的输出的神经网络的节点的数量可取决于表示语言模型的输入的神经网络的节点的数量。

[0025] 神经网络被配置为已基于包括反向传播学习算法的学习算法在学习处理中被训练。

[0026] 神经网络被配置为已使用训练数据在学习处理中被训练，其中学习处理包括同时训练声学模型、语言模型和统一模型。

[0027] 该方法还可包括：在第一识别结果的产生和第二识别结果的产生之前，使用训练

数据在学习处理中同时训练声学模型、语言模型和统一模型。

[0028] 第一识别结果的产生可包括基于注意机制执行基于神经网络的解码以确定第一识别结果。

[0029] 一个或多个实施例提供一种语音识别设备，所述语音识别设备包括：第一识别器，被配置为通过使用第一声学模型来以第一语言识别单位产生音频信号的识别结果并且通过使用被提供按照第一语言识别单位的音频信号的识别结果的第二声学模型来以第二语言识别单位产生第一识别结果；第二识别器，被配置为通过使用语言模型来以第二语言识别单位产生第二识别结果；和组合器，被配置为组合第一识别结果和第二识别结果以产生音频信号的最终识别结果。

[0030] 组合器可将第一识别结果和第二识别结果输入到统一模型中，统一模型的结果是最终识别结果。

[0031] 可在同一神经网络中表示第一识别器、第二识别器和统一模型，神经网络被配置为将表示第二声学模型的输出的神经网络的节点和表示语言模型的输出的神经网络的节点连接到表示统一模型的输入的神经网络的各节点。

[0032] 一个或多个实施例提供一种语音识别系统，所述语音识别设备包括：第一识别器，被配置为使用第一语言识别单位和声学模型产生音频信号的第一识别结果；第二识别器，被配置为通过使用语言模型来以第二语言识别单位产生音频信号的第二识别结果；和组合器，被配置为使用被配置为实现统一模型的神经网络组合第一识别结果和第二识别结果，统一模型以第二语言识别单位产生音频信号的最终识别结果并提供最终识别结果以反映在语言模型中。

[0033] 声学模型和语言模型可以是使用独立训练处理预先训练的模型，并且统一模型可以是使用使用训练数据与预先训练的声学模型和语言模型的并将最终识别结果反映在语言模型中以进一步训练语言模型的训练处理训练的模型。

[0034] 一个或多个实施例提供一种语音识别设备，所述语音识别设备包括：语音接收器，被配置为捕获用户的音频并且基于捕获的音频产生音频信号；包括一个或多个处理器中的第一处理器的语音识别器，被配置为将考虑音频信号的发音的声学模型和考虑音频信号的语言单位的连接性的语言模型的结果提供给统一模型，并且输出统一模型的结果作为音频信号的最终识别结果；和一个或多更个处理器中的第二处理器，被配置为执行预定操作并且基于输出的最终识别结果执行所述预定操作中的特定操作。

[0035] 语音识别器还可被配置为在语言模型中反映最终识别结果，以训练语言模型。

[0036] 第一处理器和第二处理器可以是所述一个或多个处理器中的相同的处理器。

[0037] 在语音识别器中，可在同一神经网络中表示声学模型、语言模型和统一模型，神经网络被配置为已通过使用包括反向传播学习算法的学习算法在学习处理中被训练。

[0038] 第一处理器可被配置为执行下面的至少一项操作：控制通过所述设备的扬声器以听觉方式或通过所述设备的显示器以文本格式输出最终识别结果，将最终识别结果翻译成另一语言，以及通过第二处理器处理用于控制所述特定操作的执行的命令。

[0039] 在下面的描述中将会部分地阐述另外和/或替代的方面，并且这些方面部分地将会通过描述而变得清楚，或者通过实施提供的实施例可学习这些方面。

附图说明

[0040] 图1是示出根据一个或多个实施例的语音识别设备的方框图。

[0041] 图2是示出由根据一个或多个实施例的语音识别设备执行的语音识别的示图。

[0042] 图3是示出根据一个或多个实施例的语音识别设备的方框图。

[0043] 图4是示出根据一个或多个实施例的语音识别方法的流程图。

[0044] 图5是示出根据一个或多个实施例的语音识别方法的流程图。

[0045] 图6是示出根据一个或多个实施例的语音识别设备的方框图。

[0046] 在附图和详细描述中，除非另外描述，否则相同的附图标号将会被理解为始终表示相同或相似的元件、特征和结构。这些元件的相对尺寸和描述可为了清楚、说明和方便而被夸大。

具体实施方式

[0047] 下面的详细描述被提供用于辅助阅读者获得对这里描述的方法、设备和/或系统的全面理解。然而，在对本公开的理解之后，这里描述的方法、设备和/或系统的各种改变、变型和等同物可随后对于本领域普通技术人员而言变得清楚。在对本公开的理解之后，对于本领域普通技术人员而言将会清楚的是，除了必须以某种次序发生的操作之外，这里描述的操作的顺序仅是示例，并且不限于这里阐述的那些顺序，而是可改变。此外，在对本公开的不同方面的理解之后，为了更加清楚和简洁，可在一些描述中省略可理解的功能和构造的描述。

[0048] 除非另外定义，否则这里使用的所有术语(包括技术和科学术语)具有与各实施例所属于的领域的普通技术人员通常所理解的含义相同的含义。还将会理解，除非在这里明确地这样定义，否则术语(诸如，在常用词典中定义的那些术语)应该被解释为具有与在相关技术和本公开的情况下的它们的含义一致的含义并且将不会在理想化或过度正式意义上被解释。

[0049] 可对实施例做出各种改变和修改，一些改变和修改将会被详细地示出在附图和详细描述中。然而，应该理解，这些实施例不被解释为局限于本公开和示出的形式并且应该被理解为包括本公开的构思和技术范围内的所有改变、等同物和替代物。

[0050] 因此，这里描述的特征可被以不同形式实现，并且不应该被解释为局限于这里描述的示例。相反地，已提供这里描述的示例，以使得本公开将会是彻底的，并且将会将本公开的范围传达给本领域普通技术人员。

[0051] 图1是示出根据一个或多个实施例的语音识别设备的方框图。

[0052] 参照图1，语音识别设备100包括例如第一识别器110、第二识别器120和组合器130。

[0053] 第一识别器110可通过使用声学模型(AM)来以语言识别单位输出输入音频信号的第一识别结果。在这种情况下，仅作为示例，并且需要注意的是，在不同实施例中存在替代方案，音频信号可通过从音频信号提取特征的一个或多个预处理过程而被转换成音频帧(例如，每秒100帧)，音频帧可被输入到语音识别设备100。这里，对可被输入到声学模型的音频信号的提及例如应该被视为提及下面的任何一项：输入音频信号、转换成数字形式的音频信号、转换成音频帧的音频信号、已被以其它方式预处理的音频信号和独立音频帧(或

其较小部分)或由这种其它预处理产生的这种独立音频帧。同样地,并且仅作为示例,对先前音频信号、当前音频信号或随后的音频信号的提及也应该被视为分别提及一个或多个先前音频帧、当前音频帧或随后的音频帧,诸如用于表示识别操作的时间顺序和未来识别操作对当前和/或先前识别结果的依赖或仅表示当前识别操作对先前识别结果的依赖。

[0054] 另外,这里,语言识别单位表示语言中的基本单位之中将要被识别的预定语言单位,诸如音素、音节、语素、词语、短语、句子、段落等。这里,仅作为示例,音素、音节、语素、词语、短语、句子和段落可分别被视为不同类型的语言单位。另外,语言单位可根据语言而不同,从而可基于每种语言的各自已知特征预先确定语言识别单位。另外,这里,被称为大于另一语言单位类型的一个语言单位类型对应于具有预定分级体系的不同语言单位类型。仅作为这种预定分级体系的示例,在一个或多个实施例中,音节语言单位类型大于音素语言单位类型,语素语言单位类型大于音节语言单位类型,词语语言单位类型大于语素语言单位类型,短语语言单位类型大于词语语言单位类型,句子语言单位类型大于短语语言单位类型,并且段落语言单位类型大于句子语言单位类型,再一次需要注意的是,这仅是示例语言单位类型的这种预定分级体系的一个示例。

[0055] 在一个或多个实施例中,语言识别单位可以是字母和/或音节单位。以下,仅为了解释的方便而使用字母或音节单位,并且语言识别单位不限于此。

[0056] 第一识别器110可将例如通过预处理转换的音频帧输入到声学模型中,并且可以/针对特定语言识别单位输出音频帧的第一识别结果。在这种情况下,第一识别结果可包括音频帧的语言识别单位,诸如字母或音节概率或状态信息。例如,第一识别结果可包括用于一个或多个不同语言识别单位类型中的每个语言识别单位类型的一个或多个语言识别单位的识别信息和对应概率。

[0057] 在一个或多个实施例中,声学模型可通常输出每个输入音频帧的作为语言识别单位的音素的概率。根据一个或多个实施例,可基于神经网络或由神经网络表示的声学模型可通过使用基于神经网络的解码方法来以字母或音节单位输出概率或状态信息。该神经网络和这里讨论的其它神经网络可包括但不限于深度神经网络(DNN)、递归神经网络(RNN)、双向递归深度神经网络(BRDNN)等。

[0058] 在一个或多个实施例中,基于神经网络的解码方法可包括注意机制。通过表示基于注意机制的输入音频数据的一种或多种声学模型解码的神经网络,可以字母或音节单位(包括这种字母或音节单位的概率)输出音频帧的识别结果。注意机制表示通过选择数据的一些部分来顺序地观察数据,而非立刻观察所有的数据。例如,在给定图像中,注意机制表示在观察图像的一部分之后观察图像的另一部分,而非立刻识别整个图像。通过使用基于注意机制的神经网络解码方法,可使由于分段而发生的信息的损失最小化。

[0059] 第二识别器120可通过使用语言模型(LM)来以语言识别单位输出第二识别结果,其中第二识别结果可包括语言识别单位,例如字母或音节概率信息或状态信息。例如,第二识别结果可包括用于一个或多个不同语言识别单位类型中的每个语言识别单位类型的一个或多个语言识别单位的识别信息和对应概率。

[0060] 语言模型可例如基于用于先前音频数据的对应识别操作的最终识别结果对音素、字母表字母、音节、词语等的先前序列建模,并且产生或输出用于当前音频数据的当前音素、字母、音节、词语等的信息。根据一个或多个实施例,语言模型可基于神经网络,并且因

此，可诸如在声学模型中以字母或音节单位输出概率或状态信息。语言模型可被提供用于示例组合器130的一次或多次先前最终语言单位识别的语言单位信息，因此语言模型可对一个或多个语言单位类型的这种序列建模以提供用于当前语言单位(诸如，用于当前字母或音节单位)的概率。在一个或多个实施例中，语音识别设备包括内存以缓存组合器130的先前最终识别结果。仅作为示例，组合器130或第二识别器120可表示一个或多个处理装置和用于缓存组合器130的这种先前最终识别结果的内存。例如，如果组合器130存储这种信息，则组合器130可在知道第一识别器110正在操作新的或下一输入音频信号的同时将这种先前结果信息提供给第二识别器120，或者独立于第一识别器110的操作，第二识别器120可独立地操作并且在组合器130获得这种最终识别结果之后立即或在某个时间自动地产生这种语言单位概率。

[0061] 组合器130可组合第一识别结果和第二识别结果，并且可以以特定语言识别单位输出音频信号的最终识别结果。例如，组合器130可将第一识别结果和第二识别结果输入到预定统一模型中，并且可以示例字母或音节单位提供所述预定统一模型的结果或输出作为最终识别结果。

[0062] 类似于声学模型和语言模型，统一模型可基于神经网络。另外，在一个或多个实施例中，声学模型、语言模型和统一模型可由一个网络(例如，单个神经网络)集成和表示。例如，在一个或多个实施例中，神经网络的表示声学模型的输出的节点和神经网络的表示语言模型的输出的节点连接到神经网络的表示统一模型的输入的节点，由此形成单个神经网络。仅作为示例，神经网络的表示统一模型的输入的节点的数量可等于神经网络的表示声学模型和语言模型的各输出的节点的数量。

[0063] 一旦以字母或音节单位输出音频信号的最终识别结果，例如，一旦确定音频帧的字母表字母或音节的概率或关于其的状态信息，组合器130可在语言模型中反映该输出结果。因此，例如，语言模型是反映来自一个或多个声学模型以及一个或多个语言模型的一个或多个帧的先前统一模型结果的识别结果的动态模型或学习模型，这增强当前音频帧的识别结果。为此，在一个或多个实施例中，前述单个神经网络可被以这种方式配置，即神经网络的表示统一模型的输出的节点的数量等于或取决于神经网络的表示语言模型的输入的节点的数量，或者被以这种方式配置，即神经网络的表示语言模型的输入的节点的数量取决于神经网络的表示统一模型的输出的节点的数量。

[0064] 如上所述，基于先前的识别操作，组合器130在语言模型中反映或已反映先前音频帧的输出最终识别结果，由此能够使第二识别器120通过考虑先前音频帧的最终识别结果来计算和输出当前音频帧的当前字母或音节概率或状态信息。因此，在一个或多个实施例中，在第一识别器110将音频信号的当前帧输入到声学模型中以产生第一识别结果的同时，第二识别器120可通过组合器130来将先前帧的最终识别结果输入到语言模型中以产生第二识别结果。

[0065] 另外，在一个或多个实施例中，声学模型、语言模型和统一模型被预先训练以例如以预定语言识别单位输出概率或状态信息。在这种情况下，可通过使用学习算法(诸如，反向传播学习算法)并且使用目标函数来在学习或训练处理中共同训练声学模型、语言模型和统一模型。例如，一个或多个实施例包括例如基于相同的训练数据同时训练声学模型和语言模型，并且还可包括统一模型在语言模型中反映最终识别结果以用于随后由语言模型

考虑。以这种方式，可通过单个训练操作训练所有的声学模型、语言模型和统一模型。然而，训练不限于此。例如，一个或多个实施例包括基于相同或不同的训练数据预先彼此分开地(例如，独立地)训练声学模型和语言模型，并且还可包括组合统一模型与声学模型和语言模型以基于声学模型和语言模型的训练结果训练统一模型。替代地，一个或多个实施例包括部分独立地并且部分依赖地训练声学模型和语言模型。

[0066] 在一个或多个实施例中，与声学模型和语言模型被彼此分开地训练并且随后组合的实施例相比，当声学模型和语言模型被例如共同地训练从而建模角色不交叠时，每个模型可更高效地执行它的功能，由此可以能够实现更准确的语音识别。例如，尽管先前的声学模型技术可能诸如通过实现的连接时间分类(CTC)已明确考虑语言单位之间的连接性，但在一个或多个实施例中，声学模型没有实现CTC。在一个示例中，声学模型可仅考虑输入音频数据中的可能的语言单位的发音。这里，可能存在一些通过声学模型考虑的隐含连接信息，诸如，可基于声学模型被设置为识别的语言单位的类型的隐含连接信息。相反，在一个或多个实施例中，仅语言模型可明确依赖这种连接信息或语言单位之间的连接。

[0067] 图2是示出由根据一个或多个实施例的语音识别设备(诸如，图1的语音识别设备)执行的语音识别的示图。这里，接收或捕获的示例语音或音频由与诸如由用户发出或来自先前的记录等的语音“My name is Steve”对应的音频信号(AS)表示。以下，虽然将参照图1的语音识别设备讨论图2，但图2应该被理解为不限于此。

[0068] 参照图2，第一识别器110以帧为单位将实际音频信号(AS)输入到声学模型中，并且针对在音频信号中考虑的每个语言单位(例如针对每个音频帧)第一识别器110输出26个字母表字母中的一个或多个的概率作为声学模型的结果。例如，音频帧可通过(诸如经由语音接收器或其他预处理器)将音频信号中的语言单位中的每个分割为单独的音频帧已经被获得。可选择地，若干音频帧可用于音频信号中的单个语言单位，或者音频信号中的两个或更多个语言单位可被包括在同一音频帧中。因此，仅作为示例，随着音频信号中的每个语言单位被分割为单独的音频帧，第一识别器110的每个输出结果被输入到组合器130。在这种情况下，在从示例26个字母表字母之中选择时，声学模型可指示每个音频帧的最高可能字母表字母，其依次可以是例如m、a、i、n、e、i、m、i、s、s、t、i和v。伴随针对每个音频帧的来自声学模型的最高可能结果中的至少一个，第一识别器110还可向组合器130提供由声学模型确定的指示的最高可能识别语言单位的相应概率和每个音频帧的其他状态信息。

[0069] 第二识别器120可例如基于第二识别器120(诸如，通过语音识别设备100的共享内存、通过从组合器130提供或通过共享的神经网络)可获得的先前最终识别结果，考虑到字母表字母的连接关系而输出字母概率，并且可将输出的语言模型结果字母概率提供给组合器130。因此，因为在语言模型中反映由组合器130产生的先前音频帧的最终识别结果，所以可考虑到先前音频帧的最终识别结果中所包括的字母表字母而输出当前音频帧的准确识别结果。

[0070] 组合器130可将由第一识别器110输出或使其可用的第一音频帧的示例26个字母表字母的概率输入到统一模型中，并且可将由第二识别器120输出或使其可用的第一音频帧的示例26个字母表字母的概率输入到统一模型中，以由统一模型输出当前最终识别结果，即字母表字母‘m’具有与第一音频帧匹配的最高概率。在这种情况下，组合器130可在语言模型中反映针对第一音频帧输出的字母‘m’和对应概率信息。

[0071] 如上所述，可由组合器130将被第一识别器110确定为具有最高概率的字母表字母“mai neim is stiv”中的每个与第二识别器120的各最高概率识别结果组合，以便将音频信号准确地识别/理解为对应于“My name is Steve”。

[0072] 图3是示出根据一个或多个实施例的语音识别设备的方框图。

[0073] 参照图3，语音识别设备300包括例如第一识别器310、第二识别器320和组合器330。

[0074] 第一识别器310可以以大于第一语言识别单位的第二语言识别单位输出第一识别结果。在这种情况下，第一语言识别单位和第二语言识别单位是如上所述的语言单位中的任何一种语言单位。例如，第一语言识别单位可以是字母或音节单位，并且第二语言识别单位可大于字母或音节单位。以下，为了解释的方便，第一语言识别单位是字母或音节单位，并且第二语言识别单位是词语单位。然而，仅作为示例讨论这些特定语言识别单位，并且实施例不限于此。

[0075] 在一个或多个实施例中，第一识别器310可包括第一声学模型和第二声学模型。在这种情况下，使用第一声学模型，第一识别器310可以以第一语言识别单位提供音频信号的识别结果，仅作为示例，第一语言识别单位可包括字母或音节单位的语言单位。另外，通过将这个识别结果(例如，与相应概率和/或状态信息一起)提供给第二声学模型，第一识别器310可以以第二语言识别单位产生第一识别结果，仅作为示例，第二语言识别单位可包括词语的语言单位。因此，按照第一语言识别单位的第一声学模型的识别结果可包括例如字母或音节概率或状态信息，并且按照第二语言识别单位的第二声学模型的第一识别结果可包括例如词语的概率或状态信息。尽管按顺序仅示出两个声学模型，但是实施例不限于此，因为可存在多于两个级别的声学建模(或多于一个级别的语言建模)，在每一级别中可存在多于一个使用的声学模型(或语言模型)，例如，并行使用和/或诸如针对个性化模型或基于个人习语的模型或基于不同方言或语言选择性使用声学模型(或语言模型)。在一个实施例中，例如，不同语言识别单位类型的声学模型的两个级别可按照声学模型的顺序，顺序地减小时间分辨率。此外，尽管声学模型的示例的两个级别对相邻等级的语言识别单位类型(例如，第二语言识别单位类型大于第一语言识别单位类型)进行建模，但是实施例不限于此，另外地或者可选择地，可通过相继的声学模型对非相邻语言识别单位类型(诸如音节和短语)进行建模。另外，在一个或多个实施例中，第一声学模型和第二声学模型由语音识别设备中的一个或多个神经网络表示，例如，神经网络的表示第一声学模型的输出的节点连接到神经网络的表示第二声学模型的输入的节点以形成单个神经网络。仅作为示例，当向第一声学模型提供表示小于第一语言识别单位类型的单个语言单位的音频帧时，第一声学模型的识别结果可包括指示第一声学模型的状态的状态信息(例如，第一声学模型未完成识别操作)，当针对剩余的特定语言单位更多个的帧被接收并且第一语言识别单位被确定时，第一声学模型可将第一识别结果输出到第二声学模型。第二声学模型和第二识别器320的语言模型可类似地输出操作的各个状态的状态信息，包括准备好下一信号、数据或帧，当前识别操作的程度和对应语言单位的识别的完成。

[0076] 如上所述，当通过神经网络来实现时，仅作为示例，模型或神经网络可实现注意机制。例如，对于注意机制，神经网络的较高/随后的级别(例如，声学或语言模型级别)的输入可以是较低/先前级别的输出的汇总，其中，通过输入的加权和来获得汇总，权重为“注意”。

为了进一步说明该示例，当示例较低级别的输出为5维向量并且较低级别根据时间顺序被操作/运行7次(例如，7个连续音频帧)时，第7操作的输出可以是5x7矩阵。这里，“注意”可以是7维向量作为权重的示例。因此，较高级别可作为输入获得(或较低级别可作为输出产生)，5维向量作为7个5维向量的加权和。

[0077] 第二识别器320可通过使用例如已反映先前音频数据、帧或信号的最终识别结果的语言模型来以第二语言识别单位输出第二识别结果。在这种情况下，按照第二语言识别单位的第二识别结果可包括例如最高可能的词的概率或不同高可能的词语的概率或状态信息。根据实施例并且如以上所讨论，该语言模型也可由神经网络表示，并且在一个或多个实施例中可被或已被训练从而以第二语言识别单位输出识别结果。

[0078] 组合器330可被配置为组合第一识别器310的第一识别结果和第二识别器320的第二识别结果从而以第二语言识别单位输出音频数据、帧或信号的最终识别结果。例如，组合器330可将第一识别结果和第二识别结果输入到例如表示为组合器330中的神经网络的预定统一模型中，所述预定统一模型产生输入音频信号的每个示例词语的最终识别结果。因此，组合器330可输出每个词语的最终识别结果。

[0079] 另外，在这里的一个或多个实施例中，声学模型、语言模型和统一模型被集成并且表示为一个神经网络。例如，神经网络的表示第二声学模型的输出的节点和神经网络的表示语言模型的输出的节点可连接到神经网络的表示统一模型的输入的节点，以形成单个神经网络。在这里的替代实施例中，一个或多个声学模型、一个或多个语言模型和/或统一模型可由分开的或集成的神经网络表示。例如，所述一个或多个声学模型中的每个声学模型可由分开的神经网络表示或者组合/集成为单个神经网络，并且神经网络可表示这种一个或多个声学模型以及仅在单个神经网络中表示语言模型或统一模型之一，其余统一模型或语言模型由分开的神经网络表示，或者一个或多个声学模型可由与共同地表示其余声学模型、语言模型和统一模型的神经网络分开的神经网络表示，再一次需要注意的是，也存在替代方案。根据实施例，如以下更详细所讨论，这种单个或分开的神经网络中的任何一个、组合或全部可由一个或多个专门控制或配置的处理装置、处理器或计算机实现。另外，这种专门控制或配置的处理装置、处理器或计算机还可被专门控制或配置为执行接收或捕获的音频的一个或多个预处理操作(诸如，以上讨论的非限制性预处理)，或者预处理可由语音识别设备或系统的替代硬件和/或一个或多个专门控制或配置的处理装置、处理器或计算机执行。

[0080] 在这种语音识别系统实施例中，这种单个或分开的神经网络中的任何一个、组合或全部可由一个或多个服务器中的一个或多个专门控制或配置的处理装置、处理器或计算机实现，其余神经网络由远处或远程装置的一个或多个专门控制或配置的处理装置、处理器或计算机实现，诸如具有用户接口的电子装置，该用户接口接收或捕获诸如用于命令或搜索请求或其它操作的用户的语音，该用户接口被配置为将接收或捕获的音频传送给所述一个或多个服务器，并且该用户接口被配置为从服务器接收由所述一个或多个服务器的神经网络实现的统一模型的输出或者由所述一个或多个服务器的一个或多个神经网络实现的一个或多个声学模型和一个或多个语言模型之一或二者的输出。电子装置还可包括这样的统一模型，所述统一模型可(例如，规律地或在特定时间)被全部或部分地更新为与一个或多个服务器的训练的统一模型对应，例如从而当这种服务器中的一个或多个不可用时，

电子装置可执行识别操作。在该示例中，当一个或多个服务器顺序地变为可用时，电子装置可向服务器通知电子装置的统一模型的任何改变。还可针对可由电子装置实施的声学模型和语言模型共同地执行这样的操作。例如，在声学模型和语言模型一起被训练的实施例中，为了从而一起识别语音，模型可被一起更新。相反，在声学模型将独立于语言模型的训练被训练或不同声学模型也被独立训练的先前方法中，模型的任何更新也将基于各个强制排列信息被独立执行。

[0081] 返回图3，一旦针对当前音频数据、帧或信号以示例词语单位输出音频信号的最终识别结果，组合器330可在语言模型中反映该输出。以这种方式，第二识别器320可其后通过考虑当前音频信号的反映的最终识别结果，来计算和输出输入到或将输入到第一识别器310的随后的音频数据、帧或信号的随后的词语的概率或关于其的状态信息。为此，并且仅作为示例，前述单个网络(例如，单个神经网络)可被以这种方式配置：神经网络的表示统一模型的输出的节点的数量等于或取决于神经网络的表示语言模型的输入的节点的数量。

[0082] 在这种情况下，如以上所讨论，一个或多个实施例包括通过使用学习或训练算法(诸如，反向传播学习算法)来在学习/训练处理中共同地并且同时/共同地训练声学模型、语言模型和统一模型。仅作为示例，一个或多个实施例包括通过使用包括主要用于用作或包括语音识别设备300的电子装置实施例的词语或短语的训练数据来训练统一模型，但训练不限于此，并且可替代地包括独立地或分开地训练声学模型和语言模型，然后通过获得和/或更新例如由用于整个网络的所述一个或多个声学模型和语言模型提供的用于各训练识别结果的权重、权重值等来训练和/或更新统一模型，以使得由统一模型执行的最后语音识别可被优化。不管声学模型和语言模型被共同地训练、与统一模型共同地训练还是分开地训练，统一模型可将不同权重施加到从声学模型和语言模型提供的各个结果，例如，从而相比于其他声学模型或语言模型，声学模型中的一个或多个的结果被给予更高的可靠性或置信权重，或者例如，相比于其他模型，语言模型被给予更高的可靠性或置信权重。权重可以是固定的或者动态的，并且在训练和/或模型的使用期间改变，并且仅作为示例可基于被应用的模型、语言单位的类型或先前语言单位的识别。权重可以简单地是不同地施加到来自模型中的一个或多个的结果的单个权重参数，或者是更复杂的矩阵权重，如对本公开的理解之后将理解的那样。

[0083] 图4是示出根据一个或多个实施例的语音识别方法的流程图。仅作为示例，可通过使用根据一个或多个实施例的语音识别设备(诸如，图1的语音识别设备100)来实现图4中示出的语音识别方法，但不限于此。类似地，虽然将会参照图1的语音识别设备100讨论图4的语音识别方法，但这仅是为了方便解释并且不应该局限于此。

[0084] 在操作410中，语音识别设备100可通过使用声学模型来以语言识别单位输出音频信号的第一识别结果。在这种情况下，音频信号可通过从音频信号提取特征的预处理过程而被转换成音频帧，并且音频信号可被以音频帧为单位输入到声学模型以输出第一识别结果。按照语言识别单位的第一识别结果可包括例如音频帧的字母或音节概率或状态信息。如以上所讨论，可通过被配置为例如以字母或音节单位提供概率或状态信息的神经网络实现声学模型。

[0085] 在操作420中，语音识别设备100可通过使用语言模型来以语言识别单位输出第二识别结果。例如，第二识别结果可包括字母或音节概率或状态信息。如以上所讨论，可通过

被配置为以字母或音节单位提供概率或状态信息的神经网络实现语言模型。在一个或多个实施例中，根据实施例，操作420可在操作410开始之后开始，操作420可在操作410之前开始，或者它们可同时开始。

[0086] 随后，在操作430中，语音识别设备100可组合音频信号的第一识别结果和第二识别结果，并且可以以语言识别单位产生或输出最终识别结果。例如，通过与声学模型和语言模型一起使用统一模型或使用集成/组合声学模型和语言模型的统一模型，语音识别设备100可组合第一识别结果和第二识别结果，并且可以示例字母或音节单位输出最终识别结果。

[0087] 类似于声学模型和语言模型，统一模型可基于神经网络。声学模型、语言模型和统一模型可由一个网络(例如，一个神经网络)集成和表示。例如，神经网络的表示声学模型的输出的节点和神经网络的表示语言模型的输出的节点可连接到神经网络的表示统一模型的各输入的节点。

[0088] 一旦产生或输出音频信号的最终识别结果，语音识别设备100还在操作430中在语言模型中反映最终识别结果，因此语言模型在操作420的随后的实现方式中通过考虑反映的当前音频帧的识别结果来产生随后的音频帧的随后的第二识别结果。

[0089] 图5是示出根据一个或多个实施例的语音识别方法的流程图。仅作为示例，可通过使用根据一个或多个实施例的语音识别设备(诸如，图3的语音识别设备300)来实现图5中示出的语音识别方法，但不限于此。类似地，虽然将会参照图3的语音识别设备300讨论图5的语音识别方法，但这仅是为了方便解释并且不应该局限于此。

[0090] 在操作510中，语音识别设备300可通过使用第一声学模型来以第一语言识别单位产生输入音频信号的识别结果。然后，在操作520中，语音识别设备300可通过使用第二声学模型来以第二语言识别单位(例如，以词语单位)产生第一识别结果。第二语言识别单位可大于第一语言识别单位。例如，第一语言识别单位可以是字母或音节单位，并且第二语言识别单位可以是词语单位。另外，第一声学模型和第二声学模型可基于神经网络。例如，神经网络的表示第一声学模型的输出的节点可连接到神经网络的表示第二声学模型的输入的节点，从而产生的第一声学模型的结果可被输入到第二声学模型。

[0091] 在操作530中，可通过使用反映或已反映先前音频信号的最终识别结果的语言模型来输出按照第二语言识别单位的第二识别结果。例如，第二识别结果可包括词语的概率或状态信息。在这种情况下，语言模型可基于神经网络，并且在一个或多个实施例中，表示语言模型的神经网络可被训练以便例如考虑到包括反映的先前音频信号、数据或帧的最终识别结果的音频信号的语言单位和/或不同语言单位类型之间的预期连接性而以第二语言识别单位输出识别结果。在一个或多个实施例中，根据实施例，操作530可在操作510或520中的任一操作开始之后、在操作510和520之前开始，或者操作530可与操作510或520中的任一操作同时开始。

[0092] 然后，语音识别设备300可组合第一识别结果和第二识别结果，并且可以以第二语言识别单位输出音频信号的最终识别结果。例如，语音识别设备300可将第一识别结果和第二识别结果输入到预定统一模型中，所述预定统一模型被配置为考虑第一识别结果和第二识别结果并且以示例性词语单位产生最终识别结果。

[0093] 类似于声学模型和语言模型，统一模型可基于神经网络。声学模型、语言模型和统

一模型可由一个神经网络集成和表示。例如，神经网络的表示第二声学模型的输出的节点和神经网络的表示语言模型的输出的节点可连接到神经网络的表示统一模型的输入的节点，由此形成单个网络。如上所述，这个单个神经网络也可表示第一声学模型，例如，神经网络的表示第一声学模型的输出的节点连接到神经网络的表示第二声学模型的输入的节点。

[0094] 接下来，一旦以词语单位产生或输出音频信号的最终识别结果，语音识别设备300可在语言模型中反映该最终识别结果。为此，前述单个网络可被以这种方式配置：神经网络的表示统一模型的输出的节点的数量等于或取决于神经网络的表示语言模型的输入的节点的数量。

[0095] 图6是示出根据一个或多个实施例的作为电子装置600的语音识别设备的方框图。

[0096] 在一个或多个实施例中，电子装置600可包括语音识别装置，诸如图1的语音识别设备100和图3的语音识别设备300之一或二者。仅作为非限制性示例，电子装置600可以是电视机、机顶盒、台式计算机、膝上型计算机、翻译机器、智能电话、平板PC、智能手表、可穿戴装置、车辆的电子控制装置等中的任何装置，并且可通过使用例如安装/包括的语音识别装置来处理用户的各种需求。然而，电子装置600不限于此，并且也可使用在语音识别的任何或所有应用中使用的其它电子装置。

[0097] 参照图6，在实施例中，电子装置600包括语音接收器610、语音识别器620和处理器630，其中语音识别器620可分别对应于图1的语音识别设备100和图3的语音识别设备300之一或二者。语音识别器620是这样的硬件：该硬件可由诸如以上讨论的特定一个或多个处理装置实现，或者可由如以上所讨论的也被配置为响应于通过语音识别器620的识别结果识别的命令或询问来控制电子装置600的其它操作(诸如，其它用户接口操作)的特定一个或多个处理装置实现，但实施例不限于此。

[0098] 语音接收器610可接收通过电子装置600中所包括的也由语音接收器610表示的麦克风等输入的用户的音频信号。替代地，在一个或多个实施例中，语音接收器610可被包括在对应语音识别系统实施例的单独装置中，诸如被配置为将接收或捕获的音频和/或当语音接收器610还被配置为实现以上讨论的接收/捕获的音频的预处理时的接收/捕获的音频的对应预处理的音频传送给语音识别系统的语音识别装置的有线或无线麦克风或控制器。如图6中所示，用户的音频信号可与词语、短语或句子相关以便被翻译成用于控制电视机、驾驶车辆等的另一语言或命令。另外，再一次仅作为示例，预处理可包括：将例如由用户输入的接收或捕获的模拟音频信号转换成数字信号，将数字信号划分成多个音频帧，并且将音频帧例如作为预处理的音频帧数据传送给语音识别器620。如上所述，在一个或多个实施例中，声学模型、语言模型和统一模型中的一个或多个可由一个或多个远程服务器实现为语音识别器620，并且电子装置600可被配置为传送例如捕获的音频并且基于所述传送的音频从声学模型和/或语言模型神经网络中的一个或多个接收合适的输出或接收表示对应统一模型的一个或多个神经网络的输出。

[0099] 因此，在不同实施例中，语音识别器620可将声学模型和语言模型的结果输入到统一模型中，并且可基于统一模型的输出结果输出音频信号的最终识别结果。

[0100] 除了这种处理器630被专门控制或配置为包括或用作语音识别器620的实施例之外，处理器630还可控制和/或执行电子装置600的另外的操作，例如包括可响应于由语音识别器620返回的最终识别结果而控制电子装置600的当前或另外的操作的操作。例如，处理

器630可通过电子装置600的扬声器等以处理器产生的语音输出由用户输入的语音的识别结果，或者可在电子装置600的显示器上以文本格式提供识别结果，这样用于起草消息或文件，或仅由电子装置600显示。另外，处理器630可被配置为控制和执行用于处理关于电子装置600的命令(例如，通电/断电、音量控制等)的操作。仅作为非限制性示例，接口640表示扬声器、电子装置600的一个或多个用户接口(诸如，显示器、键盘或触摸屏幕)和/或诸如用于与示例服务器执行以上提及的传送的电子装置600的一个或多个通信硬件模块。在一个示例中，接口640还可表示产生由语音识别器610预先处理的音频信号的麦克风。

[0101] 另外，在一个或多个实施例中，当电子装置600被配置为执行翻译时，诸如在语音接收器610从电子装置600的内存或从另一服务器/装置获得来自实时或记录的音频的音频数据的情况下，以及诸如在将最终识别结果翻译成另一语言的情况下，仅作为示例，处理器630还被配置为例如基于存储在电子装置600的内存中或通过从另一装置或服务器传送而可用的一个或多个词典将以文本格式输出的最终识别结果翻译成所述另一语言，并且诸如在电子装置600的显示器上以语音和/或以文本格式输出翻译的结果。然而，处理器630不限于此，并且可被用于电子装置600的各种应用。例如，电子装置的存储器还可存储声学模型和语言模型、可选择的声学模型和语言模型、用于训练模型的数据、以及由语音识别器620使用或产生的任何信息。存储器还可存储可执行指令，从而电子装置600的一个或多个处理器可实现上述操作中的任何一个。

[0102] 除了图1-3和6的设备、模块、元件、装置和其它组件是硬件元件(任何硬件元件可实现图2和4-5的方法)之外，仅作为示例，图2和4-5的方法可由硬件组件执行，所述硬件组件包括电子装置或系统实施例中所包括的任何以上讨论的示例硬件元件。仅作为示例，硬件组件的示例包括电阻器、电容器、电感器、电源、频率发生器、运算放大器、功率放大器、低通滤波器、高通滤波器、带通滤波器、模数转换器、数模转换器、控制器、传感器、发电机、内存、驱动器、电路和/或被配置为执行图2和4-5的上述方法中的任何方法的本领域普通技术人员已知的任何其它电子组件。在一个示例中，硬件组件由一个或多个处理装置或处理器或计算机实现。处理装置、处理器或计算机由一个或多个处理元件(诸如，逻辑门阵列、控制器和算术逻辑器件、数字信号处理器、微型计算机、可编程逻辑控制器、现场可编程门阵列、可编程逻辑阵列、微处理器或者能够以定义的方式对指令做出响应并且执行指令以实现想要的结果的本领域普通技术人员已知的任何其它装置或装置的组合)实现。在一个示例中，处理装置、处理器或计算机包括或连接到存储计算机可读代码、指令或软件的一个或多个内存，所述计算机可读代码、指令或软件由处理装置、处理器或计算机执行并且可控制处理装置、处理器或计算机执行这里描述的一个或多个方法。仅作为示例，由处理装置、处理器或计算机例如通过执行计算机执行代码、指令或软件(诸如，操作系统(OS)和在OS上运行的一个或多个软件应用)来实现的硬件组件可执行这里参照图2和4-5描述的操作。硬件组件也响应于指令或软件的执行而访问、操纵、处理、创建和存储数据。为了简单，可在这里描述的示例的描述中使用单数术语“处理装置”、“处理器”或“计算机”，但在其它示例中，使用多个处理装置、处理器或计算机，或者处理装置、处理器或计算机包括多个处理元件或多种类型的处理元件或者处理装置、处理器或计算机包括多个处理元件并且包括多种类型的处理元件。在一个示例中，硬件组件包括多个处理器，并且在另一示例中，硬件组件包括处理器和控制器。硬件组件具有不同处理结构中的任何一种或多种处理结构，其示例包括单个处

理器、独立处理器、并行处理器、远程处理环境、单指令单数据(SISD)多处理、单指令多数据(SIMD)多处理、多指令单数据(MISD)多处理和多指令多数据(MIMD)多处理。另外，以上称为设备的各种附图中示出的任何连接线或连接器旨在表示各种硬件元件之间的示例功能关系和/或物理或逻辑耦合，许多替代或另外的功能关系、物理连接或逻辑连接可存在于对应装置实施例中。

[0103] 通过如上所述的专门控制或配置的处理装置、处理器或计算机执行处理器或计算机可读代码、指令或软件以执行这里描述的操作，可执行执行这里描述的操作的图2和4-5中示出的方法。

[0104] 用于控制处理装置、处理器或计算机实现硬件组件并且执行如上所述的方法的处理器或计算机可读代码、指令或软件可被编写为计算机程序、代码段、指令或其任何组合，所述计算机程序、代码段、指令或其任何组合用于个别地或共同地指示或配置处理装置、处理器或计算机以用作用于执行由硬件组件执行的操作和如上所述的方法的机器或专用计算机。在一个示例中，处理器或计算机可读代码、指令或软件包括由处理装置、处理器或计算机直接执行的机器代码，诸如由编译器产生的机器代码。在另一示例中，处理器或计算机可读代码、指令或软件包括由处理装置、处理器或计算机使用解释器执行的高级代码，诸如利用任何编程或脚本语言(诸如，C、C++、Java、汇编程序等)实现的高级代码，利用数据结构、对象、处理、例程或其它编程元件的任何组合实现各种算法。基于这里的公开，并且仅在对该公开的理解之后，本领域普通程序员可基于公开用于执行由硬件组件执行的操作和如上所述的方法的算法的附图中示出的方框图和流程图以及说明书中的对应描述容易地编写处理器或计算机可读代码、指令或软件。

[0105] 用于控制处理装置、处理器或计算机实现诸如在图1-3和6中的任何附图中讨论的硬件组件并且执行如以上在图2和4-5中的任何附图中所述的方法的处理器或计算机可读代码、指令或软件和任何关联的数据、数据文件和数据结构被记录、存储或固定在一个或多个非暂时性计算机可读存储介质中或者被记录、存储或固定在一个或多个非暂时性计算机可读存储介质上。非暂时性处理器或计算机可读存储介质的示例包括只读存储器(ROM)、随机存取存储器(RAM)、闪存、CD-ROM、CD-R、CD+R、CD-RW、CD+RW、DVD-ROM、DVD-R、DVD+R、DVD-RW、DVD+RW、DVD-RAM、BD-ROM、BD-R、BD-R LTH、BD-RE、磁带、软盘、磁光数据存储装置、光学数据存储装置、硬盘、固态盘和能够以非暂时性方式存储处理器或计算机可读代码、指令或软件和任何关联的数据、数据文件和数据结构并且向处理装置、处理器或计算机提供处理器或计算机可读代码、指令或软件和任何关联的数据、数据文件和数据结构以使得处理装置、处理器或计算机能够执行指令的本领域普通技术人员已知的任何装置。在一个示例中，处理器或计算机可读代码、指令或软件和任何关联的数据、数据文件和数据结构分布在联网的计算机系统上，以使得由处理装置、处理器或计算机以分布式方式存储、访问和执行指令和软件和任何关联的数据、数据文件和数据结构。

[0106] 仅作为非穷举示例，并且除了诸如以上参照图1-3和6中的任何附图描述和/或被配置为实现参照图2和4-5中的任何附图描述的方法的这里的语音识别设备和电子装置实施例的可能的硬件实现方式的以上解释中的任何解释之外，也可以是：移动装置，诸如蜂窝电话、可穿戴智能装置(诸如，示例智能手表)、其它可穿戴装置、便携式个人计算机(PC)(诸如，示例膝上型计算机、笔记本、亚笔记本、上网本或超级移动PC(UMPC)、示例平板PC(平板

计算机))、平板手机、个人数字助手(PDA)、数字照相机、便携式游戏控制台、MP3播放器、便携式/个人多媒体播放器(PMP)、手持式电子书、全球定位系统(GPS)导航装置或传感器;或固定装置,诸如示例台式PC、示例TV(诸如,高清晰度电视(HDTV))、DVD播放器、Blu-ray播放器、示例机顶盒或家用电器;或能够实现无线或网络通信的任何其它移动或固定装置。在一个或多个实施例中,电子装置或电子装置系统实施例包括显示器、一个或多个扬声器、用户接口、用于存储数据或记录的音频/视频的内存或其它非暂时性介质和/或音频输入装置(诸如,一个或多个麦克风)。

[0107] 尽管本公开包括特定示例,但对于本领域普通技术人员而言将会清楚的是,在不脱离权利要求及其等同物的精神和范围的情况下,可在这些示例中做出各种形式和细节上的变化。应该仅在描述性意义上而非为了限制的目的考虑这里描述的示例。对每个示例中的特征或方面的描述应该被视为适用于其它示例中的类似特征或方面。如果描述的技术被以不同次序执行,和/或如果描述的系统、架构、装置或电路中的组件被以不同方式组合和/或由其它组件或其等同物替换或补充,则可实现合适的结果。因此,本公开的范围不由详细描述限制,而是另外由权利要求及其等同物支持,并且权利要求及其等同物的范围内的所有变化应该被解释为被包括在本公开中。

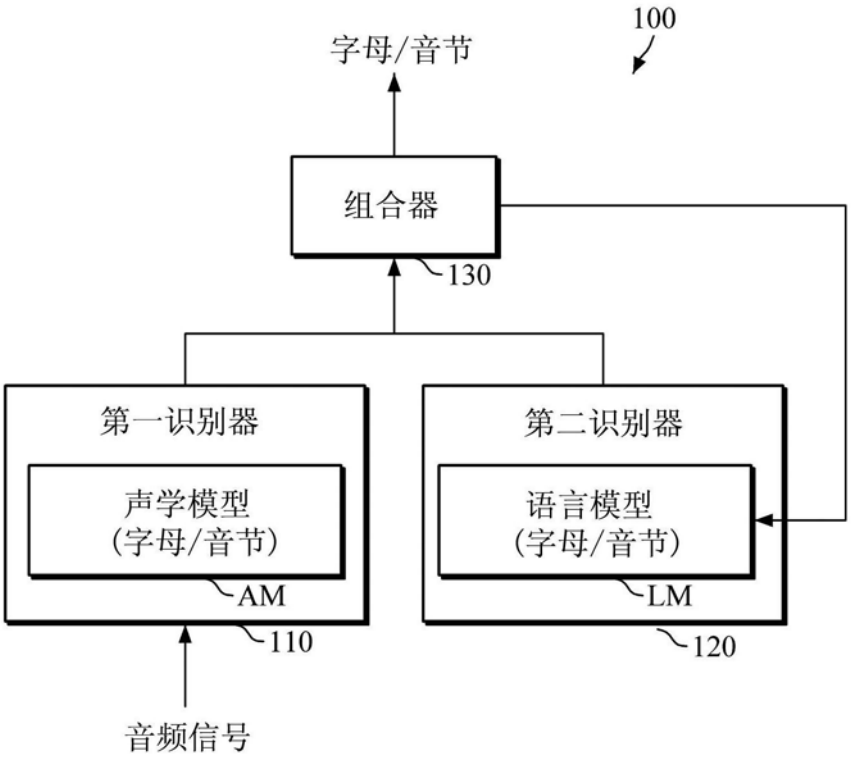

100
字母/音节
组合器
130
第一识别器
声学模型
(字母/音节)
AM
110
第二识别器
语言模型
(字母/音节)
LM
120
音频信号

图1

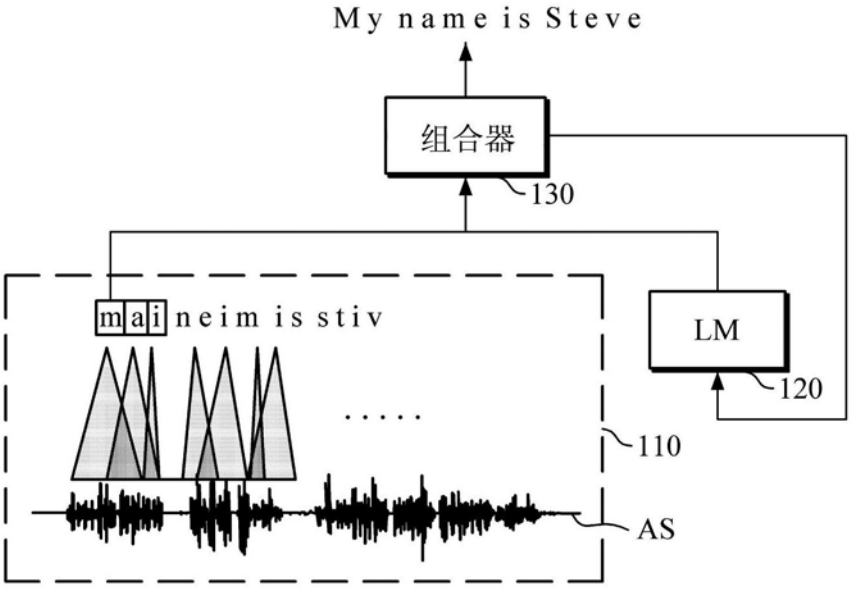

My name is Steve
组合器
130
m a i neim is stiv
LM
120
110
AS

图2

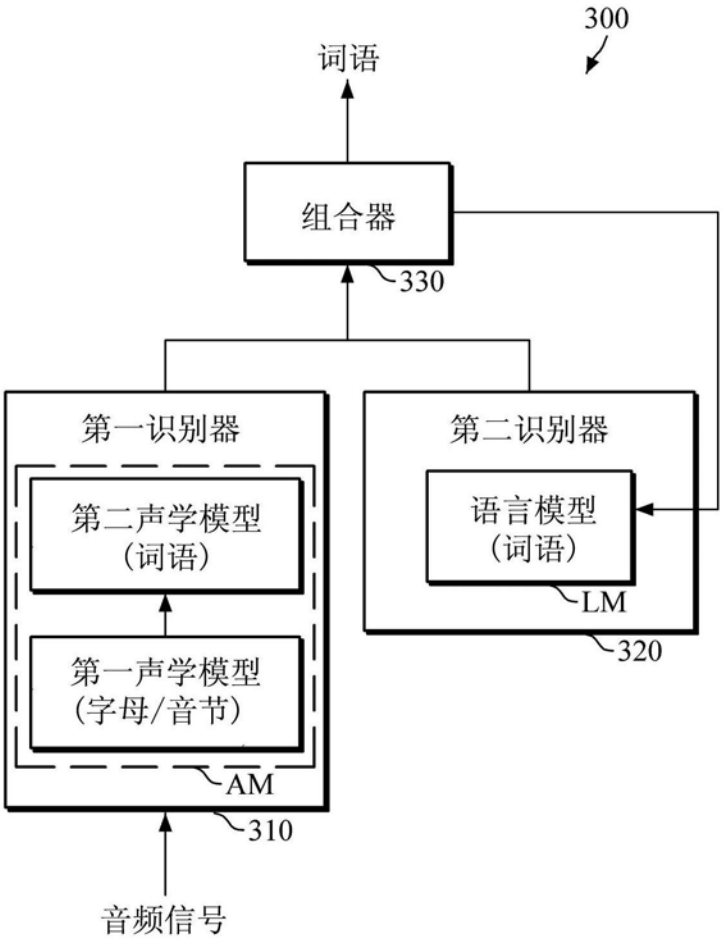

300
词语
组合器
330
第一识别器
第二声学模型
(词语)
第一声学模型
(字母/音节)
AM
310
第二识别器
语言模型
(词语)
LM
320
音频信号

图3

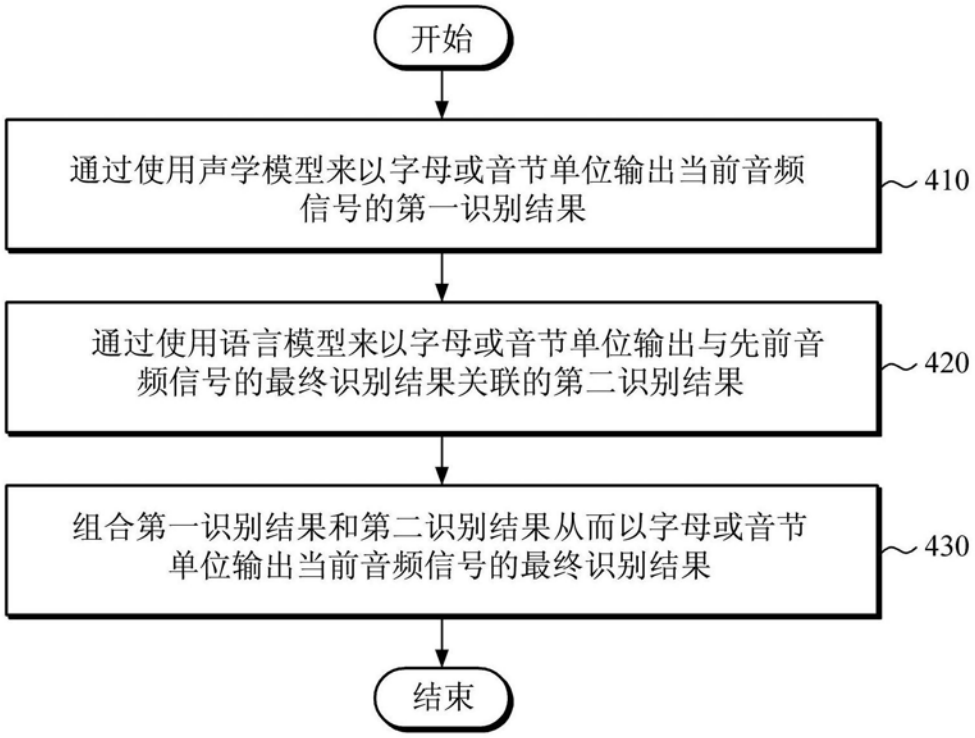

开始
通过使用声学模型来以字母或音节单位输出当前音频信号的第一识别结果
410
通过使用语言模型来以字母或音节单位输出与先前音频信号的最终识别结果关联的第二识别结果
420
组合第一识别结果和第二识别结果从而以字母或音节单位输出当前音频信号的最终识别结果
430
结束

图4

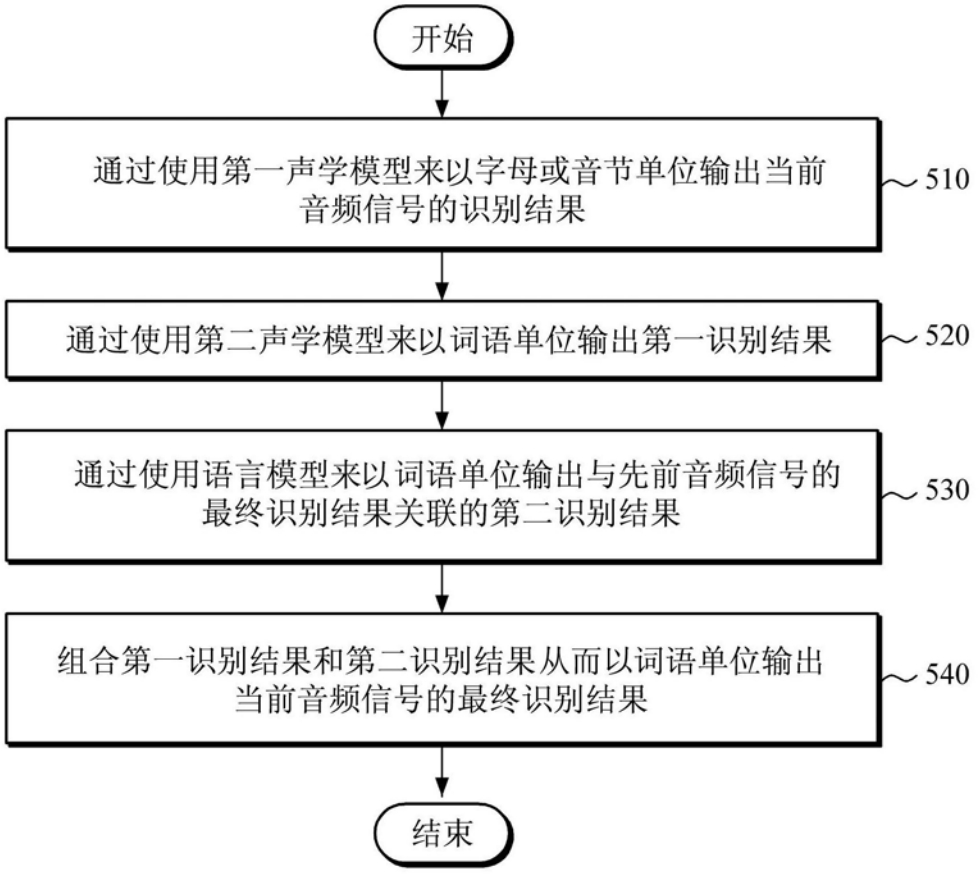

开始
通过使用第一声学模型来以字母或音节单位输出当前音频信号的识别结果
510
通过使用第二声学模型来以词语单位输出第一识别结果
520
通过使用语言模型来以词语单位输出与先前音频信号的最终识别结果关联的第二识别结果
530
组合第一识别结果和第二识别结果从而以词语单位输出当前音频信号的最终识别结果
540
结束

图5

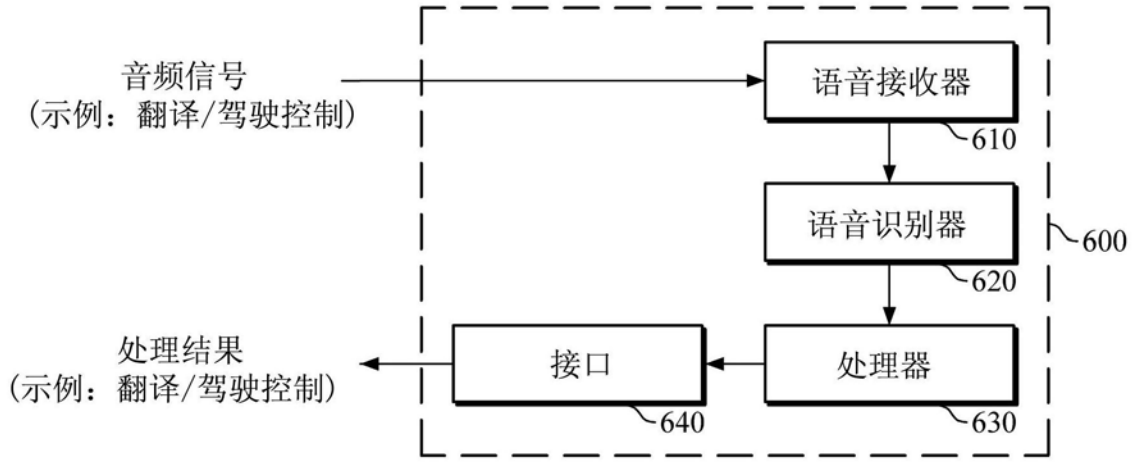

音频信号
(示例：翻译/驾驶控制)
语音接收器
610
语音识别器
620
600
处理结果
(示例：翻译/驾驶控制)
接口
640
处理器
630

图6