

说话人分离模型训练方法、两说话人分离方法及相关设备

本申请要求于 2018 年 05 月 28 日提交中国专利局，申请号为 201810519521.6，发明名称为“说话人分离模型训练方法、两说话人分离方法及相关设备”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。

技术领域

本申请涉及生物识别技术领域，具体涉及一种说话人分离模型训练方法、两说话人分离方法、终端及存储介质。

背景技术

随着音频处理技术的不断提高，从海量的数据中，如电话录音、新闻广播、会议录音等，获取感兴趣的特定人声已成为研究热点。说话人分离技术是指从多人对话中自动地将语音依据说话人进行划分，并加以标记的过程，即解决的是“什么时候由谁说”的问题。

而两说话人分离则是指将录制在同一段音轨上的两个说话人轮流讲话的录音分离，成为两段音轨，每段音轨只包含其中一个说话人的讲话录音。两说话人广泛应用于诸多领域，在广播，电视，媒体，客服中心等行业和领域有着广泛的需求。

传统的以贝叶斯信息准则（Bayesian Information Criterion, BIC）作为相似性度量的说话人分离技术，在短时对话的分离任务中能取得较好的效果，但是随着对话时长的增加，BIC 的单高斯模型不足以描述不同说话人数据的分布，因而其说话人分离的效果较差。

发明内容

鉴于以上内容，有必要提出一种说话人分离模型训练方法、两说话人分离方法、终端及存储介质，事先训练说话人分离模型，能够显著的增强模型对输入的语音数据的特征提取能力，减轻网络层次加深时而性能降低的风险；根据训练好的说话人分离模型对两说话人的语音进行分离，提高了两说话人分离的准确率，尤其是在长时对话的分离任务中能取得更佳的分离效果。

本申请的第一方面提供一种说话人分离模型训练方法，所述方法包括：

获取多个人的多个音频数据；

对每一个音频数据进行预处理；

提取预处理后的音频数据的音频特征；

将所述音频特征输入至预设神经网络模型中进行训练得到矢量特征；

获取第一说话人的第一矢量特征及第二矢量特征，并根据预设第一相似
度函数计算所述第一矢量特征与所述第二矢量特征之间的第一相似度；

获取第二说话人的第三矢量特征，并根据预设第二相似度函数计算所述
第一矢量特征与所述第三矢量特征之间的第二相似度；

将所述第一相似度及所述第二相似度输入至预设损失函数中计算损失函
数值，当损失函数值小于或等于预设损失函数阈值时，结束说话人分离模型
的训练过程，并更新所述说话人分离模型中的参数。

本申请的第二方面提供一种两说话人分离方法，所述方法包括：

1) 对待分离的语音信号进行预处理；

2) 从所述预处理后的语音信号的起始位置开始，建立左右相邻的第一滑
窗及第二滑窗，根据所述第一滑窗及第二滑窗的分割点，得到第一语音片
段及第二语音片段；

3) 将所述第一语音片段送入说话人分离模型进行特征提取得到第一语音
矢量，将所述第二语音片段送入所述说话人分离模型进行特征提取得到第二
语音矢量，其中，所述说话人分离模型利用权利要求 1 至 5 任意一项所述的
方法训练得到；

4) 计算所述第一语音矢量及所述第二语音矢量之间的距离值以得到此次分割点对应的距离值；

5) 将所述第一滑窗及所述第二滑窗同时沿时间轴方向移动预设时间段，并重复执行步骤 2) -5)，直至所述第二滑窗到达所述预处理后的语音信号的结尾；

6) 获取每次分割点对应的距离值，根据所述距离值确定局部最大值；

7) 根据所述局部最大值对应的分割点将所述待分离的语音信号进行分割，得到新的语音片段；

8) 将所述新的语音片段进行聚类为两个不同说话人的语音片段。

本申请的第三方面提供一种终端，所述终端包括处理器和存储器，所述处理器用于执行所述存储器中存储的计算机可读指令时实现所述说话人分离模型训练方法及/或所述两说话人分离方法。

本申请的第四方面提供一种非易失性可读存储介质，所述非易失性可读存储介质上存储有计算机可读指令，所述计算机可读指令被处理器执行时实现所述说话人分离模型训练方法及/或所述两说话人分离方法。

本申请所述的两说话人分离方法、终端及存储介质，事先训练说话人分离模型，能够显著的增强模型对输入的语音数据的特征提取能力，减轻网络层次加深时而性能降低的风险；根据训练好的说话人分离模型对两说话人的语音进行分离，提高了两说话人分离的准确率，尤其是在长时对话的分离任务中能取得更佳的分离效果。

附图说明

图 1 是本申请实施例一提供的说话人分离模型训练方法的流程图。

图 2 是本申请实施例二提供的两说话人分离方法的流程图。

图 3 是本申请实施例提供的根据分割点及对应的距离值确定局部最大值的示意图。

图 4 是本申请实施例三提供的说话人分离模型训练装置的功能模块图。

图 5 是本申请实施例四提供的两说话人分离装置的功能模块图。

图 6 是本申请实施例五提供的终端的示意图。

如下具体实施方式将结合上述附图进一步说明本申请。

具体实施方式

为了能够更清楚地理解本申请的上述目的、特征和优点，下面结合附图和具体实施例对本申请进行详细描述。需要说明的是，在不冲突的情况下，本申请的实施例及实施例中的特征可以相互组合。

在下面的描述中阐述了很多具体细节以便于充分理解本申请，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在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本申请保护的范围。

除非另有定义，本文所使用的所有的技术和科学术语与属于本申请的技术领域的技术人员通常理解的含义相同。本文中在本申请的说明书中所使用的术语只是为了描述具体的实施例的目的，不是旨在于限制本申请。

本申请实施例的说话人分离模型训练方法及/或两说话人分离方法应用在一个或者多个终端中。所述说话人分离模型训练方法及/或两说话人分离方法也可以应用于由终端和通过网络与所述终端进行连接的服务器所构成的硬件环境中。网络包括但不限于：广域网、城域网或局域网。本申请实施例的说话人分离模型训练方法及/或两说话人分离方法可以同时由服务器来执行，也可以同时由终端来执行；还可以是由服务器和终端共同执行，例如，本申请实施例的说话人分离模型训练方法由服务器来执行，本申请实施例的两说话人分离方法由终端来执行。

对于需要进行说话人分离模型训练方法及/或两说话人分离方法的终端，可以直接在终端上集成本申请的方法所提供的说话人分离模型训练及/或两说话人分离功能，或者安装用于实现本申请的方法的客户端。再如，本申请所提供的方法还可以以软件开发工具包（Software Development Kit, SDK）的形式运行在服务器等设备上，以 SDK 的形式提供说话人分离模型训练及/或两说话人分离功能的接口，终端或其他设备通过提供的接口即可实现两个说话人的模型的训练及/或分离。

实施例一

图 1 是本申请实施例一提供的说话人分离模型训练方法的流程图。根据不同的需求，该流程图中的执行顺序可以改变，某些步骤可以省略。

S11、获取多个人的多个音频数据。

本实施例中，所述多个音频数据的获取可包括以下两种方法：

（1）事先设置音频设备（例如，录音机等），通过所述音频设备现场录制多个人的语音，得到音频数据。

（2）从音频数据集中获取多个音频数据。

所述音频数据集为开源的数据集，例如，UBM 数据集、TV 数据集，开源的音频数据集专用于训练说话人分离模型，并对训练好的说话人分离模型的准确率进行测试。所述 UBM 数据集及 TV 数据集来自 NIST04、NIST05 及 NIST06，包括大约 500h 的音频数据，共 577 个说话人，平均每个人约有 15 句话。

S12、对每一个音频数据进行预处理。

本实施例中，当获取到多个音频数据后，需要对多个音频数据进行预处理，对音频数据进行预处理包括以下一种或多种的组合：

1）对音频数据进行降噪处理；

获取到的语音数据有可能含有各种各样的噪声，为了从含噪语音数据中提取尽可能纯净的原始语音，可以使用低通滤波器去除所述音频数据中的白噪声及随机噪声等。

2）对降噪处理后的音频数据进行语音端点检测（Voice Activity Detection、VAD），删除无效语音，得到标准语音数据样本；

本实施例中，可以采用双门限比较法进行语音端点检测，检测音频数据中的有效语音及无效语音，所述有效语音是说话人的语音，所述无效语音相对于所述有效语音而言，包括但不限于：静音。

3）对标准语音数据样本进行标注，以表明所述标准语音数据样本所属的说话人。

所述标注是指对音频数据打上身份属性标签，例如，第一音频数据所属的说话人为 A，打上身份属性标签 01，第二音频数据所属的说话人为 B，打上身份属性标签 02。

S13、提取预处理后的音频数据的音频特征。

本实施例中，可以采用 MFCC（梅尔频率倒谱系数，Mel Frequency Cepstrum Coefficient）频谱特征）等提取预处理后的音频数据的音频特征。梅尔频率倒谱系数为现有技术，本申请在此不详细赘述。

S14、将所述音频特征输入至预设神经网络模型中进行训练得到矢量特征。

本实施例中，所述预设神经网络模型采用预设层数的神经网络（Deep architecture）结构堆叠而成，所述预设层数为预先设置的网络层数。例如，预先设置 9-12 层神经网络结构进行组合成神经网络训练模型。

具体地，每一层神经网络结构包括：第一卷积层（Convolutional Layer）、第一修正线性单元（Rectified Linear Units, ReLU）、第二卷积层、第二修正

线性单元、平均层 (Average sentence)、全连接层 (Affine) 及归一化层 (Length Normalization)，其中，卷积层的卷积核为 3×3 ，步长为 1×1 ，通道数为 64。

所述预设神经网络模型对输入的音频特征进行训练的具体过程为：

1) 将所述音频特征输入至所述第一卷积层进行第一卷积处理，得到第一卷积特征；

2) 将所述第一卷积特征输入至所述第一修正线性单元进行第一修正处理，得到第一修正特征；

3) 将所述第一修正特征输入至所述第二卷积层进行第二卷积处理，得到第二卷积特征；

4) 对所述音频特征与所述第二卷积特征进行求和后输入至所述第二修正线性单元，得到第二修正特征；

5) 将所述第二修正特征依次输入至所述平均层、全连接层、归一化层进行处理，得到一维矢量特征。

所述平均层作用是作为一个 temporal pool (“temporal pool”表示沿时间轴向对矢量序列求平均值)，它将上一层前向长短期记忆网络 (Long Short-Term Memory, LSTM) 和后向 LSTM 输出的矢量序列都进行平均化，得到一个前向平均矢量和一个后向平均矢量；所述全连接层的作用是将前向平均矢量和后向平均矢量串联成一个矢量；所述归一化层可以采用归一化函数对串联后形成的矢量进行处理，得到长度为 1 的归一化后的一维矢量特征。

本实施例中，所述归一化函数可以是欧式距离函数，也可以是曼哈顿距离函数、最小绝对误差。优选地，采用欧式距离函数对经过所述全连接层处理后的矢量进行归一化处理，得到归一化的一维矢量特征。归一化处理能够对经过所述全连接层处理后的矢量进行压缩，使得经过所述全连接层处理后的矢量具有鲁棒性，从而进一步提高所述说话人分离模型的鲁棒性。另外，采用欧式距离函数进行归一化，能避免所述说话人分离模型过拟合，进而提升了所述说话人分离模型的泛化能力，对后续所述说话人分离模型中的神经网络参数的优化求解变得更稳定和快速。

S15、获取第一说话人的第一矢量特征及第二矢量特征，并根据预设第一相似度函数计算所述第一矢量特征与所述第二矢量特征之间的第一相似度。

本实施例中，所述预设第一相似度函数可以为余弦相似度函数，如下 (1-1) 所示：

$$\text{COS}(x_i, x_j) = x_i^T x_j \quad (1-1)$$

其中， x_i 代表第一说话人的第一矢量特征， x_j 代表第一说话人的第二特征矢量， $\text{COS}(x_i, x_j)$ 为计算得到的第一相似度。

S16、获取第二说话人的第三矢量特征，并根据预设第二相似度函数计算所述第一矢量特征与所述第三矢量特征之间的第二相似度。

本实施例中，所述预设第二相似度函数可以与所述预设第一相似度函数相同，也可以不相同。

优选地，所述预设第二相似度函数为 LP 范数，如下 (1-2) 所示：

$$L_p(x_i, y_i) = \left(\sum_{i=1}^n |x_i - y_i|^p \right)^{1/p} \quad (1-2)$$

其中， x_i 代表第一说话人的第一矢量特征， y_i 代表第二说话人的第三矢

量特征, $L_p(x_i, y_i)$ 为计算得到的第二相似度。

S17、将所述第一相似度及所述第二相似度输入至预设损失函数中计算损失函数值, 当损失函数值小于或等于预设损失函数阈值时, 结束说话人分离模型的训练过程, 并更新所述说话人分离模型中的参数。

本实施例中, 所述预设损失函数如下(1-3)所示:

$$L = \sum_i^N \max(S_i^{13} - S_i^{12} + \alpha, 0) \quad (1-3)$$

其中, α 为正常数, 一般取值范围在 0.05~0.2 之间。 S_i^{13} 为所述第二相似度, 即第一说话人的第一矢量特征与第二说话人的第三矢量特征之间的相似度, S_i^{12} 为所述第一相似度, 即第一说话人的第一矢量特征与第一说话人的第二矢量特征之间的相似度, L 为计算出的损失函数值。

由上述可知, 本申请通过获取多个音频数据做训练样本, 并对音频数据进行标注; 提取音频数据的音频特征后输入至预设神经网络模型中进行训练得到矢量特征, 获取第一说话人的第一矢量特征及第二矢量特征, 并根据预设第一相似度函数计算所述第一矢量特征与所述第二矢量特征之间的第一相似度; 获取第二说话人的第三矢量特征, 并根据预设第二相似度函数计算所述第一矢量特征与所述第三矢量特征之间的第二相似度; 将所述第一相似度及所述第二相似度输入至预设损失函数中计算损失函数值, 当损失函数值小于或等于预设损失函数阈值时, 结束说话人分离模型的训练过程, 并更新所述说话人分离模型中的参数。本申请基于卷积神经网络训练的说话人分离模型具有较强的特征提取能力, 减轻了网络层次加深时而性能降低的风险, 另外, 能够保证同一个人的不同的语音数据的矢量特征尽可能的相同, 不同人的语音数据的矢量特征尽可能的不同, 进而计算出的损失函数能够更快的达到收敛的条件, 节省说话人分离模型的训练时间, 提高说话人分离模型的效率。

实施例二

图2是本申请实施例二提供的两说话人分离方法的流程图。根据不同的需求, 该流程图中的执行顺序可以改变, 某些步骤可以省略。

S21、对待分离的语音信号进行预处理。

本实施例中, 所述对待分离的语音信号进行预处理的过程包括:

1) 预加重处理

本实施例中, 可以采用数字滤波器对所述待分离的语音信号进行预加重处理, 提升高频部分的语音信号。具体如下(2-1)所示:

$$S(n) = S(n) - a * S(n-1) \quad (2-1)$$

其中, $S(n)$ 为待分离的语音信号, a 为预加重系数, 一般 a 取 0.95, $\tilde{S}(n)$ 为经过预加重处理后的语音信号。

由于人类发声器官本身及采集语音信号的设备等因素易导致采集的语音信号出现混叠、高次谐波失真等问题, 通过对待分离的语音信号进行预加重处理, 可以补偿语音信号受到发音系统所抑制的高频部分, 突出高频的共振峰, 保证语音信号更均匀、平滑, 提高后续语音信号的分离效果。

2) 分帧处理

对所述待分离的语音信号根据预设分帧参数进行分帧处理, 所述预设分帧参数可以为, 例如, 帧长为 10~30ms, 优选为每隔 25 毫秒分帧, 以获得多个语音帧。如此, 对于整体的待分离的语音信号而言, 进行分帧处理后得到的是由每一帧特征参数组成的特征参数时间序列。

S22、从所述预处理后的语音信号的起始位置开始，建立左右相邻的第一滑窗及第二滑窗，根据所述第一滑窗及第二滑窗的分割点，得到第一语音片段及第二语音片段。

本实施例中，所述第一滑窗及所述第二滑窗的长度可以为 0.7-2 秒，所述第一滑窗及第二滑窗的分割点即为所述预处理后的语音信号的可能的说话人分割点。第一滑窗对应得到第一语音片段，第二滑窗对应得到第二语音片段，所述第一滑窗与所述第二滑窗的分割点也即所述第一语音片段和所述第二语音片段的分割点。

S23、将所述第一语音片段送入说话人分离模型进行特征提取得到第一语音矢量，将所述第二语音片段送入说话人分离模型进行特征提取得到第二语音矢量。

将第一语音片段送入训练好的说话人分离模型中，训练好的说话人分离模型提取第一语音片段的梅尔频谱特征（Mel Frequency Cepstrum Coefficient, MFCC），得到第一语音矢量。将第二语音片段送入训练好的说话人分离模型中，训练好的说话人分离模型提取第二语音片段的梅尔频谱特征（Mel Frequency Cepstrum Coefficient, MFCC），得到第二语音矢量。

S24、计算所述第一语音矢量及所述第二语音矢量之间的距离值以得到此次分割点对应的距离值。

本实施例中，可以采用预设距离函数计算每个第一语音矢量及每个第二语音矢量之间的距离值。所述预设距离函数为预先设置的距离函数，可以是，例如，欧式距离。本申请对采用欧式距离函数计算第一语音矢量及第二语音矢量的距离值的过程，不再做具体说明。

S25、将所述第一滑窗及所述第二滑窗同时沿时间轴方向移动预设时间段，并重复执行步骤 S22-S25，直至所述第二滑窗到达所述预处理后的语音信号的结尾。

本实施例中，所述预设时间段可以为 5ms。通过第一滑窗和第二滑窗在所述预处理后的语音信号上进行滑动，可以得到多个滑窗的分割点，从而对应得到多个第一语音片段及多个第二语音片段。也就是说，每根据所述预设时间段同时移动所述第一滑窗及所述第二滑窗一次后，即可得到一个备选的分割点，每个备选的分割点即为第一语音片段和第二语音片段的分割点，对应可以计算得到一个距离值。有多少个分割点，对应多少个距离值。

S26、获取每次分割点对应的距离值，根据所述距离值确定局部最大值。

本实施例中，所述根据所述距离值确定局部最大值的具体过程可以包括：

1) 将所有分割点对应的距离值按照分割点的时间顺序进行排列。

2) 判断 $f(n)$ 是否大于 $f(n-1)$ ，且 $f(n)$ 是否大于 $f(n+1)$ ，其中， $f(n)$ 为当前分割点对应的距离值， $f(n-1)$ 为当前分割点之前一个的分割点对应的距离值， $f(n+1)$ 为当前分割点之后一个的分割点对应的距离值。

3) 如果 $f(n) \geq f(n-1)$ 且 $f(n) \geq f(n+1)$ ，则确定 $f(n)$ 为局部最大值。

可以确定多个局部最大值，每个局部最大值对应的分割点为要寻找的分割点。

举例而言，待分离的语音信号经过第一滑窗和第二滑窗的移动后得到 10 个分割点，例如 T1、T2、T3、T4、T5、T6、T7、T8、T9、T10，每个分割点分别对应一个距离值，例如 S1、S2、S3、S4、S5、S6、S7、S8、S9、S10。将 10 个距离值按照分割点的时间顺序进行排列，如果 $S2 \geq S1$ ，且 $S2 \geq S3$ ，则 S2 为局部最大值。接着判断 S4 是否 $\geq S3$ ，且 S4 是否 $\geq S5$ ，若是，则 S4 为局部最大值，若否，则依次类推，判断其余当前距离值分别与之之前的距离值或者之后的距离值之间的大小关系，确定局部最大值。

作为一种可选的实施方式，所述根据所述距离值确定局部最大值还可以

包括：

以分割点为横轴，以分割点对应的距离值为纵轴，绘制一条平滑的曲线；
求解所述曲线上的点的切线；

将切线的斜率为零的点对应的距离值确定为局部最大值。

为了直观的表现局部最大值，可以以分割点为横轴，以分割点对应的距离值为纵轴，绘制一条平滑的曲线，如图 3 所示。求解图 3 中每个点的切线可知，S2、S4、S6、S9 对应的点的切线的斜率为零，则将 S2、S4、S6、S9 确定为局部最大值。

S27、根据所述局部最大值对应的分割点将所述待分离的语音信号进行分割，得到新的语音片段。

本实施例中，确定局部最大值后，以所述局部最大值对应的分割点为新的分割点对所述待分离的语音信号再次进行分割，从而得到新的多个语音片段。再分割的过程是指从待分离的语音信号中找寻两个不同说话人身份转变的时间点，然后根据时间点可以将语音分割成若干短语音段，经分割后的每个短语音段只会包含一个说话人的信息。

举例而言，S2、S4、S6、S9 为局部最大值，则以对应的分割点 T2、T4、T6、T9 为新的分割点将所述待分离的语音信号进行分割，得到 5 个新的语音片段，每个新的语音片段只包含一个说话人的语音信息。

S28、将所述新的语音片段进行聚类为两个不同说话人的语音片段。

聚类的过程是指将分割后的所有属于同一个说话人的小片段通过一些聚类的方法，再重新组合在一起。

本实施例中，所述聚类方法可以为 K-means 聚类或者自底向上的层次聚类（HAC），所述聚类方法为现有技术，本申请在此不再详细赘述。

由上述可知，本申请将对待分离的语音信号进行预处理后，从所述预处理后的语音信号的起始位置开始，建立左右相邻的第一滑窗及第二滑窗，根据所述第一滑窗及第二滑窗的分割点，得到第一语音片段及第二语音片段；分别将所述第一语音片段及所述第二语音片段送入说话人分离模型进行特征提取得到第一语音矢量及第二语音矢量；计算所述第一语音矢量及所述第二语音矢量之间的距离值；然后按照预设时间段移动滑窗，每移动一次滑窗即可得到两个语音片段，直至所述第二滑窗到达所述预处理后的语音信号的结尾；获取每次分割点对应的距离值，根据所述距离值确定局部最大值；根据所述局部最大值对应的分割点将所述待分离的语音信号进行分割，得到新的语音片段；将所述新的语音片段进行聚类为两个不同说话人的语音片段。本申请通过多次滑窗的过程，得到多个语音小片段，利用训练好的说话人分离模型提取语音小片段的特征后计算距离值，通过距离值之间的比较，确定局部最大值，以局部最大值对应的分割点为新的分割点再次对待分离的语音信号进行分割，即可得到两个说话人的语音片段，分离效果好。

以上所述，仅是本申请的具体实施方式，但本申请的保护范围并不局限于此，对于本领域的普通技术人员来说，在不脱离本申请创造构思的前提下，还可以做出改进，但这些均属于本申请的保护范围。

下面结合第 4 至 6 图，分别对实现上述说话人分离模型训练方法及两说话人分离方法的终端的功能模块及硬件结构进行介绍。

实施例三

图 4 为本申请说话人分离模型训练装置较佳实施例中的功能模块图。

在一些实施例中，所述说话人分离模型训练装置 40 运行于终端中。所述说话人分离模型训练装置 40 可以包括多个由程序代码段所组成的功能模块。所述说话人分离模型训练装置 40 中的各个程序段的程序代码可以存储于存储器中，并由至少一个处理器所执行，以执行（详见图 1 及其相关描述）对

说话人分离模型进行训练。

本实施例中，所述终端的说话人分离模型训练装置 40 根据其所执行的功能，可以被划分为多个功能模块。所述功能模块可以包括：获取模块 401、预处理模块 402、特征提取模块 403、训练模块 404、计算模块 405 及更新模块 406。本申请所称的模块是指一种能够被至少一个处理器所执行并且能够完成固定功能的一系列计算机可读指令段，其存储在存储器中。在一些实施例中，关于各模块的功能将在后续的实施例中详述。

获取模块 401，用于获取多个人的多个音频数据。

本实施例中，所述多个音频数据的获取可包括以下两种方法：

(1) 事先设置音频设备（例如，录音机等），通过所述音频设备现场录取多个人的语音，得到音频数据。

(2) 从音频数据集中获取多个音频数据。

所述音频数据集为开源的数据集，例如，UBM 数据集、TV 数据集，开源的音频数据集专用于训练说话人分离模型，并对训练好的说话人分离模型的准确率进行测试。所述 UBM 数据集及 TV 数据集来自 NIST04、NIST05 及 NIST06，包括大约 500h 的音频数据，共 577 个说话人，平均每个人约有 15 句话。

预处理模块 402，用于对每一个音频数据进行预处理。

本实施例中，当获取到多个音频数据后，需要对多个音频数据进行预处理，预处理模块 402 对音频数据进行预处理包括以下一种或多种的组合：

1) 对音频数据进行降噪处理；

获取到的语音数据有可能含有各种各样的噪声，为了从含噪语音数据中提取尽可能纯净的原始语音，可以使用低通滤波器去除所述音频数据中的白噪声及随机噪声等；

2) 对降噪处理后的音频数据进行语音端点检测（Voice Activity Detection、VAD），删除无效语音，得到标准语音数据样本。

本实施例中，可以采用双门限比较法进行语音端点检测，检测音频数据中的有效语音及无效语音，所述有效语音是说话人的语音，所述无效语音相对于所述有效语音而言，包括但不限于：静音。

3) 对标准语音数据样本进行标注，以表明所述标准语音数据样本所属的说话人。

所述标注是指对音频数据打上身份属性标签，例如，第一音频数据所属的说话人为 A，打上身份属性标签 01，第二音频数据所属的说话人为 B，打上身份属性标签 02。

特征提取模块 403，用于提取预处理后的音频数据的音频特征。

本实施例中，可以采用 MFCC（梅尔频率倒谱系数，Mel Frequency Cepstrum Coefficient）频谱特征等提取预处理后的音频数据的音频特征。梅尔频率倒谱系数为现有技术，本申请在此不详细赘述。

训练模块 404，用于将所述音频特征输入至预设神经网络模型中进行训练得到矢量特征。

本实施例中，所述预设神经网络模型采用预设层数的神经网络（Deep architecture）结构堆叠而成，所述预设层数为预先设置的网络层数。例如，预先设置 9-12 层神经网络结构进行组合成神经网络训练模型。

具体地，每一层神经网络结构包括：第一卷积层（Convolutional Layer）、第一修正线性单元（Rectified Linear Units, ReLU）、第二卷积层、第二修正线性单元、平均层（Average sentence）、全连接层（Affine）及归一化层（Length Normalization），其中，卷积层的卷积核为 3*3，步长为 1*1，通道数为 64。

所述训练模块 404 通过预设神经网络模型对输入的音频特征进行训练的

具体过程为：

1) 将所述音频特征输入至所述第一卷积层进行第一卷积处理，得到第一卷积特征；

2) 将所述第一卷积特征输入至所述第一修正线性单元进行第一修正处理，得到第一修正特征；

3) 将所述第一修正特征输入至所述第二卷积层进行第二卷积处理，得到第二卷积特征；

4) 对所述音频特征与所述第二卷积特征进行求和后输入至所述第二修正线性单元，得到第二修正特征；

5) 将所述第二修正特征依次输入至所述平均层、全连接层、归一化层进行处理，得到一维矢量特征。

所述平均层作用是作为一个 temporal pool (“temporal pool”表示沿时间轴向对矢量序列求平均值)，它将上一层前向长短期记忆网络 (Long Short-Term Memory, LSTM) 和后向 LSTM 输出的矢量序列都进行平均化，得到一个前向平均矢量和一个后向平均矢量；所述全连接层的作用是将前向平均矢量和后向平均矢量串联成一个矢量；所述归一化层可以采用归一化函数对串联后形成的矢量进行处理，得到长度为 1 的归一化后的一维矢量特征。

本实施例中，所述归一化函数可以是欧式距离函数，也可以是曼哈顿距离函数、最小绝对误差。优选地，采用欧式距离函数对经过所述全连接层处理后的矢量进行归一化处理，得到归一化的一维矢量特征。归一化处理能够对经过所述全连接层处理后的矢量进行压缩，使得经过所述全连接层处理后的矢量具有鲁棒性，从而进一步提高所述说话人分离模型的鲁棒性。另外，采用欧式距离函数进行归一化，能避免所述说话人分离模型过拟合，进而提升了所述说话人分离模型的泛化能力，对后续所述说话人分离模型中的神经网络参数的优化求解变得更稳定和快速。

计算模块 405，用于获取第一说话人的第一矢量特征及第二矢量特征，并根据预设第一相似度函数计算所述第一矢量特征与所述第二矢量特征之间的第一相似度。

本实施例中，所述预设第一相似度函数可以为余弦相似度函数，如下 (1-1) 所示：

$$\text{COS}(x_i, x_j) = x_i^T x_j \quad (1-1)$$

其中， x_i 代表第一说话人的第一矢量特征， x_j 代表第一说话人的第二特征矢量， $\text{COS}(x_i, x_j)$ 为计算得到的第一相似度。

所述计算模块 405，还用于获取第二说话人的第三矢量特征，并根据预设第二相似度函数计算所述第一矢量特征与所述第三矢量特征之间的第二相似度。

本实施例中，所述预设第二相似度函数可以与所述预设第一相似度函数相同，也可以不相同。

优选地，所述预设第二相似度函数为 LP 范数，如下 (1-2) 所示：

$$L_p(x_i, y_i) = \left(\sum_{i=1}^n |x_i - y_i|^p \right)^{1/p} \quad (1-2)$$

其中， x_i 代表第一说话人的第一矢量特征， y_i 代表第二说话人的第三矢量特征， $L_p(x_i, y_i)$ 为计算得到的第二相似度。

更新模块 406，用于将所述第一相似度及所述第二相似度输入至预设损

失函数中计算损失函数值，当损失函数值小于或等于预设损失函数阈值时，结束说话人分离模型的训练过程，并更新所述说话人分离模型中的参数。

本实施例中，所述预设损失函数如下（1-3）所示：

$$L = \sum_i^N \max(s_i^{13} - s_i^{12} + \alpha, 0) \quad (1-3)$$

其中， α 为正常数，一般取值范围在 0.05~0.2 之间。 s_i^{13} 为所述第二相似度，即第一说话人的第一矢量特征与第二说话人的第三矢量特征之间的相似度， s_i^{12} 为所述第一相似度，即第一说话人的第一矢量特征与第一说话人的第二矢量特征之间的相似度， L 为计算出的损失函数值。

由上述可知，本申请通过获取多个音频数据做训练样本，并对音频数据进行标注；提取音频数据的音频特征后输入至预设神经网络模型中进行训练得到矢量特征，获取第一说话人的第一矢量特征及第二矢量特征，并根据预设第一相似度函数计算所述第一矢量特征与所述第二矢量特征之间的第一相似度；获取第二说话人的第三矢量特征，并根据预设第二相似度函数计算所述第一矢量特征与所述第三矢量特征之间的第二相似度；将所述第一相似度及所述第二相似度输入至预设损失函数中计算损失函数值，当损失函数值小于或等于预设损失函数阈值时，结束说话人分离模型的训练过程，并更新所述说话人分离模型中的参数。本申请基于卷积神经网络训练的说话人分离模型具有较强的特征提取能力，减轻了网络层次加深时而性能降低的风险，另外，能够保证同一个人的不同的语音数据的矢量特征尽可能的相同，不同人的语音数据的矢量特征尽可能的不同，进而计算出的损失函数能够更快的达到收敛的条件，节省说话人分离模型的训练时间，提高说话人分离模型的效率。

实施例四

图 5 为本申请两说话人分离装置较佳实施例中的功能模块图。

在一些实施例中，所述两说话人分离装置 50 运行于终端中。所述两说话人分离装置 50 可以包括多个由程序代码段所组成的功能模块。所述两说话人分离装置 50 中的各个程序段的程序代码可以存储于存储器中，并由至少一个处理器所执行，以执行（详见图 2 及其相关描述）对两说话人的语音信号进行分离，得到两段语音片段，每段语音片段仅包含一个说话人的语音信息。

本实施例中，所述终端的两说话人分离装置 50 根据其所执行的功能，可以被划分为多个功能模块。所述功能模块可以包括：处理模块 501、第一分割模块 502、矢量提取模块 503、计算模块 504、比较模块 505、第二分割模块 506 及聚类模块 507。本申请所称的模块是指一种能够被至少一个处理器所执行并且能够完成固定功能的一系列计算机可读指令段，其存储在存储器中。在一些实施例中，关于各模块的功能将在后续的实施例中详述。

处理模块 501，用于对待分离的语音信号进行预处理。

本实施例中，所述处理模块 501 对待分离的语音信号进行预处理的过程包括：

1) 预加重处理

本实施例中，可以采用数字滤波器对所述待分离的语音信号进行预加重处理，提升高频部分的语音信号。具体如下（2-1）所示：

$$S(n) = S(n) - a * S(n-1) \quad (2-1)$$

其中， $S(n)$ 为待分离的语音信号， a 为预加重系数，一般 a 取 0.95， $\hat{S}(n)$ 为经过预加重处理后的语音信号。

由于人类发声器官本身及采集语音信号的设备等因素易导致采集的语音

信号出现混叠、高次谐波失真等问题，通过对待分离的语音信号进行预加重处理，可以补偿语音信号受到发音系统所抑制的高频部分，突出高频的共振峰，保证语音信号更均匀、平滑，提高后续语音信号的分离效果。

2) 分帧处理

对所述待分离的语音信号根据预设分帧参数进行分帧处理，所述预设分帧参数可以为，例如，帧长为 10~30ms，优选为每隔 25 毫秒分帧，以获得多个语音帧。如此，对于整体的待分离的语音信号而言，进行分帧处理后得到的是由每一帧特征参数组成的特征参数时间序列。

第一分割模块 502，用于从所述预处理后的语音信号的起始位置开始，建立左右相邻的第一滑窗及第二滑窗，根据所述第一滑窗及第二滑窗的分割点，得到第一语音片段及第二语音片段。

本实施例中，所述第一滑窗及所述第二滑窗的长度可以为 0.7-2 秒，所述第一滑窗及第二滑窗的分割点即为所述预处理后的语音信号的可能的说话人分割点。第一滑窗对应得到第一语音片段，第二滑窗对应得到第二语音片段，所述第一滑窗与所述第二滑窗的分割点也即所述第一语音片段和所述第二语音片段的分割点。

矢量提取模块 503，用于将所述第一语音片段送入说话人分离模型进行特征提取得到第一语音矢量，将所述第二语音片段送入说话人分离模型进行特征提取得到第二语音矢量。

将第一语音片段送入训练好的说话人分离模型中，训练好的说话人分离模型提取第一语音片段的梅尔频谱特征 (Mel Frequency Cepstrum Coefficient, MFCC)，得到第一语音矢量。将第二语音片段送入训练好的说话人分离模型中，训练好的说话人分离模型提取第二语音片段的梅尔频谱特征 (Mel Frequency Cepstrum Coefficient, MFCC)，得到第二语音矢量。

计算模块 504，用于计算所述第一语音矢量及所述第二语音矢量之间的距离值以得到此次分割点对应的距离值。

本实施例中，可以采用预设距离函数计算每个第一语音矢量及每个第二语音矢量之间的距离值。所述预设距离函数为预先设置的距离函数，可以是，例如，欧式距离。本申请对采用欧式距离函数计算第一语音矢量及第二语音矢量的距离值的过程，不再做具体说明。

将所述第一滑窗及所述第二滑窗同时沿时间轴方向移动预设时间段，并重复执行上述模块 (502-504)，直至所述第二滑窗到达所述预处理后的语音信号的结尾。

本实施例中，所述预设时间段可以为 5ms。通过第一滑窗和第二滑窗在所述预处理后的语音信号上进行滑动，可以得到多个滑窗的分割点，从而对应得到多个第一语音片段及多个第二语音片段。也就是说，每根据所述预设时间段同时移动所述第一滑窗及所述第二滑窗一次后，即可得到一个备选的分割点，每个备选的分割点即为第一语音片段和第二语音片段的分割点，对应可以计算得到一个距离值。有多少个分割点，对应多少个距离值。

比较模块 505，用于获取每次分割点对应的距离值，根据所述距离值确定局部最大值。

本实施例中，所述比较模块 505 根据所述距离值确定局部最大值的具体过程包括：

1) 将所有分割点对应的距离值按照分割点的时间顺序进行排列。

2) 判断 $f(n)$ 是否大于 $f(n-1)$ ，且 $f(n)$ 是否大于 $f(n+1)$ ，其中， $f(n)$ 为当前分割点对应的距离值， $f(n-1)$ 为当前分割点之前一个的分割点对应的距离值， $f(n+1)$ 为当前分割点之后一个的分割点对应的距离值。

3) 如果 $f(n) \geq f(n-1)$ 且 $f(n) \geq f(n+1)$ ，则确定 $f(n)$ 为局部最大值。

可以确定多个局部最大值，每个局部最大值对应的分割点为要寻找的分割点。

举例而言，待分离的语音信号经过第一滑窗和第二滑窗的移动后得到 10 个分割点，例如 T1、T2、T3、T4、T5、T6、T7、T8、T9、T10，每个分割点分别对应一个距离值，例如 S1、S2、S3、S4、S5、S6、S7、S8、S9、S10。将 10 个距离值按照分割点的时间顺序进行排列，如果 $S2 \geq S1$ ，且 $S2 \geq S3$ ，则 S2 为局部最大值。接着判断 S4 是否 $\geq S3$ ，且 S4 是否 $\geq S5$ ，若是，则 S4 为局部最大值，若否，则依次类推，判断其余当前距离值分别与之之前的距离值或者之后的距离值之间的大小关系，确定局部最大值。

作为一种可选的实施方式，所述比较模块 505 根据所述距离值确定局部最大值还可以包括：

以分割点为横轴，以分割点对应的距离值为纵轴，绘制一条平滑的曲线；
求解所述曲线上的点的切线；

将切线的斜率为零的点对应的距离值确定为局部最大值。

为了直观的表现局部最大值，可以以分割点为横轴，以分割点对应的距离值为纵轴，绘制一条平滑的曲线，如图 3 所示。求解图 3 中每个点的切线可知，S2、S4、S6、S9 对应的点的切线的斜率为零，则将 S2、S4、S6、S9 确定为局部最大值。

第二分割模块 506，用于根据所述局部最大值对应的分割点将所述待分离的语音信号进行分割，得到新的语音片段。

本实施例中，确定局部最大值后，以所述局部最大值对应的分割点为新的分割点对所述待分离的语音信号再次进行分割，从而得到新的多个语音片段。再分割的过程是指从待分离的语音信号中找寻两个不同说话人身份转变的时间点，然后根据时间点可以将语音分割成若干短语音段，经分割后的每个短语音段只会包含一个说话人的信息。

举例而言，S2、S4、S6、S9 为局部最大值，则以对应的分割点 T2、T4、T6、T9 为新的分割点将所述待分离的语音信号进行分割，得到 5 个新的语音片段，每个新的语音片段只包含一个说话人的语音信息。

聚类模块 507，用于将所述新的语音片段进行聚类为两个不同说话人的语音片段。

聚类的过程是指将分割后的所有属于同一个说话人的小片段通过一些聚类的方法，再重新组合在一起。

本实施例中，所述聚类方法可以为 K-means 聚类或者自底向上的层次聚类（HAC），所述聚类方法为现有技术，本申请在此不再详细赘述。

由上述可知，本申请将对待分离的语音信号进行预处理后，从所述预处理后的语音信号的起始位置开始，建立左右相邻的第一滑窗及第二滑窗，根据所述第一滑窗及第二滑窗的分割点，得到第一语音片段及第二语音片段；分别将所述第一语音片段及所述第二语音片段送入说话人分离模型进行特征提取得到第一语音矢量及第二语音矢量；计算所述第一语音矢量及所述第二语音矢量之间的距离值；然后按照预设时间段移动滑窗，每移动一次滑窗即可得到两个语音片段，直至所述第二滑窗到达所述预处理后的语音信号的结尾；获取每次分割点对应的距离值，根据所述距离值确定局部最大值；根据所述局部最大值对应的分割点将所述待分离的语音信号进行分割，得到新的语音片段；将所述新的语音片段进行聚类为两个不同说话人的语音片段。本申请通过多次滑窗的过程，得到多个语音小片段，利用训练好的说话人分离模型提取语音小片段的特征后计算距离值，通过距离值之间的比较，确定局部最大值，以局部最大值对应的分割点为新的分割点再次对待分离的语音信号进行分割，即可得到两个说话人的语音片段，分离效果好。

上述以软件功能模块的形式实现的集成的单元，可以存储在一个非易失性可读存储介质中。上述软件功能模块存储在一个存储介质中，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，双屏设备，或者网络设备）或处理器（processor）执行本申请各个实施例所述方法的部分。

实施例五

图 6 为本申请实施例五提供的终端的示意图。

所述终端 3 包括：存储器 31、至少一个处理器 32、存储在所述存储器 31 中并可在所述至少一个处理器 32 上运行的计算机可读指令 33 及至少一条通讯总线 34。

所述至少一个处理器 32 执行所述计算机可读指令 33 时实现上述说话人分离模型训练方法及/或两说话人分离方法实施例中的步骤。

示例性的，所述计算机可读指令 33 可以被分割成一个或多个模块/单元，所述一个或者多个模块/单元被存储在所述存储器 31 中，并由所述至少一个处理器 32 执行，以完成本申请说话人分离模型训练方法及/或两说话人分离方法。所述一个或多个模块/单元可以是能够完成特定功能的一系列计算机可读指令段，该指令段用于描述所述计算机可读指令 33 在所述终端 3 中的执行过程。

所述终端 3 可以是桌上型计算机、笔记本、掌上电脑及云端服务器等计算设备。本领域技术人员可以理解，所述示意图 3 仅仅是终端 3 的示例，并不构成对终端 3 的限定，可以包括比图示更多或更少的部件，或者组合某些部件，或者不同的部件，例如所述终端 3 还可以包括输入输出设备、网络接入设备、总线等。

所述至少一个处理器 32 可以是中央处理单元（Central Processing Unit, CPU），还可以是其他通用处理器、数字信号处理器（Digital Signal Processor, DSP）、专用集成电路（Application Specific Integrated Circuit, ASIC）、现场可编程门阵列（Field-Programmable Gate Array, FPGA）或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等。该处理器 32 可以是微处理器或者该处理器 32 也可以是任何常规的处理器等，所述处理器 32 是所述终端 3 的控制中心，利用各种接口和线路连接整个终端 3 的各个部分。

所述存储器 31 可用于存储所述计算机可读指令 33 和/或模块/单元，所述处理器 32 通过运行或执行存储在所述存储器 31 内的计算机可读指令和/或模块/单元，以及调用存储在存储器 31 内的数据，实现所述终端 3 的各种功能。所述存储器 31 可主要包括存储程序区和存储数据区，其中，存储程序区可存储操作系统、至少一个功能所需的应用程序（比如声音播放功能、图像播放功能等）等；存储数据区可存储根据终端 3 的使用所创建的数据（比如音频数据、电话本等）等。此外，存储器 31 可以包括高速随机存取存储器，还可以包括非易失性存储器，例如硬盘、内存、插接式硬盘，智能存储卡（Smart Media Card, SMC），安全数字（Secure Digital, SD）卡，闪存卡（Flash Card）、至少一个磁盘存储器件、闪存器件、或其他易失性固态存储器件。

所述终端 3 集成的模块/单元如果以软件功能单元的形式实现并作为独立的产品销售或使用，可以存储在一个非易失性可读存储介质中。基于这样的理解，本申请实现上述实施例方法中的全部或部分流程，也可以通过计算机可读指令来指令相关的硬件来完成，所述的计算机可读指令可存储于一非易失性可读存储介质中，该计算机可读指令在被处理器执行时，可实现上述各个方法实施例的步骤。其中，所述计算机可读指令包括计算机可读指令代码，所述计算机可读指令代码可以为源代码形式、对象代码形式、可执行文件或某些中间形式等。所述非易失性可读介质可以包括：能够携带所述计算机可读指令代码的任何实体或装置、记录介质、U 盘、移动硬盘、磁碟、

光盘、计算机存储器、只读存储器（ROM，Read-Only Memory）、随机存取存储器（RAM，Random Access Memory）、电载波信号、电信号以及软件分发介质等。需要说明的是，所述非易失性可读介质包含的内容可以根据司法管辖区内立法和专利实践的要求进行适当的增减，例如在某些司法管辖区，根据立法和专利实践，非易失性可读介质不包括电载波信号和电信号。

在本申请所提供的几个实施例中，应该理解到，所揭露的终端和方法，可以通过其它的方式实现。例如，以上所描述的终端实施例仅仅是示意性的，例如，所述单元的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式。

另外，在本申请各个实施例中的各功能单元可以集成在相同处理单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在相同单元中。上述集成的单元既可以采用硬件的形式实现，也可以采用硬件加软件功能模块的形式实现。

对于本领域技术人员而言，显然本申请不限于上述示范性实施例的细节，而且在不背离本申请的精神或基本特征的情况下，能够以其他的具体形式实现本申请。因此，无论从哪一点来看，均应将实施例看作是示范性的，而且是非限制性的，本申请的范围由所附权利要求而不是上述说明限定，因此旨在将落在权利要求的等同要件的含义和范围内的所有变化涵括在本申请内。不应将权利要求中的任何附图标记视为限制所涉及的权利要求。此外，显然“包括”一词不排除其他单元或，单数不排除复数。系统权利要求中陈述的多个单元或装置也可以由一个单元或装置通过软件或者硬件来实现。第一，第二等词语用来表示名称，而并不表示任何特定的顺序。

最后应说明的是，以上实施例仅用以说明本申请的技术方案而非限制，尽管参照较佳实施例对本申请进行了详细说明，本领域的普通技术人员应当理解，可以对本申请的技术方案进行修改或等同替换，而不脱离本申请技术方案的精神范围。

权 利 要 求 书

1. 一种说话人分离模型训练方法，其特征在于，所述方法包括：
获取多个人的多个音频数据；
对每一个音频数据进行预处理；
提取预处理后的音频数据的音频特征；
将所述音频特征输入至预设神经网络模型中进行训练得到矢量特征；
获取第一说话人的第一矢量特征及第二矢量特征，并根据预设第一相似度函数计算所述第一矢量特征与所述第二矢量特征之间的第一相似度；
获取第二说话人的第三矢量特征，并根据预设第二相似度函数计算所述第一矢量特征与所述第三矢量特征之间的第二相似度；
将所述第一相似度及所述第二相似度输入至预设损失函数中计算损失函数值，当损失函数值小于或等于预设损失函数阈值时，结束说话人分离模型的训练过程，并更新所述说话人分离模型中的参数。
2. 如权利要求 1 所述的方法，其特征在于，对所述音频数据进行预处理包括：
对所述音频数据进行降噪处理；
对降噪处理后的音频数据进行语音端点检测，删除无效语音，得到标准语音数据样本；
对标准语音数据样本进行标注，以表明所述标准语音数据样本所属的说话人。
3. 如权利要求 1 或 2 所述的方法，其特征在于，所述预设神经网络模型采用预设层数的神经网络结构堆叠而成；每一层神经网络结构包括：第一卷积层、第一修正线性单元、第二卷积层、第二修正线性单元、平均层、全连接层及归一化层，其中，卷积层的卷积核为 3×3 ，步长为 1×1 ，通道数为 64；
所述将所述音频特征输入至预设神经网络模型中进行训练得到矢量特征包括：
将所述音频特征输入至所述第一卷积层进行第一卷积处理，得到第一卷积特征；
将所述第一卷积特征输入至所述第一修正线性单元进行第一修正处理，得到第一修正特征；
将所述第一修正特征输入至所述第二卷积层进行第二卷积处理，得到第二卷积特征；
对所述音频特征与所述第二卷积特征进行求和后输入至所述第二修正线性单元，得到第二修正特征；
将所述第二修正特征依次输入至所述平均层、全连接层、归一化层进行处理，得到一维矢量特征。
4. 如权利要求 1 所述的方法，其特征在于，所述预设第一相似度函数为： $COS(x_i, x_j) = x_i^T x_j$ ，其中， x_i 代表第一说话人的第一矢量特征， x_j 代表第一说话人的第二特征矢量， $COS(x_i, x_j)$ 为计算得到的第一相似度；所述预设第二相似

度函数为: $L_p(x_i, y_i) = (\sum_{i=1}^n |x_i - y_i|^p)^{1/p}$, 其中, x_i 代表第一说话人的第一矢量特征, y_i 代表第二说话人的第三矢量特征, $L_p(x_i, y_i)$ 为计算得到的第二相似度。

5. 如权利要求 1 所述的方法, 其特征在于, 所述预设损失函数为:

$$L = \sum_i^N \max(s_i^{13} - s_i^{12} + \alpha, 0)$$
, 其中, α 为正常数, s_i^{13} 为所述第二相似度, s_i^{12} 为所述第一相似度, L 为计算出的损失函数值。

6. 一种两说话人分离方法, 其特征在于, 所述方法包括:

1) 对待分离的语音信号进行预处理;

2) 从所述预处理后的语音信号的起始位置开始, 建立左右相邻的第一滑窗及第二滑窗, 根据所述第一滑窗及第二滑窗的分割点, 得到第一语音片段及第二语音片段;

3) 将所述第一语音片段送入说话人分离模型进行特征提取得到第一语音矢量, 将所述第二语音片段送入所述说话人分离模型进行特征提取得到第二语音矢量, 其中, 所述说话人分离模型利用权利要求 1 至 5 中任意一项所述的方法训练得到;

4) 计算所述第一语音矢量及所述第二语音矢量之间的距离值以得到此次分割点对应的距离值;

5) 将所述第一滑窗及所述第二滑窗同时沿时间轴方向移动预设时间段, 并重复执行步骤 2) -5), 直至所述第二滑窗到达所述预处理后的语音信号的结尾;

6) 获取每次分割点对应的距离值, 根据所述距离值确定局部最大值;

7) 根据所述局部最大值对应的分割点将所述待分离的语音信号进行分割, 得到新的语音片段;

8) 将所述新的语音片段进行聚类为两个不同说话人的语音片段。

7. 如权利要求 6 所述的方法, 其特征在于, 所述根据所述距离值确定局部最大值包括:

将所有分割点对应的距离值按照分割点的时间顺序进行排列;

判断 $f(n)$ 是否大于 $f(n-1)$, 且 $f(n)$ 是否大于 $f(n+1)$, 其中, $f(n)$ 为当前分割点对应的距离值, $f(n-1)$ 为当前分割点之前一个的分割点对应的距离值, $f(n+1)$ 为当前分割点之后一个的分割点对应的距离值;

如果 $f(n) \geq f(n-1)$ 且 $f(n) \geq f(n+1)$, 则确定 $f(n)$ 为局部最大值。

8. 如权利要求 6 所述的方法, 其特征在于, 所述根据所述距离值确定局部最大值包括:

以分割点为横轴, 以分割点对应的距离值为纵轴, 绘制一条平滑的曲线; 求解所述曲线上的点的切线;

将切线的斜率为零的点对应的距离值确定为局部最大值。

9. 一种终端, 其特征在于, 所述终端包括处理器和存储器, 所述处理器用于执行所述存储器中存储的计算机可读指令时实现以下步骤:

获取多个人的多个音频数据；
对每一个音频数据进行预处理；
提取预处理后的音频数据的音频特征；
将所述音频特征输入至预设神经网络模型中进行训练得到矢量特征；
获取第一说话人的第一矢量特征及第二矢量特征，并根据预设第一相似度函数计算所述第一矢量特征与所述第二矢量特征之间的第一相似度；
获取第二说话人的第三矢量特征，并根据预设第二相似度函数计算所述第一矢量特征与所述第三矢量特征之间的第二相似度；
将所述第一相似度及所述第二相似度输入至预设损失函数中计算损失函数值，当损失函数值小于或等于预设损失函数阈值时，结束说话人分离模型的训练过程，并更新所述说话人分离模型中的参数。

10. 如权利要求 9 所述的终端，其特征在于，对所述音频数据进行预处理包括：

对所述音频数据进行降噪处理；
对降噪处理后的音频数据进行语音端点检测，删除无效语音，得到标准语音数据样本；
对标准语音数据样本进行标注，以表明所述标准语音数据样本所属的说话人。

11. 如权利要求 9 或 10 所述的终端，其特征在于，所述预设神经网络模型采用预设层数的神经网络结构堆叠而成；每一层神经网络结构包括：第一卷积层、第一修正线性单元、第二卷积层、第二修正线性单元、平均层、全连接层及归一化层，其中，卷积层的卷积核为 3×3 ，步长为 1×1 ，通道数为 64；

所述将所述音频特征输入至预设神经网络模型中进行训练得到矢量特征包括：

将所述音频特征输入至所述第一卷积层进行第一卷积处理，得到第一卷积特征；

将所述第一卷积特征输入至所述第一修正线性单元进行第一修正处理，得到第一修正特征；

将所述第一修正特征输入至所述第二卷积层进行第二卷积处理，得到第二卷积特征；

对所述音频特征与所述第二卷积特征进行求和后输入至所述第二修正线性单元，得到第二修正特征；

将所述第二修正特征依次输入至所述平均层、全连接层、归一化层进行处理，得到一维矢量特征。

12. 如权利要求 9 所述的终端，其特征在于，所述预设第一相似度函数为： $\text{COS}(x_i, x_j) = x_i^T x_j$ ，其中， x_i 代表第一说话人的第一矢量特征， x_j 代表第一说话人的第二特征矢量， $\text{COS}(x_i, x_j)$ 为计算得到的第一相似度；所述预设第二

相似度函数为： $L_p(x_i, y_i) = (\sum_{i=1}^n |x_i - y_i|^p)^{1/p}$ ，其中， x_i 代表第一说话人的第一矢量特征， y_i 代表第二说话人的第三矢量特征， $L_p(x_i, y_i)$ 为计算得到的第二相似度。

13. 一种终端，其特征在于，所述终端包括处理器和存储器，所述处理器用于执行所述存储器中存储的计算机可读指令时实现以下步骤：

1) 对待分离的语音信号进行预处理；

2) 从所述预处理后的语音信号的起始位置开始，建立左右相邻的第一滑窗及第二滑窗，根据所述第一滑窗及第二滑窗的分割点，得到第一语音片段及第二语音片段；

3) 将所述第一语音片段送入说话人分离模型进行特征提取得到第一语音矢量，将所述第二语音片段送入所述说话人分离模型进行特征提取得到第二语音矢量，其中，所述说话人分离模型利用权利要求 9 至 12 中任意一项所述的终端训练得到；

4) 计算所述第一语音矢量及所述第二语音矢量之间的距离值以得到此次分割点对应的距离值；

5) 将所述第一滑窗及所述第二滑窗同时沿时间轴方向移动预设时间段，并重复执行步骤 2) -5)，直至所述第二滑窗到达所述预处理后的语音信号的结尾；

6) 获取每次分割点对应的距离值，根据所述距离值确定局部最大值；

7) 根据所述局部最大值对应的分割点将所述待分离的语音信号进行分割，得到新的语音片段；

8) 将所述新的语音片段进行聚类为两个不同说话人的语音片段。

14. 如权利要求 13 所述的终端，其特征在于，所述根据所述距离值确定局部最大值包括：

将所有分割点对应的距离值按照分割点的时间顺序进行排列；

判断 $f(n)$ 是否大于 $f(n-1)$ ，且 $f(n)$ 是否大于 $f(n+1)$ ，其中， $f(n)$ 为当前分割点对应的距离值， $f(n-1)$ 为当前分割点之前一个的分割点对应的距离值， $f(n+1)$ 为当前分割点之后一个的分割点对应的距离值；

如果 $f(n) \geq f(n-1)$ 且 $f(n) \geq f(n+1)$ ，则确定 $f(n)$ 为局部最大值。

15. 一种非易失性可读存储介质，所述非易失性可读存储介质上存储有计算机可读指令，其特征在于，所述计算机可读指令被处理器执行时实现以下步骤：

获取多个人的多个音频数据；

对每一个音频数据进行预处理；

提取预处理后的音频数据的音频特征；

将所述音频特征输入至预设神经网络模型中进行训练得到矢量特征；

获取第一说话人的第一矢量特征及第二矢量特征，并根据预设第一相似度函数计算所述第一矢量特征与所述第二矢量特征之间的第一相似度；

获取第二说话人的第三矢量特征，并根据预设第二相似度函数计算所述

第一矢量特征与所述第三矢量特征之间的第二相似度；

将所述第一相似度及所述第二相似度输入至预设损失函数中计算损失函数值，当损失函数值小于或等于预设损失函数阈值时，结束说话人分离模型的训练过程，并更新所述说话人分离模型中的参数。

16. 如权利要求 15 所述的存储介质，其特征在于，对所述音频数据进行预处理包括：

对所述音频数据进行降噪处理；

对降噪处理后的音频数据进行语音端点检测，删除无效语音，得到标准语音数据样本；

对标准语音数据样本进行标注，以表明所述标准语音数据样本所属的说话人。

17. 如权利要求 15 或 16 所述的存储介质，其特征在于，所述预设神经网络模型采用预设层数的神经网络结构堆叠而成；每一层神经网络结构包括：第一卷积层、第一修正线性单元、第二卷积层、第二修正线性单元、平均层、全连接层及归一化层，其中，卷积层的卷积核为 3×3 ，步长为 1×1 ，通道数为 64；

所述将所述音频特征输入至预设神经网络模型中进行训练得到矢量特征包括：

将所述音频特征输入至所述第一卷积层进行第一卷积处理，得到第一卷积特征；

将所述第一卷积特征输入至所述第一修正线性单元进行第一修正处理，得到第一修正特征；

将所述第一修正特征输入至所述第二卷积层进行第二卷积处理，得到第二卷积特征；

对所述音频特征与所述第二卷积特征进行求和后输入至所述第二修正线性单元，得到第二修正特征；

将所述第二修正特征依次输入至所述平均层、全连接层、归一化层进行处理，得到一维矢量特征。

18. 如权利要求 15 所述的存储介质，其特征在于，所述预设第一相似度函数为： $COS(x_i, x_j) = x_i^T x_j$ ，其中， x_i 代表第一说话人的第一矢量特征， x_j 代表第一说话人的第二特征矢量， $COS(x_i, x_j)$ 为计算得到的第一相似度；所述预设第二相似度函数为： $L_p(x_i, y_i) = (\sum_{i=1}^n |x_i - y_i|^p)^{1/p}$ ，其中， x_i 代表第一说话人的第一矢量特征， y_i 代表第二说话人的第三矢量特征， $L_p(x_i, y_i)$ 为计算得到的第二相似度。

19. 一种非易失性可读存储介质，所述非易失性可读存储介质上存储有计算机可读指令，其特征在于，所述计算机可读指令被处理器执行时实现以下步骤：

1) 对待分离的语音信号进行预处理；

2) 从所述预处理后的语音信号的起始位置开始, 建立左右相邻的第一滑窗及第二滑窗, 根据所述第一滑窗及第二滑窗的分割点, 得到第一语音片段及第二语音片段;

3) 将所述第一语音片段送入说话人分离模型进行特征提取得到第一语音矢量, 将所述第二语音片段送入所述说话人分离模型进行特征提取得到第二语音矢量, 其中, 所述说话人分离模型利用权利要求 15 至 18 中任意一项所述的存储介质中的计算机可读指令被处理器执行时训练得到;

4) 计算所述第一语音矢量及所述第二语音矢量之间的距离值以得到此次分割点对应的距离值;

5) 将所述第一滑窗及所述第二滑窗同时沿时间轴方向移动预设时间段, 并重复执行步骤 2) -5), 直至所述第二滑窗到达所述预处理后的语音信号的结尾;

6) 获取每次分割点对应的距离值, 根据所述距离值确定局部最大值;

7) 根据所述局部最大值对应的分割点将所述待分离的语音信号进行分割, 得到新的语音片段;

8) 将所述新的语音片段进行聚类为两个不同说话人的语音片段。

20. 如权利要求 19 所述的存储介质, 其特征在于, 所述根据所述距离值确定局部最大值包括:

将所有分割点对应的距离值按照分割点的时间顺序进行排列;

判断 $f(n)$ 是否大于 $f(n-1)$, 且 $f(n)$ 是否大于 $f(n+1)$, 其中, $f(n)$ 为当前分割点对应的距离值, $f(n-1)$ 为当前分割点之前一个的分割点对应的距离值, $f(n+1)$ 为当前分割点之后一个的分割点对应的距离值;

如果 $f(n) \geq f(n-1)$ 且 $f(n) \geq f(n+1)$, 则确定 $f(n)$ 为局部最大值。