

流式端到端语音识别方法、装置及电子设备

本申请要求 2020 年 04 月 30 日递交的申请号为 202010366907.5、发明名称为“流式端到端语音识别方法、装置及电子设备”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

5 技术领域

本申请涉及流式端到端语音识别技术领域，特别是涉及流式端到端语音识别方法、装置及电子设备。

背景技术

语音识别技术就是让机器通过识别和理解过程把语音信号转变为相应的文本或命令的技术。其中，端到端语音识别受到了学术界和工业界越来越广泛的关注。相比于传统的基于混合系统，端到端语音识别通过一个模型联合优化声学模型，语言模型，不仅可以极大的减少系统训练复杂度，而且可以获得性能的显著提升。但是多数的端到端语音识别系统主要还是进行离线语音识别，而不能进行流式（streaming）实时语音识别。也即，能够在用户说完一句话之后进行语音识别并输出识别结果，而不能在听到声音的时候就进行识别结果的输出。

目前也有一些研究学者提出流式端到端语音识别的方案，但是效果并不明显。例如，MoCHA 模型基于注意力机制（Attention-Encoder-Decoder）的端到端语音识别系统实现流式端到端语音识别的方案。其中，在 MoCHA 中，首先可以将流式语音信息在转换成语音声学特征并输入到 Encoder（编码器）后，然后，通过 Attention 模块确定出需要进行解码输出的激活点，由 Decoder（解码器）在激活点所在的位置输出具体的识别结果（也称 token，例如，一个中文字符可以对应一个 token）。

其中，在对 Attention 模型进行训练时，通常需要以完整的一句语音作为样本，并标注出语音中的激活点所在的位置，以此完成对 Attention 模型的训练。但是，在通过 Attention 模型进行预测时，由于是进行流式的语音识别，因此，输入到模型中的是流式的语音信息，而不是对应了完整的一句话，因此，Attention 模型的方式是，为接收到的每一帧语音流计算出 Attention 系数，然后，通过与一个预置的门限值进行比较的方式，来确定激活点，例如，如果某一帧的 Attention 系数超过了门限值，即可以作为激活点，通知 Decoder 在该激活点的位置进行 token 输出。可见，在 MoCHA 方案中，训练和测试之间存在很大的不匹配性，而这种不匹配性使得 MoCHA 对于噪声的鲁棒性会比较差，从而使得基于 MoCHA 的流式端到端语音识别系统在实际任务上会面临很大的性能损失。另外，由于输入端的是持续的流式语音信号，在计算出某帧的 Attention 系数时，并不知晓未来的语音帧的情况，因此，即使当前帧的 Attention 系数大于门限值，也可能存在下一帧的 Attention 系数比当前帧更大的情况，此时，可能将下一帧作为激活点会更准确。可见，在 MoCHA 方案中，也存在激活点定位准确度比较低的问题。

因此，如何提升流式端到端语音识别系统对噪声的鲁棒性，进而提升系统性能以及准确度，成为需要本领域技术人员解决的技术问题。

发明内容

本申请提供了流式端到端语音识别方法、装置及电子设备，能够提升流式端到端语音识别系统对噪声的鲁棒性，进而提升系统性能以及准确度。

本申请提供了如下方案：

一种流式端到端语音识别方法，包括：

以帧为单位对接收到的语音流进行语音声学特征提取并进行编码；

对已完成编码的帧进行分块处理，并对同一分块中包含的需要进行编码输出的激活点数量进行预测；

根据预测结果确定需要进行解码输出的激活点所在的位置，以便解码器在所述激活点所在的位置进行解码并输出识别结果。

一种建立预测模型的方法，包括：

获得训练样本集，所述训练样本集中包括多个分块数据以及标注信息，其中，每个分块数据帧包括对语音流的多帧分别进行编码的编码结果，所述标注信息包括每个分块中包含的需要进行解码输出的激活点的数量；

将所述训练样本集输入到预测模型中进行模型的训练。

一种提供语音识别服务的方法，包括：

云服务系统接收到应用系统的调用请求后，接收所述应用系统提供的语音流；

以帧为单位对接收到的语音流进行语音声学特征提取并进行编码；

对已完成编码的帧进行分块处理，并对同一分块中包含的需要进行编码输出的激活点数量进行预测；

根据预测结果确定需要进行解码输出的激活点所在的位置，以便解码器在所述激活点所在的位置进行解码获得语音识别结果；

将所述语音识别结果返回给所述应用系统。

一种获得语音识别信息的方法，包括：

应用系统通过调用云服务系统提供的接口，向所述云服务系统提交调用请求以及待识别的语音流，由所述云服务系统以帧为单位对接收到的语音流进行语音声学特征提取并进行编码，对已完成编码的帧进行分块处理，并对同一分块中包含的需要进行编码输出的激活点数量进行预测；根据预测结果确定需要进行解码输出的激活点所在的位置后，通过解码器在所述激活点所在的位置进行解码获得语音识别结果；

接收所述云服务系统返回的语音识别结果。

一种法院自助立案实现方法，包括：

自助立案一体机设备接收语音输入的立案请求信息；

以帧为单位对接收到的语音流进行语音声学特征提取并进行编码；

对已完成编码的帧进行分块处理，并对同一分块中包含的需要进行编码输出的激活点数量进行预测；

根据预测结果确定需要进行解码输出的激活点所在的位置，以便解码器在所述激活点所在的位置进行解码并确定识别结果；

将所述识别结果录入到关联的立案信息数据库中。

一种终端设备升级方法，包括：

向终端设备提供升级建议信息；

接收到终端设备提交的升级请求后，为所述终端设备赋予使用升级后的方式进行流式语音识别的权限，所述升级后的方式进行流式语音识别包括：以帧为单位对接收到的语音流进行语音声学特征提取并进行编码，对已完成编码的帧进行分块处理，并对同一分块中包含的需要进行编码输出的激活点数量进行预测；根据预测结果确定需要进行解码输出的激活点所在的位置后，通过解码器在所述激活点所在的位置进行解码获得语音识别结果。

一种流式端到端语音识别装置，包括：

编码单元，用于以帧为单位对接收到的语音流进行语音声学特征提取并进行编码；

预测单元，用于对已完成编码的帧进行分块处理，并对同一分块中包含的需要进行编码输出的激活点数量进行预测；

激活点位置确定单元，用于根据预测结果确定需要进行解码输出的激活点所在的位置，以便解码器在所述激活点所在的位置进行解码并输出识别结果。

一种建立预测模型的装置，包括：

训练样本集获得单元，用于获得训练样本集，所述训练样本集中包括多个分块数据以及标注信息，其中，每个分块数据帧包括对语音流的多帧分别进行编码的编码结果，所述标注信息包括每个分块中包括的需要进行解码输出的激活点的数量；

输入单元，用于将所述训练样本集输入到预测模型中进行模型的训练。

一种提供语音识别服务的装置，应用于云服务系统，包括：

语音流接收单元，用于接收到应用系统的调用请求后，接收所述应用系统提供的语音流；

编码单元，用于以帧为单位对接收到的语音流进行语音声学特征提取并进行编码；

预测单元，用于对已完成编码的帧进行分块处理，并对同一分块中包含的需要进行编码输出的激活点数量进行预测；

激活点位置确定单元，用于根据预测结果确定需要进行解码输出的激活点所在的位置，以便解码器在所述激活点所在的位置进行解码获得语音识别结果；

识别结果返回单元，用于将所述语音识别结果返回给所述应用系统。

一种获得语音识别信息的装置，应用于应用系统，包括：

提交单元，用于通过调用云服务系统提供的接口，向所述云服务系统提交调用请求以及待识别的语音流，由所述云服务系统以帧为单位对接收到的语音流进行语音声学特征提取并进行编码，对已完成编码的帧进行分块处理，并对同一分块中包含的需要进行编码输出的激活点数量进行预测；根据预测结果确定需要进行解码输出的激活点所在的位置后，通过解码器在所述激活点所在的位置进行解码获得语音识别结果；

识别结果接收单元，用于接收所述云服务系统返回的语音识别结果。

一种法院自助立案实现装置，应用于自助立案一体机设备，包括：

请求接收单元，用于接收语音输入的立案请求信息；

编码单元，用于以帧为单位对接收到的语音流进行语音声学特征提取并进行编码；

预测单元，用于对已完成编码的帧进行分块处理，并对同一分块中包含的需要进行编码输出的激活点数量进行预测；

激活点位置确定单元，用于根据预测结果确定需要进行解码输出的激活点所在的位置，以便解码器在所述激活点所在的位置进行解码并确定识别结果；

信息录入单元，用于将所述识别结果录入到关联的立案信息数据库中。

一种终端设备升级装置，包括：

升级建议提供单元，用于向终端设备提供升级建议信息；

权限授予单元，用于接收到终端设备提交的升级请求后，为所述终端设备赋予使用升级后的方式进行流式语音识别的权限，所述升级后的方式进行流式语音识别包括：以帧为单位对接收到的语音流进行语音声学特征提取并进行编码，对已完成编码的帧进行分块处理，并对同一分块中包含的需要进行编码输出的激活点数量进行预测；根据预测结果确定需要进行解码输出的激活点所在的位置后，通过解码器在所述激活点所在的位置进行解码获得语音识别结果。

根据本申请提供的具体实施例，本申请公开了以下技术效果：

通过本申请实施例，在对语音流进行识别的过程中，可以针对已完成编码的帧进行分块，并对分块内包含的需要进行解码输出的激活点的数量进行预测，这样，可以根据预测结果在具体分块内确定出激活点具体所在的位置，以此指导解码器在对应的激活点位置进行解码输出识别结果即可。通过这种方式，由于不再需要将 Attention 系数与门限值进行比对的方式来确定激活点的位置，也不会受到未来帧的影响，因此，准确度能够得到提升。另外，由于对分块内包含的激活点数量的预测过程比较容易获得较高的准确度，因此，训练与预测之间的不匹配度也会比较低，从而提升流式端到端语音识别系统对噪声的鲁棒性，对系统性能产生的影响比较小。

当然，实施本申请的任一产品并不一定需要同时达到以上所述的所有优点。

附图说明

为了更清楚地说明本申请实施例或现有技术中的技术方案，下面将对实施例中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本申请的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

- 5 图 1 是本申请实施例提供的方案示意图；
 图 2 是本申请实施例提供的系统架构示意图；
 图 3 是本申请实施例提供的第一方法的流程图；
 图 4 是本申请实施例提供的第二方法的流程图；
 图 5 是本申请实施例提供的第三方法的流程图；
10 图 6 是本申请实施例提供的第四方法的流程图；
 图 7 是本申请实施例提供的第五方法的流程图；
 图 8 是本申请实施例提供的第六方法的流程图；
 图 9 是本申请实施例提供的第一装置的示意图；
 图 10 是本申请实施例提供的第二装置的示意图；
15 图 11 是本申请实施例提供的第三装置的示意图；
 图 12 是本申请实施例提供的第四装置的示意图；
 图 13 是本申请实施例提供的第五装置的示意图；
 图 14 是本申请实施例提供的第六装置的示意图；
 图 15 是本申请实施例提供的电子设备的示意图。

20 具体实施方式

下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，本领域普通技术人员所获得的所有其他实施例，都属于本申请保护的范围。

- 25 在本申请实施例中，为了提升流式端到端语音识别系统对噪声的鲁棒性，进而提升系统性能，如图 1 所示，可以在基于注意力机制的端到端语音识别系统基础上增加预测模块。该预测模块的功能是，首先可以对编码器的输出进行分块，例如，可以每 5 帧为一个分块，等等。并且，可以对每个分块中包含的需要进行解码输出的激活点（token）数量进行预测。之后，可以通过对每个分块中包含的激活点数量的预测结果，来确定激活点所在的位置，进而指导解码器在这种激活点所在的位置进行解码输出即可。例如，
30 具体实现时，由于对每个分块中包含的激活点数量进行了预测，因此，可以结合每帧对应的 Attention 系数等信息，来确定激活点的位置。具体如，假设每个分块中包括 5 帧，并且预测出某分块中一共有两个需要解码输出的激活点，则该分块中 Attention 系数最大的两帧所在的位置，就可以确定为需要进行解码输出的激活点所在的位置，之后解码器

可以在这种激活点所在的位置进行解码输出。可见，通过上述方式，使得对激活点位置的判断不再取决于人为设定的 Attention 系数门限值，而是可以以每个分块中预测出的激活点数量作为指导，在分块内，Attention 系数最大的一个或多个帧所在的位置即可作为激活点所在的位置。

- 5 当然，本申请实施例提供的方案中，对预测模块进行训练的过程，与实际利用预测模块进行测试的过程，同样可能存在不匹配之处。但是，该不匹配之处仅在于训练可以使用真实的每个分块里的预测输出数目信息（Cm），但是实际测试时候只能采用预测器的预测输出。但是由于预测每个分块里包含多少个激活点的准确性是非常高的，准确性在负责任务上也达到 95% 以上，从而使得训练和测试的不匹配性很低，从而相对于现有的
- 10 MoCHA 方案，可以获得明显的性能提升。并且，经实验表明，本申请实施例提供的流式语音识别方案与基于整句话注意力机制的离线语音识别性能基本无损。

具体实现时，本申请实施例提供的具体技术方案可以在多种应用场景中使用。例如，如图 2 所示，某云服务系统中可能会提供云语音识别服务，在该服务如果实现流式的端到端语音识别，则可以使用本申请实施例提供的方案利用流式语音识别模块来实现。

15 具体的，上述云服务系统可以提供具体的预测模型，并为用户提供云语音识别接口，多个用户可以在各自的应用系统中调用该接口，云服务系统在接收到调用之后，则可以运行相关的处理程序，实现流式的语音识别，并返回识别结果。或者，还可以在本地化的语音识别系统或者设备中使用本申请实施例提供的方案进行语音识别，例如，商场的导航机器人，法院的自助立案一体机等。

- 20 下面对本申请实施例提供的具体技术方案进行详细介绍。

实施例一

首先，该实施例一提供了一种流式端到端语音识别方法，参见图 3，包括：

S301：以帧为单位对接收到的语音流进行语音声学特征提取并进行编码；

- 在进行流式语音识别的过程中，可以以帧为单位对语音流进行语音声学特征提取，并以帧为单位进行编码，编码器会输出各帧对应的编码结果。并且，由于语音流是持续输入的，因此，对语音流进行编码的操作也可以是在持续进行。例如，假设 60ms 为一帧，则随着语音流的接收，每 60ms 的语音流作为一帧进行特征提取以及编码处理。其中，编码处理的目的是，将接收到的语音声学特征转化得到一个新的更具有区分性的高层表达，该高层表达通常可以以向量的形式存在。因此，编码器具体可以是一个多层的神经网络，神经网络的选择有多种，比如 DFSMN，CNN，BLSTM，Transformer 等等。
- 25
- 30

S302：对已完成编码的帧进行分块处理，并对同一分块中包含的需要进行编码输出的激活点数量进行预测；

在本申请实施例中，在得到编码结果后，首先可以进行分块处理，并以分块为单位进行激活点数量的预测。其中，由于一个分块可能会包括多帧，因此，在一种具体的实

现方式下，在完成对每一帧语音流的编码处理后，可以首先对编码结果进行缓存，缓存的编码结果帧数每达到一个分块对应的帧数时，便可以将当前缓存的各帧编码结果确定为一个分块，预测模块便可以对该分块中包含的需要进行编码输出的激活点数量进行预测。例如，每 5 帧对应一个分块，则可以在编码器每完成 5 帧语音流的编码后，进行一次预测处理。在可选的实施方式中，在完成对一个分块的预测后，还可以将该分块对应的各帧编码结果从缓存中进行删除。

当然，由于具体的注意力机制系统中，各帧的编码结果通常还可以用于计算 Attention 系数，并且还可以与 Attention 系数进行加权求和后，作为解码器的输入提供给解码器进行运算。因此，在具体实现时，为了避免预测模块与 Attention 模块之间在数据上相互影响或者冲突，编码器的输出可以分别提供给预测模块以及 Attention 模块，并且，预测模块与 Attention 模块可以对应不同的缓存空间，各自可以在各自的缓存空间内对解码结果数据进行处理，以分别获得对激活点数量的预测结果，以及对 Attention 系数的计算等操作。

其中，由于需要对解码结果进行分块，因此，使得语音识别的过程可能会产生一定的延迟，具体延迟的大小取决于分块的大小。例如，每 5 帧一个分块，则延迟时间就是 5 帧的时间，等等。具体实现时，可以根据系统可容忍的延迟时间，确定分块的大小。例如，在极端情况下，可以每一帧作为一个分块，等等。

关于具体的预测模块，可以通过预先训练的模型来实现。为了完成对模型的训练，可以准备训练样本集，其中包括语音流对应的编码结果，按照一定的大小进行分块，并且可以对每个分块中包含的需要输出的激活点数量进行标注。具体在对预测模型进行训练时，可以将上述样本信息以及标注信息输入到初始化的模型中，通过多轮迭代实现对模型参数的逐渐优化，直到算法收敛时结束训练。其中，如果具体的预测模型是使用神经网络等深度学习模型来实现，则对参数进行调整的过程，具体可以是对深度学习模型中各层的权重进行调整的过程。

在完成对预测模型的训练之后，只要将同一分块包含的编码结果输入到该预测模型中，该预测模型即可输出该分块中需要进行解码输出的激活点的数量信息。

需要说明的是，由于对预测模型进行训练时，使用的是真实的每个分块里的预测输出数目信息（Cm），但是实际测试时候只能采用预测器的预测输出。但是由于由于预测每个分块里包含多少个激活点的准确性是可以非常高的，因此，相对于 MoCHA 系统而言，训练和测试之间的不匹配性会很低，基本不会影响识别性能。

另外需要说明的是，对于语音流而言，其中的一个字符在语音流中存在的平均持续时间通常在 200ms 左右，也即，某用户在说话过程中，每个字的发音可能会持续 200ms 左右（当然，不同的人由于语速不同，实际的持续时间也可能会有所差别）。如果 60ms 为一帧，则同一个建模单元（例如对应中文中的一个字，或者英文中的一个单词等）的

发音可能分布在多个连续的帧中。但实际上同一个建模单元通常只需要在其中一帧上进行解码输出，并为该帧关联上周边帧的特征即可。而本申请实施例中，将多帧分成一个分块，因此，可能存在同一个建模单元所在的帧被划分到多个不同的分块中的情况。因此，为了避免同一个建模单元对应的不同帧在不同的分块中均被识别为激活点，可以在对预测模型进行训练时便考虑到该问题，也即，可以存在对应上述情况的训练样本以及对应的标注信息。这样完成训练后的模型在具体进行测试时，便可以正确应对上述情况的发生。

S303：根据预测结果确定需要进行解码输出的激活点所在的位置，以便解码器在所述激活点所在的位置进行解码并输出识别结果。

在预测出每个分块中包含的需要进行解码输出的激活点的数量后，还可以进一步确定出激活点所在的位置，以便解码器在所述激活点所在的位置进行解码并输出识别结果。其中，具体实现时，如果一个分块中仅包括一帧语音流对应的编码结果，则在对每个分块中包含的激活点数量进行预测时，预测结果要么为 0，要么为 1，因此，对分块中包含的激活点数量进行预测的问题，就演变成了对每个分块中是否包含需要进行解码输出的激活点进行预测的问题。也即，具体的预测结果可以是，当前分块中是否包含需要进行编码输出的激活点。在这种情况下，可以直接将包含所述激活点的分块所在的位置确定为所述激活点所在的位置。例如，具体的预测结果可以如表 1 所示：

表 1

帧序列	块序列	激活点数目预测结果	是否为激活点位置
0	0	0	否
1	1	0	否
2	2	0	否
3	3	1	是
4	4	0	否
.....			

可见，在以一帧为单位进行分块的情况下，对分块完成激活点数量的预测后，就可以直接确定出具体激活点所在的位置。或者，在另一种方式下，由于 Attention 系数也可以从一定程度上反映出某帧是否为激活点，或者，某帧属于激活点的概率。因此，在具体实现时，还可以分别确定各帧编码结果的注意力 Attention 系数，该 Attention 系数用于描述对应帧需要进行解码输出的概率。然后，可以根据所述 Attention 系数对所述预测结果进行验证。例如，具体实现时，可以预先设定 Attention 系数的阈值，如果某帧所在的分块预测结果显示该帧属于激活点，并且该帧的 Attention 系数也大于上述阈值，则可以

进一步提升该帧数与激活点的概率，等等。反之，如果分块预测结果显示某帧属于激活点，但是，计算出的 Attention 系数却很低，则预测模块还可以通过调整策略等方式来重新预测，例如，可以结合更多周边帧的特征进行预测，等等。当然，在该方式中，仍然用到了 Attention 系数的阈值，但是，由于仅仅是用于对分块预测结果进行验证，因此，对整个系统性能的影响不大。

另一种方式下，同一分块中可以包括多帧语音流对应的编码结果。此时，预测模块只能预测出同一分块中包括几个激活点，但是不能直接确定激活点具体在该分块中哪一帧处。因此，在具体实现，还可以结合各帧的 Attention 系数来对激活点的位置进行确定。具体的，首先可以分别确定各帧编码结果的注意力 Attention 系数。然后，可以根据分块中包含的激活点数量，将该分块包含的各帧编码结果中 Attention 系数最高的对应数量的帧所在的位置，确定为所述激活点所在的位置。也就是说，假设预测出某分块中包括两个激活点，则可以将该分块包含的各帧中 Attention 系数最高的两帧所在的位置，确定为两个激活点所在的位置。例如，在一个例子中，具体的激活点数量预测结果， Attention 系数情况，以及确定出的激活点的位置信息，可以如表 2 所示：

15

表 2

帧序列	分块序列	分块内预测出的 激活点数量	对应帧的 Attention 系数	对应帧是否属 于激活点
0	0	1	0.01	否
1			0.22	否
2			0.78	否
3			0.95	是
4			0.75	否
5	1	2	0.63	否
6			0.88	是
7			0.72	否
8			0.58	否
9			0.93	是
.....				

在上述表格中，是每 5 帧作为一个分块，则第 0 到第 4 帧被划分为一个分块，第 5 到第 9 帧被划分为一个分块，以此类推。其中，假设预测模块预测出第一个分块中包含 1 个激活点，并且计算出第 0 到第 4 帧各自的 Attention 系数分别为 0.01、0.22、0.78、0.95、0.75，此时，便可以将上述第 0 到第 4 帧中， Attention 系数最高的一帧所在的位置确定为激活点所在的位置，也即第 3 帧所在的位置为激活点，该分块中其他各帧不属于激活点，不需要进行解码输出。类似的，假设预测模块预测出第二个分块中包含 2 个激活点，

并且计算出第 4 到第 9 帧各自的 Attention 系数分别为 0.63、0.88、0.72、0.58、0.93，此时，便可以将上述第 5 到第 9 帧中，Attention 系数最高的两帧所在的位置确定为激活点所在的位置，也即第 6 帧、第 9 帧所在的位置为激活点，该分块中其他各帧不属于激活点，不需要进行解码输出。

5 可见，通过本申请实施例中的所述的方式，在对激活点的位置进行判断时，不需要将 Attention 系数与预置的门限值进行比对，而是在预测出分块中存在的激活点数量的情况下，在该分块内包括的各帧之间进行 Attention 系数的比对，将其中对应数量的较大者所在的帧作为激活点所在的位置。这样，由于训练以及测试时，都可以统一按照上述方式进行，因此，可以提高训练与测试之间的匹配度，降低对系统性能的影响。另外，由于
10 具体的 Attention 系数比对操作可以在同一分块内进行，不会受到未来帧的影响，因此，确定出的激活点位置的准确的也会比较高。

需要说明的是，具体实现时，具体分块的大小可以是预先设定的，或者，还可以预先设定初始值，在具体进行测试的过程中还可以根据实际语音流的情况进行动态调整。具体的，如前文所述，不同用户由于语速不同，同样时间长度内输入的建模单元的数量
15 （中文文字的数量等）以及密度可能是不同的。为此，在具体实现时，还可以根据已预测出的激活点的出现频度，对分块的大小进行自适应调节。例如，如果某此预测过程中发现激活点频度较高，则可以缩小分块，以缩短延迟，反之，可以扩大分块，从而使得系统的识别延迟能够跟随输入者的语速而变化。

总之，通过本申请实施例，在对语音流进行识别的过程中，可以针对已完成编码的
20 帧进行分块，并对分块内包含的需要进行解码输出的激活点的数量进行预测，这样，可以根据预测结果在具体分块内确定出激活点具体所在的位置，以此指导解码器在对应的激活点位置进行解码输出识别结果即可。通过这种方式，由于不再需要将 Attention 系数与门限值进行比对的方式来确定激活点的位置，也不会受到未来帧的影响，因此，准确度能够得到提升。另外，由于对分块内包含的激活点数量的预测过程比较容易获得较高的
25 准确度，因此，训练与预测之间的不匹配度也会比较低，从而提升流式端到端语音识别系统对噪声的鲁棒性，对系统性能产生的影响比较小。

实施例二

该实施例二提供了一种建立预测模型的方法，参见图 4，该方法具体可以包括：

S401：获得训练样本集，所述训练样本集中包括多个分块数据以及标注信息，其中，
30 每个分块数据帧包括对语音流的多帧分别进行编码的编码结果，所述标注信息包括每个分块中包括的需要进行解码输出的激活点的数量；

S402：将所述训练样本集输入到预测模型中进行模型的训练。

具体实现时，所述训练样本集中可以包括同一建模单元对应的多帧语音流被划分到不同的分块中的情况，这样可以是的同一文字等建模单元被划分到多个不同分块的情况

得到训练，从而在测试过程中遇到同样情况时能够获得准确的预测结果。

实施例三

该实施例三是针对本申请实施例提供的方案在云服务系统中应用时的场景进行介绍，具体的，该实施例三首先从云服务端的角度，提供了一种提供语音识别服务的方法，

5 参见图 5，该方法具体可以包括：

S501：云服务系统接收到应用系统的调用请求后，接收所述应用系统提供的语音流；

S502：以帧为单位对接收到的语音流进行语音声学特征提取并进行编码；

S503：对已完成编码的帧进行分块处理，并对同一分块中包含的需要进行编码输出的激活点数量进行预测；

10 S504：根据预测结果确定需要进行解码输出的激活点所在的位置，以便解码器在所述激活点所在的位置进行解码获得语音识别结果；

S505：将所述语音识别结果返回给所述应用系统。

实施例四

该实施例四是与实施例三相对应的，从应用系统的角度，提供了一种获得语音识别
15 信息的方法，参见图 6，该方法具体可以包括：

S601：应用系统通过调用云服务系统提供的接口，向所述云服务系统提交调用请求以及待识别的语音流，由所述云服务系统以帧为单位对接收到的语音流进行语音声学特征提取并进行编码，对已完成编码的帧进行分块处理，并对同一分块中包含的需要进行编码输出的激活点数量进行预测；根据预测结果确定需要进行解码输出的激活点所在
20 位置后，通过解码器在所述激活点所在的位置进行解码获得语音识别结果；

S602：接收所述云服务系统返回的语音识别结果。

实施例五

该实施例五是针对本申请实施例提供的方案在法院的自助立案一体机中的应用场景进行介绍，具体的，参见图 7，该实施例五提供了一种法院自助立案实现方法，该方法
25 可以包括：

S701：自助立案一体机设备接收语音输入的立案请求信息；

S702：以帧为单位对接收到的语音流进行语音声学特征提取并进行编码；

S703：对已完成编码的帧进行分块处理，并对同一分块中包含的需要进行编码输出的激活点数量进行预测；

30 S704：根据预测结果确定需要进行解码输出的激活点所在的位置，以便解码器在所述激活点所在的位置进行解码并确定识别结果；

S705：将所述识别结果录入到关联的立案信息数据库中。

实施例六

前述各实施例对本申请实施例提供的流式语音识别方法以及在具体场景中的应用进

行了介绍。具体实现时，对于在智能音箱等硬件设备中的应用场景，由于用户在购买具体的硬件设备时，可能尚未实现本申请实施例提供的功能，以使得比较“老”的硬件设备只能通过传统的方式进行语音识别。而在本申请实施例中，为了使得这部分“老”的硬件设备也能够通过新的方式进行流式语音识别，以提升用户的体验，可以为终端设备提供升级方案。例如，具体实现时，可以在服务端提供具体的流式语音识别的处理流程，具体的硬件设备侧只需要将采集到的用户语音流提交到服务端即可。在这种情况下，具体的语音识别过程中所需要用到的模型等只需要在服务端进行保存即可，终端设备侧通常不需要进行硬件上的改进即可实现升级。当然，由于具体在进行流式语音识别的过程中，通常会涉及到对用户数据的采集并提交到服务端，因此，具体实现时，可以首先通过服务端向具体的硬件设备推送可以进行升级的建议，如果用户需要对设备进行升级，则可以通过输入语音等方式对自己的需求进行表达，之后，便可以将具体的升级请求提交到服务端，由服务端对升级请求进行处理。其中，具体实现时，服务端还可以对具体硬件设备的状态进行判断，例如，关联的用户是否已经为获得升级后的服务而付出相应的资源，等等，如果是，则可以为其赋予通过升级后的方式进行流式语音识别的权限。这样，该硬件设备后续在与用户进行对话的过程中，便可以通过本申请实施例提供的方式进行流式语音识别。具体的，具体的流式语音识别功能可以由服务端完成，或者，对于硬件设备自身硬件资源能够支持的情况下，还可以直接将升级后的识别模型推送到具体的硬件设备，由硬件设备在本地完成流式语音识别，等等。

另外，对于具体的模型保存在服务端的情形，还可以提供“开关”功能，以使得用户可以仅在必要时使用上述功能，以达到节省资源等目的。例如，当用户仅需要与在家庭场景中使用，由于对语音识别的精度等要求不高，则可以通过发出语音指令等方式来提交关闭上述高级功能（也即本申请实施例提供的识别方式）的请求，之后，服务端可以为该用户暂时关闭该功能，如果涉及到计费等，则也可以触发计费停止。硬件设备可以退回到原来的方式进行流式语音识别，甚至等到用户说完一句话之后再进行识别可能也是能够接受的。后续如果用户需要在工作场景中使用该硬件设备，则还可以再重新开启本申请实施例中提供的高级功能，等等。

具体的，本申请实施例六提供了一种设备升级方法，参见图 8，该方法可以包括：

S801：向终端设备提供升级建议信息；

S802：接收到终端设备提交的升级请求后，为所述终端设备赋予使用升级后的方式进行流式语音识别的权限，所述升级后的方式进行流式语音识别包括：以帧为单位对接收到的语音流进行语音声学特征提取并进行编码，对已完成编码的帧进行分块处理，并对同一分块中包含的需要进行编码输出的激活点数量进行预测；根据预测结果确定需要进行解码输出的激活点所在的位置后，通过解码器在所述激活点所在的位置进行解码获得语音识别结果。

其中，所述终端设备具体可以包括智能音箱设备等。

具体实现时，还可以根据所述终端设备提交的降级请求，为所述终端设备关闭所述使用升级后的方式进行流式语音识别的权限。

其中，关于前述实施例二至实施例六中的未详述部分，可以参见实施例一中的记载，
5 这里不再赘述。

需要说明的是，本申请实施例中可能会涉及到对用户数据的使用，在实际应用中，可以在符合所在国的适用法律法规要求的情况下（例如，用户明确同意，对用户切实通知，等），在适用法律法规允许的范围内在本文描述的方案中使用用户特定的个人数据。

与实施例一相对应，本申请实施例还提供了一种流式端到端语音识别装置，参见图
10 9，该装置具体可以包括：

编码单元 901，用于以帧为单位对接收到的语音流进行语音声学特征提取并进行编码；

预测单元 902，用于对已完成编码的帧进行分块处理，并对同一分块中包含的需要进行编码输出的激活点数量进行预测；

15 激活点位置确定单元 903，用于根据预测结果确定需要进行解码输出的激活点所在的位置，以便解码器在所述激活点所在的位置进行解码并输出识别结果。

其中，所述分块中包括一帧语音流对应的编码结果；

所述预测结果包括：当前分块中是否包含需要进行编码输出的激活点；

所述激活点位置确定单元具体可以用于：

20 将包含所述激活点的分块所在的位置确定为所述激活点所在的位置。

此时，该装置还可以包括：

注意力系数确定单元，用于分别确定各帧编码结果的注意力 Attention 系数；所述 Attention 系数用于描述对应帧需要进行解码输出的概率；

验证单元，用于根据所述 Attention 系数对所述预测结果进行验证。

25 或者，所述分块中包括多帧语音流对应的编码结果；

此时，该装置还可以包括：

注意力确定单元，用于分别确定各帧编码结果的注意力 Attention 系数；所述 Attention 系数用于描述对应帧需要进行解码输出的概率；

根据激活点位置确定单元具体可以用于：

30 将同一分块中各帧的 Attention 系数进行比对，并按照大小进行排序；

根据分块中包含的激活点数量，将该分块包含的各帧编码结果中 Attention 系数最高的对应数量的帧所在的位置，确定为所述激活点所在的位置。

此时，该装置还可以包括：

分块调节单元，用于根据已预测出的激活点的出现频度，对分块的大小进行自适应

调节。

其中，所述预测单元具体可以包括：

缓存子单元，用于将所述编码结果进行缓存；

分块确定子单元，用于在加入缓存的编码结果帧数达到分块的大小时，将当前缓存
5 的各帧编码结果确定为一个分块。

具体的，该装置还可以包括：

在完成对该分块的预测处理后，将该分块的各帧编码结果从缓存中删除。

与实施例二相对应，本申请实施例还提供了一种建立预测模型的装置，参见图 10，
该装置包括：

10 训练样本集获得单元 1001，用于获得训练样本集，所述训练样本集中包括多个分块
数据以及标注信息，其中，每个分块数据帧包括对语音流的多帧分别进行编码的编码结
果，所述标注信息包括每个分块中包括的需要进行解码输出的激活点的数量；

输入单元 1002，用于将所述训练样本集输入到预测模型中进行模型的训练。

其中，所述训练样本集中包括同一建模单元对应的多帧语音流被划分到不同的分块
15 中的情况。

与实施例三相对应，本申请实施例还提供了一种提供语音识别服务的装置，参见图
11，该装置应用于云服务系统，包括：

语音流接收单元 1101，用于接收到应用系统的调用请求后，接收所述应用系统提供
的语音流；

20 编码单元 1102，用于以帧为单位对接收到的语音流进行语音声学特征提取并进行编
码；

预测单元 1103，用于对已完成编码的帧进行分块处理，并对同一分块中包含的需要
进行编码输出的激活点数量进行预测；

25 激活点位置确定单元 1104，用于根据预测结果确定需要进行解码输出的激活点所在
的位置，以便解码器在所述激活点所在的位置进行解码获得语音识别结果；

识别结果返回单元 1105，用于将所述语音识别结果返回给所述应用系统。

与实施例四相对应，本申请实施例还提供了一种获得语音识别信息的装置，参见图
12，该装置应用于应用系统，包括：

30 提交单元 1201，用于通过调用云服务系统提供的接口，向所述云服务系统提交调用
请求以及待识别的语音流，由所述云服务系统以帧为单位对接收到的语音流进行语音声
学特征提取并进行编码，对已完成编码的帧进行分块处理，并对同一分块中包含的需要
进行编码输出的激活点数量进行预测；根据预测结果确定需要进行解码输出的激活点所
在的位置后，通过解码器在所述激活点所在的位置进行解码获得语音识别结果；

识别结果接收单元 1202，用于接收所述云服务系统返回的语音识别结果。

与实施例五相对应，本申请实施例还提供了一种法院自助立案实现装置，参见图 13，该装置应用于自助立案一体机设备，包括：

请求接收单元 1301，用于接收语音输入的立案请求信息；

5 编码单元 1302，用于以帧为单位对接收到的语音流进行语音声学特征提取并进行编码；

预测单元 1303，用于对已完成编码的帧进行分块处理，并对同一分块中包含的需要进行编码输出的激活点数量进行预测；

激活点位置确定单元 1304，用于根据预测结果确定需要进行解码输出的激活点所在的位置，以便解码器在所述激活点所在的位置进行解码并确定识别结果；

10 信息录入单元 1305，用于将所述识别结果录入到关联的立案信息数据库中。

与实施例六相对应，本申请实施例还提供了一种终端设备升级装置，参见图 14，该装置可以包括：

升级建议提供单元 1401，用于向终端设备提供升级建议信息；

15 权限授予单元 1402，用于接收到终端设备提交的升级请求后，为所述终端设备赋予使用升级后的方式进行流式语音识别的权限，所述升级后的方式进行流式语音识别包括：以帧为单位对接收到的语音流进行语音声学特征提取并进行编码，对已完成编码的帧进行分块处理，并对同一分块中包含的需要进行编码输出的激活点数量进行预测；根据预测结果确定需要进行解码输出的激活点所在的位置后，通过解码器在所述激活点所在的位置进行解码获得语音识别结果。

20 另外，本申请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，该程序被处理器执行时实现前述方法实施例中任一项所述的方法的步骤。

以及一种电子设备，包括：

25 一个或多个处理器；以及与所述一个或多个处理器关联的存储器，所述存储器用于存储程序指令，所述程序指令在被所述一个或多个处理器读取执行时，执行前述方法实施例中任一项所述的方法的步骤。

其中，图 15 示例性的展示出了电子设备的架构，例如，设备 1500 可以是移动电话，计算机，数字广播终端，消息收发设备，游戏控制台，平板设备，医疗设备，健身设备，个人数字助理，飞行器等。

30 参照图 15，设备 1500 可以包括以下一个或多个组件：处理组件 1502，存储器 1504，电源组件 1506，多媒体组件 1508，音频组件 1510，输入输出（I/O）接口 1512，传感器 1514，以及通信组件 1516。

处理组件 1502 通常控制设备 1500 的整体操作，诸如与显示，电话呼叫，数据通信，相机操作和记录操作相关联的操作。处理组件 1502 可以包括一个或多个处理器 1520 来执行指令，以完成本公开技术方案提供的方法的全部或部分步骤。此外，处理组件 1502

可以包括一个或多个模块，便于处理组件 1502 和其他组件之间的交互。例如，处理组件 1502 可以包括多媒体模块，以方便多媒体组件 1508 和处理组件 1502 之间的交互。

存储器 1504 被配置为存储各种类型的数据以支持在设备 1500 的操作。这些数据的示例包括用于在设备 1500 上操作的任何应用程序或方法的指令，联系人数据，电话簿数据，消息，图片，视频等。存储器 1504 可以由任何类型的易失性或非易失性存储设备或者它们的组合实现，如静态随机存取存储器（SRAM），电可擦除可编程只读存储器（EEPROM），可擦除可编程只读存储器（EPROM），可编程只读存储器（PROM），只读存储器（ROM），磁存储器，快闪存储器，磁盘或光盘。

电源组件 1506 为设备 1500 的各种组件提供电力。电源组件 1506 可以包括电源管理系统，一个或多个电源，及其他与为设备 1500 生成、管理和分配电力相关联的组件。

多媒体组件 1508 包括在设备 1500 和用户之间的提供一个输出接口的屏幕。在一些实施例中，屏幕可以包括液晶显示器（LCD）和触摸面板（TP）。如果屏幕包括触摸面板，屏幕可以被实现为触摸屏，以接收来自用户的输入信号。触摸面板包括一个或多个触摸传感器以感测触摸、滑动和触摸面板上的手势。触摸传感器可以不仅感测触摸或滑动动作的边界，而且还检测与触摸或滑动操作相关的持续时间和压力。在一些实施例中，多媒体组件 1508 包括一个前置摄像头和/或后置摄像头。当设备 1500 处于操作模式，如拍摄模式或视频模式时，前置摄像头和/或后置摄像头可以接收外部的多媒体数据。每个前置摄像头和后置摄像头可以是一个固定的光学透镜系统或具有焦距和光学变焦能力。

音频组件 1510 被配置为输出和/或输入音频信号。例如，音频组件 1510 包括一个麦克风（MIC），当设备 1500 处于操作模式，如呼叫模式、记录模式和语音识别模式时，麦克风被配置为接收外部音频信号。所接收的音频信号可以被进一步存储在存储器 1504 或经由通信组件 1516 发送。在一些实施例中，音频组件 1510 还包括一个扬声器，用于输出音频信号。

输入输出接口 1512 为处理组件 1502 和外围接口模块之间提供接口，上述外围接口模块可以是键盘，点击轮，按钮等。这些按钮可包括但不限于：主页按钮、音量按钮、启动按钮和锁定按钮。

传感器 1514 包括一个或多个传感器，用于为设备 1500 提供各个方面的状态评估。例如，传感器 1514 可以检测到设备 1500 的打开/关闭状态，组件的相对定位，例如所述组件为设备 1500 的显示器和小键盘，传感器 1514 还可以检测设备 1500 或设备 1500 一个组件的位置改变，用户与设备 1500 接触的存在或不存在，设备 1500 方位或加速/减速和设备 1500 的温度变化。传感器 1514 可以包括接近传感器，被配置用来在没有任何的物理接触时检测附近物体的存在。传感器 1514 还可以包括光传感器，如 CMOS 或 CCD 图像传感器，用于在成像应用中使用。在一些实施例中，该传感器 1514 还可以包括加速度传感器，陀螺仪传感器，磁传感器，压力传感器或温度传感器。

通信组件 1516 被配置为便于设备 1500 和其他设备之间有线或无线方式的通信。设备 1500 可以接入基于通信标准的无线网络，如 WiFi，或 2G、3G、4G/LTE、5G 等移动通信网络。在一个示例性实施例中，通信组件 1516 经由广播信道接收来自外部广播管理系统的广播信号或广播相关信息。在一个示例性实施例中，所述通信组件 1516 还包括近场通信（NFC）模块，以促进短程通信。例如，在 NFC 模块可基于射频识别（RFID）技术，红外数据协会（IrDA）技术，超宽带（UWB）技术，蓝牙（BT）技术和其他技术来实现。

在示例性实施例中，设备 1500 可以被一个或多个应用专用集成电路（ASIC）、数字信号处理器（DSP）、数字信号处理设备（DSPD）、可编程逻辑器件（PLD）、现场可编程门阵列（FPGA）、控制器、微控制器、微处理器或其他电子元件实现，用于执行上述方法。

在示例性实施例中，还提供了一种包括指令的非临时性计算机可读存储介质，例如包括指令的存储器 1504，上述指令可由设备 1500 的处理器 1520 执行以完成本公开技术方案提供的方法。例如，所述非临时性计算机可读存储介质可以是 ROM、随机存取存储器（RAM）、CD-ROM、磁带、软盘和光数据存储设备等。

通过以上的实施方式的描述可知，本领域的技术人员可以清楚地了解到本申请可借助软件加必需的通用硬件平台的方式来实现。基于这样的理解，本申请的技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来，该计算机软件产品可以存储在存储介质中，如 ROM/RAM、磁碟、光盘等，包括若干指令用以使得一台计算机设备（可以是个人计算机，服务器，或者网络设备等）执行本申请各个实施例或者实施例的某些部分所述的方法。

本说明书中的各个实施例均采用递进的方式描述，各个实施例之间相同相似的部分互相参见即可，每个实施例重点说明的都是与其他实施例的不同之处。尤其，对于系统或系统实施例而言，由于其基本相似于方法实施例，所以描述得比较简单，相关之处参见方法实施例的部分说明即可。以上所描述的系统及系统实施例仅仅是示意性的，其中所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部模块来实现本实施例方案的目的。本领域普通技术人员在不付出创造性劳动的情况下，即可以理解并实施。

以上对本申请所提供的流式端到端语音识别方法、装置及电子设备，进行了详细介绍，本文中应用了具体个例对本申请的原理及实施方式进行了阐述，以上实施例的说明只是用于帮助理解本申请的方法及其核心思想；同时，对于本领域的一般技术人员，依据本申请的思想，在具体实施方式及应用范围上均会有改变之处。综上所述，本说明书内容不应理解为对本申请的限制。