添加音素持续时间信息到频谱模型的说话人辨认研究

刘大鹏[1,2],尾关和彦[2],朱庆生[1]
(1.重庆大学 计算机学院,重庆 400044;
2.电气通信大学 信息通信工程系,日本 东京 182-8585)

摘 要:传统的声音识别系统通过短时声音频谱信息来辨识说话人,这种方法在某些条件下具有较好的性能。但是由于有些说话人特征隐藏在较长的语音片段中,通过添加长时信息可能会进一步提高系统的性能。在文中,音素持续时间信息被添加到传统模型上,以提高说话人辨识率。频谱信息是通过短时分析获得的,但音素持续时间的提取却属于长时分析,它需要更多的语音数据。通过大量语音数据探讨了音素持续时间信息对说话人辨识的有效性,提出2种方法来解决数据量小所引起的问题。实验结果表明,当说话人的声音模型被恰当建立时,即使在语音数据量小的情况下,音素持续时间信息对说话人辨识率的提高也是有效的。

关键词:说话人声音辨识;高斯混合模型;音素持续时间信息
中图分类号:TP391.42;TN912.3 **文献标识码**:A **文章编号**:1673-629X(2007)05-0156-04

Adding Phoneme Duration Information to Spectral Model in Speaker Identification

LIU Da-peng[1,2], Kazuhiko Ozeki[2], ZHU Qing-sheng[1]
(1. Sch. of Computer Science, Chongqing University, Chongqing 400044, China;
2. Dept. of Information and Communication Engineering, The University of Electro-Communications, Tokyo 182-8585, Japan)

Abstract: Conventional speaker recognition systems use short-term spectral information to identify speakers. They perform well on some conditions. However, since a part of speaker characteristics is hidden in longer speech segments, the performance may be further improved by adding this long-term information. In this paper, phoneme duration information is added to the conventional model to improve the recognition rate. While spectral information is extracted by short-term analysis, extracting phoneme duration information requires long-term analysis. Thus phoneme duration analysis usually needs more speech data than spectral analysis does. In the first part of this work, effectiveness of phoneme duration information is investigated by using a large amount of speech data. Then two methods are presented to solve the problem caused by only using a small amount of data. Results of the experiments show that phoneme duration information is effective to improve speaker identification performance even when using a small amount of speech data, if the speaker models are built appropriately.

Key words: speaker identification; GMM; phoneme duration information

0 前 言

说话人识别的主流技术是光谱分析,并且已经被证明是一种非常有效的方法。然而由于声音是一种语音内容和说话人特征的混合体,在说话人识别的研究中,通过强调说话人的特征就可能达到提高说话人识别率和稳定性的效果[1]。另外光谱分析属于短时分析法,极易受到噪音干扰,而说话人的长时特征对噪音具有更强的抵抗性,因此添加此类信息也会提高系统的鲁壮性。很多研究者已经做过一些相关的研究,并取得了不错的结果[2~5]。例如在文献[2]中,诗律和词汇特征被添加到传统的光谱分析模型中,结果显示某些长时特征中包含着很多光谱特征的补充信息。

在文中,音素持续时间被选为长时特征添加到光谱分析模型中[6]。音素持续时间就是单个音素的发音

收稿日期:2006-08-12

作者简介:刘大鹏(1980-),男,山东莱州人,硕士研究生,研究方向为说话人识别;尾关和彦,教授,重庆大学顾问教授,研究方向为语言处理;朱庆生,教授,博士生导师,研究方向为图像及多媒体技术。

时间长短。由于光谱分析属于短时分析,从一分钟的语音数据中可以获得成千上万帧分析数据,而音素持续时间属于长时分析,只能获得 300 个左右的分析数据,同时这些数据属于不同的语音音素,因而其数据量不足以为每个音素建立标准模型。笔者提出了两种方法来解决这个问题:2 分组方法和数据直接作为模型均值法(简称为 MD 法)。在 2 分组方法中,所有的音素被分成了 2 组,每一组利用属于该组的音素建立一个高斯混合模型;在 MD 法中每个音素或语音数据建立一个高斯模型,模型的均值由属于音素的语音数据均值或数据本身决定,而协方差由其他非标准方式获得。

1　MFCC－GMM 基本模型

在基本模型中笔者采用了美倒谱系数(MFCC)提取特征,用高斯混合模型(GMM)为每个说话人训练特征[5]。在高斯混合模型中最大期望算法(EM)被使用,初始化高斯混合模型时用到了 LBG 算法。表 1 给出了基本模型中的参数配置。

表 1　基本模型参数配置

采样频率	16kHz
预加重	$1-0.97z^{-1}$
窗口类型	Hamming
窗口长度	32ms
窗口平移	8ms
MFCC 维数	48
GMM 混合度	64

实验中采用了日语数据库 ATR_DB_A。该数据库包含 5240 个单词,216 个音素平衡单词,15 个数字和 90 个句子,由 10 个男性和 10 个女性发音,并且所有的音素持续时间和停顿都在数据库中标明。在实验中 40 个训练词和 20 个测试词是从 5240 个样本中随机提取。40 个音素平衡词是从 216 个音素平衡词中每隔 5 个提取,音素平衡词是指每个音素的出现频率几乎相等。表 2 给出了 MFCC－GMM 的识别率。

表 2　MFCC－GMM 基本模型的识别率(%)

测试数据 \ 训练数据	5 个句子	40 个音素平衡词	40 个词
15 个数字	85.67	99.67	94.33
20 个单词	80.25	90.25	99.50
25 个句子	99.60	93.00	86.80

2　大量语音数据训练的音素模型

在上文中已经提到过,提取音素持续时间通常需要大量语音数据,所以首先用大量的语音数据来验证音素持续时间信息是否能提高说话人识别率。

实验显示音素持续时间符合高斯混合模型,所以为每一个音素建立 GMM 模型。音素持续时间和其出现频率的关系在附录中给出。

在实际的语音中,音素其实不能够被精确地定义和分离,例如在日语中理论上大约只有 50 个左右的音素,但是在某些说话人的语音数据中标记了 120 多个类音素的符号。例如在一些说话人发音中“u,o,u”就被合并成一个音素,不能被机器和人正确地分离。因此,不同的说话人有不同的音素数。在本次实验中,只有对于所有的说话人,其音素出现频率超过 50 次的才为其建立 GMM 模型,其他音素目前被认为是无用信息。对于每个有效音素,一个 2 维的高斯混合模型被建立,这样为每个注册的说话人都建立了一系列的音素持续时间的高斯混合模型。

从表 3 中可以发现句子中说话人的语音停顿时间长短信息对于提高说话人识别率非常重要。所以在后面的实验中句子的停顿信息都被认为是一个音素的持续时间信息。从表 3 中还可以发现当测试数据长度增加时,识别率会急剧地增加(句子的长度远大于数字和词的长度)。所以我们还实验了把 4 个单词连接作为一个测试数据,通过使用 20 个这种长单词作为测试数据,5160(5240－4×20)个单词作训练数据,识别率从 25.25%增加到 55.25%。

表 3　基于大量数据的音素模型的识别率(%)

测试数据 \ 训练数据	90 个句子	5220 个单词
15 个数字	8.33	20.33
20 个单词	10.25	25.25
25 个句子	80.80 (使用停顿) 66.00 (不使用停顿)	15.40

接下来把基本模型和大量数据训练得到的音素模型结合。设基本模型的权重时 α,音素持续时间模型权重为(1－α),通过优化 α 的值来得到最大识别率,错误降低率(ERR)如表 4 所示,其中 B1 是 5 个句子训练的基本模型,B2 是 40 个单词训练的基本模型,P1 是 90 个句子训练的音素持续时间模型,P2 是 5220 个单词的音素持续时间模型。

表 4　大量数据训练的音素持续时间模型结合到基本模型时的错误降低率(%)

测试数据 \ 训练数据	B1＋P1	B1＋P2	B2＋P1	B2＋P2
15 个数字	11.58	9.28	0	6.00
20 个单词	0	17.72	100	50
25 个句子	50	50	86.36	31.82

表 4 显示尽管基本模型的识别率已经很高了，通过添加音素持续时间信息仍然能使识别率进一步提高。

3 少量语音数据下的音素持续时间模型

在上文中已经证明音素持续时间信息确实对说话人识别率的提高有很大的帮助，它含着一些光谱信息的补充信息，然而笔者对光谱模型和音素持续时间模型采用了不同的训练数据，使得上述方法变得不太实用。所以在这部分中，将讨论如何在使用与光谱模型相同的少量数据的情况下建立音素持续时间模型。

3.1 分组方法

在此之前，曾尝试过利用每个说话人的所有数据，不分类地建立一个音素持续时间高斯混合模型，但是模型的识别率却几乎没有提高，因此考虑把音素信息分成多组。

在 2 分组模型中，所有的音素根据 LBG 算法划分成 2 组，每组建立一个高斯混合模型。当测试数据进来时，系统首先根据每个音素的持续时间将其归类，然后根据该类的模型计算其可能度。

表 5 显示，通过添加 2 分组模型到基本模型中，大约 12.16% 的错误率被降低了。

表 5 2 分组模型与基本模型结合后的错误降低率(%)

测试数据 \ 训练数据	5 个句子	40 音素平衡单词	40 个单词
15 个数字	9.28	0	0
20 个单词	1.27	0	50
25 个句子	0	42.86	6.06

3.2 MD 法

2 分组法能够描述音素持续时间特征的一些信息，但是不能精确地描述每个音素的信息特征，所以接下来考虑对每个音素建立模型。

在建立音素持续时间模型之前，先做 2 个预备实验。在第一个实验中，每个音素持续时间的数据被认为是一个高斯分布模型的均值。在辨识阶段，每个音素的可能度等于该音素的所有高斯模型的可能度的和，除以模型个数。这样就把高斯混合模型看成每个数据的高斯模型的线性结合。说话人 k 对于语音数据 x 的可能度由公式(1) 给出：

$$L_k(x) = \frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{1}{m_i}\sum_{j=1}^{m_i}P_{kij}(x)\right) \tag{1}$$

公式中 k 表示说话人，n 为测试数据中音素的个数，m_i 属于该音素的训练数据数，$L_k(x)$ 为测试数据被识别成说话人 k 的可能度，$P_{kij}(x)$ 为在说话人 k 模型中音素 i 的第 j 个高斯模型密度函数。

在第二个实验中，为每个音素建立一个高斯模型，模型的均值由属于该音素的训练数据获得。说话人 k 对于语音数据 x 的可能度由公式(2) 给出：

$$L_k(x) = \frac{1}{n}\sum_{i=1}^{n}\log(P_{ki}(x)) \tag{2}$$

当把这两个模型结合到基本模型中时，获得了相似的识别率，所以在文中只给出了第二种方法的识别率。

现在已经由训练数据获得了高斯模型的均值，接下来介绍实验中采用的 3 种方法来获得协方差。

①从大量数据的实验中发现，几乎所有音素的协方差都在 1000 到 3000 之间，所以可以对所有的音素寻找一个共用的最优协方差。

②在上文中可以得到每个音素的协方差，通过大量语音数据的协方差来估计少量语音数据情况下的协方差。这种方法可以更精确地评估每个音素的协方差，但是它的缺点是依赖于前期的大量语音数据。

③通常均值越大，协方差越大，所以可以假定协方差与均值成比例，确定最优比例系数。实验结果显示第二种方法比其他两种方法的识别率稍微高一点，下面只列出第二种方法的识别结果。

在表 6 中，当 MD 模型结合到基本模型中时，平均 26.05% 的错误率被降低了。接下来把 2 分组模型和 MD 模型结合称为混合的音素持续时间模型，然后再把它与基本模型结合。结果发现混合的音素持续时间模型的识别率进一步地提高，但是再结合到基本模型中，错误率却不再继续降低。

表 6 当 MD 模型结合到基本模型时的错误降低率(%)

训练数据 \ 测试数据	5 个句子	40 音素平衡单词	40 个单词
15 个数字	2.30	100	0
20 个单词	0	7.69	50
25 个句子	0	45.71	28.79

4 结 论

在文中音素持续时间被作为说话人特征添加到 MFCC－GMM 基本模型中。首先大量语音数据被使用来验证音素持续时间信息对说话人识别的有效性，接着验证了与光谱分析相同数据情况下的有效性。

为了解决由数据量少而引起的问题，两种方法：2 分组法和 MD 法被提出，实验证明每一种方法对说话人识别率的提高都有帮助，特别是 MD 法。因此可以得到结论：即使在数据量很小的情况下，如果模型被恰

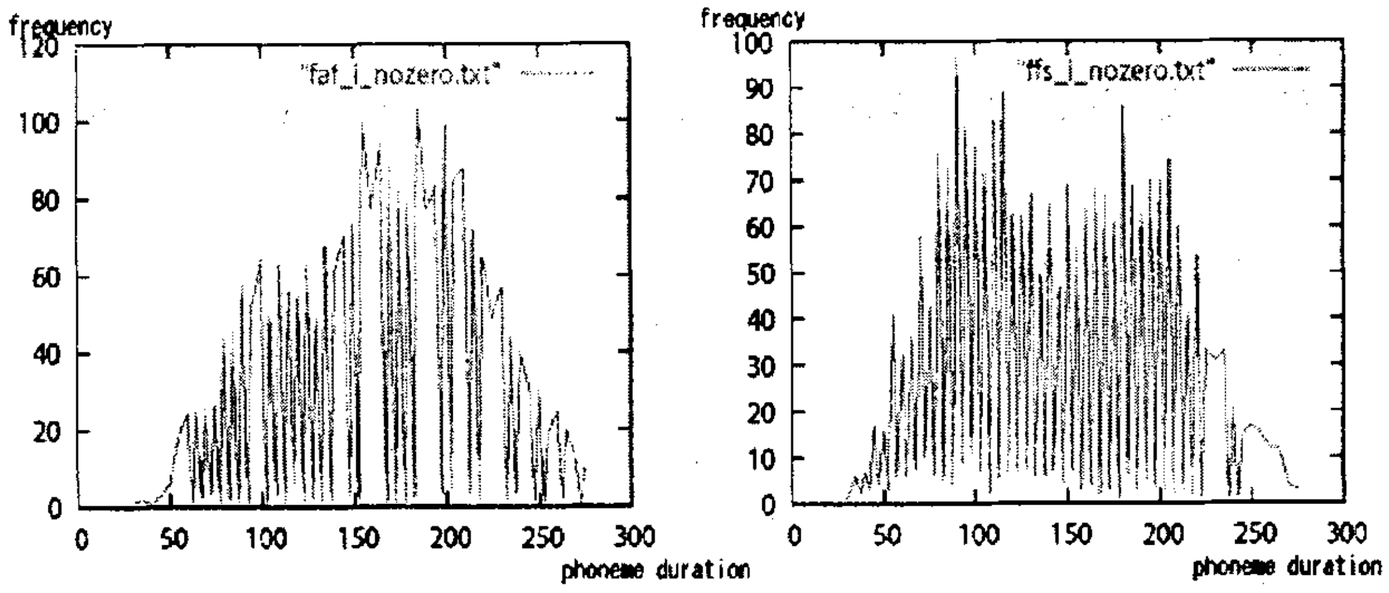
当地建立,音素的持续时间信息也可以作为光谱模型的有效补充。

在本次实验中,音素的持续时间信息是已经在语音数据库中标注好了的,今后的工作是打算实现音素持续时间的自动标注,这将使文中陈述的方法更加实用。

5 致　谢

感谢吉田健一博士对实验的帮助。同时感谢日本 JASSO 组织为文中第一作者提供留学奖学金。

附录:下图表示 5240 个日语单词中音素'i'持续时间与其出现频率的关系。



参考文献:

[1] Furui S. Digital Speech Processing, Synthesis, and Recognition [M]. [s.l.]: Marcel Dekker, Inc, 2001.

[2] Ramachandrula, Sitaram, Thippur, et al. Connected phoneme HMMs with implicit duration modelling for better speech recognition[C]//Proceedings of the International Conference on Information, Communications and Signal Processing. Singapore: IEEE, 1997: 1024 - 1028.

[3] Ghesquiere, Pieter - Jan, Van compernollt, et al. Flemish accent identificationbased on formant and duration features [C]//ICASSP. Orlando, Florida, USA: IEEE, 2002: 749 - 752.

[4] Kajarekar S, Ferrer L, Venkataraman A, et al. Speaker recognition using prosodic and lexical features[C]//In Proc. IEEE Speech Recognition and Underst anding Workshop. Virgin Islands, US: St. Thomas, 2003: 19 - 24.

[5] Chow D, Abdulla W H. Robust speaker identification based on perceptual log area ratic and gaussian mixture models[C]//Proc ICSLP. Jeju Island, South Korea: IEEE, 2004: 1761 - 1764.

[6] Miller, David R, Trischitta, et al. Statistical dialect classification based on mean phonetic features [C]//Proc ICSLP. Philadelphia: IEEE, 1996: 2025 - 2027.

(上接第 155 页)

个数据集合的顺序执行时间约为 1.46s)。

表 2　不同划分及通信速率情况下的加速比

通信时间 μs	划分策略[m,d]下的加速比 t_s/t_p		
$(t_{startup}, t_{data})$	[20,50]	[50,20]	[100,10]
(110,10)	2.83	2.93	2.48
(11,1)	4.36	4.85	4.97

由表 2 可以看出,对于通信速率较低的情况 $(t_{startup}, t_{data}) = (110\mu s, 10\mu s)$,流水线级划分过密反而造成加速比的降低。如划分策略中 $m = 100$、$d = 10$ 时,加速比降低为 2.48。当通信速率选择较高的(11,1)时,加速比随流水线划分级数的增长而增大。

需要说明的是,对于流水线这种进程间通信较频繁的并行算法,其并行的墙上时间(wall clock time)受通信速率的影响很明显,对于表 2 给出的通信时间,若再慢一个数量级,则加速比将小于 1。

5 小　结

文中对多级驱动的数字混沌编解码方案进行了并行优化,采用流水线方式对各级混沌信号分进程处理,得到了令人满意的并行加速比,为转向多进程软、硬件系统的开发提供了实验依据。在流水线效率分析中,对非均匀流水线的运行时间作了推导,并对方案中不同划分和通信速率下的加速比进行了测试。加速比分析结果表明,在进程间通信时间合理的情况下,系统的执行速度比以往的顺序执行方式有明显提高。

参考文献:

[1] Pecora L M, Carroll T L. Synchronization in chaotic system [J]. Phys Rev Lett, 1990, 64(8): 821 - 824.

[2] Frey D. Chaotic digital encoding: An approach to secure communication[J]. IEEE Transactions on Circuits and Systems - II, 1993, 40(10): 660—666.

[3] 杨世平,牛海燕,田　钢,等. 用驱动参量法实现混沌系统的同步[J]. 物理学报,2001,50(4):619 - 623.

[4] 张翌维,柯熙政,席晓莉,等. 一种多级数字混沌编码方案及其硬件实现[J]. 电子技术应用,2005,31(2):58 - 60.

[5] WILKINSON B, ALLEN M. Parallel programming: techniques and applications using networked workstations and parallel computers[M]. NJ: Prentice Hall, 1999

[6] Sharma N, Ott E. Exploring synchronization to combat channel distortions in communication with chaotic system[J]. Int J Bifurcation and Chaos, 2000, 10(4): 777 - 785.