

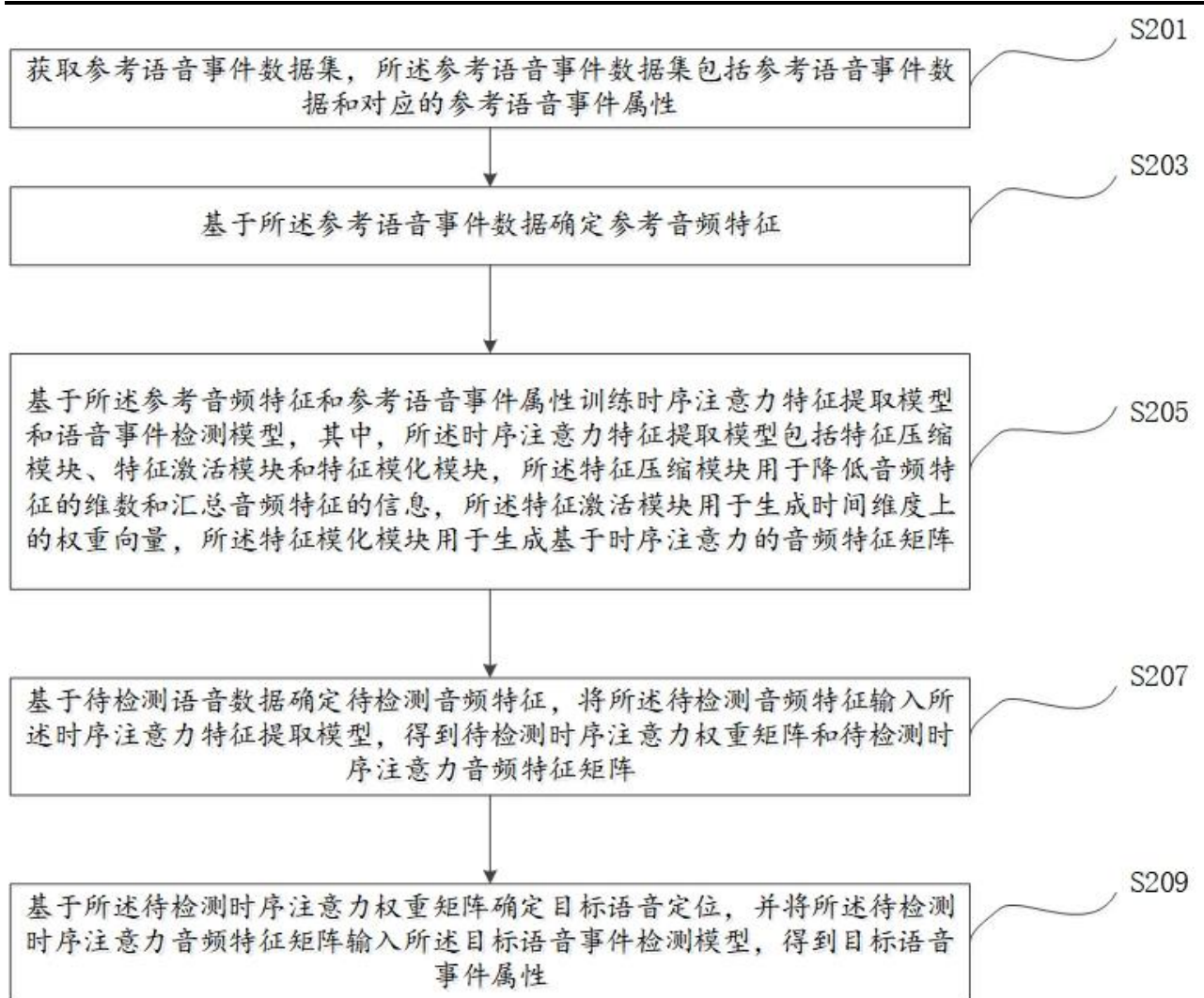
说明书摘要

本申请涉及一种基于时序注意力的语音事件检测和定位方法及装置。所述方法包括：获取参考语音事件数据集，参考语音事件数据集包括参考语音事件数据和对应的参考语音事件属性，基于参考语音事件数据确定参考音频特征，

5 基于参考音频特征和参考语音事件属性训练时序注意力特征提取模型和语音事件检测模型，基于待检测语音数据确定待检测音频特征，将待检测音频特征输入时序注意力特征提取模型，得到待检测时序注意力权重矩阵和待检测时序注意力音频特征矩阵，基于待检测时序注意力权重矩阵确定目标语音定位，并将待检测时序注意力音频特征矩阵输入语音事件检测模型，得到目标语音事件属

10 性，提高了语音事件检测的准确性以及实用性。

摘 要 附 图



权 利 要 求 书

1. 一种基于时序注意力的语音事件检测和定位方法，其特征在于，所述方法包括：

5 获取参考语音事件数据集，所述参考语音事件数据集包括参考语音事件数据和对应的参考语音事件属性；

基于所述参考语音事件数据确定参考音频特征；

10 基于所述参考音频特征和参考语音事件属性训练时序注意力特征提取模型和语音事件检测模型，其中，所述时序注意力特征提取模型包括特征压缩模块、特征激活模块和特征模化模块，所述特征压缩模块用于降低音频特征的维数和汇总音频特征的信息，所述特征激活模块用于生成时间维度上的权重向量，所述特征模化模块用于生成基于时序注意力的音频特征矩阵；

基于待检测语音数据确定待检测音频特征，将所述待检测音频特征输入所述时序注意力特征提取模型，得到待检测时序注意力权重矩阵和待检测时序注意力音频特征矩阵；

15 基于所述待检测时序注意力权重矩阵确定目标语音定位，并将所述待检测时序注意力音频特征矩阵输入所述语音事件检测模型，得到目标语音事件属性。

2. 根据权利要求 1 所述的基于时序注意力的语音事件检测和定位方法，其特征在于，所述基于所述参考语音事件数据确定参考音频特征包括：

20 搭建基础特征提取模型，基于公开语音数据集训练所述基础特征提取模型；将所述参考语音事件数据输入所述基础特征提取模型，得到参考音频特征。

3. 根据权利要求 1 所述的基于时序注意力的语音事件检测和定位方法，其特征在于，所述基于所述参考音频特征和参考语音事件属性训练时序注意力特征提取模型和语音事件检测模型包括：

将所述参考音频特征输入所述特征压缩模块，得到压缩音频特征矩阵；

25 将所述压缩音频特征矩阵输入所述特征激活模块，得到压缩时序注意力权重矩阵；

将所述压缩时序注意力权重矩阵输入所述特征模化模块，得到参考时序注

意力音频特征矩阵。

4. 根据权利要求3所述的基于时序注意力的语音事件检测和定位方法，其特征在于，所述语音事件检测模型为基于多任务学习的语音事件类型检测模型，所述基于所述参考音频特征和参考语音事件属性训练时序注意力特征提取模型和语音事件检测模型还包括：

将所述参考时序注意力音频特征矩阵输入语音事件检测模型，得到初始语音事件属性；

基于所述初始语音事件属性和参考语音事件属性确定损失函数，基于所述损失函数调整所述语音事件检测模型的参数。

5. 根据权利要求4所述的基于时序注意力的语音事件检测和定位方法，其特征在于，所述基于所述初始语音事件属性和参考语音事件属性确定损失函数如下：

$$L_{\text{main}} = - \sum_{i=1}^M \sum_{j=1}^K s_{ij} \times \log(\hat{s}_{ij})$$

其中，M为样本个数， s_{ij} 代表第i个样本属于第j种语音事件类型的真实概率， \hat{s}_{ij} 代表第i个样本属于第j种语音事件类型的预测概率；

所述辅助任务分类器执行二分类判决任务，其损失函数如下：

$$L_{\text{aux}} = - \sum_{i=1}^M \sum_{j=1}^2 y_{ij} \times \log(\hat{y}_{ij})$$

其中，M为样本个数， y_{ij} 代表第i个样本属于第j种特定语音属性的真实概率， \hat{y}_{ij} 代表第i个样本属于第j种特定语音属性的预测概率；

整个模型的损失函数为： $L_{\text{MTL}} = w_{\text{main}} \times L_{\text{main}} + w_{\text{aux}} \times L_{\text{aux}}$ ，其中，主任务分类器损失函数权重为 w_{main} ，辅助任务分类器损失函数权重为 w_{aux} 。

6. 根据权利要求1所述的基于时序注意力的语音事件检测和定位方法，其特征在于，所述基于所述待检测时序注意力权重矩阵确定目标语音定位包括：

基于所述待检测时序注意力权重矩阵与预设阈值确定目标语音起止帧索引；

基于所述目标语音起止帧索引确定目标语音定位。

7. 根据权利要求 6 所述的基于时序注意力的语音事件检测和定位方法，其特征在于，所述基于所述待检测时序注意力权重矩阵确定目标语音定位还包括：

基于相邻的所述目标语音起止帧索引确定相邻帧间距；

5 若所述相邻帧间距小于或等于参考阈值且相邻的目标语音事件属性相同，则合并相邻的所述目标语音起止帧索引。

8. 一种基于时序注意力的语音事件检测和定位装置，其特征在于，所述装置包括：

参考语音事件数据集获取模块，用于获取参考语音事件数据集，所述参考语音事件数据集包括参考语音事件数据和对应的参考语音事件属性；

10 参考音频特征确定模块，用于基于所述参考语音事件数据确定参考音频特征；

时序注意力特征提取模型和语音事件检测模型训练模块，用于基于所述参考音频特征和参考语音事件属性训练时序注意力特征提取模型和语音事件检测模型，其中，所述时序注意力特征提取模型包括特征压缩模块、特征激活模块和特征模化模块，所述特征压缩模块用于降低音频特征的维数和汇总音频特征的信息，所述特征激活模块用于生成时间维度上的权重向量，所述特征模化模块用于生成基于时序注意力的音频特征矩阵；

20 时序注意力矩阵确定模块，用于基于待检测语音数据确定待检测音频特征，将所述待检测音频特征输入所述时序注意力特征提取模型，得到待检测时序注意力权重矩阵和待检测时序注意力音频特征矩阵；

目标语音事件信息确定模块，用于基于所述待检测时序注意力权重矩阵确定目标语音定位，并将所述待检测时序注意力音频特征矩阵输入所述语音事件检测模型，得到目标语音事件属性。

25 9. 一种计算机设备，包括存储器和处理器，所述存储器存储有计算机程序，其特征在于，所述处理器执行所述计算机程序时实现权利要求 1 至 7 中任一项所述的方法的步骤。

10. 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 至 7 中任一项所述的方法的步骤。

基于时序注意力的语音事件检测和定位方法及装置

技术领域

本申请涉及语音事件检测技术领域，特别是涉及一种基于时序注意力的语音事件检测和定位方法及装置。

背景技术

语音事件检测是根据收集来的音频数据处理其中的声音信号，分析其声学特征并识别得到相应的事件类别，同时需要估计不同事件实例的开始和结束时间，并将其转换为声学环境中事件所对应的符号描述，进而感知和理解周边环境中的信息，是语音处理研究领域的一个重要分支，广泛应用于不同领域的音频分析任务。

现有技术中，语音事件检测框架通常是基于 Transformer、wav2vec2.0 等端到端的模型，广泛应用于多种语音处理下游任务。然而，目前的语音事件检测模型仅限于基于语音事件进行训练，既没有充分利用数据集中的信息，也很容易被语音信号里的其它信息干扰而导致准确性不高，同时，目前的语音事件检测技术都忽略了对语音事件的定位检测，大大降低了语音事件检测模型的实用性。

因此，相关技术中，亟需一种能够提高语音事件检测的准确性以及实用性的方式。

发明内容

基于此，有必要针对上述技术问题，提供一种能够提高语音事件检测的准确性以及实用性的基于时序注意力的语音事件检测和定位方法及装置。

第一方面，本申请提供了一种基于时序注意力的语音事件检测和定位方法。所述方法包括：

获取参考语音事件数据集，所述参考语音事件数据集包括参考语音事件数据和对应的参考语音事件属性；

基于所述参考语音事件数据确定参考音频特征；

基于所述参考音频特征和参考语音事件属性训练时序注意力特征提取模型和语音事件检测模型，其中，所述时序注意力特征提取模型包括特征压缩模块、特征激活模块和特征模化模块，所述特征压缩模块用于降低音频特征的维数和
5 汇总音频特征的信息，所述特征激活模块用于生成时间维度上的权重向量，所述特征模化模块用于生成基于时序注意力的音频特征矩阵；

基于待检测语音数据确定待检测音频特征，将所述待检测音频特征输入所述时序注意力特征提取模型，得到待检测时序注意力权重矩阵和待检测时序注意力音频特征矩阵；

10 基于所述待检测时序注意力权重矩阵确定目标语音定位，并将所述待检测时序注意力音频特征矩阵输入所述语音事件检测模型，得到目标语音事件属性。

可选的，在本申请的一个实施例中，所述基于所述参考语音事件数据确定参考音频特征包括：

搭建基础特征提取模型，基于公开语音数据集训练所述基础特征提取模型；
15 将所述参考语音事件数据输入所述基础特征提取模型，得到参考音频特征。

可选的，在本申请的一个实施例中，所述基于所述参考音频特征和参考语音事件属性训练时序注意力特征提取模型和语音事件检测模型包括：

将所述参考音频特征输入所述特征压缩模块，得到压缩音频特征矩阵；

将所述压缩音频特征矩阵输入所述特征激活模块，得到压缩时序注意力权重矩阵；
20 重矩阵；

将所述压缩时序注意力权重矩阵输入所述特征模化模块，得到参考时序注意力音频特征矩阵。

可选的，在本申请的一个实施例中，所述语音事件检测模型为基于多任务学习的语音事件类型检测模型，所述基于所述参考音频特征和参考语音事件属性训练时序注意力特征提取模型和语音事件检测模型还包括：
25 性训练时序注意力特征提取模型和语音事件检测模型还包括：

将所述参考时序注意力音频特征矩阵输入语音事件检测模型，得到初始语音事件属性；

基于所述初始语音事件属性和参考语音事件属性确定损失函数，基于所述损失函数调整所述语音事件检测模型的参数。

可选的，在本申请的一个实施例中，所述初始语音事件属性包括初始事件类型和初始语音特征，所述参考语音事件属性包括参考事件类型和参考语音特征，所述基于所述初始语音事件属性和参考语音事件属性确定损失函数如下：

$$L_{\text{main}} = - \sum_{i=1}^M \sum_{j=1}^K s_{ij} \times \log(\hat{s}_{ij})$$

其中，M为样本个数， s_{ij} 代表第i个样本属于第j种语音事件类型的真实概率， \hat{s}_{ij} 代表第i个样本属于第j种语音事件类型的预测概率；

所述辅助任务分类器执行二分类判决任务，其损失函数如下：

$$L_{\text{aux}} = - \sum_{i=1}^M \sum_{j=1}^2 y_{ij} \times \log(\hat{y}_{ij})$$

其中，M为样本个数， y_{ij} 代表第i个样本属于第j种特定语音属性的真实概率， \hat{y}_{ij} 代表第i个样本属于第j种特定语音属性的预测概率；

整个模型的损失函数为： $L_{\text{MTL}} = w_{\text{main}} \times L_{\text{main}} + w_{\text{aux}} \times L_{\text{aux}}$ ，其中，主任务分类器损失函数权重为 w_{main} ，辅助任务分类器损失函数权重为 w_{aux} 。

可选的，在本申请的一个实施例中，所述基于所述待检测时序注意力权重矩阵确定目标语音定位包括：

基于所述待检测时序注意力权重矩阵与预设阈值确定目标语音起止帧索引；

基于所述目标语音起止帧索引确定目标语音定位。

可选的，在本申请的一个实施例中，所述基于所述待检测时序注意力权重矩阵确定目标语音定位还包括：

基于相邻的所述目标语音起止帧索引确定相邻帧间距；

若所述相邻帧间距小于或等于参考阈值且相邻的目标语音事件属性相同，则合并相邻的所述目标语音起止帧索引。

第二方面，本申请还提供了一种基于时序注意力的语音事件检测和定位装置。所述装置包括：

参考语音事件数据集获取模块，用于获取参考语音事件数据集，所述参考语音事件数据集包括参考语音事件数据和对应的参考语音事件属性；

参考音频特征确定模块，用于基于所述参考语音事件数据确定参考音频特征；

5 时序注意力特征提取模型和语音事件检测模型训练模块，用于基于所述参考音频特征和参考语音事件属性训练时序注意力特征提取模型和语音事件检测模型，其中，所述时序注意力特征提取模型包括特征压缩模块、特征激活模块和特征模化模块，所述特征压缩模块用于降低音频特征的维数和汇总音频特征的信息，所述特征激活模块用于生成时间维度上的权重向量，所述特征模化模
10 块用于生成基于时序注意力的音频特征矩阵；

时序注意力矩阵确定模块，用于基于待检测语音数据确定待检测音频特征，将所述待检测音频特征输入所述时序注意力特征提取模型，得到待检测时序注意力权重矩阵和待检测时序注意力音频特征矩阵；

目标语音事件信息确定模块，用于基于所述待检测时序注意力权重矩阵确
15 定目标语音定位，并将所述待检测时序注意力音频特征矩阵输入所述语音事件检测模型，得到目标语音事件属性。

第三方面，本申请还提供了一种计算机设备。所述计算机设备包括存储器和处理器，所述存储器存储有计算机程序，所述处理器执行上述各个实施例所述方法的步骤。

20 第四方面，本申请还提供了一种计算机可读存储介质。所述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上述各个实施例所述方法的步骤。

上述基于时序注意力的语音事件检测和定位方法、装置、计算机设备和存储介质，通过获取参考语音事件数据集，参考语音事件数据集包括参考语音事件数据和对应的参考语音事件属性，基于参考语音事件数据确定参考音频特征，
25 基于参考音频特征和参考语音事件属性训练时序注意力特征提取模型和语音事件检测模型，其中，时序注意力特征提取模型包括特征压缩模块、特征激活模

块和特征模化模块，特征压缩模块用于降低音频特征的维数和汇总音频特征的信息，特征激活模块用于生成时间维度上的权重向量，特征模化模块用于生成基于时序注意力的音频特征矩阵，基于待检测语音数据确定待检测音频特征，将待检测音频特征输入时序注意力特征提取模型，得到待检测时序注意力权重矩阵和待检测时序注意力音频特征矩阵，基于待检测时序注意力权重矩阵确定目标语音定位，并将待检测时序注意力音频特征矩阵输入语音事件检测模型，得到目标语音事件属性。也就是说，在语音事件检测的过程中，通过利用大量的标签数据预训练检测模型，同时引入时序注意力，将语音事件检测和其他类型的信息结合协同训练，并实现了目标语段的定位检测，提高了语音事件检测的准确性以及实用性。

附图说明

图 1 为一个实施例中基于时序注意力的语音事件检测和定位方法的应用环境图；

图 2 为一个实施例中基于时序注意力的语音事件检测和定位方法的流程示意图；

图 3 为一个实施例中时序注意力特征提取的流程示意图；

图 4 为一个实施例中目标语音起止帧索引确定的示意图；

图 5 为一个实施例中基于时序注意力的语音事件检测和定位方法具体步骤的流程示意图；

图 6 为一个实施例中基于时序注意力的语音事件检测和定位装置的结构框图；

图 7 为一个实施例中计算机设备的内部结构图。

具体实施方式

为了使本申请的目的、技术方案及优点更加清楚明白，以下结合附图及实施例，对本申请进行进一步详细说明。应当理解，此处描述的具体实施例仅仅用以解释本申请，并不用于限定本申请。

本申请实施例提供的基于时序注意力的语音事件检测和定位方法，可以应用于如图 1 所示的应用环境中。其中，终端 102 通过网络与服务器 104 进行通信。数据存储系统可以存储服务器 104 需要处理的数据。数据存储系统可以集成在服务器 104 上，也可以放在云上或其他网络服务器上。其中，终端 102 可以但不限于各种个人计算机、笔记本电脑、智能手机、平板电脑、物联网设备和便携式可穿戴设备，物联网设备可为智能音箱、智能电视、智能空调、智能车载设备等。便携式可穿戴设备可为智能手表、智能手环、头戴设备等。服务器 104 可以用独立的服务器或者是多个服务器组成的服务器集群来实现。

在一个实施例中，如图 2 所示，提供了一种基于时序注意力的语音事件检测和定位方法，以该方法应用于图 1 中的服务器为例进行说明，包括以下步骤：

S201：获取参考语音事件数据集，所述参考语音事件数据集包括参考语音事件数据和对应的参考语音事件属性。

本申请实施例中，首先，获取用于训练模型的参考语音事件数据集，其中包括参考语音事件数据和其对应的参考语音事件属性，其中，语音事件可以是不同应用场景下的，例如医疗诊断、智能交通、智能家居等。

S203：基于所述参考语音事件数据确定参考音频特征。

本申请实施例中，对于获得的参考语音事件数据集中的参考语音事件数据，进行初步的声学特征提取，获得其相关音频特征。

S205：基于所述参考音频特征和参考语音事件属性训练时序注意力特征提取模型和语音事件检测模型，其中，所述时序注意力特征提取模型包括特征压缩模块、特征激活模块和特征模化模块，所述特征压缩模块用于降低音频特征的维数和汇总音频特征的信息，所述特征激活模块用于生成时间维度上的权重向量，所述特征模化模块用于生成基于时序注意力的音频特征矩阵。

本申请实施例中，将获得的相关音频特征输入时序注意力特征提取模型，得到与时间维度相关的参考时序注意力音频特征矩阵，之后，将参考时序注意力音频特征矩阵输入构建的语音事件检测模型，输出初始语音事件属性，基于初始语音事件属性和参考语音事件属性的差异不断调整语音事件检测模型的参

数，当差异满足训练目标时，模型训练完成，得到目标语音事件检测模型。

S207：基于待检测语音数据确定待检测音频特征，将所述待检测音频特征输入所述时序注意力特征提取模型，得到待检测时序注意力权重矩阵和待检测时序注意力音频特征矩阵。

5 本申请实施例中，基于待检测的用户语音数据确定其音频特征，确定方式同上述，之后，将获得的待检测音频特征输入时序注意力特征提取模型，输出待检测时序注意力权重矩阵和待检测时序注意力音频特征矩阵。

10 S209：基于所述待检测时序注意力权重矩阵确定目标语音定位，并将所述待检测时序注意力音频特征矩阵输入所述语音事件检测模型，得到目标语音事件属性。

 本申请实施例中，基于与时间维度相关的时序注意力权重矩阵定位目标语段在语音数据中的位置，并将待检测时序注意力音频特征矩阵输入训练好的目标语音事件检测模型中，输出目标语音事件属性，即目标语音事件类型以及其他特征信息。

15 上述基于时序注意力的语音事件检测和定位方法中，获取参考语音事件数据集，参考语音事件数据集包括参考语音事件数据和对应的参考语音事件属性，基于参考语音事件数据确定参考音频特征，基于参考音频特征和参考语音事件属性训练时序注意力特征提取模型和语音事件检测模型，其中，时序注意力特征提取模型包括特征压缩模块、特征激活模块和特征模化模块，特征压缩模块
20 用于降低音频特征的维数和汇总音频特征的信息，特征激活模块用于生成时间维度上的权重向量，特征模化模块用于生成基于时序注意力的音频特征矩阵，基于待检测语音数据确定待检测音频特征，将待检测音频特征输入时序注意力特征提取模型，得到待检测时序注意力权重矩阵和待检测时序注意力音频特征矩阵，基于待检测时序注意力权重矩阵确定目标语音定位，并将待检测时序注
25 意力音频特征矩阵输入语音事件检测模型，得到目标语音事件属性。也就是说，在语音事件检测的过程中，通过利用大量的标签数据预训练检测模型，同时引入时序注意力，将语音事件检测和其他类型的信息结合协同训练，并实现了目

标语段的定位检测，提高了语音事件检测的准确性以及实用性。

在本申请的一个实施例中，所述基于所述参考语音事件数据确定参考音频特征包括：

5 S301：搭建基础特征提取模型，基于公开语音数据集训练所述基础特征提取模型。

S303：将所述参考语音事件数据输入所述基础特征提取模型，得到参考音频特征。

在本申请的一个实施例中，首先搭建基础特征提取模型，该基础特征提取模型为 wav2vec2.0 特征提取模型，主要包括三个模块：卷积特征编码器、
10 Transformer 编码器、量化模块。其中，卷积特征编码器主要由 7 层的卷积神经网络构成，卷积层的通道数设置为 512，每个卷积层的输出通过层归一化和 GELU 激活函数输出低层次的音频表征。其中，卷积层的步长决定了输出特征的时间步数 T 。该模块将从原始音频数据中提取局部的声学特征，将原始音频特征 $X: x_1 \dots x_T$ 映射为潜在语音特征 $Z: z_1 \dots z_T$ 。Transformer 编码器以卷积特征编
15 码器的输出作为输入，随后经过 12 个 Transformer 模块。每个 Transformer 模块都包含多头注意力机制和前馈神经网络，并随后使用层归一化技术。其中，每个多头注意力使用 8 个注意力头。Transformer 模块能够抓取音频特征中的长跨度的依赖关系和全局层次的上下文信息，将潜在语音特征 $Z: z_1 \dots z_T$ 映射为上下文潜在语音特征 $C: c_1 \dots c_T$ 。量化模块通过乘积量化将所述卷积特征编码器的输出离散化为一组有限的语音特征，所述乘积量化需要首先随机采样时间步，
20 采样百分比为 p ，再对随后的 M 个时间步进行掩码操作，得到量化后的向量 $Q: q_1 \dots q_T$ 。在具体实践中， $p = 0.065, M = 10$ 。

基础特征提取模型搭建好之后，基于公开语音数据集训练所述基础特征提取模型，具体的，公开语音数据集可以选用 Librispeech-960h，训练过程包括
25 预训练与微调的过程，其中，预训练时采用自监督训练的策略，设置该模型的初始学习率为 5×10^{-4} ，随后，根据实际的训练情况可考虑使学习率线性衰减；预训练 wav2vec2.0 模型时将对比损失 L_c 作为损失函数，采用 Adam 优化算法调

整模型参数。所述对比损失函数如下，

$$L_m = -\log \frac{\exp(\text{sim}(c_t, q_t)/k)}{\sum_{\tilde{q} \sim Q_t} \exp(\text{sim}(c_t, \tilde{q})/k)}$$

其中预设缩放因子 $k = 0.1$ ， Q_t 是一个负采样的集合，表示从 q_t 的负样本集合中随机选择的向量（负样本是为了与正样本进行对比，用于训练模型并增强语义相似性的区分度）， \tilde{q} 代表从负样本集合 Q_t 中随机选择的负样本向量， \exp 是自然指数函数， sim 是预先选取的相似度计算函数。

微调过程是根据下游任务选择相应的数据集进行微调，此时使用的公开语音数据集的音频特征是有标签的。另外，在微调前需要根据特定的下游任务完成下游模型（如多任务分类器）的搭建。微调时采用有监督的训练策略，设置
10 微调时的模型学习率为 1×10^{-4} ，并且注意微调时无需训练卷积特征提取器模块，所以此时需要冻结卷积特征提取器的参数，损失函数需根据特定的下游任务（如口吃类型检测）确定，采用 Adam 优化算法

当基础特征提取模型训练完成即模型参数确认之后，将参考语音事件数据输入基础特征提取模型中，得到参考音频特征，具体应用中，参考语音事件数
15 据集可以为口吃数据集，将其中的包含语音事件的音频数据作为输入，利用完成多阶段训练的 wav2vec2.0 特征提取模型提取得到基于 wav2vec2.0 的音频特征 C。

本实施例中，通过搭建基础特征提取模型，基于公开语音数据集训练基础特征提取模型，将参考语音事件数据输入基础特征提取模型，得到参考音频特
20 征，能够提取原始语音数据的深层次表征，且基础特征提取模型在具有初步学习能力的基础上获得迁移学习的能力。

在本申请的一个实施例中，所述基于所述参考音频特征和参考语音事件属性训练时序注意力特征提取模型和语音事件检测模型包括：

S401：将所述参考音频特征输入所述特征压缩模块，得到压缩音频特征矩
25 阵。

S403：将所述压缩音频特征矩阵输入所述特征激活模块，得到压缩时序注

意力权重矩阵。

S405: 将所述压缩时序注意力权重矩阵输入所述特征模化模块, 得到参考时序注意力音频特征矩阵。

在本申请的一个实施例中, 如图 3 所示, 时序注意力特征提取模型包括特征压缩模块、特征激活模块和特征模化模块。其中, 特征压缩模块使用平均池化层和线性层, 其中平均池化层用于降低特征的维数, 也能提升计算的速度和神经网络学习的效率, 同时可以一定程度上防止过拟合。使用线性层对平均池化层的输出做线性变换, 使网络更加关注时间维度上更重要的帧。使用特征压缩模块, 可将整个 wav2vec2.0 特征矩阵的信息汇总到一个特征向量中。特征激活模块包括一个全连接神经网络层, 其中通过使用全连接层和非线性激活函数, 学习生成一个时间维度上的权重向量。这个权重向量后续将被应用于原始的 wav2vec2.0 特征矩阵的每个时间维度上, 以对不同时间帧的特征进行加权。最后一层的非线性激活函数为 Softmax, 计算得到每一个时间帧对应的目标事件发生的概率 (概率值分布在 0-1 之间) 组成的矩阵, 即基于时序注意力机制的权重矩阵。特征模化模块即是一个算术的加权操作, 将所述的基于时序注意力机制的权重矩阵与原来的 wav2vec2.0 音频特征矩阵相乘, 即得基于时序注意力的 wav2vec2.0 音频特征, 用于下游的相关分类任务。

具体的, 首先, 将参考音频特征输入特征压缩模块, 得到压缩音频特征矩阵, 即对 wav2vec2.0 音频特征矩阵C进行平均池化操作, wav2vec2.0 音频特征矩阵C是 $T \times F$ 的特征矩阵, 在池化之后得到维数为 $T \times 1$ 的压缩音频特征矩阵 C_1 。之后, 将压缩音频特征矩阵输入特征激活模块, 得到压缩时序注意力权重矩阵, 具体应用中, 特征激活模块将池化后的压缩音频特征矩阵经过若干层的全连接神经网络并通过非线性激活函数 ReLU 来学习各帧之间的权重, 具体公式如下所示。

$$C_2 = \delta (W_1 C_1)$$

其中, W_1 是该全连接层对应的权重矩阵, δ 是 ReLU 激活函数。

再将所述全连接层的输出结果输入到另一个全连接层中, 并通过一个非线

性的激活函数 Softmax 计算得到压缩时序注意力权重矩阵即 $T \times 1$ 基于时序注意力的权重矩阵 W 。具体公式如下所示。

$$W = \sigma(W_2 C_2)$$

其中， W_2 是该全连接层对应的权重矩阵， σ 是 Softmax 激活函数。

5 最后，将压缩时序注意力权重矩阵输入特征模化模块，得到参考时序注意力音频特征矩阵。具体的，指使用压缩时序注意力权重矩阵对 wav2vec2.0 音频特征矩阵 C 进行加权融合，利用矩阵的广播机制将权重矩阵 W 扩展到 $T \times F$ ，再与 wav2vec2.0 音频特征矩阵 C 逐元素相乘，最后在时序维度上融合，得到基于时序注意力的 wav2vec2.0 音频特征矩阵 A 。具体公式如下所示。

10
$$W_b = \text{broadcast}(W)$$

$$A = \text{fsum}(W_b * C)$$

其中， W_b 是基于时序注意力的权重矩阵 W 广播之后的权重矩阵， C 是基于 wav2vec2.0 的音频特征矩阵，broadcast 代表广播操作，* 代表逐元素相乘，函数 fsum 代表时序维度上的特征融合。

15 本实施例中，通过将参考音频特征输入特征压缩模块，得到压缩音频特征矩阵，将压缩音频特征矩阵输入特征激活模块，得到压缩时序注意力权重矩阵，将压缩时序注意力权重矩阵输入特征模化模块，得到参考时序注意力音频特征矩阵，能够提高特征的判断能力，有助于捕捉不同时间维度之间的关系和交互，进一步提升模型的性能。

20 在本申请的一个实施例中，所述语音事件检测模型为基于多任务学习的语音事件类型检测模型，所述基于所述参考音频特征和参考语音事件属性训练时序注意力特征提取模型和语音事件检测模型还包括：

S501：将所述参考时序注意力音频特征矩阵输入语音事件检测模型，得到初始语音事件属性。

25 S503：基于所述初始语音事件属性和参考语音事件属性确定损失函数，基于所述损失函数调整所述语音事件检测模型的参数。

在本申请的一个实施例中，首先，将参考时序注意力音频特征矩阵输入语

音事件检测模型，得到初始语音事件属性，其中，语音事件检测模型为基于多任务学习的语音事件类型检测模型，训练过程是分别通过主任务和辅助任务共同训练一个神经网络模型，获得语音事件属性并输出，具体的，将基于时序注意力的 wav2vec2.0 音频特征A馈送到全连接神经网络中，所述全连接网络的输出分别馈送到主任务与辅助任务分类器中。

所述主任务分类器包含若干线性层，所述线性层将接收到 wav2vec2.0 特征矩阵进行线性组合；将线性组合计算值通过 ReLU 激活函数；再通过 Softmax 激活函数由最后一层神经网络中的特征值计算得到特定语音事件类型的概率分布，最终输出相应的目标类型标签。具体公式如下所示。

$$S = \sigma (W_{m2} \delta (W_{m1}A))$$

其中， W_{m1} 、 W_{m2} 分别是不同全连接层对应的权重矩阵， δ 是 ReLU 激活函数， σ 是 Softmax 激活函数， $S = (s_1, s_2, \dots, s_K)$ 代表的是K种语音事件类型的概率分布。

所述辅任务分类器执行与主任务相关的二分类判决任务，模型结构与所述主任务分类器结构相似，包含若干线性层，所述线性层将接收到 wav2vec2.0 特征矩阵进行线性组合；将线性组合计算值通过 ReLU 激活函数；再通过 Sigmoid 激活函数由最后一层神经网络中的特征值计算得到区间范围为 0-1 的概率值，最终输出为特定辅助任务属性。具体公式如下所示。

$$Y = \rho (W_{a2} \delta (W_{a1}A))$$

其中， W_{a1} 、 W_{a2} 分别是不同全连接层对应的权重矩阵， δ 是 ReLU 激活函数， ρ 是 Sigmoid 激活函数， $Y = (y_1, y_2)$ 代表的是特定辅助任务属性的概率分布。

同时，基于得到的初始语音事件属性和已知的参考语音事件属性确定损失函数，并基于损失函数不断调整语音事件检测模型的参数，直到损失函数的值趋于最小并稳定时，此时确定的模型即为目标语音事件检测模型。具体应用中，损失函数一般采用交叉熵损失函数。

在本申请的一个实施例中，所述基于所述初始语音事件属性和参考语音事

件属性确定损失函数如下：

$$L_{\text{main}} = - \sum_{i=1}^M \sum_{j=1}^K s_{ij} \times \log(\hat{s}_{ij})$$

其中，M为样本个数， s_{ij} 代表第i个样本属于第j种语音事件类型的真实概率， \hat{s}_{ij} 代表第i个样本属于第j种语音事件类型的预测概率；

5 所述辅助任务分类器执行二分类判决任务，其损失函数如下：

$$L_{\text{aux}} = - \sum_{i=1}^M \sum_{j=1}^2 y_{ij} \times \log(\hat{y}_{ij})$$

其中，M为样本个数， y_{ij} 代表第i个样本属于第j种特定语音属性的真实概率， \hat{y}_{ij} 代表第i个样本属于第j种特定语音属性的预测概率；

整个模型的损失函数为： $L_{\text{MTL}} = w_{\text{main}} \times L_{\text{main}} + w_{\text{aux}} \times L_{\text{aux}}$ ，其中，
10 主任务分类器损失函数权重为 w_{main} ，辅助任务分类器损失函数权重为 w_{aux} 。

在本申请的一个实施例中，基于多任务学习的语音事件类型检测模型包括主任务和辅助任务，其中，主任务给出输入音频特征中包含的事件类型，辅助任务给出输入音频特征中包含的语音特征，例如性别等。在基于多任务学习的语音事件类型检测模型训练的过程中，主任务分类器及辅助任务分类器均采用
15 交叉熵损失函数，基于交叉熵损失函数不断调整基于多任务学习的语音事件类型检测模型的参数，直至交叉熵损失函数的值趋于稳定并达到最小，模型训练完成，具体的，主任务分类器执行K种语音事件类型的多分类任务，其中K代表特定语音事件类型的种类数，其损失函数为：

$$L_{\text{main}} = - \sum_{i=1}^M \sum_{j=1}^K s_{ij} \times \log(\hat{s}_{ij})$$

20 其中，M为样本个数， s_{ij} 代表第i个样本属于第j种语音事件类型的真实概率， \hat{s}_{ij} 代表第i个样本属于第j种语音事件类型的预测概率。

所述辅助任务分类器执行二分类判决任务，其损失函数如下：

$$L_{\text{aux}} = - \sum_{i=1}^M \sum_{j=1}^2 y_{ij} \times \log(\hat{y}_{ij})$$

其中, M 为样本个数, y_{ij} 代表第 i 个样本属于第 j 种特定语音属性的真实概率, \hat{y}_{ij} 代表第 i 个样本属于第 j 种特定语音属性的预测概率。

整个模型的损失函数为: $L_{MTL} = W_{main} \times L_{main} + W_{aux} \times L_{aux}$, 其中, 主任务分类器损失函数权重为 w_{main} , 辅助任务分类器损失函数权重为 w_{aux} , 两个权重值预先由人工设定, 通过梯度下降算法进行样本训练, 最终得到目标语音事件检测模型。

本实施例中, 通过采用多任务学习框架, 综合数据特点, 将语音事件检测和其他类型的语音特征信息结合进行协同训练, 使检测模型能更好地关注影响任务模型性能的特征, 同时能够有效避免过拟合等问题, 提高模型泛化能力。

在本申请的一个实施例中, 所述基于所述时序注意力权重矩阵确定目标语音定位包括:

S601: 基于所述待检测时序注意力权重矩阵与预设阈值确定目标语音起止帧索引。

S603: 基于所述目标语音起止帧索引确定目标语音定位。

在本申请的一个实施例中, 如图4所示, 首先, 输入时序注意力权重矩阵 W , 若权重矩阵 W 从 w_i 到 w_j 的每一个元素值都大于或等于预设阈值, 则表示所述待检测语音特征从第 i 帧到第 j 帧为目标语段的位置, 得到目标语音起止帧索引即在所述待检测语音特征中目标语段对应的若干对起止帧索引, 之后, 根据语音音频分帧选择的帧长 fs , 帧移 fl 以及起始帧索引和结束帧索引来计算时间位置, 具体计算公式如下:

$$st = fs \times (t_s - 1)$$

$$et = fs \times (t_e + v - 1) + fl$$

其中, st 指待检测语音特征中目标语段的起始时间位置, et 指待检测语音特征中目标语段的结束时间位置, v 指映射语段中前 v 个特征。

本实施例中, 通过基于时序注意力权重矩阵与预设阈值确定目标语音起止帧索引, 基于目标语音起止帧索引确定目标语音定位, 能够结合阈值判定和简单的计算就能获取目标语音的起止时间帧, 实现目标语音的定位。

在本申请的一个实施例中，所述基于所述待检测时序注意力权重矩阵确定目标语音定位还包括：

S701：基于相邻的所述目标语音起止帧索引确定相邻帧间距。

5 S703：若所述相邻帧间距小于或等于参考阈值且相邻的目标语音事件属性相同，则合并相邻的所述目标语音起止帧索引。

在本申请的一个实施例中，基于阈值判定得到的目标语音起止帧索引是分段的，可以采用平滑操作，对其进行合并，具体的，合并需要同时满足两个条件，首先，基于获得的相邻的目标语音起止帧索引确定相邻帧间距，即时序上的前一段的结束帧与后一段的起始帧之间的帧间距，判断相邻帧间距是否小于
10 或等于参考阈值，参考阈值由人工设置，为实验经验值，同时，判断前后两段即相邻的目标语音起止帧索引对应的目标语音事件属性是否相同，若相邻帧间距满足小于或等于参考阈值且目标语音事件属性相同，则将相邻的两段目标语音起止帧索引合并为一段。

本实施例中，通过对获得的多段目标语音起止帧索引进行平滑操作，能够
15 提高目标语音定位的实现效率。

下面以一个具体实施例说明本申请的基于时序注意力的语音事件检测和定位方法的具体实施步骤。如图5所示，首先，S801，获取参考语音事件数据集，所述参考语音事件数据集包括参考语音事件数据和对应的参考语音事件属性；之后，S803，基于所述参考语音事件数据确定参考音频特征；具体的，S805-S807，
20 搭建基础特征提取模型，基于公开语音数据集训练所述基础特征提取模型；将所述参考语音事件数据输入所述基础特征提取模型，得到参考音频特征。

之后，S809，基于所述参考音频特征和参考语音事件属性训练时序注意力特征提取模型和语音事件检测模型，其中，所述时序注意力特征提取模型包括特征压缩模块、特征激活模块和特征模化模块，所述特征压缩模块用于降低音频特征的维数和汇总音频特征的信息，所述特征激活模块用于生成时间维度上的权重向量，所述特征模化模块用于生成基于时序注意力的音频特征矩阵，具体的，S811-S815，将所述参考音频特征输入所述特征压缩模块，得到压缩音频

特征矩阵；将所述压缩音频特征矩阵输入所述特征激活模块，得到压缩时序注意力权重矩阵；将所述压缩时序注意力权重矩阵输入所述特征模化模块，得到参考时序注意力音频特征矩阵。

之后，所述语音事件检测模型为基于多任务学习的语音事件类型检测模型，S817-S819，将所述参考时序注意力音频特征矩阵输入语音事件检测模型，得到初始语音事件属性；基于所述初始语音事件属性和参考语音事件属性确定损失函数，基于所述损失函数调整所述语音事件检测模型的参数，确定目标语音事件检测模型。其中，所述初始语音事件属性包括初始事件类型和初始语音特征，所述参考语音事件属性包括参考事件类型和参考语音特征，S821，基于所述初始语音事件属性和参考语音事件属性确定损失函数如下：

$$L_{\text{main}} = - \sum_{i=1}^M \sum_{j=1}^K s_{ij} \times \log(\hat{s}_{ij})$$

其中，M为样本个数， s_{ij} 代表第i个样本属于第j种语音事件类型的真实概率， \hat{s}_{ij} 代表第i个样本属于第j种语音事件类型的预测概率；

所述辅助任务分类器执行二分类判决任务，其损失函数如下：

$$L_{\text{aux}} = - \sum_{i=1}^M \sum_{j=1}^2 y_{ij} \times \log(\hat{y}_{ij})$$

其中，M为样本个数， y_{ij} 代表第i个样本属于第j种特定语音属性的真实概率， \hat{y}_{ij} 代表第i个样本属于第j种特定语音属性的预测概率；

整个模型的损失函数为： $L_{\text{MTL}} = w_{\text{main}} \times L_{\text{main}} + w_{\text{aux}} \times L_{\text{aux}}$ ，其中，主任务分类器损失函数权重为 w_{main} ，辅助任务分类器损失函数权重为 w_{aux} 。

之后，S823，基于待检测语音数据确定待检测音频特征，将所述待检测音频特征输入所述时序注意力特征提取模型，得到待检测时序注意力权重矩阵和待检测时序注意力音频特征矩阵；最后，S825，基于所述待检测时序注意力权重矩阵确定目标语音定位，并将所述待检测时序注意力音频特征矩阵输入所述语音事件检测模型，得到目标语音事件属性。具体的，S827-S829，基于所述时序注意力权重矩阵与预设阈值确定目标语音起止帧索引，基于所述目标语音起

止帧索引确定目标语音定位，除此之外，S831-S833，基于相邻的所述目标语音起止帧索引确定相邻帧间距；若所述相邻帧间距小于或等于参考阈值且相邻的目标语音事件属性相同，则合并相邻的所述目标语音起止帧索引。

应该理解的是，虽然如上所述的各实施例所涉及的流程图中的各个步骤按照箭头的指示依次显示，但是这些步骤并不是必然按照箭头指示的顺序依次执行。除非本文中有明确的说明，这些步骤的执行并没有严格的顺序限制，这些步骤可以以其它的顺序执行。而且，如上所述的各实施例所涉及的流程图中的至少一部分步骤可以包括多个步骤或者多个阶段，这些步骤或者阶段并不必然是在同一时刻执行完成，而是可以在不同的时刻执行，这些步骤或者阶段的执行顺序也不必然是依次进行，而是可以与其它步骤或者其它步骤中的步骤或者阶段的至少一部分轮流或者交替地执行。

基于同样的发明构思，本申请实施例还提供了一种用于实现上述所涉及的基于时序注意力的语音事件检测和定位方法的基于时序注意力的语音事件检测和定位装置。该装置所提供的解决问题的实现方案与上述方法中所记载的实现方案相似，故下面所提供的一个或多个基于时序注意力的语音事件检测和定位装置实施例中的具体限定可以参见上文中对于基于时序注意力的语音事件检测和定位方法的限定，在此不再赘述。

在一个实施例中，如图6所示，提供了一种基于时序注意力的语音事件检测和定位装置600，包括：参考语音事件数据集获取模块601、参考音频特征确定模块603、时序注意力特征提取模型和语音事件检测模型训练模块605、时序注意力矩阵确定模块607和目标语音事件信息确定模块609，其中：

参考语音事件数据集获取模块601，用于获取参考语音事件数据集，所述参考语音事件数据集包括参考语音事件数据和对应的参考语音事件属性。

参考音频特征确定模块603，用于基于所述参考语音事件数据确定参考音频特征。

时序注意力特征提取模型和语音事件检测模型训练模块605，用于基于所述参考音频特征和参考语音事件属性训练时序注意力特征提取模型和语音事件

检测模型，其中，所述时序注意力特征提取模型包括特征压缩模块、特征激活模块和特征模化模块，所述特征压缩模块用于降低音频特征的维数和汇总音频特征的信息，所述特征激活模块用于生成时间维度上的权重向量，所述特征模化模块用于生成基于时序注意力的音频特征矩阵。

5 时序注意力矩阵确定模块 607，用于基于待检测语音数据确定待检测音频特征，将所述待检测音频特征输入所述时序注意力特征提取模型，得到待检测时序注意力权重矩阵和待检测时序注意力音频特征矩阵。

目标语音事件信息确定模块 609，用于基于所述待检测时序注意力权重矩阵确定目标语音定位，并将所述待检测时序注意力音频特征矩阵输入所述语音
10 事件检测模型，得到目标语音事件属性。

在本申请的一个实施例中，所述参考音频特征确定模块还用于：

搭建基础特征提取模型，基于公开语音数据集训练所述基础特征提取模型；
将所述参考语音事件数据输入所述基础特征提取模型，得到参考音频特征。

在本申请的一个实施例中，所述时序注意力特征提取模型和语音事件检测
15 模型训练模块还用于：

将所述参考音频特征输入所述特征压缩模块，得到压缩音频特征矩阵；

将所述压缩音频特征矩阵输入所述特征激活模块，得到压缩时序注意力权重矩阵；

将所述压缩时序注意力权重矩阵输入所述特征模化模块，得到参考时序注
20 意力音频特征矩阵。

在本申请的一个实施例中，所述语音事件检测模型为基于多任务学习的语音事件类型检测模型，所述时序注意力特征提取模型和语音事件检测模型训练模块还用于：

将所述参考时序注意力音频特征矩阵输入语音事件检测模型，得到初始语
25 音事件属性；

基于所述初始语音事件属性和参考语音事件属性确定损失函数，基于所述损失函数调整所述语音事件检测模型的参数，确定目标语音事件检测模型。

在本申请的一个实施例中，所述时序注意力特征提取模型和语音事件检测模型训练模块还用于：

基于所述初始事件类型和参考事件类型确定事件类型损失函数；

基于所述初始语音特征和参考语音特征确定语音特征损失函数；

5 基于所述事件类型损失函数和语音特征损失函数确定损失函数。

在本申请的一个实施例中，所述目标语音事件信息确定模块还用于：

基于所述时序注意力权重矩阵与预设阈值确定目标语音起止帧索引；

基于所述目标语音起止帧索引确定目标语音定位。

在本申请的一个实施例中，所述目标语音事件信息确定模块还用于：

10 基于相邻的所述目标语音起止帧索引确定相邻帧间距；

若所述相邻帧间距小于或等于参考阈值且相邻的目标语音事件属性相同，
则合并相邻的所述目标语音起止帧索引。

上述基于时序注意力的语音事件检测和定位装置中的各个模块可全部或部分通过软件、硬件及其组合来实现。上述各模块可以硬件形式内嵌于或独立于
15 计算机设备中的处理器中，也可以以软件形式存储于计算机设备中的存储器中，
以便于处理器调用执行以上各个模块对应的操作。

在一个实施例中，提供了一种计算机设备，该计算机设备可以是终端，其内部结构图可以如图 7 所示。该计算机设备包括通过系统总线连接的处理器、存储器、通信接口、显示屏和输入装置。其中，该计算机设备的处理器用于提供
20 计算和控制能力。该计算机设备的存储器包括非易失性存储介质、内存储器。
该非易失性存储介质存储有操作系统和计算机程序。该内存储器为非易失性存储
介质中的操作系统和计算机程序的运行提供环境。该计算机设备的通信接口
用于与外部的终端进行有线或无线方式的通信，无线方式可通过 WIFI、移动蜂
窝网络、NFC（近场通信）或其他技术实现。该计算机程序被处理器执行时以实
25 现一种基于时序注意力的语音事件检测和定位方法。该计算机设备的显示屏可
以是液晶显示屏或者电子墨水显示屏，该计算机设备的输入装置可以是显示屏
上覆盖的触摸层，也可以是计算机设备外壳上设置的按键、轨迹球或触控板，

还可以是外接的键盘、触控板或鼠标等。

本领域技术人员可以理解，图 7 中示出的结构，仅仅是与本申请方案相关的部分结构的框图，并不构成对本申请方案所应用于其上的计算机设备的限定，具体的计算机设备可以包括比图中所示更多或更少的部件，或者组合某些部件，或者具有不同的部件布置。

在一个实施例中，提供了一种计算机设备，包括存储器和处理器，存储器中存储有计算机程序，该处理器执行计算机程序时实现上述各方法实施例中的步骤。

在一个实施例中，提供了一种计算机可读存储介质，其上存储有计算机程序，计算机程序被处理器执行时实现上述各方法实施例中的步骤。

在一个实施例中，提供了一种计算机程序产品，包括计算机程序，该计算机程序被处理器执行时实现上述各方法实施例中的步骤。

需要说明的是，本申请所涉及的用户信息（包括但不限于用户设备信息、用户个人信息等）和数据（包括但不限于用于分析的数据、存储的数据、展示的数据等），均为经用户授权或者经过各方充分授权的信息和数据。

本领域普通技术人员可以理解实现上述实施例方法中的全部或部分流程，是可以通过计算机程序来指令相关的硬件来完成，所述的计算机程序可存储于一非易失性计算机可读取存储介质中，该计算机程序在执行时，可包括如上述各方法的实施例的流程。其中，本申请所提供的各实施例中所使用的对存储器、数据库或其它介质的任何引用，均可包括非易失性和易失性存储器中的至少一种。非易失性存储器可包括只读存储器（Read-Only Memory, ROM）、磁带、软盘、闪存、光存储器、高密度嵌入式非易失性存储器、阻变存储器（ReRAM）、磁变存储器（Magnetoresistive Random Access Memory, MRAM）、铁电存储器（Ferroelectric Random Access Memory, FRAM）、相变存储器（Phase Change Memory, PCM）、石墨烯存储器等。易失性存储器可包括随机存取存储器（Random Access Memory, RAM）或外部高速缓冲存储器等。作为说明而非局限，RAM 可以是多种形式，比如静态随机存取存储器（Static Random Access Memory, SRAM）

或动态随机存取存储器（Dynamic Random Access Memory, DRAM）等。本申请所提供的各实施例中所涉及的数据库可包括关系型数据库和非关系型数据库中至少一种。非关系型数据库可包括基于区块链的分布式数据库等，不限于此。

5 本申请所提供的各实施例中所涉及的处理器可为通用处理器、中央处理器、图形处理器、数字信号处理器、可编程逻辑器、基于量子计算的数据处理逻辑器等，不限于此。

以上实施例的各技术特征可以进行任意的组合，为使描述简洁，未对上述实施例中的各个技术特征所有可能的组合都进行描述，然而，只要这些技术特征的组合不存在矛盾，都应当认为是本说明书记载的范围。

10 以上所述实施例仅表达了本申请的几种实施方式，其描述较为具体和详细，但并不能因此而理解为对本申请专利范围的限制。应当指出的是，对于本领域的普通技术人员来说，在不脱离本申请构思的前提下，还可以做出若干变形和改进，这些都属于本申请的保护范围。因此，本申请的保护范围应以所附权利要求为准。

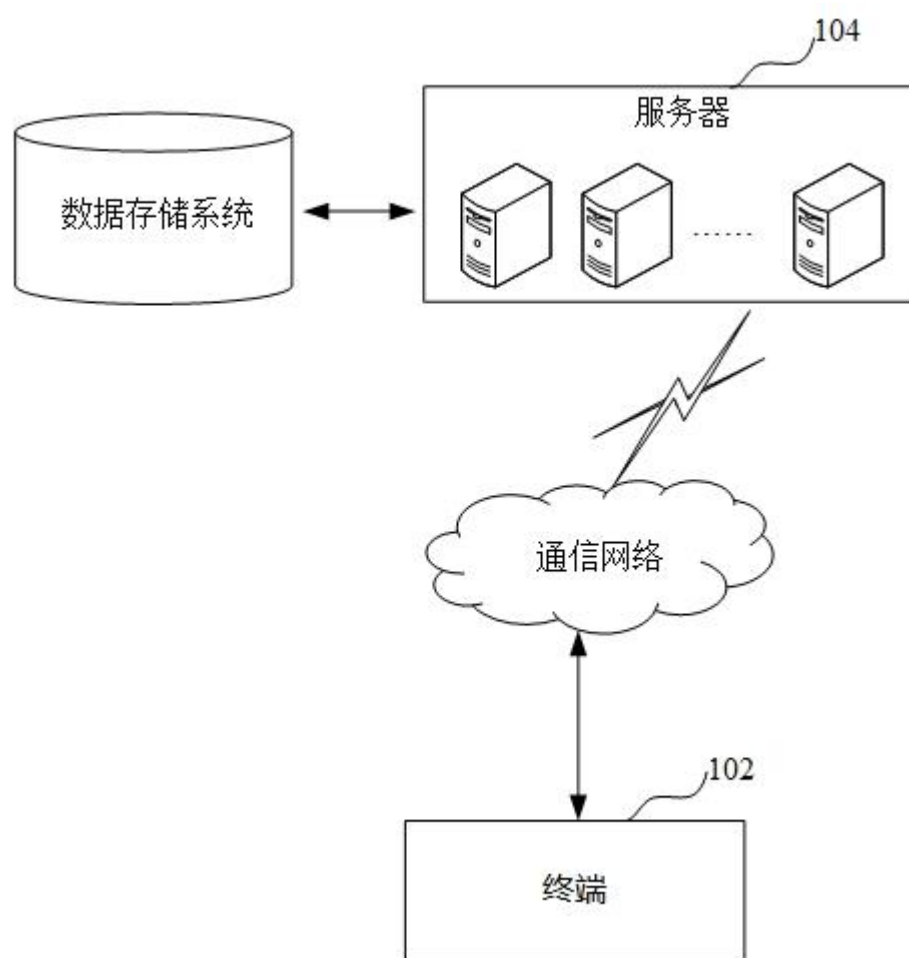


图 1

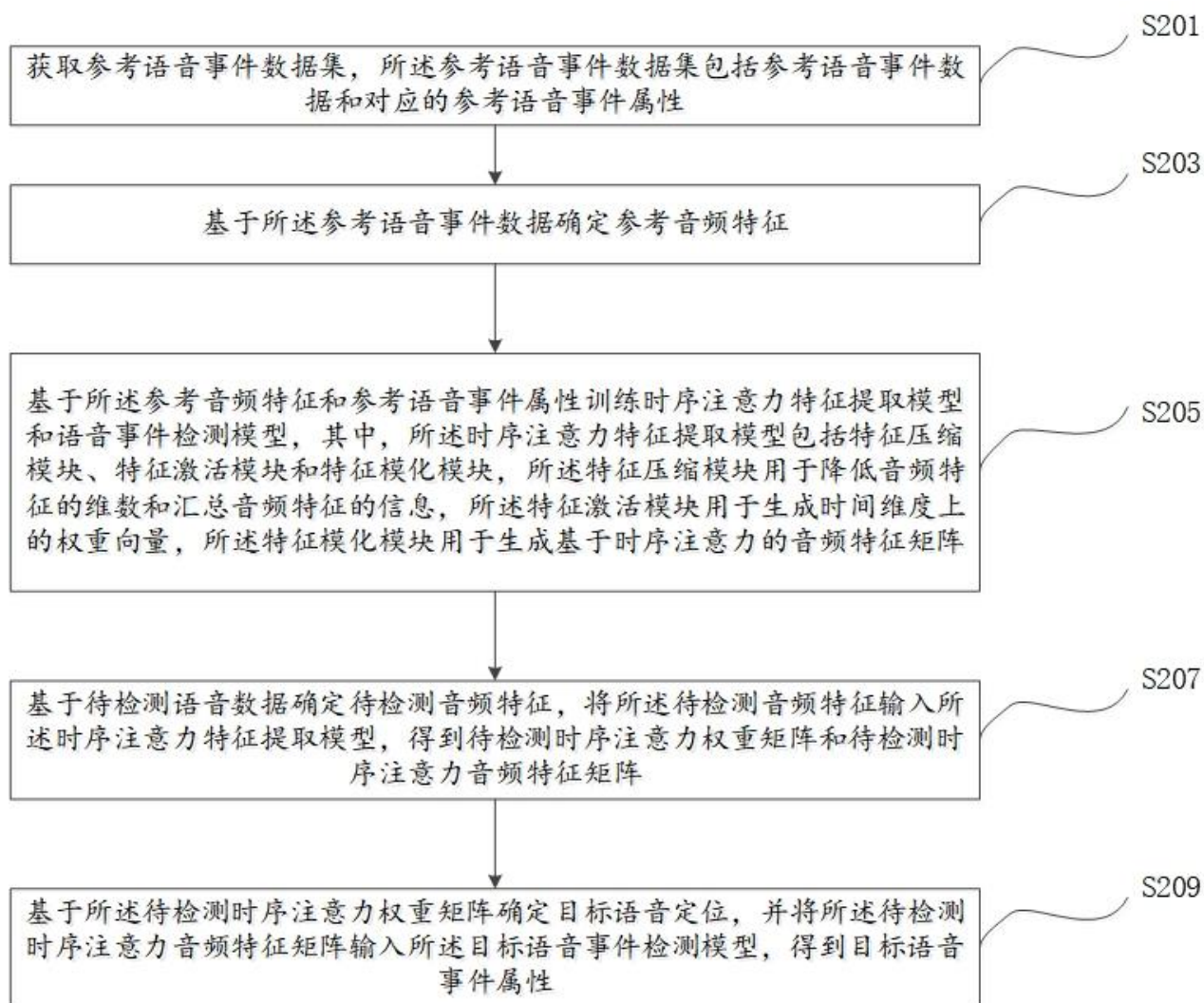


图 2

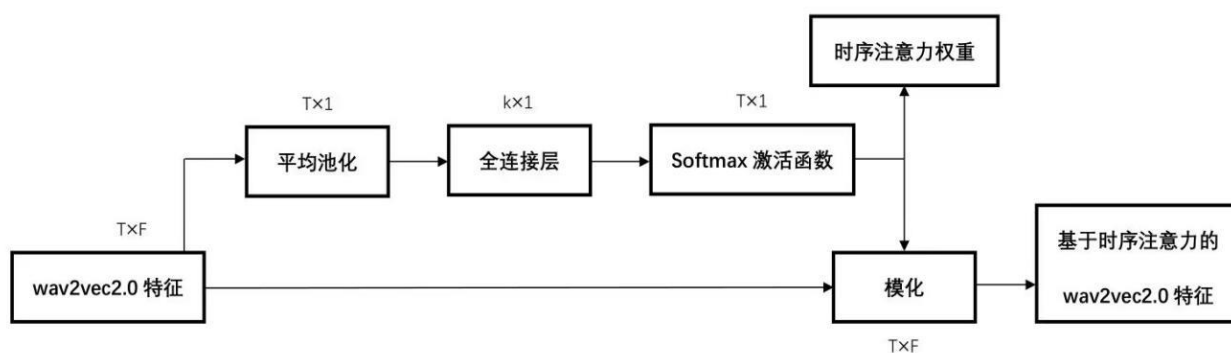


图 3

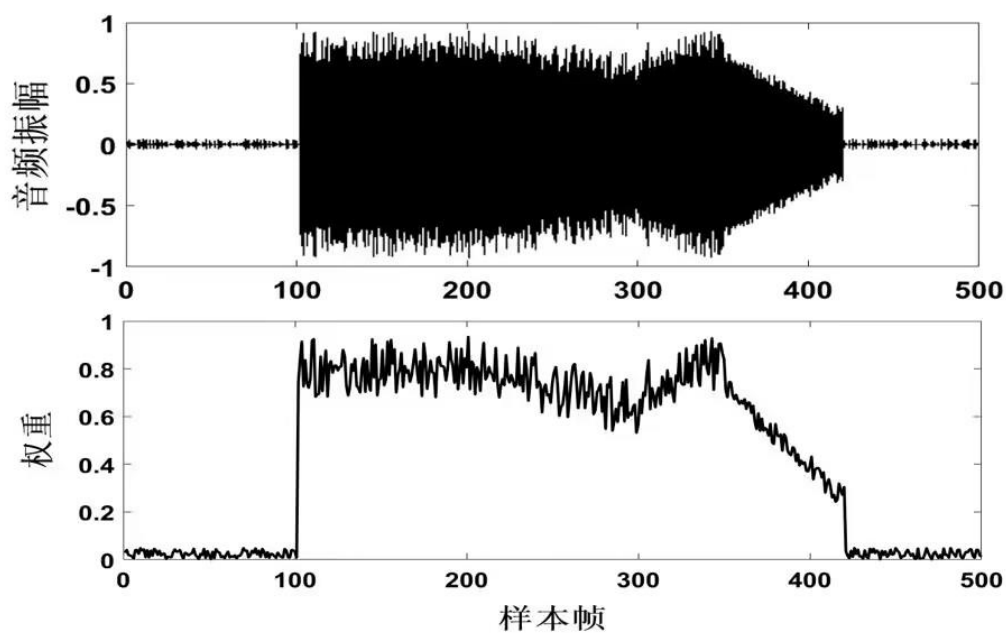


图 4

原始音频输入

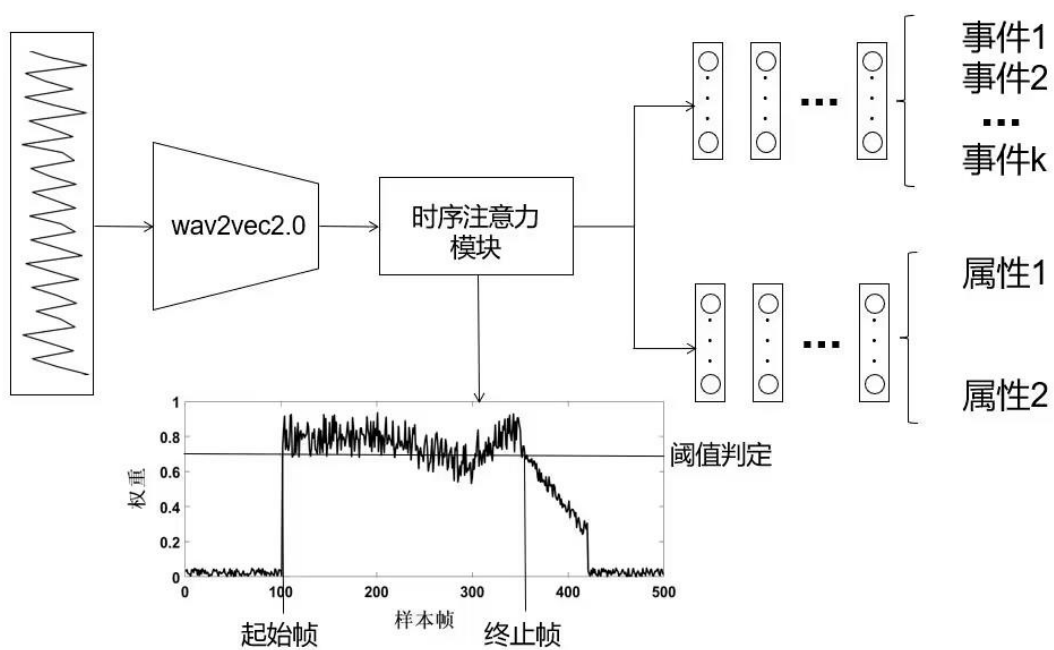


图 5

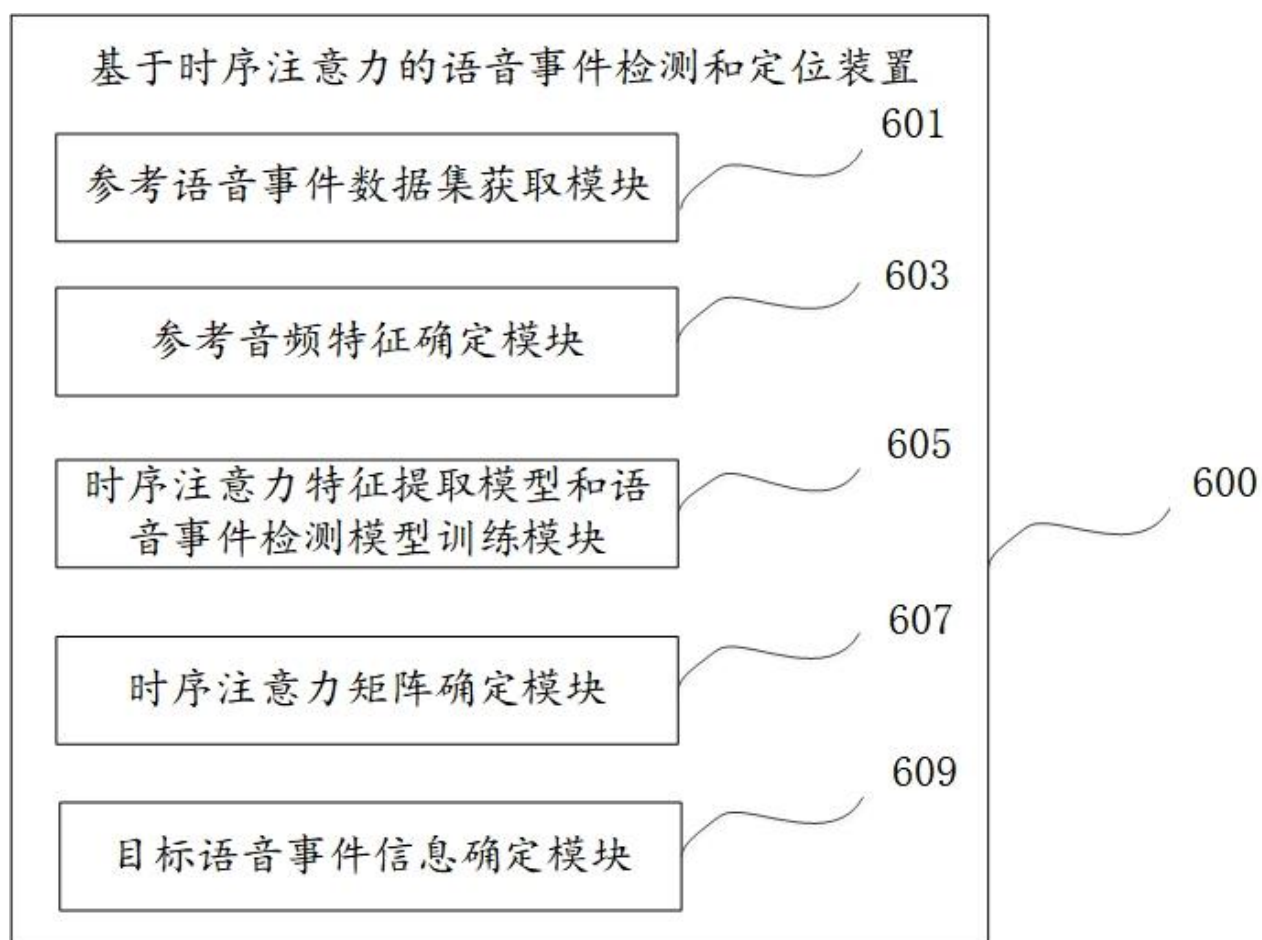


图 6

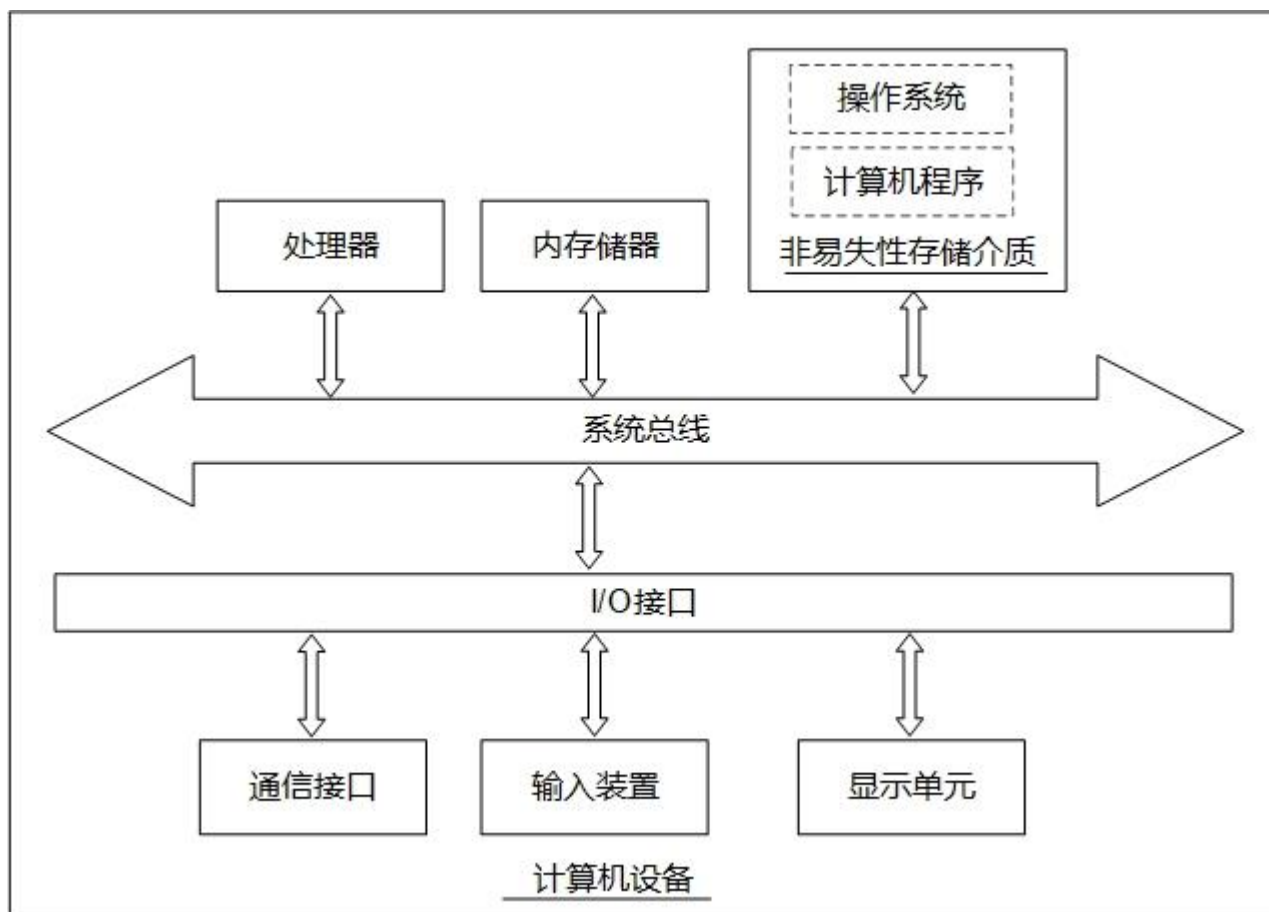


图 7