doi:10.11835/j.issn.1000-582X.2017.05.011

语谱特征的身份认证向量识别方法

冯辉宗,王芸芳
(重庆邮电大学 汽车电子实验室, 重庆 400069)

摘　要:针对采用梅尔频率倒谱系数(mel-frequency cepstrum coefficient, MFCC)作为身份认证向量(identity vector, i-vector)进行说话人识别存在语音信息不全的问题,提出一种基于语谱特征的身份认证向量识别说话人的方法。语音信号经过预加重、分帧加窗预处理之后,通过短时傅立叶变换转换成语谱图,语谱图被提交到高斯通用背景模型,在高维均值超向量空间中选择合适的低维线性子空间流型结构以构造符合正态分布的向量作为身份认证向量。这些获取的身份认证向量经过线性判别性分析实现降维并存储。最后采用对数似然比(log-likelihood ratio, LLR)方法对训练和测试阶段的 i-vector 进行评分,完成说话人识别。以 TIMIT 数据库为标准的数值实验结果表明,相比采用 MFCC 作为特征的识别方法,研究的等错误率(equal error rate, EER)更低。

关键词:语谱图;身份认证向量;说话人识别
中图分类号:TP391.42　　**文献标志码**:A　　**文章编号**:1000-582X(2017)05-088-07

An i-vector speaker recognition method based on spectrogram

FENG Huizong, *WANG Yunfang*
(Automotive Electronics Lab, Chongqing University of Posts and Telecommunications, Chongqing 400069, P.R.China)

Abstract: An i-vector speaker recognition method using spectral features was proposed to solve the problem that there is always insufficient information when the mel-frequency cepstrum coefficients (MFCC) are used as feature vectors of i-vectors. Specifically, the speech signals are pre-emphasized, framed and windowed first, and then fed to the short-time Fourier transform to obtain spectrogram. These spectrograms are submitted into Gaussian universal background model for constructing the i-vectors in an appropriate low-dimensional linear subspace flow pattern. These vectors are conformed to normal distribution and reduced by linear discriminant analysis. Finally, Log-likelihood ratio (LLR) method is used for marking i-vectors in training and testing stage to complete the speaker recognition. Standard numerical experiment results with TIMIT database show that compared with recognition method using MFCC as features, the EER(equal error rate) of the method in this paper is lower.

Keywords: spectrogram; identity vector; speaker recognition

作为一种生物特征识别技术,语音识别与指纹识别、人脸识别等技术,在模式识别领域发挥着重要作用。考虑到语音信号的采集属于非接触式、易于获取,且在保护人类隐私具有独特优势,采用语音信号对人类身

收稿日期:2016-10-21
基金项目:重庆市教育成果转化基金资助项目(KJZH14207)。
Supported by the Chongqing Education Achievement Conversion Foundation (KJZH14207).
作者简介:冯辉宗(1972-),男,博士,教授,主要从事控制理论计算机应用,汽车电子与嵌入式系统等研究,(E-mail) fenghz@cqupt.edu.cn。

份进行认证，即说话人识别，一直受到国内外专家学者的广泛关注。具体的说话人识别过程涉及声学特征的获取、识别模型的构建 2 个方面。目前在语音识别领域普遍采用的声学特征主要包括线性预测系数(linear predictive cepstrum coefficients，LPCC)、梅尔频率倒谱系数、语谱图等[1]。特别地，由于语谱图所包含的语音信息更加完备，以及图像处理技术的发展，使得语谱图在众多声学特征中的地位更加突出。文献[2]提出了一个基于语谱图的说话人识别系统，该系统利用 Radon 变换获得语谱图中有效的声学特征，并通过离散余弦变换降低特征向量的维度。文献[3]通过图像增强的方法，求取窄带语谱图上谐波成分的分布区域，然后求和得到共振峰谐波能量参数作为特征来构建语音和端点检测系统。

随着模式识别技术的不断进步，在说话人识别领域也出现了众多富有成效的识别模型，常见的有高斯混合模型(gaussian mixture models，GMMs)[4]、GMM 结合通用背景模型(gaussian mixture model-universal background model，GMM-UBM)[5]、联合因子分析(joint factor analysis，JFA)[6]等。值得一提的是，近年来在这些模型使用上述声学特征进行说话人识别的过程中，演化出一种新的声学特征向量，即身份认证向量，引起了研究者的极大兴趣。传统的 JFA 建模过程主要基于 2 个不一样的空间：由本征音空间(eigenvoice space)矩阵定义的说话人空间，以及由本征信道空间(eigenchannel space)矩阵定义的信道空间[7]。在 JFA 理论的基础上，Dehak[8]从 GMM 均值超向量(高维向量)中提取一个更为紧凑的向量，称为身份认证向量(i-vector)，它是由总体变化因子组成的向量，用来表示说话人的身份标识。身份认证向量方法采用一个新空间(total variability space，即总体变化空间)来代替本征音空间和本征信道空间，通过 EM 算法获得映射矩阵(表示从身份认证向量到高斯均值超向量的映射)，该映射矩阵不仅包含不同说话人之间的差异信息，同时也包含了信道之间的差别信息[9]。因此，i-vector 的建模过程在 GMM 均值超向量中并不严格区分说话人的影响以及信道的影响。这种建模方法基于 Dehak 的另一个研究：JFA 建模后的信道因子中不但包含了信道效应，同时也夹杂着说话人的信息[10]。

现有的 i-vector 方法采用的是 MFCC 特征，但 MFCC 降低了各维度之间的相关性，说话人特征的信息并不完整。而语谱图由于综合了时域波形和频谱图的优点，有可能弥补这些缺点。因此，提出了基于语谱特征的 i-vector 说话人识别方法。整个识别过程主要分为训练阶段和测试阶段。在训练阶段，先对输入的语音信号进行预加重、分帧加窗，以完成预处理，其中，预加重即发送端对输入信号高频分量的提升，分帧指对语音信号划分为多个短时的语音段，而加窗是为了使帧和帧之间能够平滑地过渡。预处理的语音信号通过短时傅立叶变换(short-time fourier transform，STFT)转换成语谱图；接着对语谱图进行 i-vector 建模，并将提取到的 i-vector 存入说话人模型库中。在测试阶段，也用同样的方法对语音信号进行预处理、语谱特征提取以及 i-vector 建模。最后采用对数似然比(LLR)方法对 2 个阶段的 i-vector 进行比对，完成说话人识别。

1　基于语谱特征的 i-vector 说话人识别方法

1.1　语谱图

语谱图(spectrogram)，即语音频谱图，其 x 轴表示时间，y 轴表示频率。另外，语谱图通过图像中每个点的强度或颜色隐含了第三维来表示能量值的大小，颜色越深，该点的语音能量越强[11]。其基本数学表达式为[12]

$$x'(t)=\mathrm{STFT}\{x(t)\}\equiv X(\tau,\omega)=\int_{\tau=0}^{n}x(\tau)\omega(t-\tau)\mathrm{e}^{-\mathrm{j}\omega\tau}\mathrm{d}t, \tag{1}$$

其中：$x(t)$表示语音的时间信号，t 表示时间；$\omega(t)$表示窗函数。采用的是海明(Hamming)窗。通常情况下，$X(\tau,\omega)$是 $x(\tau)\omega(t-\tau)$的傅立叶变换形式，表示信号随时间和频率变化的相位和大小。为了抑制 STFT 的相位的跳跃不连续性，在时间轴 τ 和频率轴 ω 进行相位展开。时间指数 τ 通常被认为是“慢”时间，不会在与时间 t 同样的高分辨率上进行表示。需要指出的是，实际的计算机仿真中，使用的是离散语音信号，因此采用 STFT 的离散形式，即

$$\mathrm{STFT}\{x[n]\}\equiv X(m,\omega)=\sum_{m=-\infty}^{\infty}x[m]\omega[n-m]\mathrm{e}^{-\mathrm{j}\omega m}, \tag{2}$$

其中，$x[n]$表示语音信号 $\omega[n]$表示窗函数。

1.2　i-vector 说话人识别模型

得到语谱图之后，就将语谱图中的第三维，即能量值，作为声学特征输入到 i-vector 说话人模型中进行训

练和测试。i-vector说话人建模方法基于GMM-UBM模型,其基本思想是假设说话人信息(包含说话人个性特征的信息)和信道信息(指语音信号采集通道的特征信息)同时处于GMM高维均值超向量空间中的一个低维线性子空间流型结构中[13],表示为

$$\boldsymbol{M}=\boldsymbol{m}+\boldsymbol{T}\cdot\boldsymbol{w}, \tag{3}$$

其中,$\boldsymbol{M}$ 表示GMM均值超向量;$\boldsymbol{m}$ 表示UBM超向量,该向量与说话人和信道无关[14];$\boldsymbol{T}$ 表示总体变化矩阵;$\boldsymbol{w}$ 表示空间里的一个正态分布随机向量,也称总体因子,即i-vector。在这种方法中,假定 $\boldsymbol{T}$ 作为协方差矩阵、$\boldsymbol{m}$ 作为均值向量,对 $\boldsymbol{M}$ 进行正态分布。

假设说话人 s 的声学特征为$\boldsymbol{x}_{s,t}$,其总体因子 $\boldsymbol{w}$ 的估计过程如下:

1)计算$x_{s,t}$相对于UBM均值超向量 $\boldsymbol{m}$ 的零阶统计量 $N_{c,s}$、一阶统计量$F_{c,s}$以及二阶统计量$S_{c,s}$:,用来分别表示混合权值、均值向量、协方差矩阵所对应的Baum-Welch统计量[13]

$$N_{c,s}=\sum_t\gamma_{c,s,t}, \tag{4}$$

$$F_{c,s}=\sum_t\gamma_{c,s,t}(\boldsymbol{x}_{s,t}-\boldsymbol{m}_c), \tag{5}$$

$$S_{c,s}=\mathrm{diag}\Big\{\sum_t\gamma_{c,s,t}(\boldsymbol{x}_{s,t}-\boldsymbol{m}_c)(\boldsymbol{x}_{s,t}-\boldsymbol{m}_c)^{\mathrm{T}}\Big\}, \tag{6}$$

其中:t 表示时间帧索引;$\gamma_{c,s,t}$表示UBM第 c 个高斯分量的后验概率;m_c 表示 m 中的第 c 个高斯均值分量;diag{·}表示取对角运算。

2)估计因子 w 的一阶统计量以及二阶统计量

$$\boldsymbol{L}_s=I+\boldsymbol{T}^{\mathrm{T}}\sum{}^{-1}\boldsymbol{N}_s\boldsymbol{T}, \tag{7}$$

$$E[\boldsymbol{w}_s]=\boldsymbol{L}_s^{-1}\boldsymbol{T}^{\mathrm{T}}\sum{}^{-1}F_s, \tag{8}$$

$$E[\boldsymbol{w}_s,\boldsymbol{w}_s^{\mathrm{T}}]=E[\boldsymbol{w}_s]E[\boldsymbol{w}_s^{\mathrm{T}}]+\boldsymbol{L}_s^{-1}, \tag{9}$$

其中:$\boldsymbol{L}_s$ 是临时变量;$\sum$ 表示UBM的协方差矩阵;$\boldsymbol{N}_s$ 是由 $N_{c,s}$作为主对角元拼接成的 $FC\times FC$ 维矩阵;$\boldsymbol{F}_s$ 是由$\boldsymbol{F}_{c,s}$向量拼接成的 $FC\times1$ 维向量。

3)更新 $\boldsymbol{T}$ 矩阵

$$\sum_s\boldsymbol{N}_s\boldsymbol{T}E[\boldsymbol{w}_s,\boldsymbol{w}_s^{\mathrm{T}}]=\sum_s\boldsymbol{F}_sE[\boldsymbol{w}_s]。 \tag{10}$$

4)更新协方差矩阵 $\sum$

$$\sum=\boldsymbol{N}^{-1}\sum_s\boldsymbol{S}_s-\boldsymbol{N}^{-1}\mathrm{diag}\Big\{\sum_s\boldsymbol{F}_sE[\boldsymbol{w}_s^{\mathrm{T}}]\,\boldsymbol{T}^{\mathrm{T}}\Big\}, \tag{11}$$

其中:$\boldsymbol{S}_s$ 是由$\boldsymbol{S}_{c,s}$进行矩阵对角拼接成的 $FC\times FC$ 维矩阵;$\boldsymbol{N}=\sum\boldsymbol{N}_s$。

对上述步骤反复迭代,进行6到8次后,可近似认为 $\boldsymbol{T}$ 和 $\sum$ 收敛。

由于上述步骤没有采用鉴别性准则,还需要利用线性鉴别性分析(linear discriminant analysis, LDA)对初始i-vector进行鉴别性降维。

利用EM算法从观测值中学习高斯概率LDA(gaussian probabilistic linear discriminant analysis, G-PLDA)模型。观测值为从开发集(development set)中计算得到的i-vector。其因子分析(factor analysis, FA)模型如下所示

$$\boldsymbol{w}=\boldsymbol{m}+\boldsymbol{\Phi}\cdot y+\varepsilon12, \tag{12}$$

其中:$\boldsymbol{w}$ 是i-vector;$\boldsymbol{m}$ 表示训练i-vector的均值;$\boldsymbol{\Phi}$ 是因子载荷矩阵(factor loading matrix),即本征音空间;$\boldsymbol{y}$ 是一个服从标准正态分布 $N(0,I)$的潜在因子向量;ε 为残差。

最后用对数似然比来计算得分

$$llr=\ln\frac{p(w_1,w_2\mid H_1)}{p(w_1\mid H_0)\cdot p(w_2\mid H_0)}, \tag{13}$$

其中:H_0 和 H_1 假设分别定义为

①H_0：w_1，w_2 来自不同的说话人；

②H_1：w_1，w_2 来自相同的说话人；

则 $p(x|H)$表示 w 在假设 H 下的似然度。

2　实验结果及讨论

针对所提到的基于语谱特征的 i-vector 说话人识别方法的性能进行数值仿真验证。实验中采用 TIMIT 标准语料库来对模型进行训练和测试。TIMIT 标准语料库里一共有 630 个说话人，每人有 10 句话。将进行 2 组对比实验：1)选择其中的 400 个说话人来训练背景模型，剩下的 230 个说话人，取每个人的 5 句话用来训练每个说话人的 i-vector，剩下 1 句话用来测试；2)选择其中的 530 个说话人来训练背景模型，剩下的 100 个说话人，取每个人的 9 句话用来训练每个说话人的 i-vector，剩下 1 句话用来测试，如表 1 所示。

表 1　语音数据分配

Table 1　assignment of speech data

参量	阶段	人数	句子数/人
实验(1)	训练 UBM 模型	400	10
	训练说话人模型	230	5
	测试		1
实验(2)	训练 UBM 模型	530	10
	训练说话人模型	100	9
	测试		1

基于语谱特征的 i-vector 说话人识别方法流程如图 1 所示。

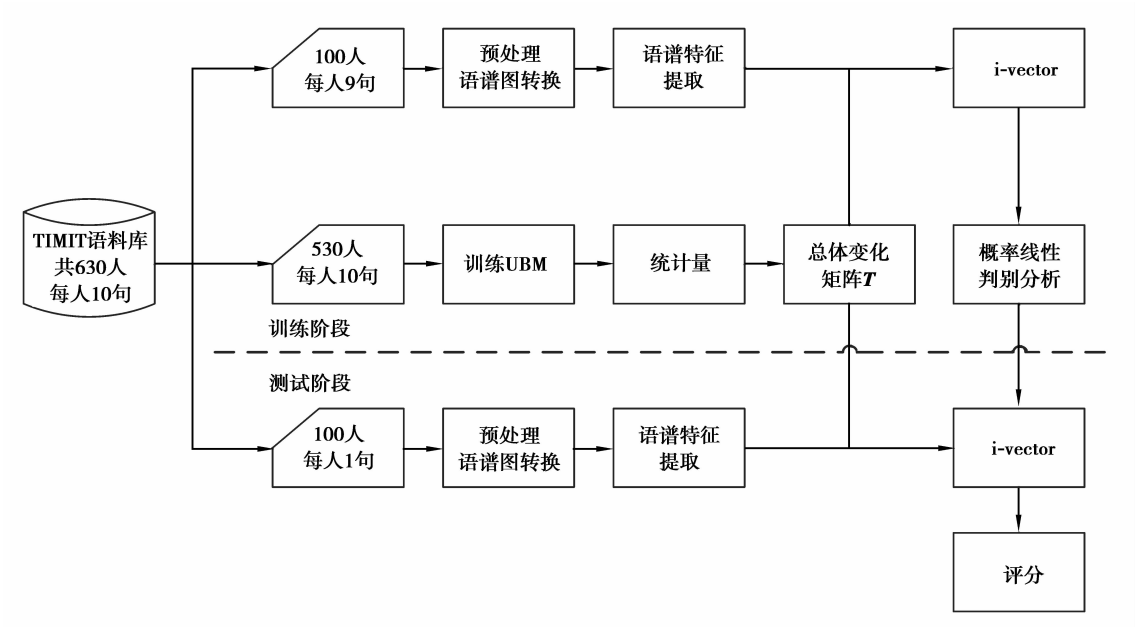


图 1　基于语谱特征的 i-vector 说话人识别方法流程图

Fig.1　flow chart of i-vector speaker recognition method based spectrogram

图 2 表示了说话人 1 对句子“She had your dark suit in greasy wash water all year.”的发音的语音信号波形图以及对应的语谱图。波形图中的 x 轴表示时间，y 轴表示振幅；语谱图中的 x 轴同样表示时间，而 y 轴表示频率。

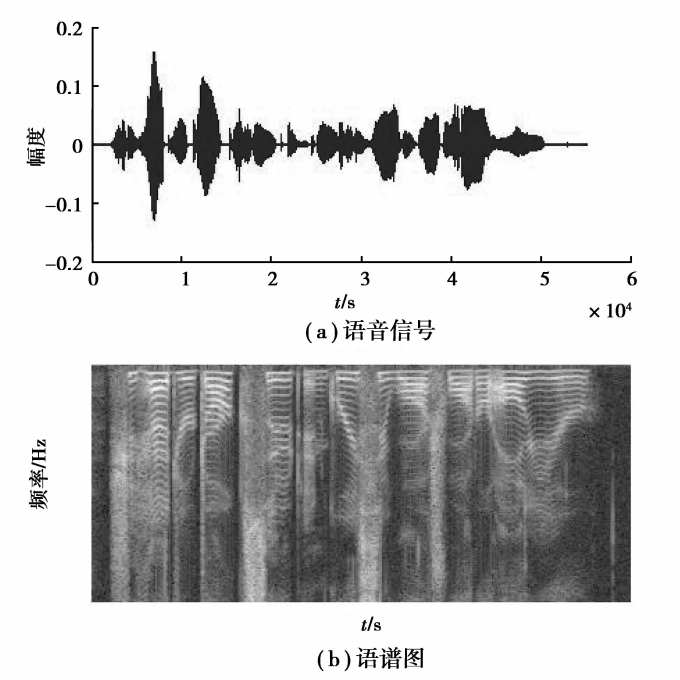


图2 说话人1语音信号及其语谱图

Fig.2 Speech signal and spectrogram of speaker 1

实验中使用的UBM混合数设置为256,高斯概率密度的方差采用对角阵。I-vector中总体变化子空间矩阵 $\boldsymbol{T}$ 的子空间维数即列数设置为400,训练时迭代次数为5次。取LDA矩阵降维后的维数为200。图3表示说话人1分别使用MFCC和语谱特征训练背景模型计算得到的i-vector特征向量。其中,实线和虚线分别表示训练和测试时提取的i-vector向量。

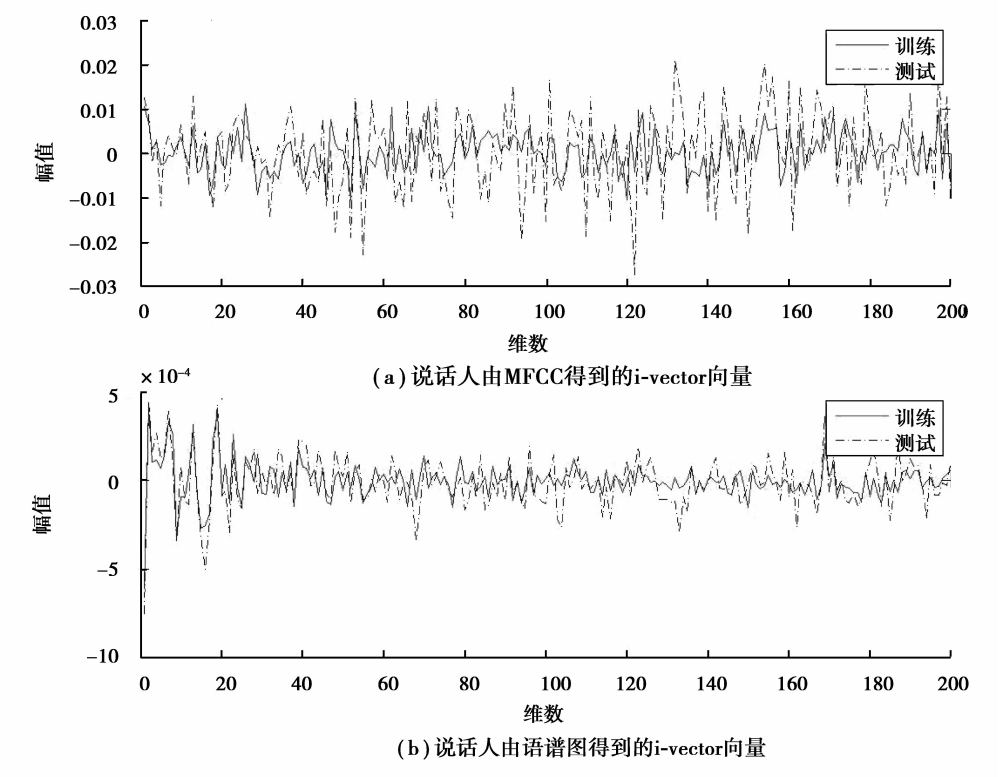
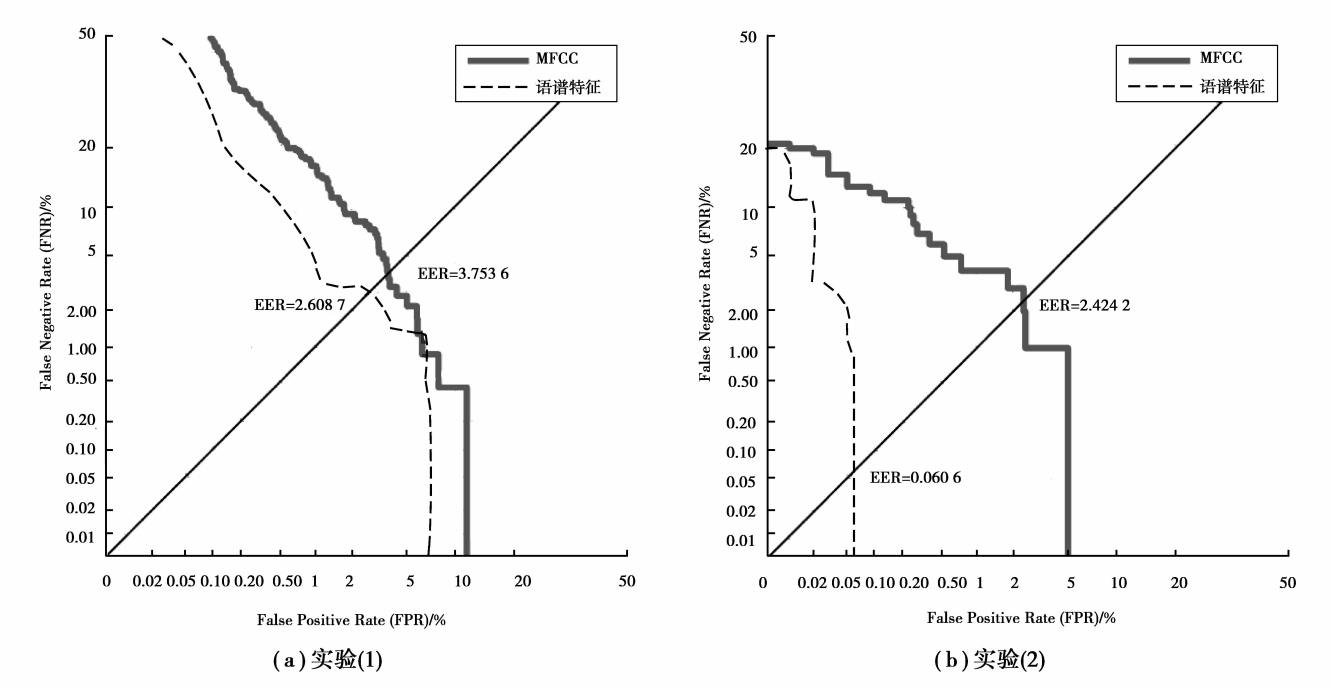


图3 说话人1由不同特征得到的i-vector向量

Fig.3 I-vectors of Speaker 1 based on MFCC and Spectrogram

本实验采用 NIST 等错误率和 DETDET(detection error trade-offs)曲线作为评测指标[15]。图 4 给出了基于不同特征下 i-vector 系统的 DET 曲线以及等错误率 EER。DET 曲线表示(false negative rate，FNR)与(positive rate，FPR)之间的关系。



(a)实验(1)　　(b)实验(2)

图 4　DET 曲线以及 EER 值

Fig.4　DET curves and EER

其中，加粗实线表示以 MFCC 为特征的 i-vector 说话人识别系统的 DET 曲线，加粗虚线表示以语谱图为特征的 i-vector 说话人识别系统的 DET 曲线，它们与斜线的交点即为对应的 EER 值，此时 FNR＝FPR。从图中可以看出，实验(1)中，以 MFCC 为特征的系统等错误率为 3.753 6%，以语谱图为特征的系统等错误率为 2.608 7%；实验(2)中，以 MFCC 为特征的系统等错误率为 2.424 2%，以语谱图为特征的系统等错误率为 0.060 6%。

实验结果表明，以语谱图为特征可以有效 i-vector 识别系统的。

3　结　论

探索性地研究了利用语谱特征作为声学特征的 i-vector 识别方法，并与基于 MFCC 作为特征的识别性能进行比较。以 TIMIT 标准语料库为训练和测试数据库的实验结果表明，语谱特征相对于 MFCC，能够更好地表征说话人的信息，基于语谱特征的 i-vector 说话人识别方法的性能比以 MFCC 作为特征的方法等错误率更低。

参考文献：

[1] Misra S, Das T K, Saha P, et al. Comparison of MFCC and LPCC for a fixed phrase speaker verification system, time complexity and failure analysis[C]// Circuit, Power and Computing Technologies (ICCPCT), 2015 International Conference on. [S.L.]: IEEE, 2015: 1-4.

[2] Ajmera P K, Jadhav D V, Holambe R S. Text-independent speaker identification using Radon and discrete cosine transforms based features from speech spectrogram[J]. Pattern Recognition, 2011, 44(10): 2749-2759.

[3] 刘红星，戴蓓蒨，陆伟.基于共振峰谐波能量的语音端点检测[J].清华大学学报：自然科学版，2008(S1)：24-30.

LIU Hongxing, Dai Peiqian, LU Wei. Speech endpoint detection based on the formant consonance energy[J]. Journal of Tsinghua Unversity, 2008(S1): 24-30. (in Chinese)

[4] Chen M G, Sui B C, Gao Y, et al. Efficient video cutout based on adaptive multilevel banded method[J]. Science China Information Sciences, 2012, 55(5): 1082-1092.

[5] Yuan D, Liang L U, Xian Y Z, et al. Studies on model distance normalization approach in text-independent speaker verification[J]. Acta Automatica Sinica, 2009, 35(5): 556-560.

[6] Verma P, Das P K. i-Vectors in speech processing applications: a survey[J]. International Journal of Speech Technology, 2015, 18(4): 529-546.

[7] Aronowitz H, Barkan O. New developments in joint factor analysis for speaker verification[C]// Twelfth Annual Conference of the International Speech Communication Association.[S.L.]. IEEE, 2011.

[8] Dehak N, Karam Z N, Reynolds D A, et al. A channel-blind system for speaker verification[C]// Acoustics, Speech and Signal Processing (ICASSP). [S.L.]: IEEE, 2011: 4536-4539.

[9] Li M, Narayanan S. Simplified supervised i-vector modeling with application to robust and efficient language identification and speaker verification[J]. Computer Speech & Language, 2014, 28(4): 940-958.

[10] Dehak N. Discriminative and generative approaches for long-and short-term speaker characteristics modeling: application to speaker verification[M]. Canada: Ecole de Technologie Superieure, 2009.

[11] Cadore J, Valverde-Albacete F J, Gallardo-Antolín A, et al. Auditory-inspired morphological processing of speech spectrograms: Applications in automatic speech recognition and speech enhancement[J]. Cognitive Computation, 2013, 5(4): 426-441.

[12] Lu W, Zhang Q. Deconvolutive short-time Fourier transform spectrogram[J]. Signal Processing Letters, IEEE, 2009, 16(7): 576-579.

[13] 栗志意,张卫强,何亮,等.基于总体变化子空间自适应的 i-vector 说话人识别系统研究[J].自动化学报,2014,40(8): 1836-1840.
LI Zhiyi, ZHANG Weiqiang, He Liang, et al. Total variability subspace adaptation based speaker recognition[J]. Journal of Automatica Sinca, 2014, 40(8): 1836-1840. (in Chinese)

[14] Li W, Fu T, Zhu J. An improved i-vector extraction algorithm for speaker verification[J]. Eurasip Journal on Audio, Speech and Music Processing, 2015, 2015(1): 1-9.

[15] Hu Y, Loizou P C. Evaluation of objective quality measures for speech enhancement[J]. Audio, Speech and Language Processing, IEEE Transactions on, 2008, 16(1): 229-238.

(编辑 侯 湘)