ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 35,No. 5,Sep. 2020,pp. 880－891
DOI:10. 16337/j. 1004-9037. 2020. 05. 009

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

# 基于i-vector的电子伪装语音鲁棒还原方法研究

郑琳琳，张雄伟，孙　蒙，李嘉康，张星昱

（陆军工程大学指挥控制工程学院，南京，210007）

**摘　要**：语音的电子伪装是指采用变声设备或语音处理软件改变说话人的个性特征，以达到故意隐藏该说话人身份的目的。电子伪装语音还原是指通过技术手段将伪装语音变回原声，这对基于语音的身份鉴别有重要意义。本文将频域和时域伪装语音的还原问题抽象为伪装因子的估计问题，通过基于i-vector的自动说话人确认方法估计伪装因子，并引入对称变换进一步提高估计效果。该方法借助于i-vector的噪声鲁棒性，提高了真实含噪场景下伪装因子的估计精度，从而改进了噪声条件下电子伪装语音的还原效果。在干净语音库TIMIT上训练i-vector并在含噪语音库VoxCeleb1上对本文方法进行测试，结果表明，伪装因子估计的错误率从基线系统的9.19％降低为4.49％，还原语音在自动说话人确认等错误率和听觉感知方面也取得了提升。

**关键词**：电子伪装语音；伪装因子估计；自动说话人确认；噪声鲁棒性

**中图分类号**：TN912　　**文献标志码**：A

## Noise Robust Restoration of Electronic Disguised Voices Based on i-vector

ZHENG Linlin, ZHANG Xiongwei, SUN Meng, LI Jiakang, ZHANG Xingyu

(College of Command and Control Engineering, Army Engineering University, Nanjing, 210007, China)

**Abstract**: Electronic voice disguise refers to hiding the identity of a speaker by voice changing equipment or voice processing software. The restoration of disguised voice refers to changing it back to its original version, which is of great significance for speaker identification. This paper first models the restoration of disguised voices as the estimation of disguising factors in both frequency and time domains. The estimation of disguising factor is made by automatic speaker verification using i-vector. Symmetric transformation is proposed to improve the performance on parameter estimation. By virtue of the noise robustness of i-vector, the proposed method improves the estimation accuracy of the disguising factor in the real noise-containing scene, thereby improving noise robust restoration effect of electronic disguised voice. Evaluation results on noisy speech library VoxCeleb1 of the trained model on clean speech library TIMIT demonstrated good performance of the approach by reducing the error rate from 9.19％ to 4.49％. The quality of the restored voice is also improved in the aspects of automatic speaker verification and auditory perception.

**Key words:** electronic voice disguise; disguising factor; automatic speaker verification; noise robustness

---

**基金项目**：国家自然科学基金(61471394,62071484)资助项目；江苏省优秀青年基金(BK20180080)资助项目。
**收稿日期**：2020-01-09；**修订日期**：2020-05-16

## 引　言

语音伪装是指通过改变说话人的个性特征,故意隐藏或伪造说话人的身份[1]。随着智能语音交互应用的不断发展,语音代表个人身份特征的场景日益广泛,伪装语音技术的出现给说话人身份的辨识带来很大的挑战。目前,利用各类变声器及变声软件可以对语音进行个性伪装,致使人耳甚至部分说话人识别系统无法辨识出说话人的身份,严重影响语音检验鉴定效果,使犯罪分子有机可乘[2-4]。因此,如何进行伪装语音的说话人身份识别已成为信息安全领域的一个重要且紧迫的课题。

语音伪装方法可以分为两种类型:人为伪装和电子伪装[5]。人为伪装是借助人本身的技能进行伪装,包括说话时采用捏鼻、咬物等方法;电子伪装是指采用电子设备或语音处理软件对说话人的原始语音进行变声伪装。电子伪装使用复杂高效的算法,以其高质量的伪装效果和便捷的实现方式,得到了越来越广泛的应用[6]。文献[7-9]的研究表明,语音经过伪装后,会明显降低说话人识别系统的准确率,而且不同的伪装方法对说话人识别的性能影响各异。实验发现,利用当前比较成熟的基于高斯混合模型(Gaussian mixture model, GMM)和通用背景模型(Universal background model, UBM)的声纹识别模型对电子伪装后的语音进行识别,等错误率(Equal error rate, EER)高达40%以上[10],几乎无法辨认出伪装者的身份。因此,在鉴别伪装者的身份之前,首先需要对伪装语音进行还原处理。电子伪装语音的还原问题可以抽象简化为伪装因子的估计问题[11]。文献[12]通过动态时间规整(Dynamic time warping, DTW)模型进行伪装因子估计,再利用矢量量化(Vector quantization, VQ)模型进行说话人识别,一定程度上缓解了VQ说话人识别系统对电子伪装语音识别率过低的问题。文献[10]利用基频比估计伪装因子,提出了一种改进的梅尔倒谱系数(Mel frequency cepstral coefficient, MFCC)提取算法,能够有效地从电子伪装的声音中还原出原始语音的MFCC。针对电子伪装后的语音,将该算法还原出来的MFCC特征输入到GMM-UBM的自动说话人确认(Automatic speaker verification, ASV)系统,说话人确认EER仅为3%~4%,明显优于未经还原的MFCC特征的40% EER。鉴于该方法的良好性能,本文将其设定为基线系统,并在其基础上进行改进研究。

实验发现,文献[10]所述的伪装因子估计方法对语音质量要求过于苛刻,对于真实情况下的含噪伪装语音的还原效果不是很理想。如何对真实含噪情况下的电子伪装语音进行还原,是一个更具挑战性的问题,对后续的伪装语音说话人身份识别具有决定性作用。鉴于此,本文将频域和时域伪装语音的还原问题抽象为伪装因子的估计问题,通过基于i-vector的自动说话人确认方法估计伪装因子,并引入对称变换进一步提高估计效果。该方法借助于i-vector的噪声鲁棒性,提高了真实含噪场景下的伪装因子估计的精度,从而改进了电子伪装语音还原的效果。在目前常用的说话人识别数据库VoxCeleb1[13]上的实验表明,利用该方法估计的伪装因子错误率为4.49%,低于基频比伪装程度估计方法的9.19%,为准确进行电子伪装语音说话人身份的辨识提供了前提条件。

## 1　电子伪装语音

电子语音伪装的目的是改变语音给予人耳的听觉感受,最直接的变化就是改变语音的音调(Pitch)。提高音调,语音变得尖锐;降低音调,语音变得低沉。随着伪装程度的加深,正常语音与伪装语音的差异增大。本节首先介绍电子伪装的工作原理,然后给出电子伪装语音伪装程度的量化表示。

### 1.1　电子伪装工作原理

音调被用来描述人对语音频率的感知量,电子语音伪装改变音调本质上是按照不同的比例因子对频谱进行压缩和扩展。语音基音频率(Fundamental frequency, FF)是指发浊音时声带振动所引起的周期性振动频率,一般用$F_0$表示,它反映了语音激励源的重要特征,是语音信号短时内较稳定的频率分量。假设原始语音音调为$p_0$、基频为$f_0$,经伪装之后的音调为$p_1$、基频为$f_1$,音调和基频存在式(1)所示的

变换关系

$$\alpha = p_1/p_0 = f_1/f_0 \tag{1}$$

式中,$\alpha$是音调变换的比例因子。根据伪装变换的方式不同,电子伪装可以分为2类:频域伪装和时域伪装。这2类伪装方式都可以通过比例因子$\alpha$来定量描述。

1.1.1 频域伪装

频域伪装是指通过直接在语音频域内拉伸或压缩频谱来提高或降低音调的伪装方式,该方式可以改变语音的音调而保持语音节奏不变。

语音频域分析最常用的方法是傅里叶分析法。因为语音波是一个非平稳过程,因此适用于平稳周期信号的标准傅里叶变换不能直接用来表示语音信号,而应该用短时傅里叶变换(Short-time Fourier transform, STFT)对语音信号的频谱进行分析。STFT首先将信号分帧,然后对每一帧语音信号进行快速傅里叶变换(Fast Fourier transform, FFT),得到频域分析结果[14]。

假设$|F(k)|$和$\omega(k)$分别代表原始语音频域分析后第$k$个频率点处的瞬时幅度和瞬时频率,$\alpha$是音调变换的比例因子。频域伪装变换根据式(2),将瞬时频率$\omega(k)$利用比例因子$\alpha$修改为$\omega'(\lfloor \alpha k \rfloor)$,即

$$\omega'(\lfloor \alpha k \rfloor) = \alpha\omega(k) \qquad 0 \leqslant k, \alpha k < N/2 \tag{2}$$

瞬时幅度$|F(k)|$利用线性插值法进行相应的拉伸或压缩变换

$$|F'(\lfloor \alpha k \rfloor)| = \mu|F(k)| + (1-\mu)|F(k+1)| \tag{3}$$

式中,$0 \leqslant k, k' < N/2$,$k = \lceil k'/\alpha \rceil$,$\mu = k'/\alpha - k$。为了简单起见,仍使用$k$作为伪装后的瞬时幅度$|F'|$和瞬时频率$\omega'$的坐标尺度,记为$|F'(k)|$和$\omega'(k)$。根据$|F'(k)|$和$\omega'(k)$可得修改后的FFT系数$F'(k)$。对$F'(k)$执行快速傅里叶逆变换(Inverse fast Fourier fransform, IFFT),即可得到频域伪装的语音信号。

1.1.2 时域伪装

时域伪装一般通过调整采样率和采用基音同步叠加(Pitch-synchronous overlap and add method, PSOLA)相结合的方法来实现。调整采样率能够改变语音信号的FF从而改变音调。但是语音信号时频结构之间的约束性使得信号的时域特性和频域特性紧密相关,只利用调整采样率生成的伪装语音往往听起来不够自然,需要采用PSOLA对语音进行进一步处理。PSOLA可以对语音的基频、时长和短时能量等韵律特征进行修改,使修改之后的语音与原来语音频谱有着基本相同的包络[15]。这种伪装方式既改变了语音的音调,又改变了语速。

PSOLA首先检测语音信号$x(t)$的音调的位置和轮廓,加窗提取基音周期函数$P(t)$。利用式(4)对语音信号进行重采样,修改基音周期函数$P(t)$,在误差最小准则下重复或丢弃部分语音帧做补偿,其中$\alpha$是伪装比例因子

$$P'(t) = P(t)/\alpha \tag{4}$$

语音经过时域伪装后,语音时长发生变化的同时音调也会得到相应升降调处理。假设原始语音语速为$v_0$,经伪装之后的语音语速为$v_1$,根据式(1),对于时域伪装语音存在如式(5)的关系

$$\alpha = p_1/p_0 = f_1/f_0 = v_1/v_0 \tag{5}$$

式中,比例因子$\alpha$不仅是时域伪装语音的音调变换比和基频变换比,还是时域伪装语音的语速变换比。

### 1.2 伪装程度的量化表示

语音学中,音调通常用12-半音法来测量,表示音调最多可提高或降低12个半音[16]。原始语音音调$p_0$与伪装后语音音调$p_1$之间存在着变换关系

$$p_1 = 2^{s/12} p_0 \tag{6}$$

式中,$s$是半音尺度因子,表示提高或降低$s$个半音。本文将半音尺度因子$s$称为伪装因子,用来量化表示电子伪装语音的伪装程度。如果伪装因子$s>0$,说明提高了$s$个半音;如果$s<0$,说明降低了$s$个半音;如果$s=0$,说明未改变音调。

根据式(1,6)可知,音调变换比例因子$\alpha$和伪装因子$s$之间的变换关系为

$$s=12\log_2(\alpha) \tag{7}$$

## 2 伪装因子估计基线方法

电子伪装是按照不同的比例因子对频率分量进行缩放,从而改变语音的音调。考虑到伪装前后基频的变化能反映频率分量整体的缩放程度,Wang等[10]提出用基频比来估计伪装因子,进而还原语音,其原理如图1所示。该方法根据待测语音与注册语音的基频比来估计伪装因子,利用估计出的伪装因子修正待测语音的MFCC,从而得到还原后的MFCC特征。将提出的方法作为特征还原工具应用于GMM-UBM说话人识别系统的前端,可提高电子伪装语音伪装者的识别准确率。在TIMIT语音库上的实验表明,估计所得的伪装比例因子$\alpha'$与真实的比例因子$\alpha$较接近,平均误差也很小,最大错误率在1.6%到7.7%之间,说明基频比估计作为伪装还原的手段是可行的。

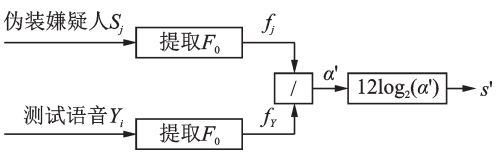



图1 利用基频比确定伪装因子原理图

Fig.1 Estimation of disguising factor by the ratio of fundamental frequencies

本文将文献[10]中提出的利用基频比估计伪装程度的算法作为基线系统。在训练阶段,提取伪装嫌疑人$S_j$的基频$f_0$的平均值$f_j$;在测试阶段,计算待检测语音$Y_i$的基频$f_0$的平均值$f_Y$,通过式(8,9)估计出伪装因子$s'$

$$\alpha'=f_Y/f_j \tag{8}$$

$$s'=12\log_2(\alpha') \tag{9}$$

基于基频比的伪装因子估计方法首先利用简化逆滤波跟踪法(Simplified inverse filter tracking, SIFT)提取基频[17],考虑到每条语句两端发音的不稳定性,舍弃基频序列的前15%和后15%数据,保留中间的70%数据用来计算基频平均值[10]。

然而,基频提取准确度与语音质量有很大关系,当待测语音中含有环境噪声或者捏鼻、捂嘴等人为伪装时,基频提取会产生较大误差。对比实验发现,该基线系统对语音质量要求过于苛刻,对于真实情况下的含噪语音的伪装因子估计结果不是很理想,相关实验结果将在第4节给出。真实含噪场景下的电子伪装语音还原是一个更具有实际应用价值的问题,对于推动电子伪装语音身份鉴定技术的应用和发展具有重要作用。因此,本文借助于ASV系统的噪声鲁棒性,利用ASV系统估计伪装因子,并引入对称变换进一步提高估计精度。

## 3 基于说话人确认的伪装因子估计方法及改进

本节首先介绍基于说话人确认的伪装因子估计方法,然后引入对称变换提高自动说话人确认估计伪装因子的精度。该方法以目前发展比较成熟的基于i-vector的说话人确认模型为基础,通过概率线性判别分析(Probabilistic linear discriminant analysis, PLDA)最优得分时的自变量取值来估计伪装因子,从而实现电子伪装语音的还原。

### 3.1 基于说话人确认的伪装因子估计

说话人确认是说话人识别任务的一种,旨在利用语音信号中能反映说话人生理和行为的特征来判

断两段语音是否来自同一个说话人。近年来，说话人确认方法的性能得到了显著提高，如Reynolds在实验室环境中使用TIMIT语音数据库对630个人进行实验，识别率近乎达到100%[18]。现实使用中，说话人确认被应用于访问控制、交易认证和军事侦察等诸多涉及逻辑和物理访问的真实身份验证场景[19]。

基于GMM-UBM和i-vector的说话人确认方法是目前发展比较成熟且被广泛采用的说话人确认模型，原理如图2所示。该模型首先对提取的语音信号的特征（如MFCC等）在大量语料上训练一组GMM-UBM作为通用背景。在注册和测试阶段，从待测语音$S$中提取特征，并将这些特征作为观测值对训练好的GMM-UBM做最大后验概率估计（Maximum a posteriori, MAP），得到高斯超矢量，并进一步提取说话人的特征i-vector，用$\lambda$表示[20]。通过对比注册语句和测试语句所提取的i-vector的相似程度，即可完成2条语句是否来自同一个说话人的判决任务。

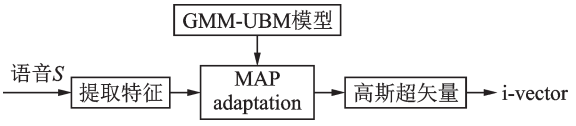



图2 基于GMM-UBM的i-vector提取方法

Fig.2 The i-vector extraction by GMM-UBM

基于说话人确认系统的伪装因子估计方法如图3所示。该方法通过遍历伪装因子的理论取值范围，对待测伪装语音进行逐一还原，然后说话人确认系统对每条还原语音与伪装嫌疑人的语音进行打分，得分最高的还原语音对应的伪装因子即认为是正确的伪装因子。本文中说话人确认模型选择了通过GMM-UBM提取的i-vector，具体步骤如下：

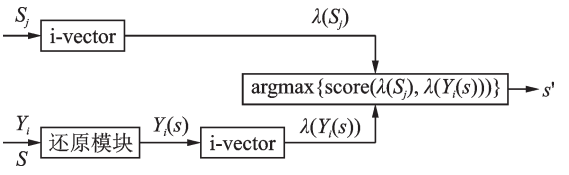



图3 基于说话人确认的伪装因子估计方法

Fig.3 Estimation of disguising factor by automatic speaker verification

（1）训练阶段，利用伪装嫌疑人$S_j$的正常语音进行注册，通过说话人确认中的特征提取部分计算得到该说话人的注册特征$\lambda_j$；

（2）测试阶段，待测语音$Y_i$是经过电子伪装的语音信号，但伪装因子未知，根据电子伪装语音的变声规律，利用伪装因子的理论取值$s(3\leqslant|s|\leqslant11, s\in\mathbf{Z})$对待测语音$Y_i$的频谱特征进行还原，而后经过Griffin_Lim算法[21]得到还原语音$Y_i(s)$；

（3）利用说话人确认分别提取每个因子的还原语音$Y_i(s)$的特征，与伪装嫌疑人$S_j$的注册特征$\lambda_j$计算得分，按照式（10），分数最高的还原语音对应的伪装因子即为估计所得的伪装因子$s'$。

$$s'=\arg\max_{s}\{\mathrm{score}(\lambda(S_j),\lambda(Y_i(s)))\}\qquad -11\leqslant s\leqslant 11, s\in\mathbf{Z} \tag{10}$$

### 3.2 基于对称变换的伪装因子估计

语音经过升调（$s>0$）电子伪装后，频率范围拉伸，原始语音的高频部分会被丢弃。因此，升调电子伪装语音的还原过程需要将频率范围压缩，并且需要将高频部分的数据额外补全。然而，高频部分的频谱数据补全过程中会存在误差[22]，导致升调电子伪装语音的还原语音与原始语音存在一定差距，还原语音和原始语音的频谱对比图在第4节中给出。所以第3.1节介绍的方法对升调电子伪装语音的伪装因子估计存在潜在误差。为提高升调电子伪装语音的伪装因子估计精度，本节通过引入伪装因子的对称变换，对基于说话人确认的伪装因子估计方法进行改进，如图4所示。具体步骤如下所述：

（1）注册阶段，通过将伪装因子遍历取值范围$3\leqslant s\leqslant 11$来修改伪装嫌疑人$S_j$的语音，加上该说话人的正常语音，共得到10组语音，所以注册阶段可得到该说话人的10个模型（每组语音注册1个模型），如图4左半部分所示；

(2)测试阶段,仅利用伪装因子的降调理论取值范围对待测语音 $Y_i$ 进行还原,即 $-11\leqslant s\leqslant -3$,得到还原语音 $Y_i(s)$,如图4右半部分所示;

(3)计算伪装嫌疑人 $S_j$ 的正常语音模型与还原语音 $Y_i(s)$ 的得分 $\text{score}_{-11}$~$\text{score}_{-3}$,伪装嫌疑人 $S_j$ 的9个升调伪装语音模型与待测语音 $Y_i$ 的得分 $\text{score}_3$~$\text{score}_{11}$,比较上述18个得分,最高分对应的伪装因子就是估计所得的伪装因子 $s'$,如图4中间虚线框所示。

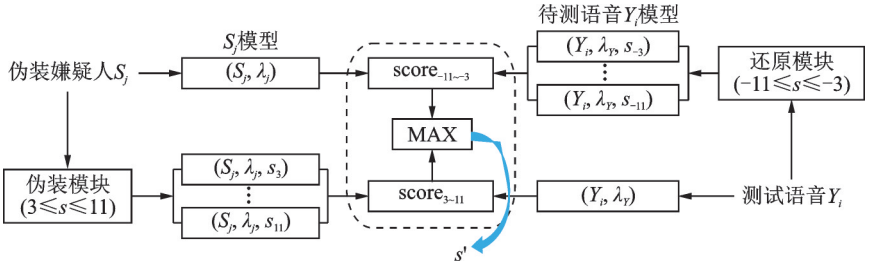



图4 利用对称变换改进基于说话人确认的伪装因子估计方法

Fig.4 Improving estimation of disguising factor based on automatic speaker verification by symmetric transform

## 4 仿真实验及结果分析

### 4.1 实验设置

4.1.1 电子伪装语音生成

实验用的电子伪装语音由SoundStretch音频处理软件产生。SoundStretch可以对音频文件执行实现变速不变调(Rate)、变调不变速(Pitch)、变速同时变调(Tempo)3个操作。由于Rate处理对说话人确认系统以及人耳辨识干扰不大,这里只考虑基于频域伪装的Pitch处理和基于时域伪装的Tempo处理作为变声手段。当伪装程度过小或过大时,伪装效果不明显或不能辨别出语义特征,对说话人确认系统以及人耳辨识系统的威胁很小。因此,本文考虑了18种伪装程度的伪装语音,对应伪装因子取值范围为+3~+11以及-3~-11。

4.1.2 数据集

实验用的数据集包括TIMIT和VoxCeleb1两个语音数据集。

由德州仪器(Texas Instruments, TI)、麻省理工学院(Massachusetts Institute of Technology, MIT)和斯坦福研究院合作构建的声学-音素连续语音语料库TIMIT是一个评价语音识别和说话人识别常用的权威语音库,包括630人8个不同地区的美国方言录制的音频信息。该语音库采用16 kHz采样率、16位量化和RIFF/WAV格式,每段录音的时长约为3 s。实验利用该语音库训练GMM-UBM模型。

VoxCeleb1是一个视听数据集,含有语音数据和视频数据,其中语音部分由从上传到YouTube的采访视频中提取的语音短片组成,带有真实噪声,且噪声出现时间点无规律。说话者覆盖到了不同年龄、性别、口音;语音的场景也非常丰富,包括红毯走秀、室外场馆、室内录影棚等,属于完全真实的英文语音[13]。本文随机选取该数据集中100位说话人,每人11条语音,其中10条用来注册1条用来测试。测试语音利用SoundStretch音频处理程序进行不同程度的伪装处理,得到18组伪装因子为+3~+11以及-3~-11的频域电子伪装语音和18组伪装因子为+3~+11以及-3~-11的时域电子伪装语音,每组含有100位说话人各1条待测电子伪装语音。

### 4.2 伪装因子估计结果

4.2.1 基线系统估计伪装因子结果

在含噪语音库VoxCeleb1的频域伪装数据集上利用基频比估计伪装因子,实验结果如表1所示。

**表1 VoxCeleb1频域伪装数据集上利用基频比估计伪装因子的实验结果**

**Table 1 Performance on the estimation of disguising factor using $F_0$-ratio on VoxCeleb1 with frequency-domain disguise**

| 参数 | 伪装因子估计结果 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $s$ | −11 | −10 | −9 | −8 | −7 | −6 | −5 | −4 | −3 |
| $s'_{\text{mean}}(s)$ | −8.71 | −8.20 | −7.81 | −6.85 | −6.23 | −5.37 | −4.41 | −3.55 | −2.56 |
| $E_{\text{mean}}(s)$ | 2.29 | 1.80 | 1.19 | 1.15 | 0.77 | 0.63 | 0.59 | 0.45 | 0.44 |
| $(E_{\text{mean}}(s)/s)/\%$ | 20.83 | 18.00 | 13.25 | 14.32 | 10.96 | 10.50 | 11.73 | 11.22 | 14.54 |
| $\text{Var}(s)$ | 17.94 | 16.90 | 13.00 | 15.93 | 11.12 | 9.63 | 9.22 | 8.72 | 8.63 |
| $s$ | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| $s'_{\text{mean}}(s)$ | 3.00 | 3.85 | 4.78 | 5.60 | 6.73 | 7.79 | 8.51 | 9.07 | 10.39 |
| $E_{\text{mean}}(s)$ | 0.00 | 0.15 | 0.22 | 0.40 | 0.27 | 0.21 | 0.49 | 0.93 | 0.61 |
| $(E_{\text{mean}}(s)/s)/\%$ | 0.00 | 3.63 | 4.44 | 6.63 | 3.92 | 2.64 | 5.40 | 9.27 | 5.55 |
| $\text{Var}(s)$ | 8.53 | 10.29 | 11.72 | 18.03 | 21.39 | 22.90 | 25.31 | 31.76 | 24.83 |

表1中，$s$是真实伪装因子，$s'_{\text{mean}}(s)$是每组数据估计出的伪装因子的平均值，$E_{\text{mean}}(s)=\left|s'_{\text{mean}}(s)-s\right|$是平均误差，$E_{\text{mean}}(s)/s$是平均误差率，$\text{Var}(s)$是每组实验数据的方差。

从表1中可以看出，随着伪装程度增大，伪装因子估计误差也呈增大趋势，最大错误率高达20.83%。表1的实验结果还表明，估计所得的伪装因子平均错误率达9.27%，平均方差为15.88，估计偏差远大于干净语音库上的实验结果。基于基频比的伪装因子估计方法在VoxCeleb1的时域电子伪装数据集上也得到了类似的结果，此处不再赘述。

4.2.2 利用说话人确认估计伪装因子的实验结果

利用GMM-UBM和i-vector的说话人确认方法对VoxCeleb1频域伪装数据集和时域伪装数据集分别进行伪装因子估计实验，实验结果在表2和表3中给出。利用说话人确认系统估计的伪装因子在频域伪装数据集上的平均错误率为12.26%、平均方差为5.05，在时域伪装数据集上的平均错误率为14.13%、平均方差为6.44。

**表2 VoxCeleb1频域伪装数据集上利用ASV估计伪装因子实验结果**

**Table 2 Performance on the estimation of disguising factor using ASV on VoxCeleb1 with frequency-domain disguise**

| 参数 | 伪装因子估计结果 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $s$ | −11 | −10 | −9 | −8 | −7 | −6 | −5 | −4 | −3 |
| $s'_{\text{mean}}(s)$ | −10.91 | −10.18 | −9.20 | −8.24 | −7.30 | −6.28 | −5.14 | −4.32 | −3.43 |
| $E_{\text{mean}}(s)$ | 0.09 | 0.18 | 0.20 | 0.24 | 0.30 | 0.28 | 0.14 | 0.32 | 0.43 |
| $(E_{\text{mean}}(s)/s)/\%$ | 0.82 | 1.80 | 2.22 | 3.00 | 4.29 | 4.67 | 2.80 | 8.00 | 14.33 |
| $\text{Var}(s)$ | 0.08 | 0.29 | 0.34 | 0.28 | 0.37 | 0.32 | 1.90 | 1.66 | 0.33 |
| $s$ | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| $s'_{\text{mean}}(s)$ | 2.76 | 3.12 | 3.87 | 4.67 | 5.67 | 6.57 | 6.6 | 7.81 | 8.96 |
| $E_{\text{mean}}(s)$ | 0.24 | 0.88 | 1.13 | 1.33 | 1.33 | 1.43 | 2.4 | 2.19 | 2.04 |
| $(E_{\text{mean}}(s)/s)/\%$ | 8.00 | 22.00 | 22.60 | 22.17 | 19.00 | 17.88 | 26.67 | 21.90 | 18.55 |
| $\text{Var}(s)$ | 2.64 | 2.87 | 4.99 | 5.62 | 6.16 | 7.39 | 19.36 | 19.71 | 16.60 |

**表3 VoxCeleb1时域伪装数据集上利用ASV估计伪装因子实验结果**

**Table 3 Performance on the estimation of disguising factor using ASV on VoxCeleb1 with time-domain disguise**

| 参数 | 伪装因子估计结果 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $s$ | −11 | −10 | −9 | −8 | −7 | −6 | −5 | −4 | −3 |
| $s'_{\text{mean}}(s)$ | −10.94 | −10.02 | −9.35 | −8.23 | −7.19 | −6.19 | −5.28 | −4.44 | −3.47 |
| $E_{\text{mean}}(s)$ | 0.06 | 0.02 | 0.35 | 0.23 | 0.19 | 0.19 | 0.28 | 0.44 | 0.47 |
| $(E_{\text{mean}}(s)/s)/\%$ | 0.55 | 0.20 | 3.89 | 2.88 | 2.71 | 3.17 | 5.60 | 11.00 | 15.67 |
| $\text{Var}(s)$ | 0.06 | 4.74 | 0.31 | 0.28 | 0.31 | 0.31 | 0.34 | 0.39 | 0.37 |
| $s$ | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| $s'_{\text{mean}}(s)$ | 2.74 | 3.07 | 3.74 | 4.56 | 5.12 | 6.15 | 6.67 | 7.40 | 8.18 |
| $E_{\text{mean}}(s)$ | 0.26 | 0.93 | 1.26 | 1.44 | 1.88 | 1.85 | 2.33 | 2.60 | 2.82 |
| $(E_{\text{mean}}(s)/s)/\%$ | 8.67 | 23.25 | 25.20 | 24.00 | 26.86 | 23.13 | 25.89 | 26.00 | 25.64 |
| $\text{Var}(s)$ | 2.93 | 2.77 | 5.65 | 7.67 | 11.85 | 11.67 | 18.06 | 21.10 | 27.17 |

当$s<0$时,估计的伪装因子偏差较小,方差最大值仅为4.74,说明基于说话人确认的伪装因子估计方法对降调伪装语音的效果较好。当$s>0$时,估计的伪装因子方差仍明显小于基线系统,但偏差较大(≈20%)。我们做出正常语音、升调伪装语音以及升调伪装语音的还原语音的频谱图进行对比,如图5所示。正如3.1节所指出的,对升调语音进行还原时,高频部分不能被有效恢复。虽然还原后的语音不影响人耳听觉效果(人耳听觉对低频信息敏感,对高频信息不太敏感),但丢失了大量高频信息,对说话人确认方法的性能造成了较大影响,从而影响了伪装因子的准确估计。

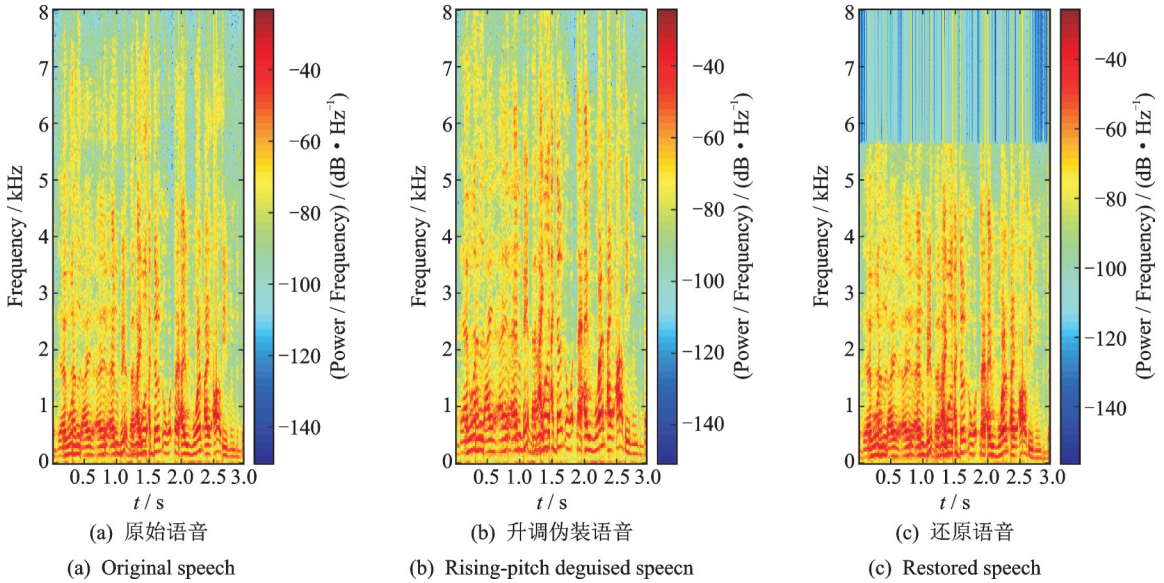



图5 正常语音、伪装语音及还原语音的频谱图

Fig.5 Spectrum of normal speech, disguised speech and restored speech

4.2.3 利用对称变换改进说话人确认估计伪装因子的实验结果

本文利用VoxCeleb1伪装数据集对经过对称变换改进后的基于说话人确认的伪装因子估计方法进行了测试,实验结果在表4和表5中给出。可以看出,与3.1节中的伪装因子估计方法相比,改进后的方法对升调伪装语音的伪装因子估计的错误率仅为6%,最大方差为0.74,准确率明显提高。同时也可以发现,改进模型对降调伪装语音的伪装因子识别率较4.2.2节中的结果有所下降,除伪装因子−11外,识别准确率仍明显优于基线系统。经计算,利用改进模型估计的伪装因子在频域伪装数据集上的平均错误率为4.49%、平均方差为6.19,在时域伪装数据集上的平均错误率为3.17%、平均方差为3.75,均明显优于基线系统。

**表4 VoxCeleb1频域伪装数据集上利用对称变换改进伪装因子估计的实验结果**

**Table 4 Performance on estimation of disguising factor using ASV and symmetric transform on VoxCeleb1 with frequency-domain disguise**

| 参数 | 伪装因子估计结果 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $s$ | −11 | −10 | −9 | −8 | −7 | −6 | −5 | −4 | −3 |
| $s'_{\text{mean}}(s)$ | −9.40 | −9.12 | −8.61 | −7.46 | −6.27 | −5.97 | −5.10 | −4.29 | −3.30 |
| $E_{\text{mean}}(s)$ | 1.6 | 0.88 | 0.39 | 0.54 | 0.73 | 0.03 | 0.1 | 0.29 | 0.3 |
| $(E_{\text{mean}}(s)/s)/\%$ | 14.55 | 8.80 | 4.33 | 6.75 | 10.43 | 0.50 | 2.00 | 7.25 | 10.00 |
| $\text{Var}(s)$ | 30.54 | 21.17 | 10.72 | 14.11 | 17.66 | 5.85 | 2.73 | 2.43 | 1.85 |
| $s$ | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| $s'_{\text{mean}}(s)$ | 3.18 | 4.09 | 5.10 | 5.97 | 7.07 | 8.09 | 8.94 | 9.99 | 10.71 |
| $E_{\text{mean}}(s)$ | 0.18 | 0.09 | 0.10 | 0.03 | 0.07 | 0.09 | 0.06 | 0.01 | 0.29 |
| $(E_{\text{mean}}(s)/s)/\%$ | 6.00 | 2.25 | 2.00 | 0.50 | 1.00 | 1.13 | 0.67 | 0.10 | 2.64 |
| $\text{Var}(s)$ | 0.23 | 0.48 | 0.51 | 0.41 | 0.55 | 0.74 | 0.68 | 0.47 | 0.25 |

**表5 VoxCeleb1时域伪装数据集上利用对称变换改进伪装因子估计的实验结果**

**Table 5 Performance on estimation of disguising factor using ASV and symmetric transform on VoxCeleb1 with time-domain disguise**

| 参数 | 伪装因子估计结果 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $s$ | −11 | −10 | −9 | −8 | −7 | −6 | −5 | −4 | −3 |
| $s'_{\text{mean}}(s)$ | −10.06 | −9.81 | −8.95 | −7.64 | −7.00 | −5.87 | −4.96 | −4.44 | −3.34 |
| $E_{\text{mean}}(s)$ | 0.94 | 0.19 | 0.05 | 0.36 | 0 | 0.13 | 0.04 | 0.44 | 0.34 |
| $(E_{\text{mean}}(s)/s)/\%$ | 8.55 | 1.90 | 0.56 | 4.50 | 0.00 | 2.17 | 0.80 | 11.00 | 11.33 |
| $\text{Var}(s)$ | 18.54 | 8.69 | 8.05 | 11.01 | 3.58 | 5.81 | 5.22 | 0.43 | 1.88 |
| $s$ | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| $s'_{\text{mean}}(s)$ | 3.18 | 4.09 | 5.10 | 5.97 | 7.07 | 8.09 | 8.94 | 9.99 | 10.71 |
| $E_{\text{mean}}(s)$ | 0.18 | 0.09 | 0.10 | 0.03 | 0.07 | 0.09 | 0.06 | 0.01 | 0.29 |
| $(E_{\text{mean}}(s)/s)/\%$ | 6.00 | 2.25 | 2.00 | 0.50 | 1.00 | 1.13 | 0.67 | 0.10 | 2.64 |
| $\text{Var}(s)$ | 0.23 | 0.48 | 0.51 | 0.41 | 0.55 | 0.74 | 0.68 | 0.47 | 0.25 |

对上述3种伪装因子估计方法的结果进行综合比较,可以看出,本文利用说话人确认估计伪装因子的错误率明显低于基线系统,对于降调电子伪装语音的估计结果与理论值误差很小,但对于升调电子

伪装语音效果略差。改进后的基于对称变换的伪装因子估计方法的误差对于升调电子伪装语音保持在较低的水平,对于降调电子伪装语音误差增大,但总体伪装因子估计均值明显优于基线系统,说明本文提出的利用伪装因子对称变换改进的基于说话人确认的伪装因子估计方法是有效的。

此外,本文中的i-vector自动说话人确认模型是在干净语音库TIMIT上训练的,而测试集是含噪语音库Voxceleb1。在训练集和测试集噪声条件不匹配的情况下,基于说话人确认的伪装因子估计方法的实验效果仍明显优于基线系统。因此,本文改进的电子伪装语音还原方法不仅具有噪声鲁棒性,还具有较好的泛化性能。

### 4.3 说话人确认性能

伪装因子的估计过程本质上就是电子伪装语音的还原过程,得到了伪装因子,就能相应地得到还原语音。利用基频比方法估计得到伪装因子后,对电子伪装语音的基频进行逆变换,就可以对应得到基于基频比方法的还原语音;利用本文提出的基于说话人确认的伪装因子估计方法得到伪装因子后,相应地也可以从$N$句预还原语音中找出正确的还原语音。基于不同的伪装因子估计方法,可以将电子伪装语音还原方法分为以下几种:基于基频比的基线还原方法、基于说话人确认的还原方法以及利用对称变换改进的基于说话人确认的还原方法。

为了进一步评测语音还原的效果,引入了说话人确认中的EER作为另一种客观指标。不同的电子伪装语音还原方法的说话人确认系统性能如表6,7所示。从表中可以看出,电子伪装语音对说话人确认系统影响很大,不采取任何还原措施的电子伪装语音说话人确认系统EER高达40%以上。基于基频比的还原方法得到的还原语音说话人确认性能得到一定的改善,但EER仍高于24%。基于说话人确认的还原方法和利用对称变换改进的基于说话人确认的还原方法得到的还原语音说话人确认性能得到明显改善,EER低于20%。

**表6 频域伪装数据集上不同还原方法得到的还原语音说话人确认性能对比**

**Table 6 Comparison of recognition performance of restored speech speakers using different methods on frequency-domain disguise**

| 参数 | 伪装因子估计结果 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $s$ | −11 | −10 | −9 | −8 | −7 | −6 | −5 | −4 | −3 |
| 未还原语音 | 0.51 | 0.51 | 0.50 | 0.54 | 0.56 | 0.54 | 0.51 | 0.49 | 0.43 |
| 基频比还原法 | 0.29 | 0.28 | 0.28 | 0.29 | 0.30 | 0.29 | 0.31 | 0.33 | 0.29 |
| ASV还原法 | 0.09 | 0.07 | 0.07 | 0.08 | 0.08 | 0.07 | 0.08 | 0.08 | 0.08 |
| 对称ASV还原法 | 0.08 | 0.07 | 0.08 | 0.06 | 0.08 | 0.07 | 0.07 | 0.08 | 0.09 |
| $s$ | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| 未还原语音 | 0.41 | 0.49 | 0.56 | 0.57 | 0.55 | 0.52 | 0.51 | 0.53 | 0.52 |
| 基频比还原法 | 0.26 | 0.27 | 0.28 | 0.34 | 0.35 | 0.34 | 0.34 | 0.32 | 0.31 |
| ASV还原法 | 0.12 | 0.12 | 0.13 | 0.13 | 0.14 | 0.18 | 0.21 | 0.22 | 0.22 |
| 对称ASV还原法 | 0.15 | 0.12 | 0.13 | 0.14 | 0.14 | 0.18 | 0.18 | 0.19 | 0.18 |

利用对称变换改进的基于说话人确认的还原方法对于升调伪装与降调伪装的伪装因子估计准确率类似,但对于升调伪装的EER尚存在差距,值得进一步探索其原因。

## 5 结束语

语音技术的发展给人们带来了极大的便利,然而电子伪装语音技术的出现给说话人识别带来了极

**表7 时域伪装数据集上不同还原方法得到的还原语音说话人确认性能对比**

**Table 7 Comparison of recognition performance of restored speech speakers using different methods on time-domain disguise**

| 参数 | 伪装因子估计结果 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $s$ | −11 | −10 | −9 | −8 | −7 | −6 | −5 | −4 | −3 |
| 未还原语音 | 0.52 | 0.52 | 0.53 | 0.56 | 0.57 | 0.54 | 0.51 | 0.48 | 0.43 |
| 基频比还原法 | 0.24 | 0.26 | 0.27 | 0.28 | 0.30 | 0.29 | 0.30 | 0.33 | 0.29 |
| ASV还原法 | 0.09 | 0.07 | 0.08 | 0.08 | 0.08 | 0.07 | 0.08 | 0.08 | 0.08 |
| 对称ASV还原法 | 0.09 | 0.07 | 0.07 | 0.07 | 0.07 | 0.07 | 0.07 | 0.08 | 0.07 |
| $s$ | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| 未还原语音 | 0.40 | 0.49 | 0.55 | 0.58 | 0.56 | 0.53 | 0.52 | 0.53 | 0.53 |
| 基频比还原法 | 0.27 | 0.28 | 0.28 | 0.32 | 0.35 | 0.36 | 0.34 | 0.32 | 0.30 |
| ASV还原法 | 0.12 | 0.12 | 0.12 | 0.14 | 0.14 | 0.17 | 0.18 | 0.20 | 0.20 |
| 对称ASV还原法 | 0.15 | 0.12 | 0.13 | 0.14 | 0.14 | 0.18 | 0.18 | 0.19 | 0.18 |

大挑战,电子伪装语音的身份识别成为目前语音处理和信息安全领域非常有实用意义的研究问题。本文针对当前伪装程度估计方法在真实含噪数据集上不理想的问题,提出了一种基于对称变换和ASV的电子伪装语音还原方法,能够有效估计含噪电子伪装语音的伪装因子,错误率仅为4.49%,明显低于利用基频比确定伪装因子的方法,为深入开展电子伪装语音的说话人身份识别任务奠定了基础。

**参考文献:**


[1] PERROT P, AVERSANO G, CHOLLET G. Voice disguise and automatic detection: Review and perspectives[C]//Proceedings of Progress in Nonlinear Speech Processing. Heidelberg, Berlin: Springer, 2007: 101 - 117.

[2] 张雄伟, 苗晓孔, 曾歆, 等. 语音转换技术研究现状及展望[J]. 数据采集与处理, 2019, 34(5): 753-770.
ZHANG Xiongwei, MIAO Xiaokong, ZENG Xin, et al. Voice conversion: The state of the art and prospects[J]. Journal of Data Acquisition and Processing, 2019, 34(5): 753-770.

[3] 王永全, 施正昱, 张晓. 基于DC-CNN的电子伪装语音还原研究[J]. 计算机科学, 2019, 46(8): 183-188.
WANG Yongquan, SHI Zhengyu, ZHANG Xiao. Study on restoration of electronic disguised voice based on DC-CNN[J]. Computer Science, 2019, 46(8): 183-188.

[4] 张桂清, 金怡珠, 刘红伟, 等. 电子伪装语音的变声规律研究[J]. 证据科学, 2010, 18(4): 503-509.
ZHANG Guiqing, JIN Yizhu, LIU Hongwei, et al. Study on changing rules of electronic disguised voice[J]. Evidence Science, 2010, 18(4): 503-509.

[5] WU H, WANG Y, HUANG J. Identification of electronic disguised voices[J]. IEEE Transactions of Information Forensics and Security, 2014, 9(3): 489-500.

[6] WANG Y, SU Z, Detection of voice transformation spoofing based on dense convolutional network[C]//Proceedings of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, United Kingdom: IEEE, 2019: 2587-2591.

[7] SRIVASTAVA B M L, VAUQUIER N, SAHIDULLAH M, et al. Evaluating voice conversion-based privacy protection against informed attackers[C]//Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain: IEEE, 2020: 2802-2806.

[8] FARRÚS M. Voice disguise in automatic speaker recognition[J]. ACM Computing Surveys, 2018, 51(4): 1-22.

[9] TAN T. The effect of voice disguise on automatic speaker recognition[C]//Proceedings of IEEE International Congress on Image and Signal Processing (CISP'10). Yantai, China: IEEE, 2010: 3538-3541.

[10] WANG Y, WU H, HUANG J. Verification of hidden speaker behind transformation disguised voices[J]. Digital Signal Processing, 2015, 45: 84-95.

[11] PERROT P, CHOLLET G. The question of disguised voice[J]. Journal of the Acoustical Society of America, 2008, 123(5): 3878.

[12] 李燕萍, 陶定元, 林乐. 基于DTW模型补偿的伪装语音说话人识别研究[J]. 计算机技术与发展, 2017, 27(1): 93-96.
LI Yanping, TAO Dingyuan, LIN Le. Study on electronic disguised voice speaker recognition based on DTW model compensation[J]. Computer Technology and Development, 2017, 27(1): 93-96.

[13] NAGRANI A, CHUNG J S, ZISSERMAN A. Voxceleb: A large-scale speaker identification dataset[C]//Proceedings of Annual Conference of the International Speech Communication Association. Stockholm, Sweden: [s.n.], 2017: 20-24.

[14] LIANG H, LIN X, ZHANG Q, et al. Recognition of spoofed voice using convolutional neural networks[C]//Proceedings of 2017 IEEE Global Conference on Signal and Information Processing. Montreal, Canada: IEEE, 2017: 293-297.

[15] JANG K Y, KIM J J, BAE M J, Pitch alteration technique in a speech synthesis system[C]//Proceedings of 2000 Digest of Technical Papers. International Conference on Consumer Electronics. Los Angles, USA: [s.n.], 2000: 332-333.

[16] WU H, WANG Y, HUANG J, Blind detection of electronic disguised voice[C]//Proceedings of 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Vancouver, Canada: IEEE, 2013: 3013-3017.

[17] 陶定元. 电子伪装语音下的说话人识别方法研究[D]. 南京:南京邮电大学, 2016.
TAO Dingyuan. Study on speaker recognition under electronic disguised voices[D]. Nanjing: Nanjing University of Posts and Telecommunications, 2016.

[18] REYNOLDS D A, QUATIERI T F, DUNN R B. Speaker verification using adapted Gaussian mixture models[J]. Digital Signal Processing, 2000, 10(1/2/3): 19-41.

[19] SREENIVAS S T, SEYED R S, ABHIMANYU S G, et al. Speaker identification features extraction methods: A systematic review[J]. Expert Systems Applications, 2017, 90(1): 250-271.

[20] DEHAK N, KENNY P J, DEHAK R, et al. Front-end factor analysis for speaker verification[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2010, 19(4): 788-798.

[21] STURMEL N, DAUDET L. Iterative phase reconstruction of wiener filtered signals[C]//Proceedings of 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Kyoto, Japan: IEEE, 2012: 101-104.

[22] 张雄伟,郑昌艳,曹铁勇,等. 骨导麦克风语音增强技术研究现状及展望[J]. 数据采集与处理, 2018, 33(5): 769-778.
ZHANG Xiongwei, ZHENG Changyan, CAO Tieyong, et al. Blind enhancement of bone-conducted microphone speech: Review and prospects[J]. Journal of Data Acquisition and Processing, 2018, 33(5): 769-778.

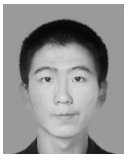
**作者简介:**

**郑琳琳**(1992-),女,硕士研究生,研究方向:语音处理与网络安全,E-mail:zlinlinc21@163.com。

**张雄伟**(1965-),男,教授,研究方向:语音与图像处理、智能信息处理。

**孙蒙**(1984-),**通信作者**,男,副教授,研究方向:智能语音处理、机器学习,E-mail: sunmengccjs@163.com。

**李嘉康**(1993-),男,博士研究生,研究方向:语音处理与网络安全。



**张星昱**(1994-),男,博士研究生,研究方向:语音处理与网络安全。

(编辑:张彤)