文章编号:1004-9037(2013)02-0239-05

应用似然比框架的法庭说话人识别

王华朋[1,2] 杨 军[1] 许 勇[1]

(1. 中国科学院噪声与振动重点实验室(声学研究所),北京,100190;
2. 中国刑警学院声像资料检验技术系,沈阳,110854)

摘要:为了检验元音倒谱特征在法庭说话人识别中的性能,提出了使用元音稳定段美尔倒谱系数(Mel-frequency cepstral coefficients, MFCC)作为识别特征的基于似然比的法庭说话人识别方法,并使用 45 人电话对话录音中元音/a/作为样本进行了测试。实验结果表明,该方法不仅能正确识别说话人,而且能根据当前嫌疑人样本和问题语音样本的差异,量化该语音样本作为证据的力度,为法庭提供科学合理的证据评估结果。与人工提取共振峰特征相比,自动特征提取的引入提高了工作效率,使识别系统的性能获得了大幅提升。

关键词:MFCC;似然比;法庭说话人识别;证据力度

中图分类号:TP391.42 **文献标志码**:A

Forensic Speaker Recognition in Likelihood Ratio Framework

Wang Huapeng[1,2], *Yang Jun*[1], *Xu Yong*[1]

(1. Key Laboratory of Noise and Vibration Research, Institute of Acoustics, Chinese Academy of Sciences, Beijing, 100190, China; 2. Audio and Video Materials Inspection Technology Department, China Criminal Police University, Shenyang, 110854, China)

Abstract: To test the performance of vowel cepstrum in forensic speaker recognition, a forensic speaker identification method based on likelihood ratio and Mel-frequency cepstral coefficients (MFCC) features is presented. The method is tested in vowel /a/ of 45 peoples′ telephone dialog recordings and shows high identification ratio. Experimental results show that the method can identify the speaker, and quantify the evidence strength according to the acoustic difference between the questioned recording and the suspect′s recording, thus providing the scientific and reasonable evaluation results to court. Compared with the manual method in formants extraction and pitch extraction, the auto extraction of features increases the efficiency and performance of forensic speaker identification system.

Key words: MFCC; likelihood ratio; forensic speaker identification; evidence strength

引 言

法庭说话人识别,其最主要的任务就是比对犯罪现场或犯罪过程中获得的罪犯的语音样本和嫌疑人的语音样本,提取足够的稳定语言特征或者说话人个体相关的语音特征,利用这些语音特征加以识别或确认。目前,在国内,绝大多数的法庭说话人识别案件中,都希望语音鉴定专家给出“是同一人”或“不是同一人”这样明确的结论,法官也都习惯于使用类似的证据。但是,由于受各种主客观条件的限制,如:录音的环境及条件,语音证据提取、保存条件与方法,检验鉴定的时间间隔以及检验设备、检验方法的局限等等,罪犯样本和嫌疑人样本之间或多或少都会存在一定程度的差异,这就决定了鉴定人认定同一或否定排除要达到 100% 的确认几乎是不可能的。在 DNA、指纹、声纹、笔迹、足迹等法庭证据的同一认定上都出现过错误。出现这些问题的原因主要是对样本之间的辨证关系认识不足,对证据力度的评估缺乏科学有效的方

基金项目:国家自然科学基金(11074279)资助项目。

收稿日期:2011-11-07;**修订日期**:2012-03-29

法。随着语音证据在法庭上使用的次数越来越多，国际上对法庭语音证据评估方法有了新的发展，对证据也有了全新的认识。DNA 就率先采取了新的证据评估方法，即基于似然比的证据评估方法，将其引入其他的法庭证据领域，可以评估证据对鉴定结论支持力度的大小，该方法在国内外获得了广泛的认同[1-2]。基于似然比的证据评估方法是逻辑上和法律上都正确的法庭证据评估方法，也是向法庭提供证据强度评估的科学方法[3-4]。但是，利用似然比的方法来研究语音证据，目前还处于初始阶段，本文就尝试提取语音的美尔倒谱系数(Mel-frequency cepstral coefficients, MFCC)作为特征参数，利用似然比进行说话人识别。

1 似然比计算

目前，在国际法庭说话人识别研究中，似然比是最重要的组成部分，因为它可以量化证据对鉴定结论支持的力度。似然比可以表示成在一个给定的假设条件下观测到犯罪证据(罪犯和嫌疑人样本间的声学差异)的概率和在完全相反的假设条件下观测到犯罪证据概率的比值，例如，似然比可以表示成，在罪犯和嫌疑人样本为同一人语音的假设条件下和非同一人语音的假设条件下，观测到罪犯和嫌疑人语音样本之间声学差异(证据)的概率的比值[5]。似然比的分子，用来估计在罪犯样本和嫌疑人样本来自同一人的假设条件下，获得当前样本间匹配程度的概率；似然比的分母，用来估计在罪犯样本和嫌疑人样本来自不同人的假设条件下，获得当前样本间匹配程度的概率。因此，它们的比率就是当前语音证据支持同一人的假设和支持不是同一人假设的相对强度，强度的大小反映在似然比的幅度上。似然比的值和 1 之间的相对距离，反映了证据强度的大小。似然比的值和 1 之间的差值越大，说明证据对结论的支持力度越大；似然比的值越是逼近 1，说明当前的证据有效性越低，因为它既不能为是同一人的假设提供强力的支持，也不能为不是同一人的假设提供强力的支持[6]，因此当前语音作为证据来讲，作用是很小的，其较小的证据强度不能帮助法官做出判断。如果似然比的值等于 1，那说明该证据对两个相反的假设支持的力度是一样的，故不具有证据意义。似然比和 1 的大小关系表明，当前的语音证据支持是同一人的假设还是非同一人的假设，似然比的值并不是真相的二值表示。也就是说，对于嫌疑人样本和罪犯样本是不是由同一人产生的这一问题，似然比并没有给出“是”或“否”的回答，它只是量化了当前语音证据对鉴定结论支持的强度。如果用 P 来表示概率，E 表示证据，H 代表假设，那么似然比可写成下面的形式

$$\mathrm{LR}=\frac{P(E\mid H)}{P(E\mid \overline{H})}$$

在法庭说话人识别中，似然比的分子量化了罪犯样本和嫌疑人样本之间相似的程度，其分母量化了罪犯样本和嫌疑人样本在参考人群里的典型性。如果罪犯样本和嫌疑人样本越相似，它们来自同一人的可能性就越大，似然比的值也就会越大。然而，这个结果还需要样本的典型性来平衡。这两个样本越是典型，它们就越可能是从人群中随机抽取的，似然比的值就会越低。因此，似然比的值是样本的相似性和典型性相互作用的结果，贝叶斯理论明确指明，相似性和典型性对证据评估来说都是必不可少的。事实上，在实际工作中，经常会忽视样本特征的典型性，认为仅仅相似性对证据同一认定就足够了，这是不正确的做法。比如，在比较两个语音样本时，显然对它们之间的相似性很感兴趣，但是，仅仅靠相似性来评价证据是不够的，样本特征的典型性也应该被考虑在内[7]。

在非自动的法庭说话人识别中，因为语音特征经常是多维的，在理论上，可以先计算出每一个语音特征的似然比，然后把这些似然比组合成一个全局的似然比。似然比可以进行非常简单的组合，这也是贝叶斯方法的显著优点之一。如果特征之间是相互独立的，全局似然比就是单个特征似然比的乘积。这个忽略了变量间相关性的方法被称为朴素贝叶斯方法。

如果提取的特征为单个特征，即使用单变量计算似然比，则可采用 Lindley 提出的公式，见文献[8]。如果特征为多变量，则可采用 Aitken 和 Lucy[9]提出的多变量核密度的似然比计算方法。本文因使用的特征变量为线性预测系数，为多变量特征，故采用多变量核密度的方法来计算似然比。

2 美尔倒谱特征

数字化的语音信号是声道频率特性和激励信号源两者的共同结果，后者对于某帧信号而言常带有一定的随机性[10]。说话人的个性特征很大程度上体现在说话人的发音声道变化上，即声道频率特

性。

有必要采用一定的方法将这两者有效地分开，这种方法就是同态滤波。滤波的过程是先将卷积处理化为乘积，然后作对数处理，使之成为可分离的相加成分，结果就形成了倒谱 $c(n)=\hat{h}(n)+\hat{i}(n)$。因为 $\hat{h}(n)$ 描述了说话人的声道分量，所以是非常有效的说话人个性特征参数。

将一帧中的语音信号 $s(n)=h(n)*i(n)$（$*$ 表示卷积）处理为其倒谱 $c(n)$ 的过程，如图 1 所示。可以先用离散傅里叶变换（Discrete Fourier transform，DFT）计算 $s(n)$ 的短时傅里叶变换，变换的结果使在 B 点得到了声道冲激响应和音源激励的傅里叶变换的乘积。再取这一乘积的幅度的对数，在 C 点就得到了声道冲激响应和音源激励的傅里叶变换的对数之和。最后对其进行逆傅里叶变换，得到的信号称为 $s(n)$ 的倒谱 $c(n)$，也称为倒谱系数，它是声道分量的倒谱 $\hat{h}(n)$ 和音源激励分量的倒谱 $\hat{i}(n)$ 之和。

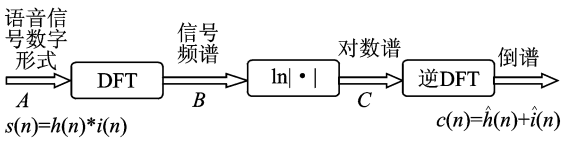

图 1　倒谱的计算流程图

如果不是直接对语音信号的对数谱作逆 DFT，而是先经过一定频率坐标的尺度弯折 $T\omega(\cdot)$，频率坐标在 1 000 Hz 以下的采用线性的频率弯折，频率坐标在 1 000 Hz 以上的采用对数的频率弯折；然后再进行逆 DFT，这样得到的特征称为美尔倒谱系数。

3　实验结果分析

为了验证本文结果的稳定性，本文采用两个 45 人电话数据库，是由相同的 45 人在不同时间段录制的电话对话录音，该录音从磁带采集进入电脑之后，以 16 位的 PCM 格式声音文件保存，采样频率为 11 025 Hz。对话中采用的是汉语普通话，参与者年龄都在 19～23 岁之间。由一位引导人询问他们相关的个人信息，他们依次作答。因电话的日益普及，越来越多的案件中会出现电话号码或者银行卡号码，因此，录音语料中多次提及了电话号码，本文选取了中国人都喜欢用的幸运数字“8”作为分析的对象，提取其中的元音/a/进行分析。选取“8”作为分析对象还有另外一个好处，因为人们经常对该数字进行重读，因此其中的元音/a/发音比较饱满，能较好地反映个体的声道特征。剑桥大学 Jong[11] 的研究结果也表明，/iː，ɑː，ɔː/和其他的单元音相比，具有更好的稳定性。

本文对数据库中的元音/a/进行手工标注，标注出该元音稳定段，然后用 Hamming 窗进行加窗，窗的数据长度为 256 个采样点。对同一个人的语音，前一半元音/a/的 MFCC 特征和后一半元音/a/的 MFCC 特征进行比较；对于不同的说话人，所有该对话中标注的元音特征都用来和其他每一个人的所有标注元音特征进行比较，是属于交叉比较验证，可充分验证该方法的性能。因此，每一个数据库共有 45 次同一说话人比较，$45\times(45-1)/2=990$ 次不同说话人比较。

在似然比的讨论中，似然比经常用以 10 为底的对数值表示，因为在对数域，越大的正数能为是同一人的假设提供越大的支持力度，越大的负数对不同人的假设提供越大的支持力度。例如，对数似然比＋1 表示在当前的语音证据条件下，它们来自同一人的概率是来自不同人概率的 10 倍；对数似然比－1 表示，在当前的语音证据条件下，它们来自不同人的概率是来自同一人概率的 10 倍，以此类推。图 2 是 14 阶 MFCC 作为识别特征时的 Tippett 图[12]，左上较粗的曲线表示不同说话人的对数 10 似然比大于等于 x 轴刻度的样本所占的比率；右上较细的曲线表示同一说话人对数 10 似然比小于等于 x 轴刻度的样本所占的比率。图中的竖线为识别阈值，似然比的识别阈值为 1，取对数后为 0，最理想的情况是，左上方的粗线和右上方的细线与表示阈值的竖线都没有交点，同一说话人和不同说话人都达到 100％的识别率。表 1 为基于 MFCC 特征的不同说话人似然比分布表，表 2 为基于 MFCC 特征的相同说话人似然比分布表。表中的第一行为 MFCC 的阶次，本文计算了从 20 阶到 10 阶取不同阶次的部分识别结果情况。综合以上同源识别和非同源识别的结果，MFCC 取 12～16 阶时，识别效果达到最优，考虑系数稳定性，可取 14 阶为最优阶。表左侧部分是似然比取值大小的分布区间，越往下对相应的假设支持力度越大，表中的内容为所有似然比的值在相应区间内分布的个数，最后一行为错误否定或者错误认定的错误率。从表 1 可以看出，绝大多数的不同说话人

数据对是不同说话人这一假设的支持力度都在 $1/10^{-5}=100\ 000$ 以上，即当前的罪犯样本和嫌疑人样本之间的声学差异，对不同说话人假设的支持力度是对相同说话人假设支持力度的 100 000 倍，这无疑是非常强的证据力度。在 990 个不同说话人数据比对中，错误率仅有 0.007 1。表 2 是相同说话人数据似然比结果分布情况，MFCC 取 14 阶时，错误率仅有 0.111 1，也具有很高的识别性能。从似然比的分布情况也可以看出，同一说话人的似然比值和表 1 的分布不同，主要分布在力度较大的区域，且分布相对比较均匀，这说明同一说话人说相同语料时也存在着差异，但这种差异要比不同说话人之间的差异小。因此，该方法能对不同说话人提供强力的证据，对相同说话人认定的证据力度相对较小，这在一定程度上可避免冤假错案，保护公民权益。

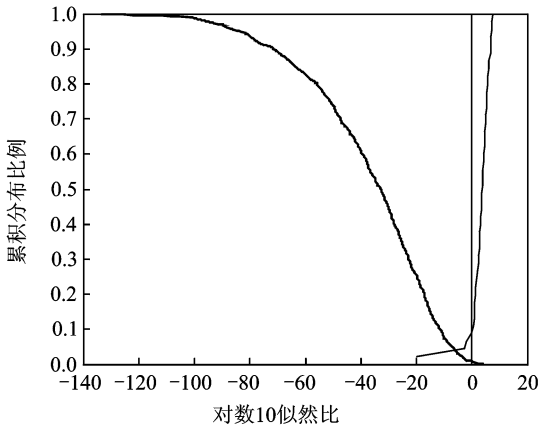

图 2 基于 14 阶 MFCC 的 Tippett 图

表 1 基于 MFCC 的不同说话人似然比分布表

似然比取值区间	20 阶	18 阶	16 阶	14 阶	12 阶	10 阶
>1	7	6	9	7	10	12
$10^{-1}\sim1$	3	5	5	5	7	10
$10^{-2}\sim10^{-1}$	5	5	2	6	6	14
$10^{-3}\sim10^{-2}$	6	6	6	7	15	12
$10^{-4}\sim10^{-3}$	5	1	5	6	10	23
$10^{-5}\sim10^{-4}$	4	6	7	10	8	14
$<10^{-5}$	960	961	956	949	934	905
错误率	0.007 1	0.006 1	0.009 1	0.007 1	0.010 1	0.012 1

表 2 基于 MFCC 的相同说话人似然比分布表

似然比取值区间	20 阶	18 阶	16 阶	14 阶	12 阶	10 阶
<1	5	4	3	5	5	6
$1\sim10$	1	4	4	1	5	7
$10\sim10^{2}$	4	2	4	7	6	4
$10^{2}\sim10^{3}$	5	3	6	7	6	7
$10^{3}\sim10^{4}$	5	9	6	5	9	10
$10^{4}\sim10^{5}$	11	7	5	7	5	6
$>10^{5}$	14	16	17	13	9	5
错误率	0.111 1	0.088 9	0.066 7	0.111 1	0.111 1	0.133 3

4 结束语

通过对两个 45 人不同时间段录制的数据库中元音/a/的测试，在基于似然比的证据强度评估框架下，使用 MFCC 特征作为识别参数，取得了很高的正确识别率，比起传统人工提取共振峰特征的方法，大大提高了工作效率，减少了人工的参与。经过对不同阶次 MFCC 向量的识别性能的计算，综合识别性能和参数稳定性，本文推荐采用 14 阶的 MFCC 作为识别特征向量。结果表明，似然比方法是法庭说话人识别的一个科学有效的方法，能大大提高识别的准确率，并且量化了证据的强度。本文仅使用了每个说话人元音/a/的数据，还可以使用更多的元音和更多的特征进行特征融合，获得一个全局证据力度，进一步提高识别结果的可靠度，这也是下一步工作的研究方向。

参考文献：

[1] Morrison G S, Thiruvaran T. Estimating the precision of the likelihood-ratio output of a forensic-voice-comparison system [C]//Proceedings of Odyssey 2010: The Language and Speaker Recognition Workshop. Brno, Czech Republic: [s. n.], 2010.

[2] Morrison G S. Forensic voice comparison. Expert evidence[M]. Sydney, Australia: Thomson Reuters, 2010:99.

[3] Bonastre J F. Mistral: Open source platform for biometrics authentification, version 1.3[EB/OL]. http://mistral.univ-avignon.fr/. [2013-03-07].

[4] Morrison G S. Forensic voice comparison and the paradigm shift [J]. Science & Justice, 2009,49(4): 298-308.

[5] Morrison G S. Measuring the validity and reliability of forensic likelihood-ratio systems[J]. Science & Justice, 2011,51(3):91-98.

[6] Kinoshita Y, Osanai T. Within speaker variation in diphthongal dynamics: What can we compare? [C]//Proceedings of the 11th Australasian International Conference on Speech Science & Technology.

New Zealand, Australia: Australasian Speech Science & Technology Association, 2006:112-117.

[7] Rose P. Technical forensic speaker recognition: Evaluation, types and testing of evidence[J]. Computer Speech & Language, 2006,20(2):159-191.

[8] Rose P. Forensic Speaker Identification[M]. UK: Taylor & Francis,2002.

[9] Aitken C G G, Lucy D. Evaluation of trace evidence in the form of multivariate data[J]. Journal of the Royal Statistical Society: Series C (Applied Statistics), 2004,53(1):109-122.

[10] 何继爱,达正花,唐艳娟.基于 AR 模型的盲源分离方法[J].数据采集与处理,2011,26(2):162-166.
He Jiai, Da Zhenghua, Tang Yanjuan. Blind separation based on AR model [J]. Journal of Data Acquisition & Processing, 2011,26(2):162-166.

[11] Jong De G, McDougall K, Hudson T, et al. The speaker discriminating power of sounds undergoing historical change: A formant-based study[C]//Proceedings of ICPhS Saarbrücken. Germany: [s. n.], 2007:1813-1816.

[12] Rose P. Accounting for correlation in linguistic-acoustic likelihood ratio-based forensic speaker discrimination [C]//IEEE Odyssey-The Speaker and Language Recognition Workshop. San Juan: IEEE, 2006.

作者简介:王华朋(1979-),男,博士研究生,研究方向:语音信号处理、法庭科学,E-mail:huapeng. wang@gmail. com;杨军(1968-),男,研究员,研究方向:声学、信号处理;许勇(1973-),男,副研究员,研究方向:信号处理。