DOI: 10.13245/j.hust.150226

融合 MFCC 和 LPCC 的语音感知哈希算法

黄羿博[a] 张秋余[a,b] 袁占亭[a] 杨仲平[b]

(兰州理工大学 a 电气工程与信息工程学院; b 计算机与通信学院, 甘肃 兰州, 730050)

摘要 为了提高语音感知哈希算法的鲁棒性和识别小范围篡改定位的能力,利用人类听觉模型提出了一种语音感知哈希算法.该算法基于人类听觉特性,首先对倒谱系数 MFCC 算法每帧的滤波器数量进行控制,得到每帧语音的梅尔频率倒谱参数;其次对自适应梅尔倒谱系数 MFCC 参数和语音 LPCC 系数进行融合,并采用分块方法对特征矩阵进行处理,对特征块进行 2DNMF 分解运算,降低特征矩阵的复杂度;最后对分解后的系数矩阵进行哈希构造,得到语音感知哈希串,利用哈希匹配实现语音认证.结果表明:该算法可以有效提高哈希认证的鲁棒性,并能够实现语音小范围篡改定位功能.

关键词 语音识别; 信息安全技术; 语音感知哈希; 自适应倒谱系数; 篡改定位

中图分类号 TP309.2 **文献标志码** A **文章编号** 1671-4512(2015)02-0124-05

The hash algorithm of speech perception based on the integration of adaptive MFCC and LPCC

Huang Yibo[a] *Zhang Qiuyu*[a,b] *Yuan Zhanting*[a] *Yang Zhongping*[b]

(a College of Electrical and Information Engineering; b School of Computer and Communication, Lanzhou University of Technology, Lanzhou 730050, China)

Abstract A new kind of hash algorithm of speech perception was proposed with the help of human auditory model in order to improve the robustness of speech perception hash algorithm and the ability to identify tampering with a small scale positioning. Firstly, the method controlled the number of filters of the MFCC (Mel frequency cepstral coefficients) algorithm based on human auditory characteristics, and obtained the Mel frequency cepstral parameters of each frame of speech. Secondly, it fused the AMFCC (adaptive Mel frequency cepstral coefficients) parameter and the LPCC (linear prediction cepstrum coefficients), and dealed with characteristic matrix using blocking. At the same time, it processed 2DNMF (two-dimensional nonnegative matrix factorization) decomposition algorithm to characteristic blocks, so the complexity of the characteristic matrix was reduced. Finally, coefficient matrixes after decomposition were hashed to get the hash string of speech perception. The voice authentication could be realized by hash matching. The results show that the algorithm is able to enhance the robustness of hash authentication and achieve tamper localization of small range.

Key words speech recognition; information security technology; perceptual speech hash; adaptive cepstral coefficients; tamper localization

语音内容认证技术[1]是实现对语音数据完整性、真实性保护的有效技术手段,目前常见的语音认证方法包括语音水印、语音身份签名、语音声纹[2]等.语音感知哈希技术是一种新的语音识别

收稿日期 2014-06-07.

作者简介 黄羿博(1982-),男,博士研究生,E-mail: huangyibo1982@163.com.

基金项目 国家自然科学基金资助项目(61363078); 甘肃省自然科学基金资助项目(1212RJZA006, 1310RJYA004).

技术，通过语音特征的哈希映射，实现语音的内容认证、说话人身份认证功能[3-4]. 语音感知哈希提取系统主要包括语音的特征提取和哈希构造两个部分.

语音特征提取直接影响认证的效果和效率. 提取方法包括时域特征和频域特征，如共振峰频率、线性预测系数(LPC)、线谱对(LSP)、线性预测倒谱系数(LPCC)、基于人耳听觉特性的梅尔倒谱系数(MFCC)[5]. 哈希构造分为特征值摘要和哈希重构. 特征值摘要能够有效降低数据规模，减少时间开销. 常见的方法有主成分分析(PCA)、独立成分分析(ICA)、奇异值分解(SVD)、矢量量化(VQ)和非负矩阵分解(NMF)[6]等.

本研究结合人类语音特性，提出了自适应梅尔倒谱系数(AMFCC)，并融合 AMFCC 和 LPCC 两种特征参数作为语音特征. 两种参数体现了不同的语音特性，MFCC 利用了人耳听觉的非线性特性，更具有鲁棒性；LPCC 体现了人类声道个体差异的特性. 并对特征矩阵进行分块，结合二维非负矩阵分解对特征矩阵块进行映射. 当认证时，使用动态时间规整(DTW)算法，实现篡改定位功能. 实验结果表明：本文算法具有很好的认证性能，并有篡改定位的功能.

1　语音特征值的提取

1.1　自适应 MFCC 的语音特征提取算法

当实际提取语音特征时，多采用 MFCC 作为特征矢量来使用[7-8]，MFCC 是从对频率高低的非线性心理感觉角度反映了语音短时幅度谱的特征，特征表征更为准确，抗噪性能也更强.

语音信号是通过梅尔标度频率域提取出的信号倒谱参数得到梅尔倒谱系数的. 梅尔标度描述了人耳对频率感知的非线性特性，它与实际频率之间的关系可近似表示为

$$\mathrm{Mel}(f) = 2\ 595\log(1+f/700)\quad(0\leqslant f\leqslant F_{\mathrm{n}}),\tag{1}$$

式中：f 为语音实际频率；F_{n} 为输入语音信号的 Nyquist 频率，$F_{\mathrm{n}}=f_{\max}/2$，$f_{\max}$ 为语音信号的采样频率.

在 MFCC 算法中，三角滤波器组包含的梅尔滤波器的个数 N 和组内各滤波器的中心频率是固定不变的. 这种设计方法没有充分考虑不同的说话人说出的不同语义的语音特征，不能有效地区别不同说话人的 MFCC 参数. 为了提高语音特征提取的准确性和鲁棒性，考虑将 MFCC 的三角滤波器组构成一个动态数量控制的滤波器组，这个滤波器组可以根据说话人发出声音频率的大小，得到不同数量的滤波器组.

基音频率 f_{p} 可以作为语音特征来使用，表征了说话人发出浊音时声带震动产生的周期，描述了语音中频率最低的震动，可以很好地描述说话人各自的声带特征. 通常基音周期 T_{p} 估计采用语音的自相关函数，

$$R_{\mathrm{w}}(k)=\sum_{m=0}^{N-a-1}S_{\mathrm{w}}(m)S_{\mathrm{w}}(m+a).\tag{2}$$

式(2)表示经过分帧加窗的语音信号 $S_{\mathrm{w}}(m)$ 和延迟 a 点后语音的相似性，N 为帧长. 可以得到 $R_{\mathrm{w}}(k)\approx T_{\mathrm{p}}$，基音周期 T_{p} 为基音频率的倒数，即 $f_{\mathrm{p}}=1/T_{\mathrm{p}}$. 若用基音周期内的采样点数表征基音周期，则 T_{p} 取决于基音频率 f_{p} 和语音采样频率 f_{r}，即 $T_{\mathrm{p}}=f_{\mathrm{r}}/f_{\mathrm{p}}$.

根据上述方法得到每帧语音信号的基音频率 f_{p}，$\mathrm{Mel}(f_{\mathrm{p}})$ 和 $\mathrm{Mel}(F_{\mathrm{n}})$，则自适应梅尔滤波器组的滤波器个数为 $\mathrm{floor}(x)=\mathrm{Mel}(F_{\mathrm{n}})/\mathrm{Mel}(f_{\mathrm{p}})$. 得到的滤波器个数将梅尔标度频率进行等分并以各等分在实际频率域中对应点的频率作为滤波器的中心频率，设计的梅尔滤波器组为 $\{H_i(k), i=1,2,\cdots,N\}$.

1.2　线性预测倒谱系数

本研究采用全极点模型对线性预测系数进行递推，构成线性预测倒谱系数，即

$$\begin{cases} c_1 = a_1; \\ c_n = a_n + \sum_{k=1}^{n-1} c_n a_{n-k}\,\dfrac{k}{n} \quad (1<n\leqslant p); \\ c_n = \sum_{k=1}^{n-1} c_n a_{n-k}\,\dfrac{k}{n} \quad (n>p), \end{cases}\tag{3}$$

式中 a_p 为 p 阶线性预测系数.

2　二维非负矩阵分解

非负矩阵分解是多元统计分析方法[9]，是基于“部分”的分解方法，符合人类大脑对事物的感知过程[10]. NMF 可以对高维数据矩阵进行降维处理，具有收敛速度快，存储空间小的特点[11]. 文献[12]提出的 2DNMF 最早使用在图像处理领域. 本文采用 2DNMF 分解语音的 AMFCC 和 LPCC 系数特征矩阵.

将非负矩阵 $\boldsymbol{X}=[A_1, A_2, \cdots, A_m]_{p\times q_m}$ 进行非负矩阵分解，找出非负矩阵 $\boldsymbol{L}_{p\times d}$ 和非负矩阵 $\boldsymbol{H}_{d\times q_m}$，即

$$\boldsymbol{X}_{p\times q_m}\approx \boldsymbol{L}_{p\times d}\boldsymbol{H}_{d\times q_m},\tag{4}$$

式中 d 为分解的阶数. 对 $\boldsymbol{H}_{d\times q_m}$ 进行转置后,继续进行非负矩阵分解,使其满足

$$\boldsymbol{H}'_{q\times d_m} \approx \boldsymbol{R}_{q\times g}\boldsymbol{C}_{g\times d_m}, \tag{5}$$

式中:g 为分解的阶数;$\boldsymbol{R}_{q\times g}=[R_1,R_2,\cdots,R_g]$为列基矩阵;$\boldsymbol{C}_{g\times d_m}=[C_1,C_2,\cdots,C_m]$为系数矩阵.

3 动态时间规整算法及其实现

3.1 规整算法

动态时间规整算法(DTW)可以实现语音特征序列的最佳匹配. 本文使用 DTW 算法评价两段语音的距离. 不同的识别系统会有不同的结果[13].

AMFCC 和 LPCC 特征参数的距离融合为

$$d(i,j)=w_m\Big(\sum_{i,j=1}^{i=I+1,j=J+1}|\boldsymbol{t}_m(:,i)-\boldsymbol{r}_m(:,j)|^2\Big)+w_1\Big(\sum_{i,j=1}^{i=I+1,j=J+1}|\boldsymbol{t}_1(:,i)-\boldsymbol{r}_1(:,j)|^2\Big), \tag{6}$$

式中:w_m 和 w_1 为 AMFCC 和 LPCC 的权重系数,$w_m+w_1=1$;$\boldsymbol{t}_m$ 和 $\boldsymbol{r}_m$ 为 AMFCC 的特征矩阵;$\boldsymbol{t}_1$ 和 $\boldsymbol{r}_1$ 为 LPCC 的特征矩阵;$d(i,j)$为两段语音的距离;$(:,i)$表示只对列进行控制.

3.2 算法的实现

结合提出的算法,对语音特征系数进行分块处理,解决了对语音特征系数进行分解后无法进行篡改定位的问题. 图 1 是算法结构图,图中:$x(t)$为原始语音信息;$y(t)$为待匹配的语音信息.

a. 分块. 得到特征系数矩阵后,不会直接对系数矩阵进行 2DNMF,而是先将系数矩阵进行分块,将系数矩阵分为 m 个小矩阵,即

$$\boldsymbol{C}_{g\times d_m}=[\boldsymbol{L}_1,\boldsymbol{L}_2,\cdots,\boldsymbol{L}_{\mathrm{m}}]. \tag{7}$$

b. 量化. 计算小矩阵 $\boldsymbol{L}$ 每列的元素之和,即

$$s(i)=\sum_{j=1}^{r}H_{ij}\quad(1\leqslant i\leqslant k), \tag{8}$$

式中 H_{ij} 为第 i 行第 j 列特征系数. 对形成的系数和行矩阵进行量化,形成语音段的哈希值,即

$$h(i)=\begin{cases}1 & (s(i)>\hat{s},\ 1\leqslant i\leqslant k);\\ 0 & (\text{其他}),\end{cases} \tag{9}$$

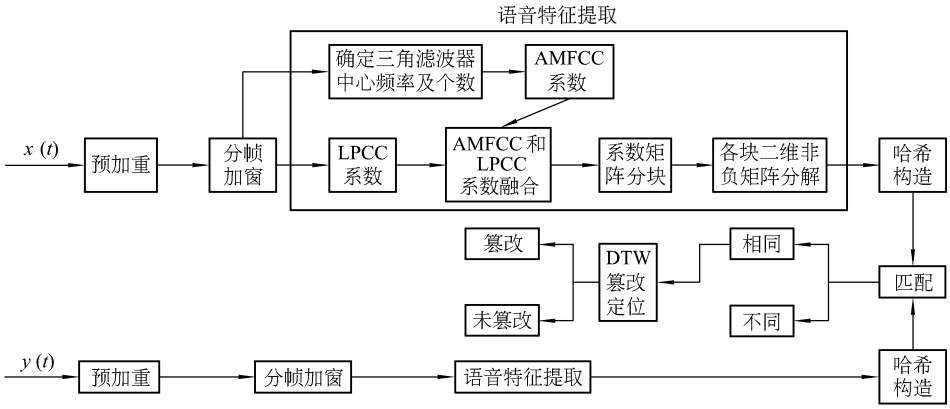

图 1 本文算法结构图

式中 $\hat{s}$ 为中值.

4 仿真实验与分析

实验采用 Matlab 2010b 仿真实现. 使用 TIMIT 语音库和录制的语音共 1 189 段,时长 4 s. 采样率为 16 kHz,比特率为 256 Kbit/s,采样精度为 16 bit,文件格式为 wav. 帧长为 20 ms,帧移 10 ms.

4.1 区分性

区分性[14]用于评价算法针对区分不同或者相同的人读出不同语音内容的可靠性. 实验结果表明:不同语音的误码率值的概率分布与标准正态分布的概率曲线几乎重叠,故本文算法所得哈希距离值近似服从正态分布,即感知不同的语音生成不同的哈希值.

从表 1 可以看出:误识率随阈值 τ 的增大而增加. 当 $\tau=0.20$ 时,碰撞概率为 1×10^{16} 个语音碰撞 5 个片段;当 $\tau=0.25$ 时,碰撞概率为 1×10^{12} 个语音片段碰撞 10 个片段;当 $\tau=0.30$ 时,碰撞概率为 1×10^{8} 个语音片段碰撞 2 个语音片段. 本文的判定阈值为 $\tau=0.35$ 时,1×10^{5} 个语音片段碰撞 2 段语音,因此本文算法能够准确地识别出认证的语音片段.

4.2 鲁棒性

语音感知哈希鲁棒性用于评价相同语音经受过不同的内容保持操作后的可靠性. 通过表 2 可以看出:本文提出的 AMFCC 算法,针对回声攻击的鲁棒性不如其他三种算法,但对高斯白噪声攻击的鲁棒性得到了提高,本文算法的鲁棒性整

表 1　算法的阈值误识率

τ	LPCC+NMF	MFCC+NMF	文献[11]	本文算法
0.20	$2.956\ 4\times10^{-14}$	$1.178\ 3\times10^{-13}$	$1.731\ 4\times10^{-13}$	$5.255\ 6\times10^{-16}$
0.25	$1.924\ 9\times10^{-10}$	$5.066\ 0\times10^{-10}$	$6.754\ 1\times10^{-10}$	$9.690\ 1\times10^{-12}$
0.30	$2.721\ 5\times10^{-7}$	$5.098\ 3\times10^{-7}$	$6.254\ 8\times10^{-7}$	$1.992\ 1\times10^{-8}$
0.35	$8.495\ 7\times10^{-5}$	$1.220\ 9\times10^{-4}$	$1.398\ 0\times10^{-4}$	$1.992\ 1\times10^{-5}$

表 2　受到攻击后语音认证匹配通过率　%

参数	LPCC+NMF	MFCC+NMF	文献[11]	本文算法
−50%音量	97.4	98.4	96.5	98.8
+50%音量	96.4	97.4	93.1	97.8
回声攻击	88.4	85.3	83.3	78.5
重采样	100	100	100	99.8
+50 dB 噪声	95.6	97.9	94.6	98.5
4 kHz 低通滤波	100	100	100	100
1%篡改攻击	100	100	100	100

体优于其他三种算法.

根据上述攻击得到的误拒率，绘制了误拒率和误识率曲线，结果显示：从受到内容保持操作的内容相同语音中提取的感知哈希值，BER 的分布多数在阈值 0.35 以下，只有极少数超过阈值，与误识率曲线交叉.

以上实验结果表明：本算法具有较高的鲁棒性，同时具有良好的区分性和鲁棒性，可以准确地对经过内容保持操作的语音进行识别.

4.3　篡改定位

通过表 2 中恶意攻击 1%，5%和 10%篡改恶意攻击数据可以看出：目前的语音感知哈希算法都具有较强的鲁棒性，但针对小范围篡改攻击的敏感性还很弱. 本文提出的算法可以实现语音篡改定位.

对语音受到 1%，5%和 10%篡改攻击的定位情况进行了实验研究，结果表明：本文提出的算法能够定位到篡改内容区域，可以实现语音小范围篡改定位.

5　结论

a. 本文语音感知哈希区分性和抗碰撞实验结果表明：在区分性阈值范围内，错误识别语音的最高概率为 $1.992\ 1\times10^{5}$，对比对照的三种算法，抗碰撞性有了很大的提高，能够满足工程实际需要.

b. 鲁棒性实验表明：本文算法相比其他三种算法，鲁棒性有了一定提高，但针对回声攻击鲁棒性要差于其他三种算法，这与回声攻击会影响滤波器个数选择有关. 由于算法针对个别攻击的识别率不是 100%，误识率-误拒率曲线并没有明显的阈值区间，因此在实际过程中，选择合理的阈值就很重要，本文选择 $\tau=0.35$.

c. 目前的算法针对小范围恶意篡改不具有敏感性. 本文算法根据特征矩阵结构特点，解决了经过矩阵分解的特征值不能够篡改定位的问题. 可以实现语音受到 1%篡改攻击后，篡改语音定位功能，篡改定位的精确性，可以由特征矩阵分块大小来控制.

参　考　文　献

[1] Tomar V S, Rose R C. Efficient manifold learning for speech recognition using locality sensitive hashing [J]. Acoustics, Speech and Signal Processing, 2013, 5: 6995-6999.

[2] Pathak M A, Raj B, Rane S D, et al. Privacy-preserving speech processing: cryptographic and string-matching frameworks show promise[J]. Signal Processing Magazine, 2013, 30(2): 62-74.

[3] Chen Ning, Xiao Haidong. Perceptual audio hashing algorithm based on Zernike moment and maximum-likelihood watermark detection [J]. Digital Signal Processing, 2013, 23(4): 1216-1227.

[4] 项世军，杨建权. 基于约束随机分块的 NMF 图像哈希算[J]. 电子与信息学报, 2011, 33(2): 337-341.

[5] 王宏志，徐玉超，李美静. 基于 Mel 频率倒谱参数相似度的语音端点检测算法[J]. 吉林大学学报：工学版, 2012, 42(5): 1331-1335.

[6] Jiao Y, Li Q, Niu X. Compressed domain perceptual hashing for melp coded speech[C]// Proc of 2008 International Conference on Intelligent Information Hiding and Multimedia Signal Processing. Harbin: IEEE, 2008: 410-413.

[7] Afzal H, Sheeraz M, Mark A G. A novel approach for MFCC feature extraction[C]// Proc of Signal Pro-

cessing and Communication Systems, Gold Coast: IEEE, 2010: 1-5.

[8] Deshmukh A A, Pramod K M. Combined LPC and MFCC features based technique for isolated speech recognition [C] // Proc of ISCAS 2013. Beijing: IEEE, 2013: 437-441.

[9] 孙锐,陈军,高隽. 基于显著性检测与 HOG-NMF 特征的快速行人检测方法[J]. 电子与信息学报, 2013, 35(8): 1921-1926

[10] 薛二娟,鲍长春,李汝玮. 基于二维非负矩阵分解的 1 Kb/s WI 语音编码算法[J]. 电子学报, 2010, 7(38): 1574-1579.

[11] 仵博,陈鑫,郑红燕,等. 基于非负矩阵分解更新规则的部分可观察马尔科夫决策过程信念状态空间降维算法[J]. 电子与信息学报, 2013, 12(35): 2901-2907.

[12] Zhang Daoqing, Chen Songcan, Chou Zhihua. Two-dimensional non-negative matrix factorization for face representation and recognition [C] // ICCV05 Workshop on Analysis and Modeling of Face and Gestures. Beijing: Springer, 2005: 350-363.

[13] Aida-Zade K R, Ardil C, Rustamov S S. Investigation of combined use of mFCC and LPC features in speech recognition systems[J]. World Academy of Science, Engineering and Technology, 2008, 7(2): 72-78

[14] Chen Ning, Wan Wanggen. Robust speech hash function[J]. ETRI Journal, 2010, 2(32): 345-347.