第 35 卷第 2 期　　声　学　学　报　　Vol. 35, No. 2
2010 年 3 月　　ACTA　ACUSTICA　　Mar., 2010

献给马大猷教授 95 华诞

长时语音特征在说话人识别技术上的应用*

张建平　李明　索宏彬　杨琳　付强　颜永红
(中国科学院声学研究所，中科信利语音实验室　北京　100190)

2010 年 2 月 2 日收到
2010 年 2 月 11 日定稿

摘要　本文除介绍常用的说话人识别技术外，主要论述了一种基于长时时频特征的说话人识别方法，对输入的语音首先进行 VAD 处理，得到干净的语音后，对其提取基本时频特征。在每一语音单元内把基频、共振峰、谐波等时频特征的轨迹用 Legendre 多项式拟合的方法提取出主要的拟合参数，再利用 HLDA 的技术进行特征降维，用高斯混合模型的均值超向量表示每句话音时频特征的统计信息。在 NIST06 说话人 1side-1side 说话人测试集中，取得了 18.7% 的等错率，与传统的基于 MFCC 特征的说话人系统进行融合，等错率从 4.9% 下降到了 4.6%，获得了 6% 的相对等错率下降。

PACS 数：43.72, 43.60

Long span prosodic features for speaker recognition

ZHANG Jianping　LI Ming　SUO Hongbin　YANG Lin　FU Qiang　YANG Yonghong
(*ThinkIT Speech Lab, Chinese Academy of Sciences* Beijing　100190)

Received Feb. 2, 2010
Revised Feb. 11, 2010

Abstract　In this paper, we first give an introduction about speaker recognition techniques. Then a novel speaker verification method based on long span prosodic features is proposed. After speech is pre-processed by a voice activity detection module, and basic prosody features are extracted for each speech unit, we carried out an approximation of the pitch, formant, time domain energy and harmonic energy contours by taking the leading terms in a Legendre polynomial expansion. HLDA is used to reduce the feature dimension and mean supervector in each individual Gaussian is used to represent the distribution of the time-frequency features. Experiments on NIST06 show that the proposed method can reduce the EER from 4.9% to 4.6% when fusing with the traditional MFCC-featured system.

引言

目前经典的说话者识别系统都是基于频谱参数上的，如梅尔倒谱系数 (MFCC)，线性感知系数 (PLP) 等，随着识别任务环境复杂度的增加，出现了大量基于频谱参数上面的建模及信道补偿方法 [3−5]，这些系统都能取得良好的识别性能。但由于说话者的语音中含有丰富的能代表说话者身份的信息，如何从语音中提取出除频谱参数以外的能代表话者身份的特征参数一直是研究者所关注的领域。近几年来，如韵律，词汇，音素等高层特征参数逐渐被应用于话者识别系统中 [7−8]，一方面是这些高层特征参数在声学环境变化时能够保持一定的稳定性，不易受信道及噪声的影响，另一方面是这些时频特征还能反映说话者的说话风格，习惯等个性特征 [9] 。由于特征参数上的互补性，将本文的时频特征系统与基于底层声学参数系统融合后可以提升整体识别性能。

本文提出了一种基于长时时频特征 (基频、时域能量、共振峰、谐波频域能量等特征的 phone 级别的长时轨迹拟合参数) 的话者确认系统 PRO-GSV，对所提取的基本时频特征进行前端预处理后，通过能量进行分段，在每一个小段内部把这些时频特征的轨迹用多项式拟合的方法提取出拟合参数，再利用 HLDA[10] 的技术进行特征降维，用高斯混合模型的均值超向量表示每句话音时频特征的统计信息，利

* 本论文工作由国家科技支撑计划 (2008BAI50B00) 、国家自然科学基金 (10925419, 90920302, 10874203, 60875014) 资助项目。

用 SVM 支持向量机训练每个说话人的模型，最后对模型得分进行 ZTNORM 归一化。

1 长时时频特征的提取与处理技术

1.1 特征提取算法

本方法框架上采用 Kenny[7] 提出的架构，把连续有基频值的浊音段提取出来，在每一个段内，通过能量曲线的谷点，切分出类似音字 (phoneme) 的单元，然后，在每一个单元内部，把基频曲线，时域能量曲线，分别利用 6 阶多项式拟合得到 6 维的参数，与单元的长度一起构成 13 维的特征．因此，每一个单元提取出一帧特征，这种时频特征的帧数大大减少。但是由于描述的是长时的信息，可以描述帧与帧之间的联系，而短时的 MFCC 系数差分特征恰恰描述不了这种长时的变化趋势，因此这种长时时频特征的系统与基于 MFCC 的系统融合在一起会进一步提高整体系统的性能.

本方法的不同之处在于，不仅仅拟合基频和时域能量曲线，还拟合了前 4 个共振峰曲线和前 10 个谐波能量的曲线。因为，我们认为，每个人的共振峰频率长时变化趋势也可以在一定程度上反应说话人的信息，而且每个谐波能量的曲线变化信息不仅仅反应了被共振峰调制的谐波能量变化趋势，也反应了共振峰的强弱，这恰恰弥补了共振峰频率不能提供的幅度信息。在实验中，我们融合谐波能量曲线特征和共振峰频率曲线特征在一起来提高系统的性能。

1.2 说话人模型建模方法

本文采用了 GSV 系统 [1] 的建模方法。首先，根据每一个音字单元拟合曲线的参数得到一帧特征，提取出来的特征和 MFCC 的特征都是一帧一帧的，并且对数处理了以后，经过 CMN 和 CVN[2] 等鲁棒性处理，特征服从高斯分布，可以用高斯模型来建模，然后把高斯模型的均值超向量提取出来用 SVM 分类器进行建模，最后通过 ZTnorm[6] 等归一化方法做归整，得到得分。

1.3 利用异方差线性区分分析技术进行降维

混合高斯模型 (Gaussian Mixture Model, GMM) 是一种参数化的模型其参数为均值、方差及混合系数。其中方差可以采用对角阵或者全矩阵的形式。由于采用全矩阵的方差将导致模型参数个数大幅增加在训练数据不足时会产生过学习的问题同时出于计算量的考虑实际应用中方差采用的是对角化的矩阵形式。然而能够采用对角化矩阵的前提假设是参数矢量各维之间相互独立。而实际上参数的各维之间具有比较强的相关性并不满足对角化方差的前提假设。当继续采用对角化方差的 GMM 模型结构时就会带来负面影响造成系统性能下降。采用全矩阵的方差形式又将要求增加训练数据量。异方差线性区分分析技术 (Heteroscedastic Linear Discriminant Analysis HLDA) 技术 [10] 就是被用来解相关通过参数空间变换使参数满足对角化方差的前提假设。 HLDA 与 LDA 的区别在于取消了各类数据类内方差相同的限制也就是异方差。具体的做法是采用最大似然准则训练变换矩阵通过变换矩阵对坐标轴的旋转与拉伸使原始参数空间变换到更适合 GMM 建模的新空间。这样通过 HLDA 解相关使参数优化更适合对角化方差的 GMM 模型结构。

2 实验和结果

2.1 实验数据

本文中用于评测系统性能的数据集选用 NIST 2006 SRE 1side-1side 数据集，这个数据集是 NIST 评测最核心的数据集，所有参赛单位都必须提交系统在这个数据集上的结果。在这个数据集中，训练集包括 816 个语音文件，其中 354 个文件由男性说话人所说， 463 个文件由女性说话人所说。每个语音文件从自然电话对话语音中截取，并由单一的说话人所说，时长为 5 分钟 (包括静音) 。测试集共包括 3735 个测试语音文件，每个语音文件从电话自然对话语音中截取，并由单一的说话人所说，时长为 5 分钟 (包括静音) 。

本次实验中， UBM 训练数据采用 nist04.05 训练数据集， LFA 训练数据来自 nist04 年 8side 的数据， norm 归一化的集子是 nist05 数据集．测试集采用 NIST2006 说话人识别评测集中的男说话人部分，建模方法为 GSV 超矢量建模方法。

2.2 基于长时时频特征的实验结果及融合

从结果中，可以发现，基于长时时频特征建模的系统与传统的基于 MFCC 特征的系统具有一定的互补性，融合以后，系统性能有 6% 的相对提高.

表 1 基于长时时频特征的 PRO-GSV 系统与基于 MFCC 的系统融合性能

系统	等错率 (%)
MFCC-GSV	4.92
PRO-GSV	18.7
融合	4.61

3 结论

本文提出了一种基于长时时频特征 (基频，时域能量，共振峰，谐波频域能量等特征的 phone 级别的长时轨迹拟合参数) 的话者确认系统 PRO-GSV，对所提取的基本时频特征进行前端预处理后，通过能量进行分段，在每一个小段内部把这些时频特征的轨迹用多项式拟合的方法提取出拟合参数，再利用 HLDA 的技术进行特征降维，用高斯混合模型的均值超向量表示每句话音时频特征的统计信息，利用 SVM 支持向量机进行建模. 在 NIST2006 说话人 1side-1side 男说话人测试集中，取得了 18.7% 的等错率，与基于 MFCC 的 GSV 系统进行融合，等错率从 4.9% 下降到了 4.6%, 获得了 6% 的相对等错率下降。

参 考 文 献

1 Campbell W M, Sturim D E, Reynolds D A. Support vector machines using GMM supervectors for speaker verification. *IEEE SIGNAL PROCESSING LETTERS*, 2006; **13**(5)

2 Reynolds D A, Rose R C. An integrated speech-background model for robust speaker identification. ICASSP-92 pp. II-185 - II-188

3 Pelecanos J, Sridharan S. Feature warping for robust speaker verification. In: Proc. ISCA Workshop on Speaker, Recognition - 2001

4 Campbell W M, Sturim D, Reynolds D A, Solomonoff A. SVM based speaker verification using a GMM supervector kernel and NAP variability compensation. ICASSP, 2006: 97—100

5 Kenny P, Boulianne G, Ouellet P, Dumouchel P. Joint factor analysis versus eigenchannels in speaker recognition. *IEEE Transactions. on Audio, Speech, and Language*, 2007; **15**(4): 1435—1447

6 Auckenthaler R, Carey M, Lloyd-Thomas H. Score normalization for text-independent speaker verification systems. *Digital Signal Processing*, 2000; **10**: 42—54

7 Dehak Najim, Demouchel Pierre, Kenny Patrick. Modeling prosodic feature with joint factor analysis for speaker verification. *IEEE Trans. Audio Speech and Language Processing*, 2007

8 Baker Brendan, Vogt Robbie, Sridharan Sridha. Gaussian mixture modeling of broad phonetic and syllable events for text-independent speaker verification. In: Proc. Interspeech2005, Lisbon, Portugal, 2005: 2429—2432

9 Zeng Yumin, Wu Huayu, Gao Rong. Pitch synchronous analysis method and criterion based speaker identification. In: Proc. ICNC2007

10 Kumar N. Investigation of silicon auditory models and generalization of linear discriminant analysis for improved speech recognition. Ph.D. thesis, Johns Hopkins University, Baltimore, 1997