(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 103530432 A
(43)申请公布日 2014.01.22

(21)申请号 201310439113.7

(22)申请日 2013.09.24

(71)申请人 华南理工大学
地址 510640 广东省广州市天河区五山路381号

(72)发明人 王梓里 李艳雄 李广隆

(74)专利代理机构 广州市华学知识产权代理有限公司 44245
代理人 蔡茂略

(51)Int.Cl.
G06F 17/40(2006.01)
G10L 13/00(2006.01)

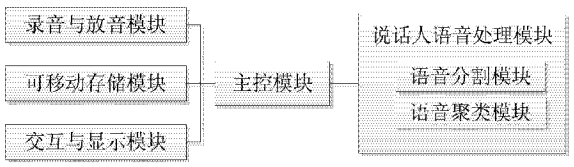
权利要求书2页 说明书6页 附图3页

(54)发明名称

一种具有语音提取功能的会议记录器及语音提取方法

(57)摘要

本发明公开了一种具有说话人语音提取功能的会议记录器，包括主控模块、录音与放音模块、可移动存储模块、交互与显示模块和说话人语音处理模块，其中说话人语音处理模块包含说话人分割模块和说话人聚类模块。主控模块将会议语音流传输至说话人分割模块，说话人分割模块检测上述语音流中说话人改变点，根据这些改变点将语音流分成多个语音段；说话人聚类模块利用谱聚类算法对分割出来的语音段进行说话人聚类，把相同说话人的语音段按顺序拼接在一起，得到说话人个数以及各个说话人的语音。本发明的会议记录器及语音提取方法，可以从会议语音中自动提取出各个说话人的语音，功能全面且使用方便。



CN 103530432 A

1. 一种具有语音提取功能的会议记录器，包括主控模块、录音与放音模块、可移动存储模块、交互与显示模块，其特征在于，还包括说话人语音处理模块，说话人语音处理模块包含说话人分割模块和说话人聚类模块，其中

说话人分割模块：主控模块将会议语音流传输至说话人分割模块，说话人分割模块检测上述会议语音流中说话人改变点，根据这些改变点将语音流分成多个语音段；

说话人聚类模块，利用谱聚类算法对说话人分割模块分割出来的语音段进行说话人聚类，把相同说话人的语音段按顺序拼接在一起，得到说话人个数以及各个说话人的语音。

2. 根据权利要求 1 所述的具有语音提取功能的会议记录器，其特征在于，所述的说话人分割模块，包含静音段和语音段检测模块、音频特征提取模块、说话人改变点检测模块和语音段分割模块，其中

静音段和语音段检测模块，利用基于门限判决的静音检测算法从上述读入的语音流中找出静音段和语音段；

音频特征提取模块，将上述语音段按顺序拼接成一个长语音段，并从长语音段中提取音频特征；

说话人改变点检测模块，利用上述提取出来的音频特征，根据贝叶斯信息准则，判断长语音段中相邻数据窗之间的相似度来检测说话人改变点；

语音段分割模块，根据上述说话人改变点，把语音流分割成多个语音段，且每个语音段只包含一个说话人。

3. 根据权利要求 2 所述的具有语音提取功能的会议记录器，其特征在于，静音段和语音段检测模块中，所述的基于门限判决的静音检测算法包含以下顺序的步骤：

（1）对读入的语音流进行分帧，并计算每帧语音的能量，得到语音流的能量特征矢量；

（2）计算能量门限；

（3）将每帧语音的能量与能量门限比较，低于能量门限的帧为静音帧，否则为语音帧，将相邻的静音帧按顺序拼接成一个静音段，将相邻的语音帧按顺序拼接成一个语音段。

4. 根据权利要求 2 所述的具有语音提取功能的会议记录器，其特征在于，音频特征提取模块中，所述的音频特征包括梅尔频率倒谱系数及其一阶差分。

5. 根据权利要求 1 所述的具有语音提取功能的会议记录器，其特征在于，所述录音与放音模块，包括麦克风、扬声器和音频处理芯片。

6. 根据权利要求 1 所述的具有语音提取功能的会议记录器，其特征在于，所述交互与显示模块，包括一个触摸屏及其控制电路，提供具有控制功能的用户交互界面，利用触摸屏与用户交互。

7. 根据权利要求 1 所述的具有语音提取功能的会议记录器，其特征在于，所述可移动存储模块，采用 SD 卡实现对数据的存储。

8. 一种语音提取方法，包含以下顺序的步骤：

（1）读入语音流：读入记录有多说话人语音的语音流；

（2）通过说话人语音处理模块对读入的语音流进行处理，其中说话人语音处理模块包括说话人分割模块和说话人聚类模块；

（3）通过说话人分割模块检测上述语音流中说话人改变点，根据这些改变点将语音流分成多个语音段；

(4)说话人聚类模块利用谱聚类算法对说话人分割模块分割出来的语音段进行说话人聚类,把相同说话人的语音段按顺序拼接在一起,得到说话人个数以及各个说话人的语音。

9. 根据权利要求 8 所述的语音提取方法,其特征在于,所述的步骤(3)具体包含以下步骤:

a、说话人分割模块包含静音段和语音段检测模块、音频特征提取模块、说话人改变点检测模块和语音段分割模块;

b、静音段和语音段检测模块利用基于门限判决的静音检测算法从上述读入的语音流中找出静音段和语音段;

c、音频特征提取模块,将上述语音段按顺序拼接成一个长语音段,并从长语音段中提取音频特征;

d、说话人改变点检测模块,利用上述提取出来的音频特征,根据贝叶斯信息准则,判断长语音段中相邻数据窗之间的相似度来检测说话人改变点;

e、语音段分割模块,根据上述说话人改变点,把语音流分割成多个语音段,且每个语音段只包含一个说话人。

10. 根据权利要求 9 所述的语音提取方法,其特征在于,步骤 b 中,所述的基于门限判决的静音检测算法包含以下顺序的步骤:

(1)对读入的语音流进行分帧,并计算每帧语音的能量,得到语音流的能量特征矢量;

(2)计算能量门限;

(3)将每帧语音的能量与能量门限比较,低于能量门限的帧为静音帧,否则为语音帧,将相邻的静音帧按顺序拼接成一个静音段,将相邻的语音帧按顺序拼接成一个语音段;

步骤 c 中,所述的音频特征包括梅尔频率倒谱系数(Mel Frequency Cepstral Coefficients,MFCCs)及其一阶差分(Delta-MFCCs)。

一种具有语音提取功能的会议记录器及语音提取方法

技术领域

[0001] 本发明涉及音频处理领域，特别涉及一种具有语音提取功能的会议记录器及语音提取方法。

背景技术

[0002] 目前市场上的会议记录器只是具有简单的录音、回放、转存等功能，没有说话人语音内容分析与理解的功能。使用者在做作会议记录时，如果需要针对某一个特定的说话人讲话进行汇总与处理，必须听完整个录音，人工进行识别是否为同一说话人。为了节省时间，快进播放又会存在漏掉有用信息的风险。通过手工对语音数据进行标注和提取，对使用者来说，是极为不方便的。

[0003] 因此，人们希望会议记录器除了能录音、放音等功能外，还能对会议记录内容进行内容分析与理解，特别希望会议记录器能根据会议语音资料自动地从所有与会人员中提取出每个说话人的语音。

发明内容

[0004] 本发明的目的在于克服现有技术的缺点与不足，提供一种具有语音提取功能的会议记录器，其不仅具有录音、放音、转存功能，而且还可以自动提取各个说话人的语音。

[0005] 本发明的另一目的在于提供一种语音提取方法，其能分析说话人的个数以及对各个说话人的语音进行分类。

[0006] 本发明的目的通过以下的技术方案实现：一种具有语音提取功能的会议记录器，包括主控模块、录音与放音模块、可移动存储模块、交互与显示模块，还包括说话人语音处理模块，说话人语音处理模块包含说话人分割模块和说话人聚类模块，其中

[0007] 说话人分割模块：主控模块将会议音流传输至说话人分割模块，说话人分割模块检测上述语音流中说话人改变点，根据这些改变点将语音流分成多个语音段；

[0008] 说话人聚类模块，利用谱聚类算法对说话人分割模块分割出来的语音段进行说话人聚类，把相同说话人的语音段按顺序拼接在一起，得到说话人个数以及各个说话人的语音。

[0009] 所述的说话人分割模块，包含静音段和语音段检测模块、音频特征提取模块、说话人改变点检测模块和语音段分割模块，其中

[0010] 静音段和语音段检测模块，利用基于门限判决的静音检测算法从上述读入的语音流中找出静音段和语音段；

[0011] 音频特征提取模块，将上述语音段按顺序拼接成一个长语音段，并从长语音段中提取音频特征；

[0012] 说话人改变点检测模块，利用上述提取出来的音频特征，根据贝叶斯信息准则，判断长语音段中相邻数据窗之间的相似度来检测说话人改变点；

[0013] 语音段分割模块，根据上述说话人改变点，把语音流分割成多个语音段，且每个语

音段只包含一个说话人。

[0014] 静音段和语音段检测模块中，所述的基于门限判决的静音检测算法包含以下顺序的步骤：

[0015] (1)对读入的语音流进行分帧，并计算每帧语音的能量，得到语音流的能量特征矢量；

[0016] (2)计算能量门限；

[0017] (3)将每帧语音的能量与能量门限比较，低于能量门限的帧为静音帧，否则为语音帧，将相邻的静音帧按顺序拼接成一个静音段，将相邻的语音帧按顺序拼接成一个语音段。

[0018] 音频特征提取模块中，所述的音频特征包括梅尔频率倒谱系数(Mel Frequency Cepstral Coefficients,MFCCs)及其一阶差分(Delta-MFCCs)。梅尔频率倒谱系数及其一阶差分是业内公知的特征。

[0019] 所述录音与放音模块，包括麦克风、扬声器和音频处理芯片。

[0020] 所述交互与显示模块，包括一个触摸屏及其控制电路，提供具有控制功能的用户交互界面，利用触摸屏与用户交互。

[0021] 所述可移动存储模块，采用 SD 卡实现对数据的存储。

[0022] 本发明的另一目的通过以下的技术方案来实现：一种语音提取方法，包含以下顺序的步骤：

[0023] (1)读入语音流：读入记录有多说话人语音的语音流；

[0024] (2)通过说话人语音处理模块对读入的语音流进行处理，其中说话人语音处理模块包括说话人分割模块和说话人聚类模块；

[0025] (3)通过说话人分割模块检测上述语音流中说话人改变点，根据这些改变点将语音流分成多个语音段；

[0026] (4)说话人聚类模块利用谱聚类算法对说话人分割模块分割出来的语音段进行说话人聚类，把相同说话人的语音段按顺序拼接在一起，得到说话人个数以及各个说话人的语音。

[0027] 所述的步骤(3)具体包含以下步骤：

[0028] a、说话人分割模块包含静音段和语音段检测模块、音频特征提取模块、说话人改变点检测模块和语音段分割模块；

[0029] b、静音段和语音段检测模块利用基于门限判决的静音检测算法从上述读入的语音流中找出静音段和语音段；

[0030] c、音频特征提取模块，将上述语音段按顺序拼接成一个长语音段，并从长语音段中提取音频特征；

[0031] d、说话人改变点检测模块，利用上述提取出来的音频特征，根据贝叶斯信息准则，判断长语音段中相邻数据窗之间的相似度来检测说话人改变点；

[0032] e、语音段分割模块，根据上述说话人改变点，把语音流分割成多个语音段，且每个语音段只包含一个说话人。

[0033] 步骤 b 中，所述的基于门限判决的静音检测算法包含以下顺序的步骤：

[0034] (1)对读入的语音流进行分帧，并计算每帧语音的能量，得到语音流的能量特征矢量；

[0035] （2）计算能量门限；

[0036] （3）将每帧语音的能量与能量门限比较，低于能量门限的帧为静音帧，否则为语音帧，将相邻的静音帧按顺序拼接成一个静音段，将相邻的语音帧按顺序拼接成一个语音段；

[0037] 步骤 c 中，所述的音频特征包括梅尔频率倒谱系数及其一阶差分。

[0038] 本发明与现有技术相比，具有如下优点和有益效果：

[0039] A、使用方便、节省时间：本发明所述的会议记录器通过录音与放音模块采集语音数据之后，可以对语音资料自动进行处理，将各个说话人区别开来，并将各个说话人的语音进行归类、存储，使用者可以根据自己的需要直接选择特定说话人及特定说话人的语音。

[0040] B、功能全面：本发明的会议记录器同时具有一般会议记录器的功能，如录音、放音、转存，另外其可移动存储模块可以将别处获得的语音数据拷贝到本会议记录器进行分析处理。

附图说明

[0041] 图 1 为本发明所述的一种具有说话人语音提取功能的会议记录器的结构框图；

[0042] 图 2 为图 1 所述会议记录器的工作流程图；

[0043] 图 3 为本发明所述的语音提取方法的流程图。

具体实施方式

[0044] 下面结合实施例及附图对本发明作进一步详细的描述，如图 1、2，一种具有说话人语音提取功能的会议记录器，如图 1，包括主控模块、录音与放音模块、可移动存储模块、交互与显示模块，还包括说话人语音处理模块，说话人语音处理模块包含说话人分割模块和说话人聚类模块，其中

[0045] 录音与放音模块，包括麦克风、扬声器和音频处理芯片；

[0046] 交互与显示模块，包括一个触摸屏及其控制电路，提供具有控制功能的用户交互界面，利用触摸屏与用户交互；

[0047] 可移动存储模块，采用 SD 卡实现对数据的存储；

[0048] 录音与放音模块，负责语音资料的录入与播放；

[0049] 主控模块，发出指令，控制各个模块之间的协调工作，主控模块采用基于三星 S5PV210 处理器的微电脑处理平台，搭载嵌入式 Linux 系统；

[0050] 说话人分割模块，主控模块将读入记录有多个说话人语音的语音流传输至说话人分割模块，说话人分割模块检测上述语音流中说话人改变点，根据这些改变点将语音流分成多个语音段，说话人分割模块具体包含静音段和语音段检测模块、音频特征提取模块、说话人改变点检测模块和语音段分割模块，其中

[0051] 静音段和语音段检测模块，利用基于门限判决的静音检测算法从上述读入的语音流中找出静音段和语音段，其中基于门限判决的静音检测算法包含以下顺序的步骤：

[0052] （1）对读入的语音流进行分帧，并计算每帧语音的能量，得到语音流的能量特征矢量；

[0053] （2）计算能量门限；

[0054] （3）将每帧语音的能量与能量门限比较，低于能量门限的帧为静音帧，否则为语音帧，将相邻的静音帧按顺序拼接成一个静音段，将相邻的语音帧按顺序拼接成一个语音段；

[0055] 音频特征提取模块，将上述语音段按顺序拼接成一个长语音段，并从长语音段中提取音频特征，音频特征包括梅尔频率倒谱系数及其一阶差分；

[0056] 说话人改变点检测模块中，所述的利用贝叶斯信息准则确定说话人改变点的方法具体包括以下步骤：

[0057] （1）将经过静音检测得到的各个语音段按顺序拼接成一个长语音段，将长语音段切分成数据窗，窗长为 2 秒，窗移为 0.1 秒。对每个数据窗进行分帧，帧长为 32 毫秒，帧移为 16 毫秒，从每一帧语音信号中提取 MFCCs 与 Delta-MFCCs 特征，MFCCs 与 Delta-MFCCs 的维数 M 都取 12，每个数据窗的特征构成一个特征矩阵 F，特征矩阵 F 的维数 d=2M 为 24；

[0058] （2）计算两个相邻数据窗（x 和 y）之间的 BIC 距离，BIC 距离计算公式如下：

[0059] $$\Delta BIC = (n_x + n_y)\ln\left(\left|\det\left(\mathrm{cov}(F_z)\right)\right|\right) - n_x \ln\left(\left|\det\left(\mathrm{cov}(F_x)\right)\right|\right) -$$

[0060] $$n_y \ln\left(\left|\det\left(\mathrm{cov}(F_y)\right)\right|\right) - \alpha\left(d + \frac{d(d+1)}{2}\right)\ln(n_x + n_y)$$

[0061] 其中，z 是将数据窗 x 和 y 合并之后得到的数据窗，n_x 和 n_y 分别是数据窗 x 和 y 的帧数，F_x、F_y 和 F_z 分别是数据窗 x、y 和 z 的特征矩阵，$\mathrm{cov}(F_x)$、$\mathrm{cov}(F_y)$ 和 $\mathrm{cov}(F_z)$ 分别是特征矩阵 F_x、F_y 和 F_z 的协方差矩阵，det（•）表示求矩阵的行列式值，α 是惩罚系数且实验取值为 2.0；

[0062] （3）如果 BIC 距离 ΔBIC 大于零，则这两个数据窗被视为属于两个不同的说话人（即它们之间存在说话人改变点），否则这两个数据窗被视为属于同一个说话人并将它们合并；

[0063] （4）不断地滑动数据窗判断两个相邻数据窗之间的 BIC 距离是否大于零，并保存说话人改变点，直到长语音段的所有相邻数据窗之间的 BIC 距离都被判断完为止；

[0064] 语音段分割模块，根据上述说话人改变点，把语音流分割成多个语音段，且每个语音段只包含一个说话人；

[0065] 说话人聚类模块中，所述的谱聚类方法具体包括以下步骤：

[0066] （1）从每帧语音中提取梅尔频率倒谱系数及其一阶差分的音频特征，MFCCs 和 Delta-MFCCs 的维数 M，每个语音段的特征构成一个特征矩阵 F_j，特征矩阵 F_j 的维数 d=2M；

[0067] （2）根据各个特征矩阵 F_j 得到所有待聚类语音段的特征矩阵集合 $F=\{F_1,\ldots,F_J\}$，J 为语音段总个数，再根据 F 构造亲和矩阵 $A \in R^{J\times J}$，A 的第（i，j）个元素 A_{ij} 定义如下：

[0068] $$A_{ij} = \begin{cases} \exp\left(\dfrac{-d^2(F_i,F_j)}{2\sigma_i\sigma_j}\right) & i \neq j,\ 1 \le i,j \le J \\ 0 & i = j,\ 1 \le i,j \le J \end{cases}$$

[0069] 其中，$d(F_i,F_j)$ 是特征矩阵 F_i 与 F_j 之间的欧式距离，σ_i 或 σ_j 表示尺度参数，定

义为第 i 或 j 个特征矩阵 F_i 或 F_j 与其它 J-1 个特征矩阵之间的欧式距离矢量的方差，所述 T 表示将多人会话语音分成的总帧数，i、j 表示语音段的编号；

[0070] (3)构造对角矩阵 D，它的第 (i,i) 个元素等于亲和矩阵 A 的第 i 行所有元素之和，再根据矩阵 D 和 A 构造归一化的亲和矩阵 $L=D^{-1/2}AD^{-1/2}$；

[0071] (4)计算亲和矩阵 L 的前 K_{max} 个最大的特征值 $\lambda_1,\ldots,\lambda_{K_{max}}$ 及其特征值矢量 $v_1,\ldots,v_{K_{max}}$，其中 v_k 为列向量且 $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{K_{max}}$，根据相邻特征值之间的差值估计说话人个数 K：

[0072] $$K = \arg\max_{i\in[1,\ K_{max}-1]}\left(\lambda_i - \lambda_{i+1}\right)$$

[0073] 根据估计出来的说话人个数 K，构造矩阵 $V=[v_1, v_2, \ldots, v_K] \in R^{J\times K}$，式中：$1 \leqslant k \leqslant K_{max}$；

[0074] (5)归一化矩阵 V 的每一行，得到矩阵 $Y \in R^{J\times K}$，Y 的第 (j,k) 个元素 Y_{jk}：

[0075] $$Y_{jk} = \frac{V_{jk}}{\sqrt{\left(\sum_{k=1}^{K} V_{jk}^2\right)}} \qquad 1 \leq j \leq J;$$

[0076] (6)将矩阵 Y 中的每一行当作空间 R^K 中的一个点，利用 K 均值算法聚类成 K 类；

[0077] (7)当矩阵 Y 的第 j 行被聚类在第 k 类中，则特征矩阵 F_j 所对应的语音段判为第 k 类即第 k 个说话人；

[0078] (8)根据上述聚类结果，得到说话人个数、各个说话人的语音时长及各个说话人的语音段数。

[0079] 如图 2，一种具有说话人语音提取功能的会议记录器的工作流程如下所示：

[0080] 1)会议记录器开机，进行系统初始化；

[0081] 2)通过交互与显示模块，会议记录器显示交互界面；

[0082] 3)使用者通过交互界面选择是否进行录音动作：

[0083] 若录音，则主控模块控制录音与放音模块开始录音，并将录音资料存储在可移动存储模块中，结束后返回交互界面；

[0084] 若不录音，则使用者通过交互界面选择已录文件，然后主控模块控制说话人语音处理模块即说话人分割模块和说话人聚类模块，对说话人的语音进行分割、聚类处理，提取出各个说话人的语音；

[0085] 4)然后交互界面提示使用者选择是否播放原始语音：

[0086] 若是，则播放原始语音；

[0087] 若否，则进一步提示是否某说话人语音：若是，则选择此人并播放其语音；若否，则返回到交互界面。

[0088] 一种语音提取方法，如图 3，包含以下顺序的步骤：

[0089] (1)读入语音流：读入记录有多说话人语音的语音流；

[0090] (2)通过说话人语音处理模块对读入的语音流进行处理，其中说话人语音处理模

块包括说话人分割模块和说话人聚类模块；

[0091] （3）通过说话人分割模块检测上述语音流中说话人改变点，根据这些改变点将语音流分成多个语音段，具体包含以下步骤：

[0092] a、说话人分割模块包含静音段和语音段检测模块、音频特征提取模块、说话人改变点检测模块和语音段分割模块；

[0093] b、静音段和语音段检测模块利用基于门限判决的静音检测算法从上述读入的语音流中找出静音段和语音段，其中基于门限判决的静音检测算法包含以下顺序的步骤：

[0094] （1）对读入的语音流进行分帧，并计算每帧语音的能量，得到语音流的能量特征矢量；

[0095] （2）计算能量门限；

[0096] （3）将每帧语音的能量与能量门限比较，低于能量门限的帧为静音帧，否则为语音帧，将相邻的静音帧按顺序拼接成一个静音段，将相邻的语音帧按顺序拼接成一个语音段；

[0097] c、音频特征提取模块，将上述语音段按顺序拼接成一个长语音段，并从长语音段中提取音频特征，音频特征包括梅尔频率倒谱系数及其一阶差分；

[0098] d、说话人改变点检测模块，利用上述提取出来的音频特征，根据贝叶斯信息准则，判断长语音段中相邻数据窗之间的相似度来检测说话人改变点；

[0099] e、语音段分割模块，根据上述说话人改变点，把语音流分割成多个语音段，且每个语音段只包含一个说话人；

[0100] （4）说话人聚类模块利用谱聚类算法对说话人分割模块分割出来的语音段进行说话人聚类，把相同说话人的语音段按顺序拼接在一起，得到说话人个数以及各个说话人的语音。

[0101] 上述实施例为本发明较佳的实施方式，但本发明的实施方式并不受上述实施例的限制，其他的任何未背离本发明的精神实质与原理下所作的改变、修饰、替代、组合、简化，均应为等效的置换方式，都包含在本发明的保护范围之内。

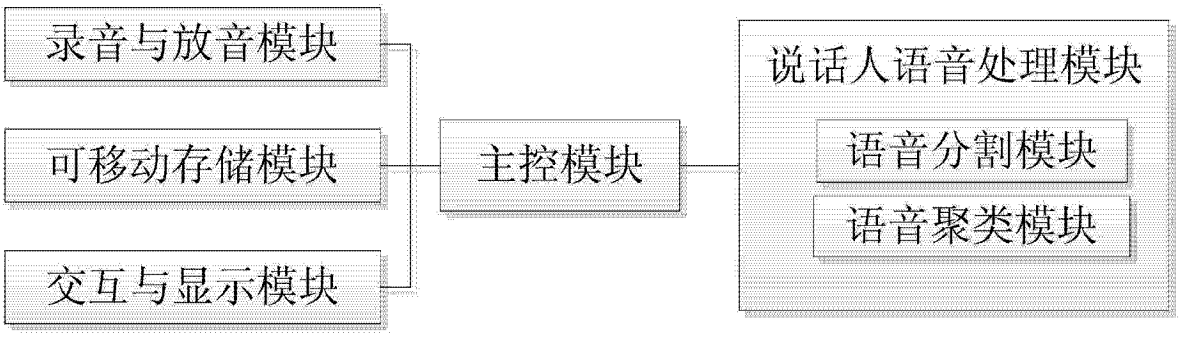


图 1

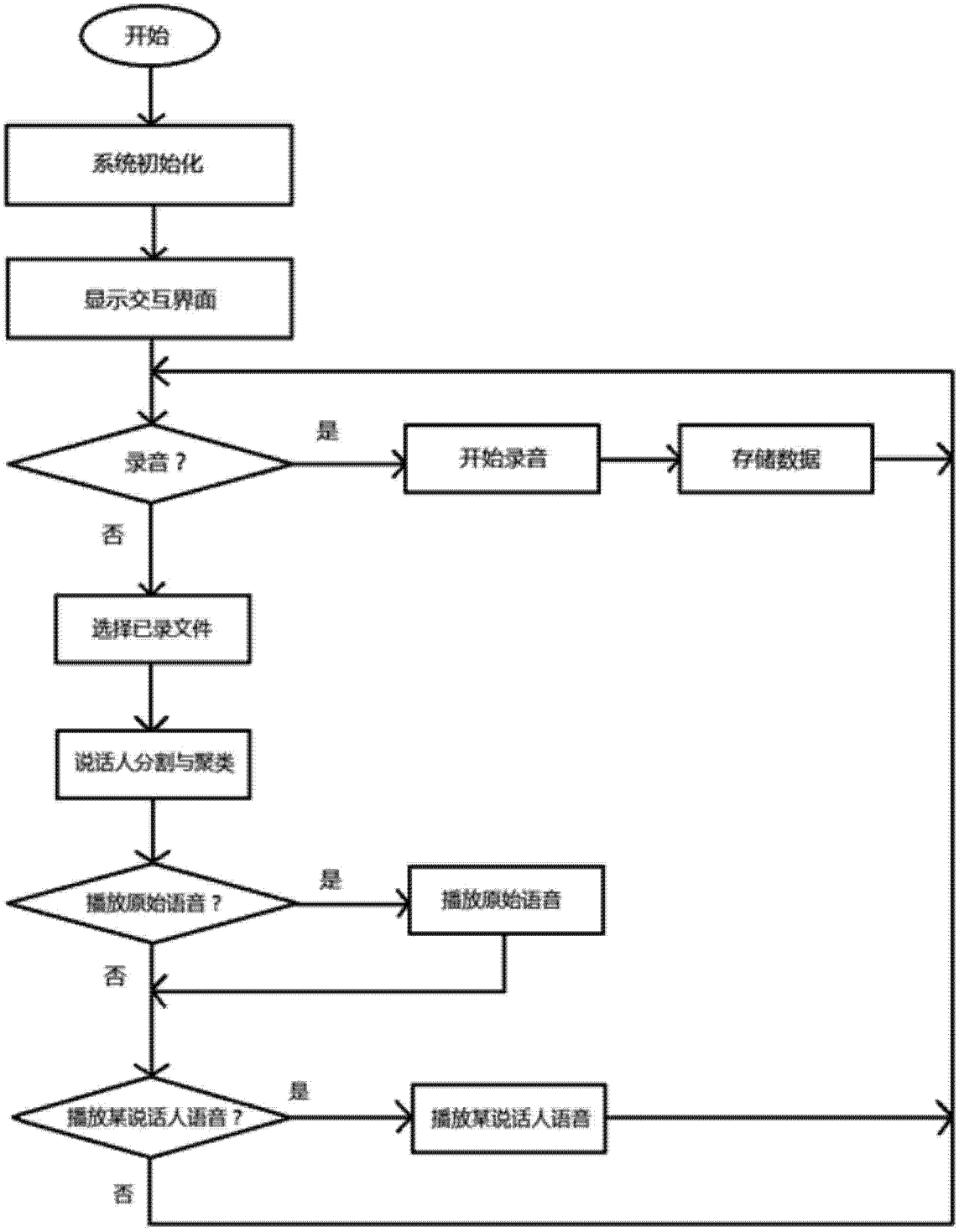

开始
系统初始化
显示交互界面
录音？
是
开始录音
存储数据
否
选择已录文件
说话人分割与聚类
播放原始语音？
是
播放原始语音
否
播放某说话人语音？
是
播放某说话人语音
否

图 2

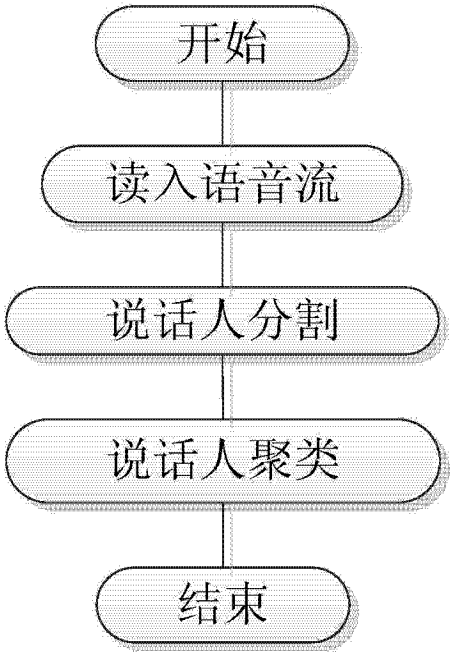


图 3