文章编号:1004-9037(2014)02-0248-06

基于分数倒谱变换的取证语音拼接特征提取与分析

钟 巍 孔祥维 尤新刚 王 波

(大连理工大学信息安全研究中心,大连,116024)

摘要:针对语音取证中相同采样率的语音拼接识别进行研究,着重分析了拼接对噪声特征的影响,提出了基于分数倒谱变换的拼接帧检测算法和语音拼接联合识别模型。实验结果表明,在分数阶因子 a 为 0.2 时,分数倒谱变换的拼接帧过零率检测算法优于普通倒谱域方差法,在分数阶因子 a 为 1.2 时,分数倒谱变换的拼接帧高频方差检测算法优于普通倒谱域方差法。

关键词:语音取证;分数倒谱变换;语音拼接

中图分类号:TP391 **文献标识码**:A

Splicing Feature Extraction and Analysis Based on Fractional Cepstrum Transform in Voice Forensics

Zhong Wei, *Kong Xiangwei*, *You Xingang*, *Wang Bo*

(Information Security Research Center, Dalian University of Technology, Dalian, 116024, China)

Abstract: Voice splicing recognition of the same sampling rate is conducted, focusing on analyzing the influence of stitching on the noise characteristics. In addition, this paper also presents algorithm for mosaic frame detection based on fractional cepstrum transform(FRCT), proposing a model of voice stitching and joint identification . Experimental results show that the cross-zero ratio performance of the FRCT method is much better than the MFCC in noise moment when fractional factor a is about 0.2. Furthermore, when fractional factor a is about 1.2 , the high frequency variance performance of the FRCT method is also better than the MFCC.

Key words: speech forensics; fractional cepstrum transform; joint speech

引 言

在语音取证中,如何判断语音证据是否为两段语音拼接而成,即对语音证据进行拼接检测是语音取证研究的热点和难点[1-4]。当采用一些语音编辑软件如 Cooledit,DC Live forensics 对语音信号进行增加删除等编辑操作时,一般会对编辑处的间断点会进行平滑处理,对周围样点的频谱影响较小,从频谱上看编辑软件将多段语音很好地拼接在一起,从时域波形上看无法判别,从听觉效果看几乎听不出来。因此,由于编辑而带来的新频谱成分检测将变得很困难。

目前国外最新的研究方法是通过分析蕴含在语音信号中的 50 Hz 交流电频率信号的相位是否连续来作出判断[5-6],对语音信号进行编辑会导致 50 Hz 交流电信号波形的相位不连续[7-8],而国内的录音证据通常为在电池供电情况下进行录音,录音证据中通常不包括交流电频率,无法进行判决[9]。而文献[10-11]针对不同采样频率下的语音信号拼接进行研究。

语音信号中不仅蕴含语义信息而且还包含说话人个性特征[1]。说话人特征可分为说话人短时特征和说话人长时特征,说话人短时特征指的是相邻多帧语音共同蕴含的说话人个性特征,说话人长时特征是指整段语音或多个语音共同的说话人特征。对于语音取证而言,如果能够提取相邻各帧的短时说话人特征,就能够有效地进行语音拼接识别。这一点与语音识别和说话人识别又有所区别,语音识别研究的是语音的内容,说话人识别研究的

基金项目:国家自然科学基金面上项目(60971095)资助项目。

收稿日期:2012-05-21;**修订日期**:2012-09-29

是说话人长期个性特征,而语音拼接中的说话人特征提取既要排除语义特征又要剔除对拼接点不敏感的说话人长期特征,即提取说话人相邻帧共性特征。

篡改语音的拼接处通常发生在语音间歇段,在语音间歇段噪声占主要成分,由于拼接点前后的噪声来自不同的场合或相隔了较长的时间间隙,其方差、自相关特性、频谱特征在拼接处就可能会发生跳变。如果拼接点前后的噪声皆为平稳随机噪声,通过检测上述特征的跳变点就能确定篡改语音的拼接点,但是实际录音环境下的噪声一般为非平稳噪声,其方差等参数呈时变特征,采用上述特征进行检测易造成虚警。因此,必须深入分析拼接对噪声特征的影响和寻找新的噪声拼接特征参数[11-12]。

提取倒谱特征是语音识别、说话人识别中常用的特征参数[13-14],而在分数变换域上[4]提取倒谱特征既可以通过旋转时频角度最大限度地分离信号与噪声,又可以寻找提取与拼接特征关联的特征参数。

1　噪声帧拼接点检测

1.1　噪声特征分析

当拼接发生时,拼接点所在的噪声帧为

$$r(t)=n_1(t)g(t)+n_2(t)(1-g(t))=g(t)[n_1(t)-n_2(t)]+n_2(t) \tag{1}$$

式中:$n_1(t)$,$n_2(t)$ 为两段互不相关、不同方差$(\sigma_{n1}^2,\sigma_{n2}^2)$、不同功率谱密度的噪声,假定其皆为零均值平稳高斯噪声,则有 $n_3(t)=n_1(t)-n_2(t)$,也为平稳高斯噪声。$g(t)$ 为拼接处理,如果当前帧为拼接点所在之帧,即帧内 $g(t)$ 出现跳变,设 t_0 时刻为编辑点,则式(1)方差为

$$\sigma_n^2=E[r^2(t)]=E([g(t)[n_1(t)-n_2(t)]+n_2(t)][g(t)[n_1(t)-n_2(t)]+n_2(t)])=g^2(t)(\sigma_{n1}^2+\sigma_{n2}^2)-2g(t)\sigma_{n2}^2+\sigma_{n2}^2 \tag{2}$$

假定 $g(t)$ 为矩形函数

$$g(t)=\begin{cases}0 & t<t_0\\ 1 & t_0<t<t_1\\ 0 & t>t_1\end{cases}$$

,即有 $g^2(t)=g(t)$,则上式可为

$$\sigma_n^2=g(t)(\sigma_{n1}^2-\sigma_{n2}^2)+\sigma_{n2}^2 \tag{3}$$

上式表明,在 t_0 时刻跳变处,方差发生了变化,从 σ_{n1}^2 变为 σ_{n2}^2。因此,对于平稳随机噪声而言,可以通过检测其方差等特性的跳变来寻找跳变点。由于噪声干扰等原因,跳变点变化不明显,而且也存在虚警现象,实际中只凭方差进行跳变点检测误差太大。

1.2　拼接帧的分数倒谱特征分析

分数傅里叶变换在时频域上具有旋转能力,通过改变分数阶值,可以使拼接点出现尖锐的跃变特性。Almeida 所定义的 a 阶分数傅里叶变换算法如下[1]

$$S_a(u)=F^a(s(t))=\begin{cases}A(a)\int_{-\infty}^{+\infty}\exp\left(\mathrm{j}\pi\dfrac{(u^2+t^2)\cos a-2ut}{\sin a}\right)s(t)\mathrm{d}t & a\neq n\pi\\ s(t) & a=2n\pi\\ s(-t) & a=(2n\pm1)\pi\end{cases} \tag{4}$$

式中:$a=\dfrac{p\pi}{2}$,$0<|p|<2$,$A(a)=\sqrt{\dfrac{1-\mathrm{j}\cos a}{2\pi}}$。

假定当前帧为篡改点所在帧,则其前后帧皆为未篡改帧,假定前后帧都为噪声帧,则前后帧分别为 $n_1(t)$,$n_2(t)$。对式(4)进行分数傅里叶变换

$$R_a(u)=A(a)\int_{-\infty}^{+\infty}\exp\left(\mathrm{j}\pi\frac{(u^2+t^2)\cos a-2ut}{\sin a}\right)\cdot[g(t)(n_1(t)-n_2(t))+n_2(t)]\mathrm{d}t=A(a)\int_{-\infty}^{+\infty}\exp\left(\mathrm{j}\pi\frac{(u^2+t^2)\cos a-2ut}{\sin a}\right)\cdot g(t)(n_1(t)-n_2(t))\mathrm{d}t+N_2(a,u) \tag{5}$$

上式中 $N_2(a,u)$ 为噪声 $n_2(t)$ 的 a 阶分数谱。则当前帧分数功率谱为

$$P_r^a(u)=E[R_a(u)R_a*(u)]$$

将式(5)代入

$$P_r^a(u)=P_1+P_2+P_3+|N(a,u)|^2 \tag{6}$$

令 $\tau=t-t'$ 代入上式,化简整理可得

$$P_1=|A(a)|^2\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\exp\left(\mathrm{j}\pi\frac{(\tau^2-2\tau t^2)\cos a-2u\tau}{\sin a}\right)\cdot g(t)g(t+\tau)(R_{n1}(\tau)+R_{n2}(\tau))\mathrm{d}t\mathrm{d}\tau$$

由随机过程理论可知,理想情况下,噪声自相关函数在 $\tau=0$ 时不为 0,$\tau\neq0$ 时为 0。则上式可进一步化简

$$P_1=|A(a)|^2(\sigma_{n1}^2+\sigma_{n1}^2)\int_{-\infty}^{+\infty}g^2(t)\mathrm{d}t \tag{7}$$

$$P_2=P_3=|A(a)|^2\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\exp\left(\mathrm{j}\pi\frac{(t^2-t'^2)\cos a-2u(t-t')}{\sin a}\right)\cdot$$

$$g(t)(R_{n2}(t-t'))\mathrm{d}t\mathrm{d}t' = |A(a)|^2\sigma_{n2}^2\int_{-\infty}^{+\infty}g(t)\mathrm{d}t \tag{8}$$

则式(6)可为

$$P_r^a(u) = |A(a)|^2(\sigma_{n1}^2+\sigma_{n2}^2)\cdot \left(\int_{-\infty}^{+\infty}g^2(t)\mathrm{d}t + 2\frac{\sigma_{n2}^2}{\sigma_{n1}^2+\sigma_{n2}^2}\int_{-\infty}^{+\infty}g(t)\mathrm{d}t + \frac{\sigma_{n2}^2}{\sigma_{n1}^2+\sigma_{n2}^2}\right) \tag{9}$$

分析(7-9)可看出,拼接函数受到了噪声的乘性干扰,因此,可以考虑采用分数倒谱把噪声和拼接函数在分数倒谱域进行分离。同时利用分数傅里叶变换使拼接函数能量更集中,有

$$\hat{r}_{\mathrm{acep}}(t) = \mathrm{Frft}^{-a}(\log[\mathrm{Frft}^a(R(r(t),r(t+\tau)))]) \tag{10}$$

式中:$R(r(t),r(t+\tau))$ 为对信号进行相关运算,$\mathrm{Frft}^a()$,$\mathrm{Frft}^{-a}()$ 为 a 阶及 $-a$ 阶分数傅里叶变换。

具体步骤如下:

(1)首先采用能量及过零率联合检测的方法对当前帧进行有无声判决。如果是噪声帧则进行如下处理。

(2)计算当前帧的相关函数,并对之进行 a 阶分数傅里叶变换。

(3)对分数幅度谱进行对数变换;

(4)进行 $-a$ 阶分数傅里叶变换,得到 $\hat{r}_{\mathrm{acep}}$,并进行中值滤波

$$\hat{r}_{\mathrm{acep}}(n) = \mathrm{mid}(\hat{r}_{\mathrm{acep}}(n-2),\hat{r}_{\mathrm{acep}}(n-1),\hat{r}_{\mathrm{acep}}(n),\hat{r}_{\mathrm{acep}}(n+1),\hat{r}_{\mathrm{acep}}(n+2)) \tag{11}$$

从式(9)可看出,当前帧如果发生拼接,则第1,2项不为零,导致分数倒谱域上出现跳变,因此有计算 $\hat{r}'(n)=\hat{r}_{\mathrm{acep}}(n)-\hat{r}_{\mathrm{acep}}(n-1)$,求其极值,如果下式成立

$$\max[\hat{r}'(n)] > \Delta \tag{12}$$

则认为该帧为两段语音拼接处所在之帧。Δ 为预设门限。

2 语音帧间拼接特性

对于语音拼接而言,通常将同一说话人两段不同阶段时的语音进行拼接,如何提取语音中蕴含的不同阶段说话人特征即提取说话人相邻帧共性特征成为语音取证、语音篡改分析的难点。用于语音取证的说话人相邻帧共性特征不同于普通说话人特征,而普通说话人识别需要提取长期稳健的说话人个性特征,这一点与相邻帧共性特征相矛盾。目前可沿用的特征包括说话人情感特征、高频倒谱特征、E250 能量等。本文采用分数倒谱域上的高频方差和过零率特征进行分析。

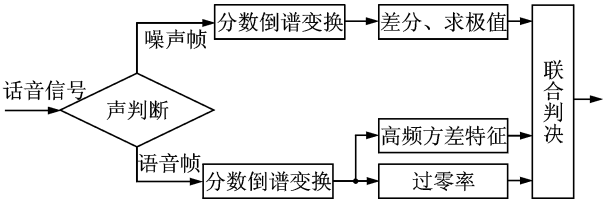


图 1 语音拼接识别

Fig. 1 Speech forensics identification

3 实验分析

Diamond cut DC live forensics 是一款优秀的专业音频恢复(修复)、音频编辑、音频增效软件。为了分析上节提出的在分数倒谱变换域进行拼接帧检测方法的性能,本节对采用 DC live forensics 7.5 编辑过的语音信号进行拼接检测和分析。DC live forensics 工具不仅实现了对语音信号的无缝拼接,而且还在拼接处进行了滤波,防止了带外辐射。

在本文的所有实验中,语音信号为在实验室内录制的信号,采样频率是 8 kHz,采样位数 8 bit。

图 2 给出了 DC live forensics 7.5 环境下原始语音和编辑语音的波形和频谱对比从图 2 中编辑语音的波形和频谱中很难发现其是否发生了编辑以及无法确定编辑的位置。

图 3 给出了分数倒谱变换域($a=0.8$)上原始语音与编辑语音特性变化曲线,编辑点在噪声帧大约 410 帧处,图中可发现,在 410 帧左右的编辑点附近的噪声帧的倒谱特性不连续,编辑点以外的其他相邻噪声帧的分数倒谱特性则呈连续特性,并且波动不大。

图 4 中原始语音和拼接语音如图 3。图 4 中给出了普通倒谱域高频段方差特性比较,图中无法明显区分原始语音与拼接语音。

图 5 给出了在同一原始语音的同一处嵌入不同的一句话。针对分数倒谱变换域上低频谱(小于 $L/4$)的过零率特性,从图中可看出,拼接发生时,过零率曲线发生较大的变化。改变不同的 a 值,发现 $a=0.2$ 时更加明显。

图 6 则给出了普通倒谱域上幅度谱过零率变化曲线,图 6 中无法区分拼接帧。

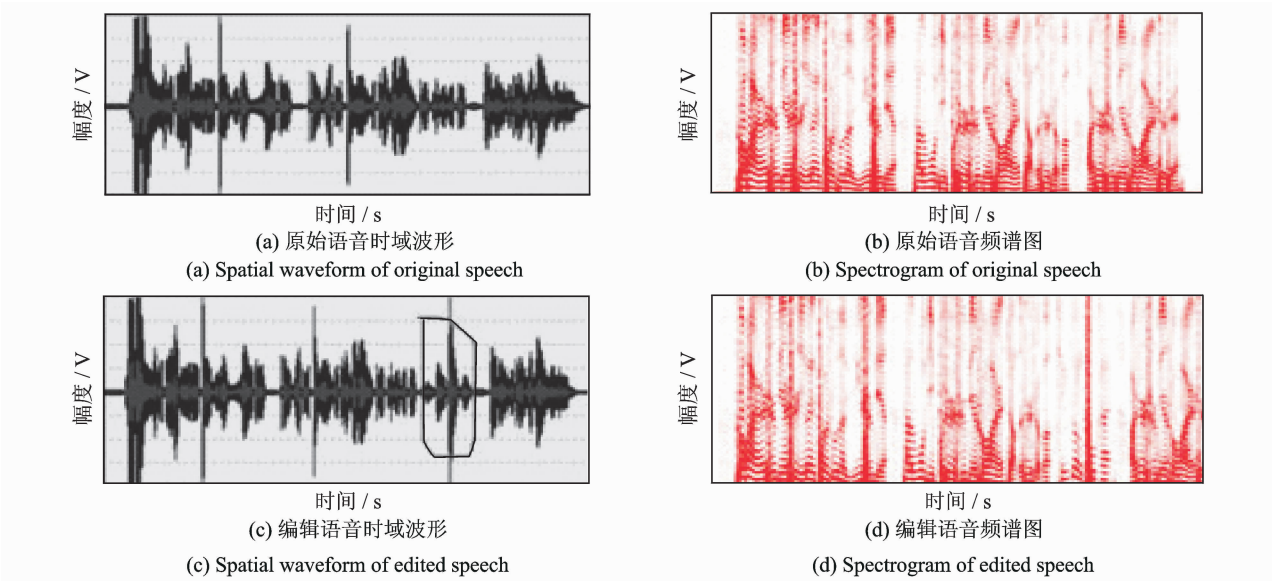


图 2　原始语音和编辑语音时域波形和频谱图

Fig. 2　Spatial waveform and spectrogram of original speech and edited speech

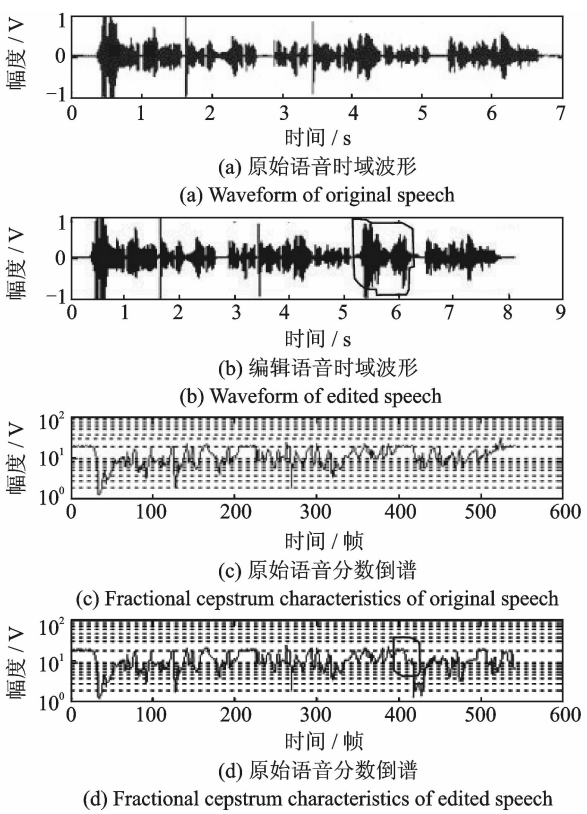


图 3　分数倒谱变换域上特性比较

Fig. 3　Characteristics comparison in spatial and fractional cepstrum transform

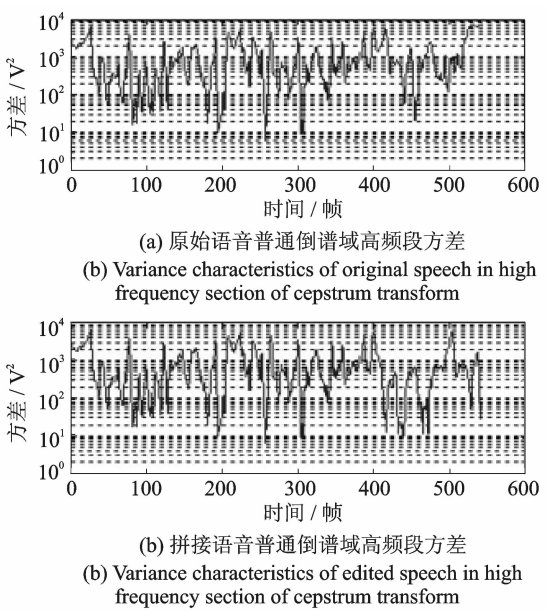


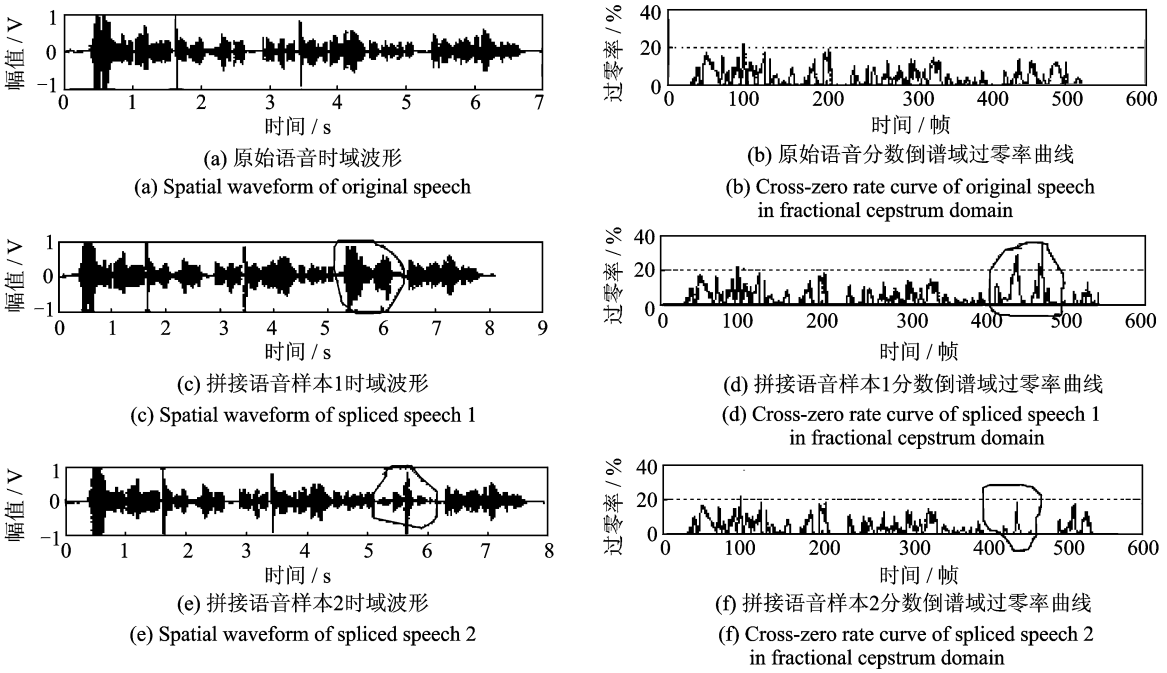
图 4　普通倒谱域高频段方差特性比较

Fig. 4　Variance characteristics comparison in high frequency section of cepstrum transform

图 7 给出了 $a=1.2$ 时分数倒谱域方差特性曲线，从图中可看出，不包含编辑点的噪声帧方差呈相同的常数，而编辑点附近的分数倒谱方差出现跳变。

分析图(2,4,7)，分数倒谱变换域上高频方差特性能够识别拼接点所在的噪声帧。图(5-6)显示，采用分数倒谱域上过零率检测能够检测出拼接语音帧。

大量的男女语音实验显示，在 $a=0.2$ 左右，分数倒谱域上过零率方法效果更好一些，$a=1.2$ 左右，分数倒谱变换域上高频方差方法效果较好。



(a) 原始语音时域波形
(a) Spatial waveform of original speech

(b) 原始语音分数倒谱域过零率曲线
(b) Cross-zero rate curve of original speech in fractional cepstrum domain

(c) 拼接语音样本1时域波形
(c) Spatial waveform of spliced speech 1

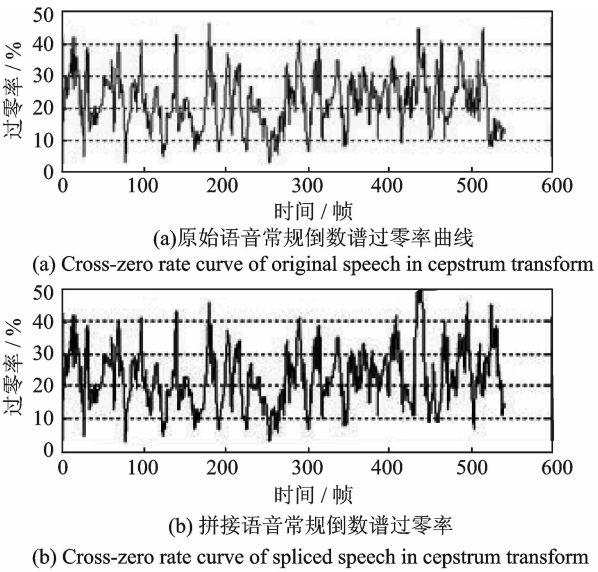
(d) 拼接语音样本1分数倒谱域过零率曲线
(d) Cross-zero rate curve of spliced speech 1 in fractional cepstrum domain

(e) 拼接语音样本2时域波形
(e) Spatial waveform of spliced speech 2

(f) 拼接语音样本2分数倒谱域过零率曲线
(f) Cross-zero rate curve of spliced speech 2 in fractional cepstrum domain

图 5 在同一语音上拼接同一说话人不同阶段语音的分数倒谱域过零率变化曲线

Fig. 5 Cross-zero rate curve in fractional cepstrum domain of spliced speech

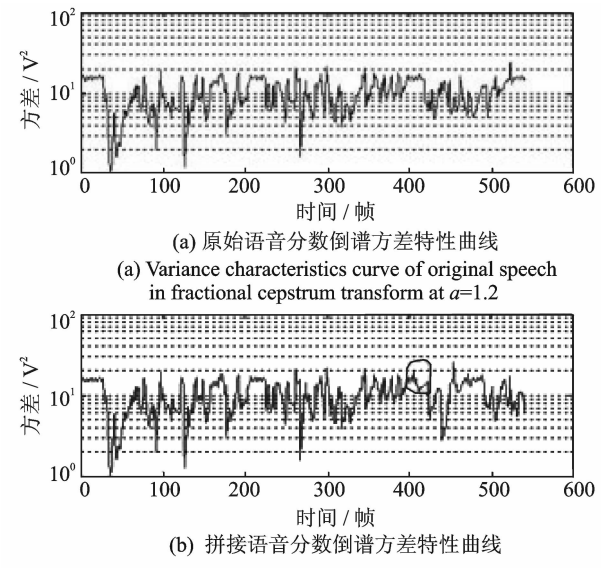


(a)原始语音常规倒数谱过零率曲线
(a) Cross-zero rate curve of original speech in cepstrum transform

(b) 拼接语音常规倒数谱过零率
(b) Cross-zero rate curve of spliced speech in cepstrum transform

图 6 常规倒谱域上的过零率比较

Fig. 6 Cross-zero rate comparison in cepstrum transform



(a) 原始语音分数倒谱方差特性曲线
(a) Variance characteristics curve of original speech in fractional cepstrum transform at a=1.2

(b) 拼接语音分数倒谱方差特性曲线
(b) Variance characteristics curve of spliced speech in fractional cepstrum transform at a=1.2

图 7 $a = 1.2$ 时分数倒谱方差特性曲线

Fig. 7 Variance characteristics curve in fractional cepstrum transform at $a = 1.2$

4 结束语

目前对语音编辑识别、语音拼接识别仅限于不同采样率的语音信号，对于同一采样率的语音信号目前方法基本没有有效的方法。本文针对语音取证中相同采样率的语音拼接识别进行了初步研究，着重分析了拼接对噪声特征的影响，提出了基于分数倒谱变换的拼接帧检测算法，同时也给出了采用说话人相邻帧特征用于拼接帧识别。实验素材是通过 DC Live Forensics 编辑后的实验室录制语音。实验结果表明，分数倒谱变换的拼接帧检测算法优于普通倒谱域方差法。用于检测语音的分数倒谱域上过零率检测法也优于常规倒谱域过零率法。需要说明的是，本文的研究仅针对语音拼接点在噪声帧时有效，如果拼接点发生在语音帧，还需

要基于分数阶因子[1,4,14]采用比如神经网络的方法进行进一步探讨,在分数域上进行语音与噪声的最大分离或者采用说话人识别的方法进行特征提取或模式识别[14]。

参考文献:

[1] Campbell J P,Shen W,Campbell W M, et al. Forensic speaker recognition[J]. IEEE Signal Processing Magazine, 2009,26(2):95-103.

[2] Maher R C. Audio forensic examination(authenticity, enhancement, and interpretation)[J]. IEEE Signal Processing Magazine,2009,26(2):84-94.

[3] Gupta S,Cho S,Kuo C,et al. Current developments and future trends in audio authentication[J]. IEEE Multimedia, 2012,19(1):50-59.

[4] Han K J,Omar M K,Pelecanos J,et al. Forensically inspired approaches to automatic speaker recognition [C]// IEEE International Conference on Acoustics, Speech and Signal Processing. USA: IEEE Press, 2011:5160-5163.

[5] Liu Yuming,Yuan Zhiyong,Markham,et al. Wide-area frequency as a criterion for digital audio recording authentication[C]// IEEE Power and Energy Society General Meeting. Detroit, USA: IEEE Press, 2011:1-7.

[6] Chang Fengcheng, Huang H C. Electrical network frequency as a tool for audio concealment process[C] // International Conference on Intelligent Information Hiding and Multimedia Signal Processing. Darmstadt,Germany:IEEE Press, 2010:75-178.

[7] Alan J C. Detection of copies of digital audio recordings for forensic purposes [D]. Buckinghamshire, UK:The Open University,2006:1-255.

[8] Grigoras C. Application of ENF analysis method in authentication of digital audio and video recordings [J]. Jounal of Audio Engineering Society, 2009, 57 (9):643-661.

[9] Nicolalde R, Daniel P. Audio authenticity: Detecting ENF discontinuity with high precision phase analysis [J]. IEEE Transactions on Information Forensics and Security, 2010,5(3):534-543.

[10] 姚秋明,柴佩琪,宣国荣,等. 基于期望最大化算法的音频取证中的篡改检测[J]. 计算机应用,2006,26 (11):1598-2601.
Yao Qiuming,Chai Peiqi, Xuan Guorong,et al. Audio resampling detection in audio forensics based on EM algorithm[J]. Journal of Computer Applications, 2006,26(11):1598-2601.

[11] 丁琦,平西建. 针对语音变换的语音篡改检测[J]. 数据采集与处理,2012,27(1):57-62.
Ding Qi,Ping Xijian. Speech tampering detection for voice transformation[J]. Journal of Data Acquisition and Processing, 2012,27(1):57-62.

[12] Han K J,Omar M K,Pelecanos J,et al. Forensically inspired approaches toautomatic speaker recognition [C]// IEEE International Conference on Acoustics, Speech and Signal Processing. Prague Czech Republic:IEEE Press, 2011:5160-5163.

[13] 余华,黄程韦,金赟,等. 基于粒子群优化神经网络的语音情感识别[J]. 数据采集与处理,2011,26(1):58-62.
Yu Hua,Huang Chengwei,Jin Yun,et al. Speech emotion recognition based on particle swarm optimizer neural network[J]. Journal of Data Acquisition and Processing, 2011,26(1):58-62.

[14] 李嘉,黄程韦,余华. 语音情感的维度特征提取与识别[J]. 数据采集与处理,2012,27(3):389-393.
Li Jia, Huang Chengwei, Yu Hua. Dimensional feature extraction and recognition of speech emotion[J]. Journal of Data Acquisition and Processing, 2012,27 (3):389-393.

作者简介:钟　巍(1973-),男,博士研究生,研究方向:数字音频取证,E-mail:zww110221@163.com;孔祥维(1963-),女,教授,博士生导师,研究方向:数字音频取证;尤新刚(1963-),男,教授,研究方向:多媒体通信及信息安全;王波(1981-),男,讲师,研究方向:数字媒体取证。