(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 102982809 A
(43)申请公布日 2013.03.20

(21)申请号 201210528629.4

(22)申请日 2012.12.11

(71)申请人 中国科学技术大学
地址 230026 安徽省合肥市包河区金寨路96号

(72)发明人 陈凌辉 戴礼荣 凌震华

(74)专利代理机构 中科专利商标代理有限责任公司 11021
代理人 宋焰琴

(51)Int.Cl.
G10L 25/30(2013.01)

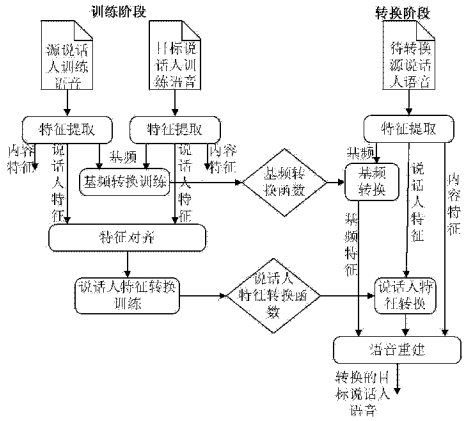
权利要求书 2 页 说明书 9 页 附图 5 页

(54)发明名称

一种说话人声音转换方法

(57)摘要

本发明公开了一种说话人声音转换方法，包括训练阶段和转换阶段，训练阶段包括：从源说话人和目标说话人的训练语音信号中分别提取基频特征、说话人特征和内容特征；根据所述基频特征构建基频转换函数；根据所述说话人特征构建说话人转换函数。转换阶段包括：从源说话人的待转换语音信号中提取基频特征和频谱特征；使用训练阶段得到的基频转换函数和说话人转换函数对从所述待转换语音信号中提取出的基频特征和说话人特征进行转换，得到转换后的基频特征和说话人特征；根据所得到的转换后的基频特征、说话人特征和待转换语音信号中的内容特征合成目标说话人的语音。本发明易于实现且转换后的音质和相似度较高。



CN 102982809 A

1. 一种说话人声音转换方法，用于把源说话人所说的话的语音信号进行转换，使转换后的语音听起来是不同于源说话人的目标说话人所说的，其特征在于，该方法包括训练阶段和转换阶段，其中，

所述训练阶段包括：

步骤 A1、从源说话人和目标说话人的训练语音信号中分别提取基频特征和频谱特征，所述频谱特征包括说话人特征和内容特征；

步骤 A2、根据源说话人和目标说话人的训练语音信号的基频特征，构建从源说话人的语音到目标说话人的语音的基频转换函数；

步骤 A3、根据步骤 A1 提取的源说话人和目标说话人的说话人特征构建说话人转换函数；

所述转换阶段包括：

步骤 B1、从源说话人的待转换语音信号中提取基频特征和频谱特征，所述频谱特征包括说话人特征和内容特征；

步骤 B2、分别使用训练阶段得到的基频转换函数和说话人转换函数，对从步骤 B1 中从所述待转换语音信号中提取出的基频特征和说话人特征进行转换，得到转换后的基频特征和说话人特征；

步骤 B3、根据步骤 B2 得到的转换后的基频特征和说话人特征，以及步骤 B1 提取的待转换语音信号中的内容特征，合成目标说话人的语音。

2. 如权利要求 1 所述的说话人声音转换方法，其特征在于，所述步骤 A2 统计源说话人和目标说话人的训练语音信号的基频特征在对数域分布的均值和方差，根据所统计的均值和方差构建从源说话人的语音到目标说话人的语音的基频转换函数。

3. 如权利要求 2 所述的说话人声音转换方法，其特征在于，所述基频转换函数为线性变换函数。

4. 如权利要求 1 所述的说话人声音转换方法，其特征在于，所述步骤 A1 和步骤 B1 的提取语音信号的基频特征和频谱特征的方法包括：

步骤 a1、基于语音信号的源 - 滤波器结构，将语音信号以 20 ～ 30ms 进行分段，每一段作为一帧，并对每一帧的语音信号提取基频和频谱参数；

步骤 a2、使用一个神经网络来分离所述频谱参数中的说话人特征和内容特征，该神经网络结构采用上下对称的共 2K-1 层多层（K 为自然数）网络结构，包括：最下层为输入层，从该层输入待分离的声学特征；最上层为输出层，该层输出重构出的声学特征；中间 2K-3 个隐层，每层若干个节点，模拟神经单元的处理过程。从输入层到从下至上的第 K 个隐层为编码网络，用于从输入的语音声学特征中提取出高层的信息；从下至上的第 K 个隐层为编码层；编码层的网络节点分为两部分，一部分与说话人相关，另一部分与内容相关，它们的输出分别对应说话人特征和内容特征；从下至上的第 K 个隐层以上的隐层为解码网络，用于从高层的说话人特征和内容特征中重建出声学频谱参数。

5. 如权利要求 4 所述的说话人声音转换方法，其特征在于，所述步骤 a2 包括在一语音信号数据库上对所述神经网络进行训练，以使其具备从声学特征中提取和分离说话人特征和内容特征的能力。

6. 如权利要求 5 所述的说话人声音转换方法，其特征在于，所述对所述神经网络进行

训练的步骤包括：

步骤 b1、通过预训练来初始化所述神经网络的网络权值；

步骤 b2、对所述神经网络的编码层的每个节点的输出特征，采用一个区分性准则来统计其在不同说话人之间和不同内容之间的区分性，将不同说话人间区分性大而不同内容之间区分性小的节点作为说话人相关节点，其余的节点作为内容相关节点；

步骤 b3、设计特定的区分性目标函数来精细调整该神经网络的权值，使该神经网络具备从声学特征中分离说话人信息和内容信息的能力。

7. 如权利要求 5 所述的说话人声音转换方法，其特征在于，所述步骤 b1 采取无监督的学习模式，使用贪婪算法来逐层训练该神经网络；

8. 如权利要求 7 所述的说话人声音转换方法，其特征在于，所述步骤 b1 包括：

在输入层，输入特征服从高斯分布，则在输入的各维上加入适量的高斯噪声，并采用最小均方误差准则来训练；在第一层以上各层，输入特征服从二值分布，因此以一定的概率，将输入特征的某些维置零，并使用最小交叉熵准则来训练；经过预训练得到一个 K 层叠加的自动编码器后，将其向上翻转，便得到了上下对称的自动编码器结构。

9. 如权利要求 6 所述的说话人声音转换方法，其特征在于，所述步骤 b2 采用 Fisher′ s ratio 准则作为区分性准则。

10. 如权利要求 9 所述的说话人声音转换方法，其特征在于，所述步骤 b3 包括：

设计具有对比竞争机制的区分性目标函数，使用误差后向传播算法来精细调整所述神经网络的网络权值，使该神经网络具备从声学特征中分离说话人信息和内容信息的能力。

11. 如权利要求 5 所述的说话人声音转换方法，其特征在于，其中所述的语音信号数据库是通过下列步骤制作的：

步骤 c1、建立一个语料库，使该语料库中包括多个句子；

步骤 c2、录制多个说话人朗读所述语料库中的句子的语音信号，构建语音信号数据库，并对该语音信号数据库中的语音信号进行预处理，以去除语音信号中的不正常部分；

步骤 c3、使用隐马尔科夫模型来对进行预处理的杨这语音信号数据库中的语音信号行切分，切分后的每一段作为一个帧，由得到各语音信号的帧一级的说话人标注信息和内容标注信息；

步骤 c4、对所述语音数据库的各语音信号进行随机组合，构造神经网络的训练数据。

一种说话人声音转换方法

技术领域

[0001] 本发明属于信号处理技术领域，具体涉及在不改变语音信号中内容信息的前提下，将一个说话人的语音信号通过转换处理，改变为能够被感知为另一个说话人的语音信号，特别是一种将语音信号中的说话人信息和内容信息进行分离的说话人声音转换方法。

背景技术

[0002] 在如今的信息时代，人机交互一直是计算机领域的研究热点，高效智能的人机交互环境已经成为了当前信息技术的应用和发展的迫切需求。众所周知，语音是人类交流的最重要、最便捷的途径之一。语音交互将是人际交互中最为“友好”的。基于语音识别、语音合成及自然语言理解的人机语音对话技术是世界公认的一个难度很大，极富挑战性的高技术领域，但是其应用前景十分光明。

[0003] 作为人机交互的核心技术之一，语音合成近年来在技术和应用方面都取得了长足进展。目前，基于大语料库的合成系统合成的语音在音质和自然度方面都取得了不错的效果，因此大家对语音合成系统提出了更多的需求——多样化的语音合成，包括多个发音人、多种发音风格、多种情感以及多语种等。而现有的语音合成系统大多是单一化的，一个合成系统一般只包括一到两个说话人，采用朗读或者新闻播报风格，而且针对某个特定的语种。这种单一化的合成语音大大限制了语音合成系统的在实际中的应用，包括教育、娱乐和玩具等。为此，多样化语音合成方面的研究逐渐成为近期语音合成研究领域的主流方向之一。

[0004] 实现一个多说话人、多种发音风格、多种情感的语音合成系统，最直接的方法就是录制多个人、多种风格的音库，并分别构建各个发音人、各个风格的个性化语音合成系统。由于针对每个发音人、每种风格、每种情感制作一个特定的语音库的工作量过大，因此这种方法在实际中并不可行。在这一背景下，说话人声音转换技术被提出。说话人声音转换技术就是试图把一个人（源说话人）说的话（的语音）进行转换（对基频、时长、谱参数等包含说话人特征信息的参数进行调整），使它听起来好像另一个人（目标说话人）说出来的一样。与此同时，保持源说话人表达的意思不变。说话人声音转换技术通过录制少量的说话人的语音信号进行训练，调整源说话人的语音得到目标说话人的合成语音，从而快速实现个性化语音合成系统。

[0005] 实现一个说话人声音转换系统，最主要的挑战在于转换语音的相似度和音质。作为当前的一种主流的说话人声音转换方法——基于联合空间高斯混合模型的说话人声音转换方法，由于使用了统计建模的框架，相对来说具有很好的鲁棒性和推广性，但是该方法只是一个典型的机器学习中的特征映射的方法，并没有利用语音信号特有的一些特性（说话人信息和内容信息共存），而且统计建模带来了诸多问题，如对数据量的依赖，建模精度不够，统计模型对声学参数原有的信息的破坏，均导致转换语音的效果急剧下降。而另一种主流的语音合成技术，基于共振峰的频谱弯折方法，则利用到了语音信号中的说话人共振峰结构这一主要反映说话人信息的特征，在转换时尽可能的保留语音信号中的细节成分，保证了转换语音的音质，但是由于共振峰的提取和建模很难，就使得这一类方法需要很多

人工的干预，而且鲁棒性较差。

[0006] 总的来说，传统的说话人语音转换方法，由于其对语音信号中特定说话人的声音信息缺乏有效表达及有效建模，对建模数据要求高，所构建的转换方法往往包含了对语音信号内容的转换，因此转换后的语音音质和相似度目前不能达到令人满意的程度。

发明内容

[0007] （一）要解决的技术问题

[0008] 本发明所要解决的技术问题是现有的说话人语音转换方法的语音音质较差和相似度不高的问题。

[0009] （二）技术方案

[0010] 本发明提出一种说话人声音转换方法，用于把源说话人所说的话的语音信号进行转换，使转换后的语音听起来是不同于源说话人的目标说话人所说的，其特征在于，该方法包括训练阶段和转换阶段，其中，

[0011] 所述训练阶段包括：

[0012] 步骤 A1、从源说话人和目标说话人的训练语音信号中分别提取基频特征和频谱特征，所述频谱特征包括说话人特征和内容特征；

[0013] 步骤 A2、根据源说话人和目标说话人的训练语音信号的基频特征，构建从源说话人的语音到目标说话人的语音的基频转换函数；

[0014] 步骤 A3、根据步骤 A1 提取的源说话人和目标说话人的说话人特征构建说话人转换函数；

[0015] 所述转换阶段包括：

[0016] 步骤 B1、从源说话人的待转换语音信号中提取基频特征和频谱特征，所述频谱特征包括说话人特征和内容特征；

[0017] 步骤 B2、分别使用训练阶段得到的基频转换函数和说话人转换函数，对从步骤 B1 中从所述待转换语音信号中提取出的基频特征和说话人特征进行转换，得到转换后的基频特征和说话人特征；

[0018] 步骤B3、根据步骤B2得到的转换后的基频特征和说话人特征，以及步骤B1提取的待转换语音信号中的内容特征，合成目标说话人的语音。

[0019] 根据本发明的一种具体实施方式，所述步骤 A1 和步骤 B1 的提取语音信号的基频特征和频谱特征的方法包括：

[0020] 步骤 a1、基于语音信号的源 - 滤波器结构，将语音信号以 20 ～ 30ms 进行分段，每一段作为一帧，并对每一帧的语音信号提取基频和频谱参数；

[0021] 步骤 a2、使用一个神经网络来分离所述频谱参数中的说话人特征和内容特征，该神经网络结构采用上下对称的共 2K-1 层多层（K 为自然数）网络结构，包括：最下层为输入层，从该层输入待分离的声学特征；最上层为输出层，该层输出重构出的声学特征；中间 2K-3 个隐层，每层若干个节点，模拟神经单元的处理过程。从输入层到从下至上的第 K 个隐层为编码网络，用于从输入的语音声学特征中提取出高层的信息；从下至上的第 K 个隐层为编码层；编码层的网络节点分为两部分，一部分与说话人相关，另一部分与内容相关，它们的输出分别对应说话人特征和内容特征；从下至上的第 K 个隐层以上的隐层为解码网

络，用于从高层的说话人特征和内容特征中重建出声学频谱参数。

[0022] 根据本发明的一种具体实施方式，步骤 a2 包括在一语音信号数据库上对所述神经网络进行训练，以使其具备从声学特征中提取和分离说话人特征和内容特征的能力，所述对所述神经网络进行训练的步骤包括：

[0023] 步骤 b1、通过预训练来初始化所述神经网络的网络权值；

[0024] 步骤 b2、对所述神经网络的编码层的每个节点的输出特征，采用一个区分性准则来统计其在不同说话人之间和不同内容之间的区分性，将不同说话人间区分性大而不同内容之间区分性小的节点作为说话人相关节点，其余的节点作为内容相关节点；

[0025] 步骤 b3、设计特定的区分性目标函数来精细调整该神经网络的权值，使该神经网络具备从声学特征中分离说话人信息和内容信息的能力。

[0026] 根据本发明的一种具体实施方式，所述的语音信号数据库是通过下列步骤制作的：

[0027] 步骤 c1、建立一个语料库，使该语料库中包括多个句子；

[0028] 步骤 c2、录制多个说话人朗读所述语料库中的句子的语音信号，构建语音信号数据库，并对该语音信号数据库中的语音信号进行预处理，以去除语音信号中的不正常部分；

[0029] 步骤 c3、使用隐马尔科夫模型来对进行预处理的杨这语音信号数据库中的语音信号行切分，切分后的每一段作为一个帧，由得到各语音信号的帧一级的说话人标注信息和内容标注信息；

[0030] 步骤 c4、对所述语音数据库的各语音信号进行随机组合，构造神经网络的训练数据。

[0031] （三）有益效果

[0032] 本发明的说话人声音转换方法具有以下优点：

[0033] 1、本发明首次提出了使用深层神经网络来实现语音信号中说话人信息和内容信息的分离，以满足不同语音信号处理任务的需求，如语音识别、说话人识别与转换。

[0034] 2、本发明在进行说话人声音转换时，仅考虑说话人的因素，排除了内容因素的干扰，使得说话人声音转换更易于实现，转换后的音质和相似度得以大幅度提高。

[0035] 3、本发明采用的分离器只需要训练一次，训练好后能够对任意说话人语音提取说话人特征和内容特征，一次训练多次使用，无需重复训练模型。

附图说明

[0036] 图 1 是本发明的的说话人声音转换方法的流程图；

[0037] 图 2 是本发明的特征提取步骤的框图；

[0038] 图 3 是本发明的用于特征分离的神经网络结构示意图；

[0039] 图 4 是本发明的神经网络训练流程图；

[0040] 图 5 是本发明中数据库制作的流程图；

[0041] 图 6 是本发明中倒谱特征在不同说话人和不同内容之间的区分性的示意图；

[0042] 图 7 是本发明中提取出的说话人特征和内容特征在不同说话人和不同内容之间的区分性的示意图。

具体实施方式

[0043] 为使本发明的目的、技术方案和优点更加清楚明白，以下结合具体实施例，并参照附图，对本发明作进一步的详细说明。

[0044] 从生理学的角度来讲，已有学者的工作证实，人脑在感知语音信号时，对说话人信息的感知和对说话内容的感知分别是在大脑皮层的不同区域完成的。这说明人脑在高层对说话人和内容信息做了分解，语音信号中的信息是可分离的，说话人信息和内容信息的分离对语音信号处理的意义很重大，分离出来的信息可分别用于说话人识别，语音识别以及其他的一些针对性的应用。

[0045] 本发明从说话人声音转换的本质出发，即保持说话人所说的话的内容不变，而仅改变说该句话的说话人的信息。基于这一考虑，对语音信号中的信息进行分离，得到说话人特征和内容特征，以便对说话人的成分进行操作。本发明中所说的“说话人特征”指的是反应说话人特性、区别不同说话人的特征，“内容特征”指的是反应语音信号所要表达的意思的的特征。

[0046] 对此，本发明使用一种基于深层神经网络的技术，在高层将语音信号的声学特征分解为说话人特征和内容特征，以而使说话人声音转换得以更完美和简单的实现，达到音质和相似度大幅提升的转换语音信号。

[0047] 图 1 是本发明的的说话人声音转换方法的流程图。如图所示，本发明的方法总体上包含两个阶段 :训练阶段和转换阶段。下面依次介绍 :

[0048] （一）训练阶段

[0049] 训练阶段主要包括三个步骤 :

[0050] 步骤 A1 :特征提取。

[0051] 该步骤从源说话人和目标说话人的训练语音信号中分别提取特征，所述特征包括基频特征和频谱特征，频谱特征在本发明中分为说话人特征和内容特征。

[0052] 步骤 A2 :基频转换函数训练。

[0053] 该步骤根据源说话人和目标说话人的训练语音信号的基频特征，构建从源说话人的语音到目标说话人的语音的基频转换函数。

[0054] 根据一种具体实施方式，该步骤统计源说话人和目标说话人的训练语音信号的基频特征在对数域分布的均值和方差，根据所统计的均值和方差构建从源说话人的语音到目标说话人的语音的基频转换函数。

[0055] 由于每个说话人的基频特征参数在对数域呈高斯分布，因此对于基频转换，本发明中优选为仅使用对数域的简单线性变换进行。

[0056] 步骤 A3 :频谱转换函数训练。

[0057] 该步骤根据从源说话人和目标说话人的训练语音信号中提取的频谱特征中的说话人特征构建说话人转换函数。

[0058] 前述说话人转换的要求保持说话内容不变而只改变说话人信息。因此，本发明只需要训练说话人特征的转换函数（说话人转换函数）即可。

[0059] 由于在录制源说话人和目标说话人的语音信号时，无法做到不同说话人进行同一句话的录音时保持完全相同的时长，因此需要一些规整手段来将不同时长的句子规整

到相同的时长以便进行有监督的特征转换学习（特征对齐），本发明采用动态时间规整(dynamic time warping)算法来进行时长规整，说话人特征转换的建模可以使用线性回归模型或者联合空间高斯混合模型等方法来实现。

[0060] （二）转换阶段

[0061] 转换阶段包括三个步骤：

[0062] 步骤 B1：特征提取。

[0063] 与训练阶段相仿，该步骤从源说话人的待转换语音信号中提取特征，所述特征包括基频特征和频谱特征，频谱特征分为说话人特征和内容特征。

[0064] 步骤 B2：特征转换。

[0065] 分别使用训练阶段得到的基频转换函数和说话人转换函数，对从步骤 B1 中从所述待转换语音信号中提取出的基频特征和说话人特征进行转换，得到转换后的基频特征和说话人特征。

[0066] 对于基频转换，具体的，训练阶段在训练集上统计出源、目标说话人语音信号的基频在对数域的均值 μ_x、μ_y 和方差 σ_x^2、σ_y^2，基频转换时转换函数形如下式所示：

[0067] $$\log f_0^y = \mu_y + \frac{\sigma_y}{\sigma_x}(\log f_0^x - \mu_x)$$

[0068] 而对于说话人特征的转换，假设有源和目标说话人对应时间对齐的说话人特征 $X = \{x_1, x_2, \ldots x_T\}$ 和 $Y = \{y_1, y_2, \ldots y_T\}$ 作为训练数据。本发明采用两种方案。一种方案是使用线性回归模型 $F(x_t) = Ax_t + b$ 作为频谱转换函数，其中的参数可有下式计算得到：

[0069] $[A, b] = YX^T(XX^T)^{-1}$

[0070] 另外一种方案，基于联合空间高斯混合模型的方法，需要使用联合特征 $Z = [X^T, Y^T]^T$ 来训练一个高斯混合模型，他以如下形式来描述联合特征空间的分布：

[0071] $P(z) = \sum_m w_m N(z; \mu_m, \Sigma_m)$，

[0072] 其中

[0073] $$\boldsymbol{\mu}_m = \begin{bmatrix} \boldsymbol{\mu}_m^{(x)} \\ \boldsymbol{\mu}_m^{(y)} \end{bmatrix}, \boldsymbol{\Sigma}_m = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(xx)} & \boldsymbol{\Sigma}_m^{(xy)} \\ \boldsymbol{\Sigma}_m^{(yx)} & \boldsymbol{\Sigma}_m^{(yy)} \end{bmatrix}$$

[0074] 从中，导出转换函数：

[0075] $$F(\boldsymbol{x}_t) = \sum_m h_m(\boldsymbol{x}_t)[\boldsymbol{\mu}_m^{(y)} + \boldsymbol{\Sigma}_m^{(yx)}\boldsymbol{\Sigma}_m^{(xx)-1}(\boldsymbol{x}_t - \boldsymbol{\mu}_m^{(x)})]$$

[0076] 式中 $h_m(\boldsymbol{x}_t) = \dfrac{w_m \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_m^{(x)}, \boldsymbol{\Sigma}_m^{(xx)})}{\sum_j w_j \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_j^{(x)}, , \boldsymbol{\Sigma}_m^{(xx)})}$ 为后验概率。

[0077] 步骤 B3：语音合成。

[0078] 该步骤根据步骤 B2 得到的转换后的基频特征和说话人特征，以及步骤 B1 提取的待转换语音信号中的内容特征，合成目标说话人的语音。

[0079] 本发明使用基于源－滤波器结构的合成器，需要输入激励（即基频）和声道响应（频谱参数）来生成待转换的语音。因此首先需要从转换的说话人特征和待转换的说话人语音信号的内容特征中重建出转换的说话人频谱参数（频谱参数重建过程见下文所述），进而通过合成器来生成转换的语音。本发明采用 STRAIGHT 分析合成器来进行语音生成。

[0080] （三）特征提取

[0081] 以上对本发明的方法进行了整体性的介绍，下面对于所述方法中采用的特征提取步骤进行详细的说明。

[0082] 如前所述，本发明所述特征提取包括基频特征、说话人特征和内容特征的提取。本发明中基频特征提取采用传统的基频提取方法。说话人特征和内容特征的特征提取方法是本发明核心所在。

[0083] 3.1 基本步骤

[0084] 图 2 是本发明的特征提取步骤的框图。如图 2 所示，特征提取步骤具体分为两步骤：

[0085] 步骤 a1：声学特征提取。

[0086] 基于语音信号的源 - 滤波器结构，考虑到语音信号的短时平稳性和长时非平稳性，将语音信号以 20-30ms 进行分段，每一段本发明称作一帧。对每一帧语音信号，使用现有的语音分析算法（如 STRAIGHT 等）从语音信号中提取基频和频谱参数（如线谱对、Mel 倒谱等）。

[0087] 步骤 a2：说话人特征和内容特征提取。

[0088] 考虑到说话人之间的差异主要体现在声道结构上，在声学特征上，即主要反映在频谱参数中。因此，本发明主要考虑从频谱特征分离出说话人相关特征和内容相关特征。另外，本发明考虑到说话人特征是一种超音段长时的特征，为有效提取语音信号中的说话人相关特征，使其与内容相关特征更好地分离，本发明将连续多帧的特征拼接成一个称之为超音段特征输入到特征分离器中。具体的特征分离方法如下：

[0089] 3.2 特征分离算法

[0090] 本发明使用一个深层的神经网络来分离声学频谱参数中的说话人特征和内容特征。图 3 是本发明的用于特征分离的神经网络结构示意图。如图 3 所示，该神经网络结构采用上下对称的共 2K-1 层多层（K 为自然数）网络结构，包括：最下层为输入层，从该层输入待分离的声学特征；最上层为输出层，该层输出重构出的声学特征；中间 2K-3 个隐层，每层包括若干个节点，模拟神经单元的处理过程。

[0091] 从输入层到从下至上的第 K 个隐层为编码网络（或称编码器），用于从输入的语音声学特征中提取出高层的信息，从下至上的第 K 个隐层为编码层；编码层的网络节点分为两部分，一部分与说话人相关，另一部分与内容相关，它们的输出分别对应说话人特征和内容特征。从下至上的第 K 个隐层以上的隐层为解码网络（或称解码器），它的功能与编码网络相反，用于从高层的说话人特征和内容特征中重建出声学频谱参数。

[0092] 本发明采用的图 3 所示的深层神经网络是对人的神经系统处理语音信号的一个模拟，需要对其进行训练，从而使其具有所需要的能够从声学特征中实现提取和分离说话人特征和内容特征这一特定的能力。图 3 所示深层神经网络的训练是在本发明提出的数据库制作方法所设计的语音信号数据库上进行，本发明提出的数据库制作方法见本发明数据库制作部分。

[0093] 图 4 是本发明中神经网络训练的具体流程图。训练过程分为三步骤：

[0094] 步骤 b1：预训练。

[0095] 由于深层神经网络的优化比较困难，在训练之前需要通过预训练来初始化网络权值。本发明采取一种无监督的学习模式，使用贪婪算法来逐层训练网络，快速的得到模

型的初始参数。在每一层的训练中，可以使用消除噪声干扰的自动编码器 (De-noising auto-encoder) 来初始化网络权值，即在输入特征上加上一定的噪声掩盖，使得神经网络的训练能够更加鲁棒，并且防止过训练。具体的，在输入层，输入特征服从高斯分布，则在输入的各维上加入适量的高斯噪声，并采用最小均方误差准则来训练。而在第一层以上各层，输入特征服从二值分布，因此以一定的概率，将输入特征的某些维置零，并使用最小交叉熵 (cross-entropy) 准则来训练。经过预训练得到一个 K 层叠加的自动编码器后，将其向上翻转，便得到了上下对称的自动编码器结构。

[0096] 步骤 b2 ：编码层调整。

[0097] 经过预训练之后的神经网络，已经具备了一定的高层信息提取能力，在编码层，某些节点能反映出较强的说话人区分能力，另外一些节点则能反映较强的内容区分能力。这一步将使用一些客观的准则来将这些节点挑选出来，其输出分别作为对应的特征。这里可以使用一些区分性准则，如 Fisher′ s ratio，来挑选。具体的，在所述语音信号数据库的训练集上，对编码层的每个节点的输出特征，均用该准则来统计其在不同说话人之间和不同内容之间的区分性，将不同说话人间区分性大而不同内容之间区分性小的节点作为说话人相关节点，其余的节点作为内容相关节点。

[0098] 步骤 b3 ：精细调整。

[0099] 本发明需要从输入的声学频谱参数中分离出说话人相关和内容相关的特征，并能将其应用到说话人声音转换中去。对此，要设计特定的区分性目标函数来训练该网络，使其具备本发明所期望的这种能力。要达到这种要求，需要在输入训练样本中引入对比竞争的手段。在如图 3 所示的网络结构中，在输入层，每次同时并行输入两个样本 x_1 和 x_2，他们分别在编码输出层生成说话人特征 c_{s1}、c_{s2} 和内容特征 c_{c1}、c_{c2}，然后通过解码网络，重建出输入的声学特征$\hat{\boldsymbol{x}}_1$和$\hat{\boldsymbol{x}}_2$。因此，训练网络的目标函数中包含如下的三部分 ：

[0100] 重建误差 ：一方面，由于说话人声音转换应用的需要，要从高层特征中重建恢复出声学频谱参数，解码网络需要具有很好的恢复重建的能力，该能力将会直接影响合成语音的质量。因此，在训练目标函数中需要对重建误差加以限制。另一个方面，加入重建误差的限制也是为了保证编码输出的说话人特征和内容特征中信息的完整性。本发明中采用如下形式的误差形式 ：

[0101] $$L_r = \sum_{i\in\{0,1\}} |\boldsymbol{x}_1 - \hat{\boldsymbol{x}}_2|^2$$

[0102] 说话人特征代价 ：为了使说话人特征对说话人具有很强的区分性，而对内容不具有区分性，可以设计这样一种准则，使相同说话人之间的说话人特征误差尽量小，而不同说话人之间的误差尽量大，这种准则可以表示为下式 ：

[0103] $L_{sc} = \delta_s * E_s + (1-\delta_s) * \exp(-\lambda_s E_s)$

[0104] 其中，$E_s = |c_{s1} - c_{s2}|^2$，$\delta_s$ 是输入的两个样本的说话人标注，$\delta_s = 1$ 表示两个输入它们来自同一个说话人，而 $\delta_s = 0$ 则表示来自不同的两个说话人。

[0105] 内容特征代价 ：与说话人特征误差类似，可以构造内容特征的区分性代价函数 ：

[0106] $L_{cc} = \delta_c * E_s + (1-\delta_c) * \exp(-\lambda_c E_c)$

[0107] 综合上述三种代价，可以得到最终用于的精细调整的目标函数 ：

[0108] $L_{cc} = \alpha L_r + \beta L_{sc} + \zeta L_{cc}$

[0109] α、β 和 ζ 调整这三种代价比重的权值，神经网络的训练目标是调整网络权值使得该目标函数尽量小，训练时本发明使用误差反向传播算法，利用带冲量的梯度下降算法来更新网络权值。

[0110] （四）说话人语音信号库的制作

[0111] 本发明中所使用的神经网络需要大量的训练数据来进行，需要包含很多的说话人，每个说话人也需要录制充足内容的语料。

[0112] 所要特别指出的是，神经网络所需要的大量训练数据，并不是图 1 中所示训练过程的源说话人或目标说话人数据。实际应用中，获得图 1 中所示训练过程的源说话人或目标说话人的大量数据不切实际或要求过高，但获得本处所述神经网络所需要的大量训练数据是可行的，符合实际要求。

[0113] 图 5 是本发明中数据库制作的流程图。分为四个步骤：

[0114] 步骤 c1：建立一个语料库，使该语料库中包括多个句子。

[0115] 考虑到要设计一种鲁棒的分离网络，需要其能处理所有的人以及所有的内容，本发明中设计一个音素均衡的语料库，而且句子数不能太多，通常在 100 句以内，以便采集大量的说话人数据。所谓音素均衡是指语料中包含所有的音素，而且各音素的数量相对均衡。

[0116] 步骤 c2：录制多个说话人朗读所述语料库中的句子的语音信号，构建语音信号数据库，并对该语音信号数据库中的语音信号进行预处理，以去除语音信号中的不正常部分。

[0117] 考虑到要使网络具有区分说话人的能力，需要录制大量说话人的数据来训练网络。在录音阶段，由于成本等方面的原因，无法找到如此多的播音员来录制音库，只能采集业余人员的录音，这就使得录制的语音质量参差不齐，因此，录制完成后，需要对录制的语音做一些预处理，如能量规整、信道均衡、喷麦现象的处理等等，保证训练语料的质量。

[0118] 步骤 c3：使用隐马尔科夫模型来对进行预处理的杨这语音信号数据库中的语音信号行切分，切分后的每一段作为一个帧，由得到各语音信号的帧一级的说话人标注信息和内容标注信息。

[0119] 从上文可知，在神经网络训练的精细调整阶段，是有监督的学习过程，需要知道输入每帧训练数据的说话人标注信息和内容标注信息。因此，需要对语音信号数据库中的语音信号做帧一级的标注，即进行音段的切分。具体的，可以采用一个现有的用作语音合成的上下文相关的隐马尔可夫模型来实现音段切分。在切分之前，先用每个说话人的录音数据使用最大似然线性回归算法将该模型自适应到该说话人的声学空间，再使用自适应得到的模型对该说话人的录音数据利用维特比算法进行解码，得到模型各状态的边界信息。

[0120] 步骤 c4：对所述语音数据库的各语音信号进行随机组合，构造神经网络的训练数据。

[0121] 根据上文描述，神经网络的训练数据有四类：相同说话人相同内容、相同说话人不同内容、不同说话人相同内容和不同说话人不同内容。由于有很多的说话人特征和内容特征属性，在训练阶段，本发明在训练数据中随机挑选组合，输入到网络进行训练。

[0122] （五）具体实施例

[0123] 根据上文所述方法，作为本发明实施方式举例，本发明搭建了一个说话人声音转换系统。首先，本发明设计了包含 100 句话的音素平衡的语料，募集了 81 个说话人（其中包含 40 个男性和 41 个女性说话人）来录音，经过处理后形成最终的训练语料库。录音的

语音文件是单声道、16kHz 采样率的。在这 81 个说话人的数据中，我们随机挑选 60 人（30 个男性、30 个女性）的数据作为训练神经网络的训练集，另外 10 人（5 个男性和 5 个女性）的数据作为训练神经网络训练的验证集，余下的 11 人的数据作为测试集，测试说话人声音转换的效果。在提取声学特征时，我们采用 25ms 的汉明窗对波形信号进行分帧处理，并以 5ms 的帧移来移动短时窗，每帧提取一个基频和一组 24 维的 Mel 倒谱参数作为声学特征。

[0124] 在训练用于特征分离的神经网络阶段，网络的输入向量为当前帧与其前后各 5 帧共 11 帧拼成的超音段特征，共 264 维，由于输出只需要重建出输入的当前帧，因此，输出层为 24 维。另外，网络包含 7 个隐层，其中节点数分别为 500、400、300、200、300、400、500，在中间的那一层，我们使前 100 个节点的输出为说话人特征，剩下的 100 个节点的输出为内容特征。在预训练阶段，我们采用 4 个层叠的自动编码器的形式来初始化网络权值，节点数分别为 :264-500、500-400、400-300 和 300-200，自底向上，每一个自动编码器的输出作为下一个自动编码器的输入，通过无监督学习的形式初始化网络权值，最后将网络权值翻转，得到整个网络的初始化权值，需要注意的是，第一层翻转到整个网络的最上面一层的时候，由于输出只有 24 维，只需要将输入层当前帧对应的权值翻转上去即可。另外，在中间层翻转之前，需要计算每个节点输出在不同说话人之间和不同内容之间的区分性（上文中提到的 Fisher's ratio)，并以此来对节点和网络权值进行重排。预训练之后，按照上文所述的方法进行精细调整，在这个过程中，需要在验证集上对目标函数的权值进行调整，得到最优值。

[0125] 训练好特征分离器之后，便可以进行搭建说话人声音转换系统了，我们在测试集上任意挑选两个说话人来，选择其中 50 句话作为训练数据，按上文提取需要的特征，训练基频、说话人特征的转换函数（本实施方式举例中使用直接的线性回归模型），剩下的 50 句话作为测试数据来验证说话人声音转换的效果。

[0126] 我们使用 Fisher's ratio 来度量提取出的不同特征在不同说话人之间和不同内容之间的区分性。Fisher's ratio 度量的是特征类内距离和类间距离的比值，该比值越大，说明特征在此种分类方法下更加具有区分性。图 6 和图 7 分别是 Mel 倒谱系数和分离出的特征在不同说话人（实线）和不同内容（虚线）之间的区分性。可见，输入的声学特征中，除了低维在内容上显示较强的区分性外，其余维并没有很强的区分性。而提取出的特征（前 100 维为说话人特征，剩下 100 维为内容特征）经过训练，对不同的分类体现出所期望的区分性。而在说话人转换实验上，直接用目标说话人的说话人特征加上源说话人的内容特征合成出的语音，倒谱误差为 4.39dB，而用线性变换过的源说话人的说话人特征和其内容特征合成的语音倒谱误差为 5.64dB，从主观听感上已经逼近目标说话人的语音。

[0127] 以上所述的具体实施例，对本发明的目的、技术方案和有益效果进行了进一步详细说明，应理解的是，以上所述仅为本发明的具体实施例而已，并不用于限制本发明，凡在本发明的精神和原则之内，所做的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内。

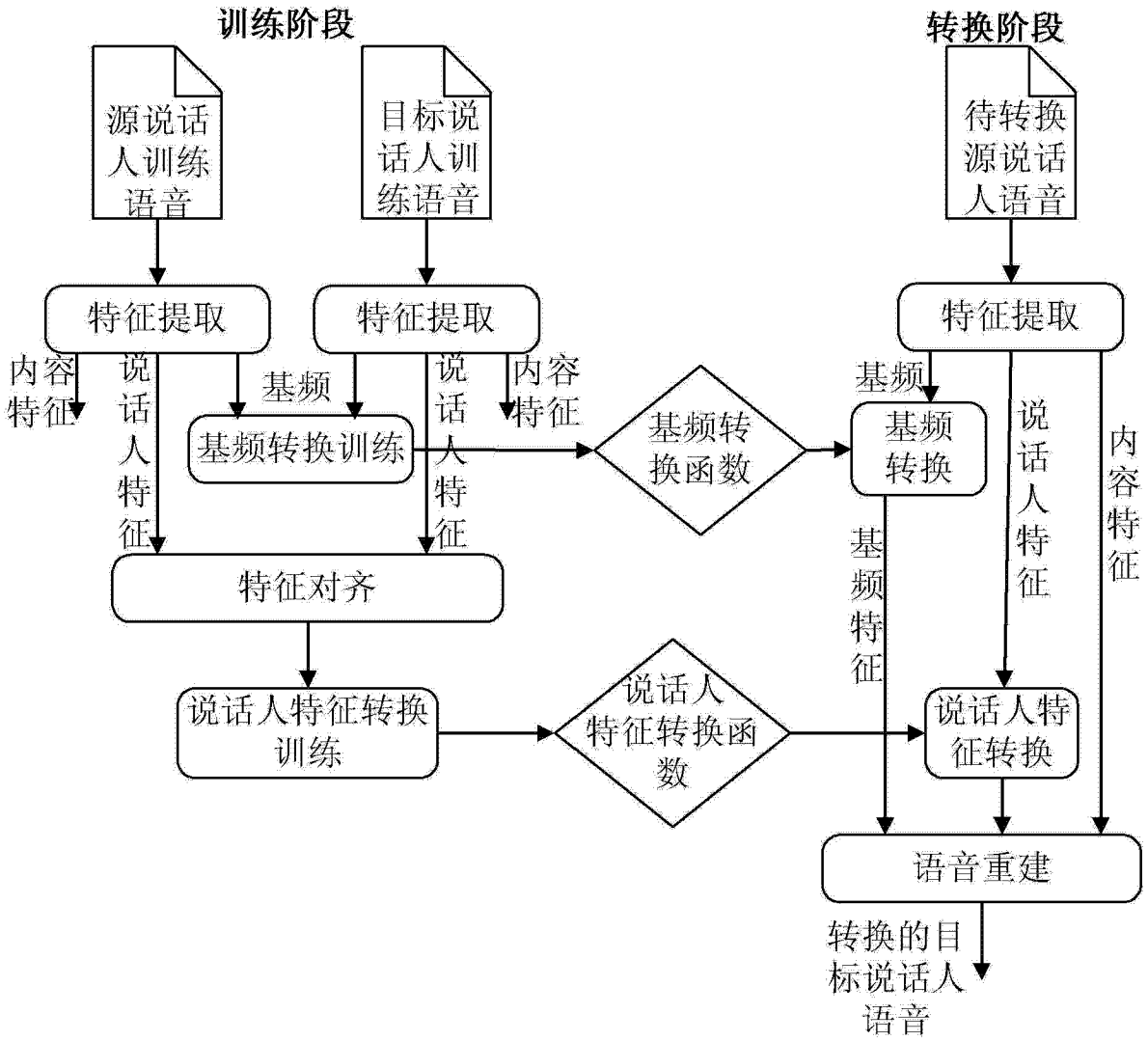

训练阶段
转换阶段
源说话人训练语音
目标说话人训练语音
待转换源说话人语音
特征提取
特征提取
特征提取
内容特征
说话人特征
基频
基频转换训练
说话人特征
内容特征
基频转换函数
基频
基频转换
说话人特征
内容特征
基频特征
特征对齐
说话人特征转换训练
说话人特征转换函数
说话人特征转换
语音重建
转换的目标说话人语音

图 1

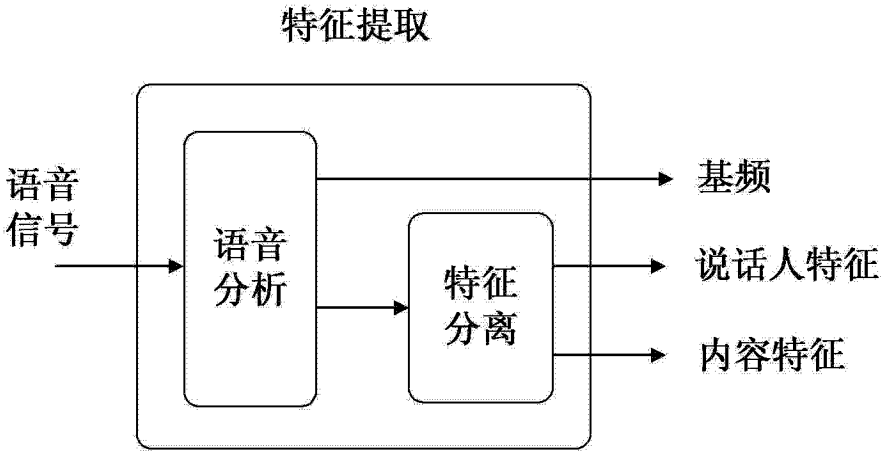

特征提取
语音信号
语音分析
特征分离
基频
说话人特征
内容特征

图 2

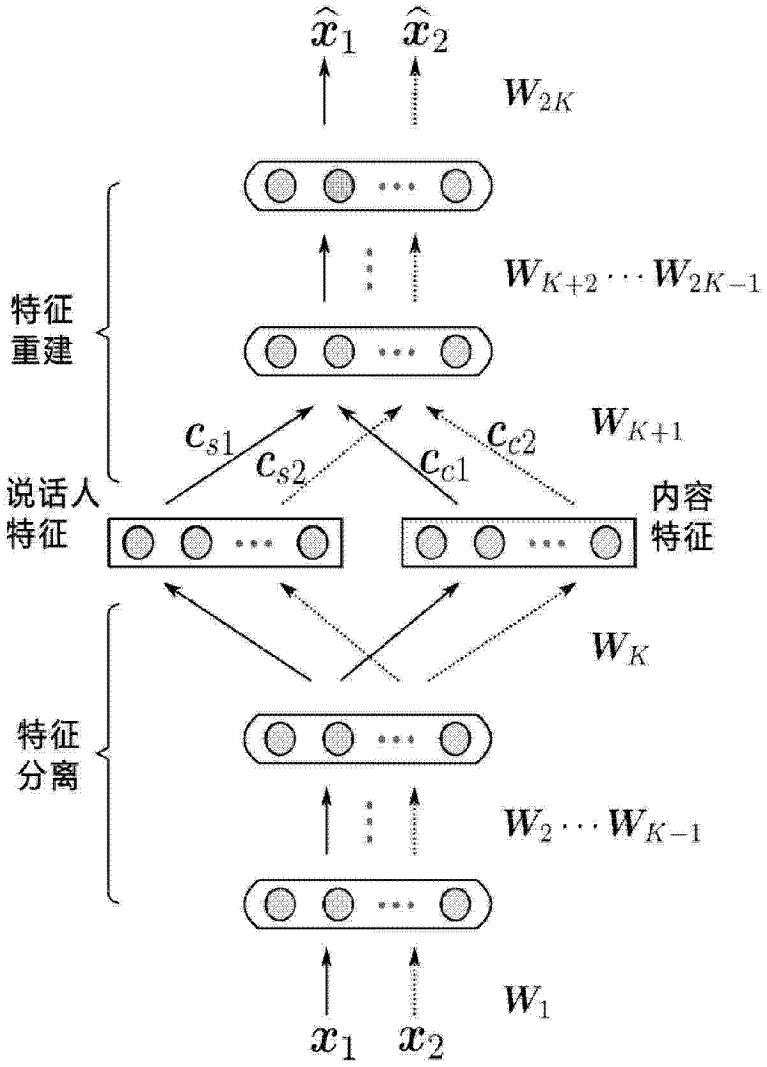

$\hat{x}_1$
$\hat{x}_2$
W_{2K}
$W_{K+2} \cdots W_{2K-1}$
特征
重建
c_{s1}
c_{s2}
c_{c1}
c_{c2}
W_{K+1}
说话人
特征
内容
特征
W_K
特征
分离
$W_2 \cdots W_{K-1}$
W_1
x_1
x_2

图 3

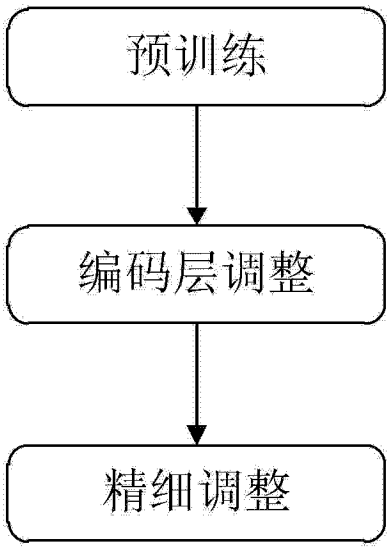

预训练
编码层调整
精细调整

图 4

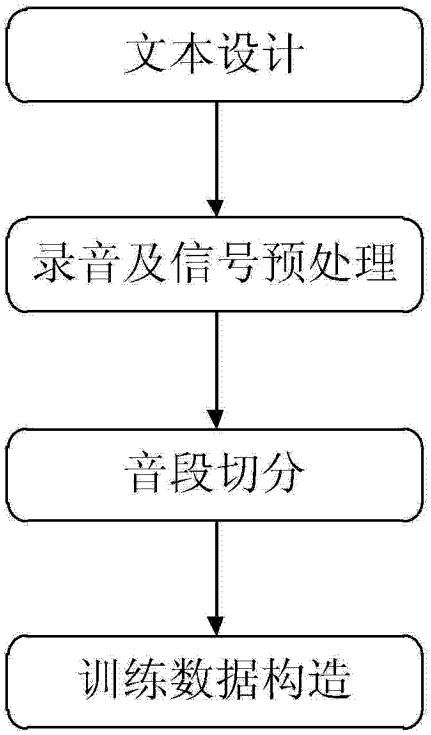


图 5

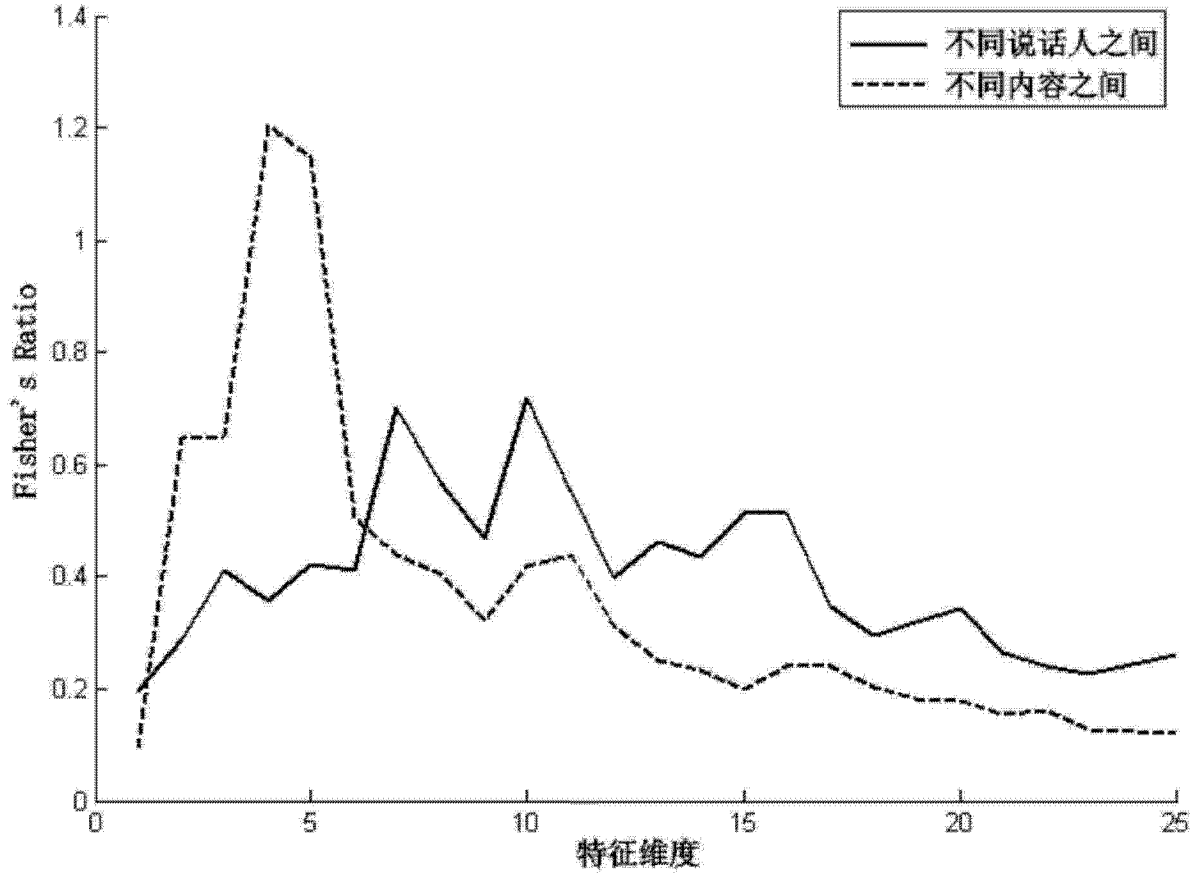

不同说话人之间
不同内容之间
Fisher's Ratio
特征维度
0
0.2
0.4
0.6
0.8
1
1.2
1.4
0
5
10
15
20
25

图 6

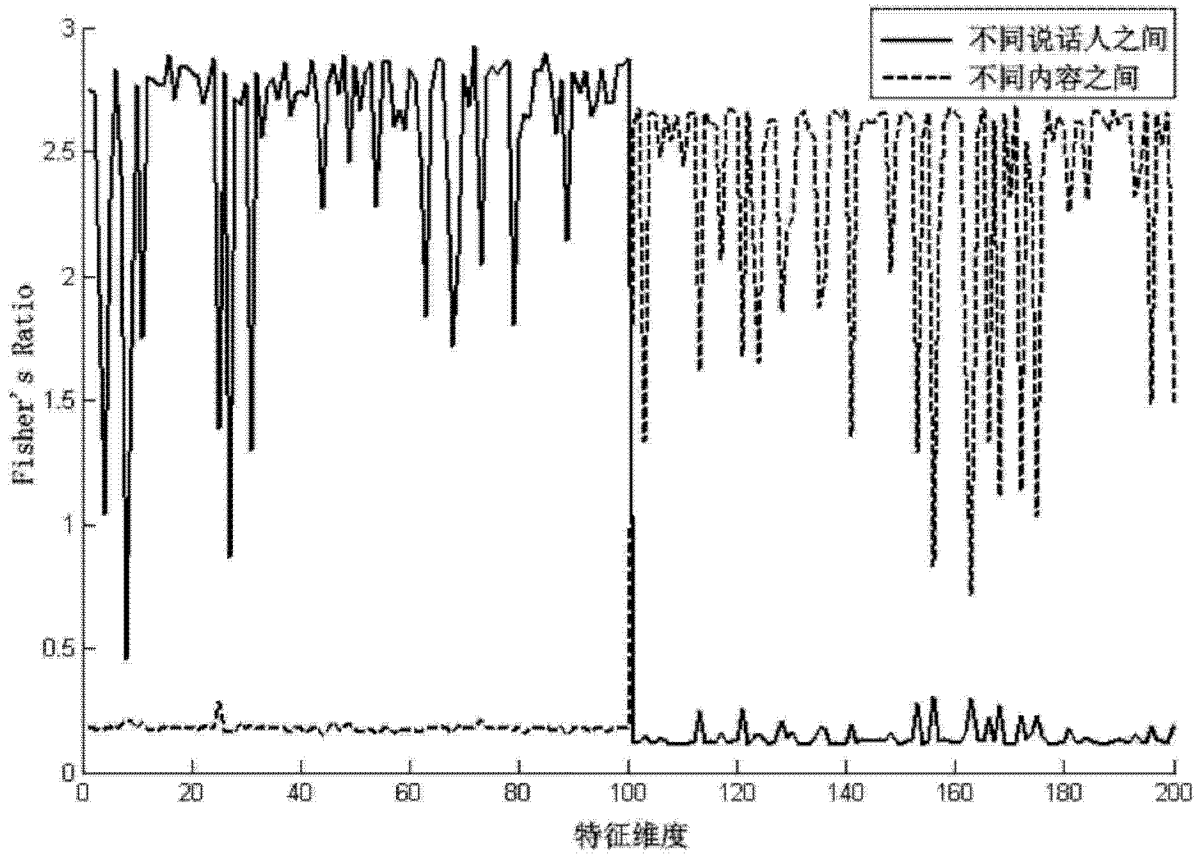

不同说话人之间
不同内容之间
Fisher's Ratio
特征维度
0
0.5
1
1.5
2
2.5
3
0
20
40
60
80
100
120
140
160
180
200

图 7