第35卷 第10期 计 算 机 工 程 2009年5月
Vol.35 No.10 Computer Engineering May 2009
·人工智能及识别技术· 文章编号：1000—3428(2009)10—0188—02 文献标识码：A 中图分类号：TN912.34

基于单字音特征提取的说话人识别方法

张 燕[1,2]，唐振民[2]，李燕萍[2]
(1. 金陵科技学院信息技术学院，南京 210006；2. 南京理工大学模式识别与智能系统实验室，南京 210094)

摘 要：证实普通话可以分解为辅音音素和单元音音素通过过度音的连接，提出一种单字音特征提取方法。该方法在传统的帧特征提取基础上，对相关帧进行二次处理，得到单字语音中的多个代表帧，将代表帧进行拼接作为单字的特征矢量。这种特征提取方法能更好地表现说话人单字发音中相邻语音帧之间的连续性。仿真实验表明该方法在说话人识别系统的应用中达到较高的识别率，使识别时间进一步缩短。
关键词：说话人识别；特征提取；单字音特征

Speaker Recognition Method Based on Character Extracting of Single Word

ZHANG Yan[1,2], TANG Zhen-min[2], LI Yan-ping[2]
(1. School of Information Technology, Jinlin Institute of Technology, Nanjing 210006;
2. Laboratory of Pattern Recognition and Intelligent System, Nanjing University of Science and Technology, Nanjing 210094)

【Abstract】 In this paper, a novel method of extracting speaker character is proposed, which is based on the single word. Compared with conventional methods, this method aims at interrelated frame, effectively finding multi-representative frames in single word utterance, and puts together the character vectors of these frames to constitute the character vector of this single word, which is adopted to calculate the codebook and model matching in training and recognition process. Simulation experiment indicates that this method can acquire higher recognition rate in Speaker Recognition(SR) system.
【Key words】 Speaker Recognition(SR); character extracting; single word character

1 概述

说话人识别(Speaker Recognition, SR)就是从说话人的一段语音中提取出说话人的个性特征，它包括训练和识别2个过程，识别系统要求有效提取语音信号中表征话者信息的基本特征，特征参数的选择对识别性能的效果有非常重要的作用，该特征应该能够有效区分不同说话人且对同一说话人的发音变化保持稳定。由于目前提取的低层参数均同时含有语音信息和说话人信息，不能将两者有效地分离[1]。而高层的特征，如发音习惯、节奏、习惯用语、情感等又很难定量描述。因此，目前主要采用低层参数(最主要的是 Mel 倒谱系数[2-3]和线性预测倒谱系数[4-5])进行模板匹配、概率统计处理或使用辨别分类器，包括人工神经网络和支持向量机等。在这些方法中，各帧的特征矢量是独立地参与处理，从而忽略了连续发音中相邻帧特征矢量的相互联系，这种方式损失了语音片断中的相关性及其包含的话者信息。本文的思路是从单字音的角度对汉语普通话进行特征提取，在此基础上根据连续帧特征矢量的相关性为单字音寻找到几个代表帧后利用 DTW 算法[6]进行规整，将各帧矢量均值拼接得到单字音特征。最后利用矢量量化[7]方法对该特征进行训练和识别。

2 帧特征的提取

目前最为常用的语音特征为 MFCC(Mel 倒谱系数)和 LPCC(线性预测倒谱系数)及其推导出的其他参数，本文采用的是 MFCC。MFCC 利用 Mel 频率尺度三角滤波法来对信号频谱进行滤波，这种方式更符合人耳的听觉特性。一般地，MFCC 参数的计算过程的具体步骤如下：

(1)将信号进行分帧、预加重和加窗处理，然后进行短时傅里叶变换计算语音信号的频率谱，将实际频率尺度转换为 Mel 频率尺度。

(2)设置三角滤波器组，每个滤波器的中心频率在 Mel 频率轴上等间隔取点，每个滤波器的上限和下限分别为前后滤波器的中心频率。

(3)根据语音信号幅度谱求出每个滤波器的输出。

(4)对所有滤波器输出作对数变换后进行离散余弦变换，DCT 将上一步获得的 Mel 频谱变换到时域。

经过上述步骤可以得到各帧的 MFCC 系数序列，处于前列的系数主要反映了说话人声道特性，用这些系数组成语音信号的特征矢量，就可以建立说话人的模型参考集。MFCC 模拟了人的听觉特性，是目前说话人识别领域最有效的特征。

3 单字音特征提取和规整

众所周知，汉语具有相对稳定的音节结构，音节由声母和韵母组成，而韵母主要由元音音素构成。汉语韵母主要分为单韵母和复韵母。一般来说，单韵母由单独的元音音素构成，而复韵母由 2 个或 2 个以上单元音音素组合而成(鼻韵母中还包含鼻音音素)，因此单元音音素是汉语韵母的基本组成元素。如果能够将复韵母中的元音部分进行分解，并将分解后的音素映射到单元音音素的特征空间中，就能够将几十种汉语韵母缩小为几种汉语单元音音素来处理，从而有利于单

作者简介：张 燕(1969－)，女，副教授，主研方向：音频信息处理；唐振民，教授、博士生导师；李燕萍，博士研究生
收稿日期：2008-12-03 **E-mail：**sandson6@163.com

字音特征的提取。将汉语韵母分解为单元音音素的映射表如表 1 所示。

表 1　汉语韵母-单元音音素映射表

韵母	音素	韵母	音素	韵母	音素	韵母	音素	韵母	音素
a	a	ui	u-ê-i	en	e-n	ia	i-a	uai	u-a-i
o	o	ao	a-u	in	i-n	ie	i-ê	uan	u-a-n
e	e	ou	e-u	un	u-e-n	iao	i-a-u	uen	u-e-n
i	i	iu	i-e-u	ün	ü-n	ian	i-a-n	uang	u-a-ng
u	u	ie	i-ê	ang	a-ng	iang	i-a-ng	ueng	u-e-ng
ü	ü	üe	ü-ê	eng	e-ng	iong	i-u-ng	üan	ü-a-n
ai	a-i	er	e-r	ing	i-ng	ua	u-a	ê	ê
ei	ê-i	an	a-n	ong	u-ng	uo	u-o		

汉语单字音由声母和韵母组成，相对于拼写式的语言含有的音素更少，这使单字特征的提取成为可能。在单字发音过程中，擦音和元音的发声时间较长，声道形状也比较稳定，相邻帧的特征更为接近，文中将这些帧称为稳定帧。相对来说，发爆破音和各音素的转换阶段，声道形状发生较大的转变，相邻帧的特征会发生较大的跳变现象，称为过渡帧。用具有代表性的稳定帧可以表示单字发音过程中的声道变化，这些被选定的帧在文中统称为代表帧。

经过逐帧的 MFCC 特征提取，得到特征序列 $\boldsymbol{C}_1, \boldsymbol{C}_2, \cdots, \boldsymbol{C}_n$，$n$ 为帧总数，定义相邻帧特征向量的夹角为

$$\theta_i = \arccos[\frac{\boldsymbol{C}_i \cdot \boldsymbol{C}_{i+1}}{|\boldsymbol{C}_i| \cdot |\boldsymbol{C}_{i+1}|}] \tag{1}$$

给定一个夹角的阈值可以将稳定帧和过渡帧分割开，此时连续 n 个短时帧被划分为稳定帧和过渡帧相间隔的帧序列。由于瞬时特性，过渡帧一般不会超过 2 帧的变化时间。将单字音中连续的稳定帧特征取均值代表该稳定阶段的总的特性。最后按顺序将稳定帧特征均值向量排列起来就可以得到单字发音过程中特征变化过程，作为单字特征。

由于同一话者在不同时刻即使发同一音，在持续时间和瞬时速度上都会产生差异，声道变化也会有所不同，各次发音提取的代表帧个数也会出现差别，因此利用动态时间规整(DTW)技术较好地解决单字特征的对齐问题。

动态时间规整 DTW 是将时间规整和距离测度结合起来的一种非线性规整技术，设参考模板是 M 帧，待识模板有 N 帧，长度不同($M \neq N$)，寻找建立一个时间规整函数 $m = w(n)$ 满足 $D = \min\limits_{w(n)} \sum\limits_{n=1}^{N} d[n, w(n)]$，将待识矢量的时间轴非线性地映射到参考模板的时间轴上，其中，$d[n, w(n)]$ 是第 n 帧待识矢量和第 m 帧参考矢量之间的距离测度；D 是相应于最优时间规整下 2 个模板矢量的匹配路径。利用该技术可以得到动态规划匹配距离和最优规整函数。针对每个汉语单字发音选择出的不同个数的帧特征矢量(包括可以反映音素特性的稳定帧和过渡帧)，利用 DTW 技术实现矢量之间的映射后，就可以直接利用矢量量化方法实现说话人特征的提取和识别。

4　数字音说话人识别系统结构

数字音说话人识别系统将本文提出的基于单字音特征提取方法和矢量量化(VQ)相结合进行汉语数字音集合的说话人识别实验。主要包括单字音分割、帧特征提取、单字特征提取和规整、码本建立和匹配判决等模块。单字音分割采用了常用的能量-过零率算法将数字音分割进行分别处理，为更好地反映本文提出方法的有效性，数据采集时选择了易于分割的连续数字音。经过帧特征提取模块可以获得各帧特征矢量，单个数字音的特征提取和规整完全按照本文第 3 节描述的方法实现。训练过程中利用矢量量化方法计算出每个说话人的每个数字发音的码本中心；识别时则对待识数字特征计算与每个说话人的每个单字特征码本中心的距离，以距离最小的码本中心所属的说话人为该数字识别结果。最后，对连续语音中的多个数字的识别结果进行融合给出判决。系统结构如图 1 所示。

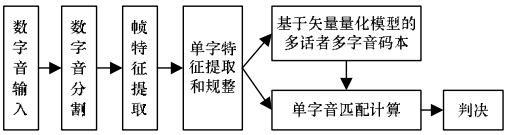

图 1　数字音说话人识别系统结构

5　仿真实验及讨论

本实验所用的语音数据均使用 NT2 型电容传声器和 MAYA44.V3 专业数字音频卡采集，其采样频率为 22.1 kHz，采用 24 bit 量化。录音在普通实验室安静环境下进行，所录数据包括 100 个说话人(64 男 36 女)的单字和连续数字发音，一部分用于训练，另一部分用于测试。语音信号先进行时域归一化处理，短时帧长 25 ms，帧交叠 30%，预加重系数选择 0.96，时域计算加矩形窗，频域加汉明窗。

5.1　代表帧数量的选择对识别性能的影响

对于单字音来说，通过改变相邻特征矢量的角度阈值可以控制代表帧的数量，较多的代表帧更倾向于单字音的细节，较少的代表帧则更多地表达单字音的帧特征总体走向。因此代表帧数量的选择将在很大程度上影响单字特征的提取和识别效率。根据不同代表帧数量进行了实验，实验中训练数据均按照给定的数字音序列进行采集，采用单字音作为测试数据，结果如表 2 所示。

表 2　代表帧数量与识别性能的关系

帧数量	识别率/(%)	时间/s
1	82.51	0.44
3	95.21	0.62
5	99.37	0.77
8	99.81	0.85
10	98.93	0.92
15	99.35	0.99
25	99.44	1.08

数据显示，当选择的代表帧数量大于 3 个时就能够取得较好的识别效果，这与所选择汉字组成的音素较少和发音比较稳定有关。随着代表帧数进一步增加，识别率并没有出现进一步显著的提高，这表明不需要过多地关注单字特征精细结构。从运行时间的角度看，较少的代表帧数量意味着较少的分类和匹配运算量。综合考虑识别率和运行时间得出结论：当代表帧数量在 5 附近时系统达到较好的性能。

5.2　数字音特征系统性能

该实验将本文提出的单字音特征提取方法和矢量量化方法相结合实现的数字音说话人识别系统和 LPCC 系数、MFCC 系数长时统计方法[8]进行了对比。训练数据均按照给定的数字音序列进行采集，保证训练过程中系统能够准确地对说话人不同字音进行分类。单字特征-矢量量化方法选择的代表帧数量为 5。识别数据分为单字音识别和连续数字音识别，作为识别使用的连续语音允许连读口音等自然语言方式发音。实验结果如表 3 所示。

表 3　单字特征法和 LPCC 和 MFCC 法识别系统性能比较

特征类型	单字音识别率/(%)	平均识别时间/s	连续音识别率/(%)	平均识别时间/s
LPCC 特征法	85.31	1.25	75.22	4.44
MFCC 特征法	98.61	1.14	83.59	3.51
单字 MFCC 特征法	99.37	0.77	88.49	2.53

(下转第 192 页)