基于参数转换的语音深度伪造及其对声纹认证的威胁评估

苗晓孔, 孙 蒙, 张雄伟, 李嘉康, 张星昱

陆军工程大学 指挥控制工程学院 智能信息处理实验室 江苏 南京 210007

摘要 声纹认证系统作为一种生物认证或识别机制, 在人们的日常生活中得已经到了广泛应用。但目前该系统在实际应用中容易受到欺骗攻击, 还存在一定的风险。语音转换通常是指将一个人的声音个性化特征参数通过“修改变换”, 使之听起来像另外一个人的声音, 同时保持说话内容信息不变的技术, 用语音转换可生成特定目标说话人的语音, 并在听觉感知上难以区分转换语音和目标语音。但是对于声纹认证系统来说, 听觉上感知的相似有时还不足以欺骗认证系统。本文通过分析语音转换和声纹认证过程中所提取共同特征向量——梅尔倒谱, 通过采用改进深度残差的双向长短时记忆网络对联合动态特征的梅尔倒谱实现更准确转换, 同时改变损失函数优化转换网络性能并引入全局均值滤波滤除转换过程中产生的倒谱杂波, 进而整体提升转换语音的质量。在提升语音转换相似度的同时保证主观感知不下降, 并将转换后的语音用于欺骗两个广为采用的声纹认证系统, 欺骗实验表明, 该系统能够成功地欺骗这些认证系统, 并且具有很高的成功率。

关键词 语音转换, 声纹认证, 对抗攻击, 深度学习
中图法分类号 TP391.9 **DOI 号** 10.19363/J.cnki.cn10-1380/tn.2020.11.05

Deep Speech Forgery Based on Parameter Transformation and Threat Assessment to Voiceprint Authentication

MIAO Xiaokong, SUN Meng, ZHANG Xiongwei, LI Jiakang, ZHANG Xingyu

College of Command and Control Engineering Intelligent Information Processing Laboratory, Army Engineering University, Nanjing 210007, China

Abstract Automatic Speaker verification (ASV) system, as a biometric authentication or recognition mechanism, has been widely used in people's daily life. However, the system is vulnerable to deception attack in practical application, and the system also faces different potential risks. Voice conversion (VC) usually refers to the technology of "modifying and transforming" a person's voice characteristics to make it sound like another person's voice, while keeping the speech content information unchanged. VC could generate the voice of a specific target speaker, and it is difficult to distinguish the converted voice and the target voice in auditory perception. But for the speaker verification system, the auditory similarity is not enough to cheat the authentication system. This paper analyzes Mel cepstrum, a common feature vector extracted in speech conversion and speaker verification, and realizes more accurate conversion of Mel cepstrum with joint dynamic features by using a two-way long and short-time memory network with improved depth residuals. At the same time, the loss function is changed to optimize the performance of the conversion network and the global mean filter is introduced to filter out the cepstrum clutter generated in the conversion process and improve the quality of the converted voice as a whole. At the same time, the similarity of speech conversion is improved and the subjective perception is not decreased. And the converted voice is used to cheat two different speaker verification systems. Experiments show that the system can successfully cheat these authentication systems, and has a high success rate.

Key words voice conversion, voiceprint authentication, anti-attack, deep learning

1 引言

声纹认证系统是指用于验证一句语音是否属于说话人本身, 进而对说话人身份进行确认的系统[1]。经历了几十年的发展研究, 目前这一生物特征识别技术已被广泛应用到实际生活中, 例如: 智能手机解锁、智能家居调用、企业安防系统开发等 [2-6]。传统的声纹认证系统通常使用的是高斯混合模型和通

通讯作者: 孙蒙, 博士, 副教授, Email:sunmengccjs@163.com。
本课题得到江苏省自然科学基金(No.BK20180080)资助。
收稿日期: 2019-12-30; 修改日期: 2020-03-05; 定稿日期: 2020-09-22

用背景模型, 随着能够将高维统计信息从通用背景模型映射到低维表示的 i-vector 技术的提出, 基于 i-vector 的声纹认证系统得到了进一步发展。近些年, 随着深度学习技术的不断发展, 基于深度学习方法的声纹认证系统无论是在认证准确性还是在认证复杂度上相较于传统方法都有了较大的进步。虽然声纹认证技术在不断的进步和发展, 但是随着技术的进步, 这些系统所面临来自未经授权的欺骗性访问的风险和威胁也逐步增多。一般来说欺骗性访问的攻击类型主要分为两大类: 一种是物理欺骗, 包括模拟, 重放; 另一种是逻辑欺骗, 主要包括语音转换和语音的合成[1]。而在这些欺骗类型中, 以语音转换和语音合成的逻辑欺骗技术进展最为迅速, 且其具有很强的可塑性和适应性, 因此也是当前声纹认证系统所面临的主要挑战之一。

本文主要针对语音转换技术对声纹认证系统的欺骗效果展开研究。语音转换研究的相关工作最早可追溯到 20 世纪六七十年代, 至今已经有五十多年的研究历史。语音转换的方法也由最开始的高斯混合模型[7-8]、频率弯曲[9-11]、非负矩阵分解[12-13]等, 逐步发展到现在以不同转换网络为主的基于深度学习的语音转换方法, 不同的网络模型, 如: 双向长短时记忆网络(Bidirectional Long Short-Term Memory, BLSTM), 全卷积神经网络(Fully Convolutional Network, FCN)和生成对抗网络(Generative Adversarial Network, GAN)等。但当前的语音转换或语音合成技术, 主要是以提高人类听觉感知质量和相似度为主, 使得转换或合成的语音与目标语音听起来更加相近。但听觉上相似的语音有时还不足以欺骗声纹认证系统, 因为声纹认证系统的“感知”方式与人们听觉感知存在很大区别[14]。本文从声纹认证系统和语音转换系统过程中提取共同的中间特征入手, 通过引入改进的深度残差双向长短时记忆网络对联合动态特征的梅尔倒谱进行转换, 同时采用结构相似性度量作为损失函数进而优化转换网络性能, 并引入全局均值滤波去除转换过程中产生的倒谱杂波, 从而提升转换语音的整体质量。

本文除了对转换语音进行主观测试外, 还将转换语音用于欺骗基于 i-vector[15]和 x-vector[16]这两个目前主流的声纹认证系统。实验结果表明, 基于本文所提出方法进行转换的语音无论是在主观评价得分还是客观攻击测试等方面都具有良好的表现, 且与最近公布的基线相比具有一定程度的提升。

2 相关研究技术

2.1 声纹认证系统

声纹认证系统是通过将说话人的语音生物特征与预先存储的说话人模型进行比较, 从给定的话语中识别出说话人的过程[17]。

声纹认证系统的简要实现框架如图 1 所示。从图中可以看出, 该认证系统主要分为注册阶段和匹配阶段两个部分。注册阶段对预处理后的语音进行特征提取, 然后训练出一个能够识别不同说话人特征之间的模型; 测试阶段则是用训练好的模型对新输入的语音进行验证。如果测试语音与注册语音特征相匹配则接受身份验证, 反之则拒绝身份验证。近些年, 多项研究结果表明, 在提取的特征中, 基于梅尔倒谱的特征提取方法比其他方法得到了更多的应用。此外, 其他梅尔倒谱变体以及梅尔倒谱融合的方法也备受关注。

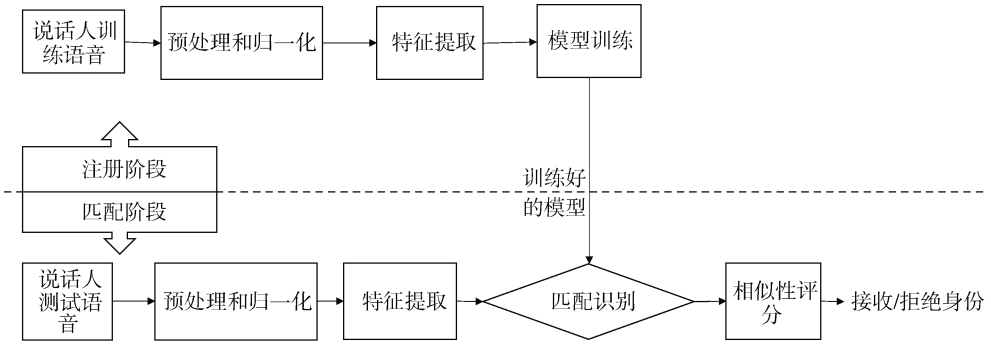

图 1 声纹认证系统框图

Figure 1 The block diagram of speaker verification system

2.2 BLSTM 语音转换技术

BLSTM 是双向递归神经网络(Recurrent neural network, RNN)的一种改进, 它可以用循环连接对一定数量的上下文信息进行建模, 原则上来说, RNN 网络能够将以前输入的整个历史逐帧映射到输出。然而, 在双向 RNN 网络的误差反向传播和参数优化

过程中, 当对长距离上下文进行传输时, 反向传播梯度的累积或衰减会随着时间爆炸或消失。为了有效解决这一问题, BLSTM 引入了长期和短期记忆结构, 这种结构可以在线性存储单元中存储多个时间步骤的信息, 并且可以学习与回归任务相关的最佳上下文信息[18]。由于 BLSTM 是由许多长短期记忆细胞和双 RNN 组成, 所以其能够同时考虑前向和后向序列信息, 因此采用 BLSTM 对语音信息进行处理, 可以更加有效利用声学特征的前后相关性, 实现更好的声学特征处理。

图 2 展示的是基于 BLSTM 的语音转换框图。可以看出, 通常的语音转换系统包含了训练和转换两个阶段。训练阶段, 首先对源说话人和目标说话人的语音进行特征参数提取, 然后对提取到的特征(非周期成分, 基音周期, 声道谱)进行预处理, 对源和目标的声道谱进行动态时域规整(Dynamic Time Warping, DTW), 最后将这些对齐的声道谱特征送入模型进行训练, 得到最终的转换模型。转换阶段, 对待转换源语音进行分析和特征参数提取, 然后用训练阶段获取的转换模型对声道谱进行特征转换, 最后将转换后的特征用于语音合成得到转换语音。这种方法可以得到较高质量的转换语音。

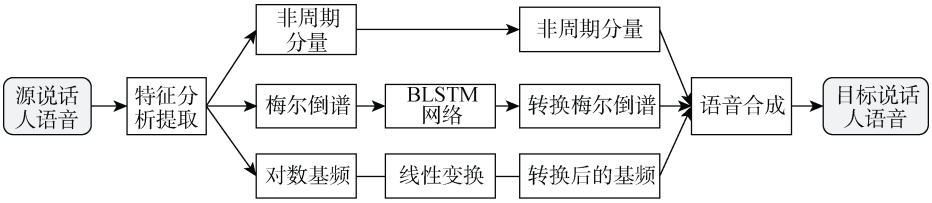

图 2 基于 BSLTM 的语音转换框图

Figure 2 The VC block diagram based BLSTM

3 本文所提方法

3.1 总体框架

本文提出的语音转换方案的具体流程如图 3 所示, 其中的深色模块为本方法的主要改进和创新点。

由图 3 可以看出, 语音转换通常分为训练和转换两个阶段。训练阶段主要通过对源说话人和目标说话人的语音进行特征参数提取, 然后对提取到的特征(非周期成分、基音周期、梅尔倒谱)进行预处理。

在预处理阶段本文将源语音的梅尔倒谱的一阶差分动态特征作为拓展维度与原始的梅尔谱融合再送入转换模型, 而目标语音的只提取了其梅尔倒谱。然后对源和目标的梅尔倒谱进行 DTW, 最后将这些对齐的梅尔倒谱特征送入改进后的 BLSTM 模型进

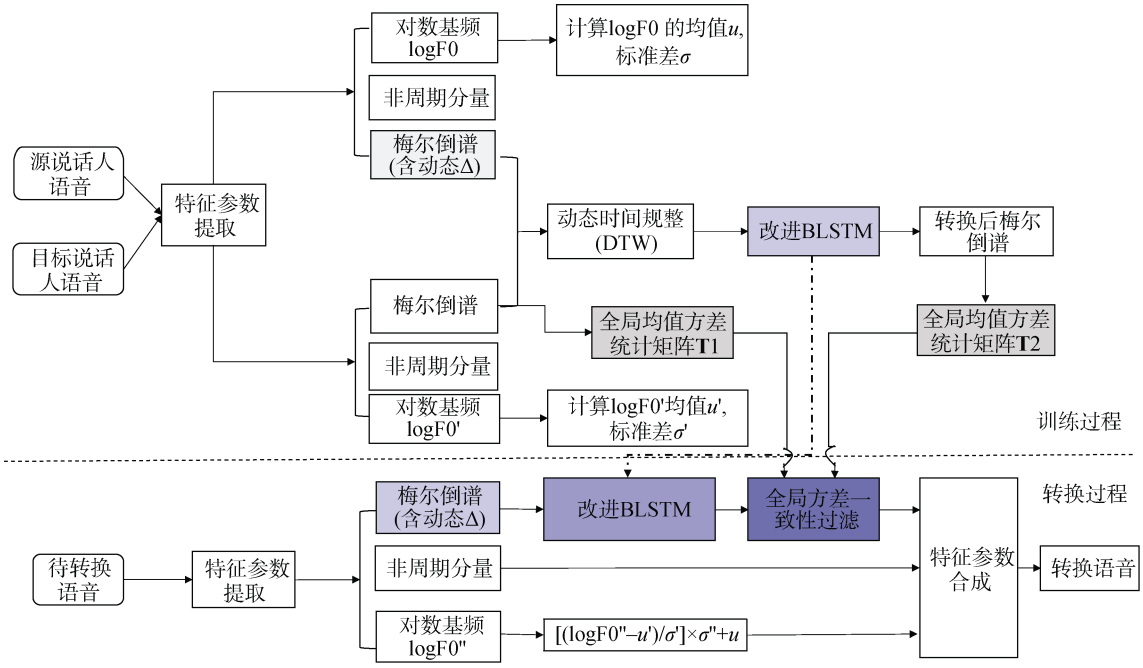

图 3 本文所提语音转换方案框图

Figure 3 The flow chart of proposed VC scheme

行训练, 得到最终的转换映射函数。转换阶段, 对待转换源语音进行特征参数提取, 然后用训练阶段得到的转换映射函数对梅尔倒谱进行特征转换, 最后将转换后的特征用于语音合成得到转换语音。

3.1.1 网络结构的改进

传统的 BLSTM 网络多采用直联多层的网络结构来构建模型, 本文将残差网络引入其中, 能够解决网络退化和梯度消失等问题, 同时在构建多层残差的 BLSTM 网络基础上, 本文还将隐藏层的非线性激活函数 ReLu 改为性能较优得到 PReLu[19], 具体的网络结构层如图 4 所示。从图中可以看出, 除了相邻的每一层之间构建了残差连接外, 还对非相邻层进行了残差连接, 进一步改进在误差反向传播过程中的第一层网络参数的调整和优化。

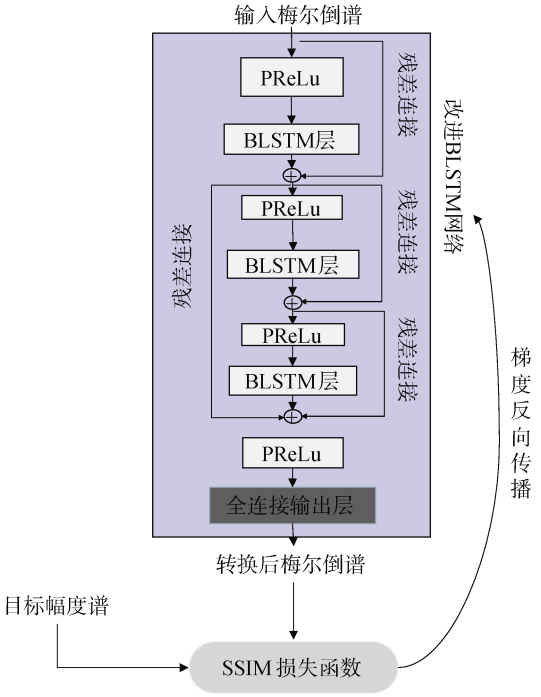

图 4 改进深度 BLSTM 网络结构

Figure 4 The improved deep BLSTM network structure

图 4 中输出层采用 ReLu 是考虑到后面损失函数的输入变量的非负性, 下文将对损失函数详细介绍。

3.1.2 SSIM 函数构建

SSIM 损失函数最初是图像中用于度量两幅图像结构相似性的指标。它能够很好地刻画图像像素之间的时空关系, 相比于最小均方误差更符合人类的视觉评价机制。本文将其应用于语音中的梅尔倒谱图, 其目的在于能够产生出更符合人类听觉的转换音频, 同时提高转换语音的客观评价指标。所以本文在网络的损失函数上采用 SSIM 代替原有的均方误差(Mean-square error, MSE)损失函数的方法。

假设 X 和 Y 分别表示输入源语音和目标语音的梅尔倒谱, 其转换映射函数为 f, 则转换语音梅尔倒谱 $\hat{Y}$ 可表示为:

$$\hat{Y}=f\left(X\right), \tag{1}$$

则 $\hat{Y}$ 与 Y 之间的 SSIM 可表示为:

$$SSIM\left(\hat{Y},Y\right)=\frac{1}{M}\sum_{i=1}^{M}SSIM\left(x_i,\hat{y}_i\right), \tag{2}$$

上式中, M 表示共包含的局部梅尔倒谱图的块个数, 图像块分别以 x_i 和 $\hat{y}_i$ 为像素中心, i 表示像素块的索引。因为 SSIM 度量要求信号是非负的, 所以在 BLSTM 网络的最后一层输出时, 为了保证信号的非负性, 采用了 ReLu 的非线性激活单元。SSIM 的评价依据是其数值越高, 表示转换的梅尔倒谱越接近目标梅尔倒谱, 而由于损失函数越小越好, 所以以 SSIM 作为损失函数可以表示如下:

$$J_{SSIM}\left(\theta\right)=-\frac{1}{N}\sum_{n=1}^{N}SSIM\left(\hat{Y}_n,\ Y_n\right), \tag{3}$$

上式中, N 表示训练样本总数量。n 表示数量索引。通过最小化 J 实现网络参数的更新。θ 表示网络参数。

3.2 全局方差一致性滤波

滤波是信号处理中一种常见的去噪方法, 如果信号本身含有噪声, 滤波是可以在一定程度上提升转换语音的质量, 但如果在网络转换过程中, 由于转换特征参量的不精确, 会造成后期语音参数化合成中的“噪声”则难以被滤除。因此, 本文将全局方差一致性滤波模型引入到 BLSTM 转换网络中, 通过统计训练阶段转换语音和目标语音的均值方差, 构建一个统计滤波。然后将其应用于转换阶段的后处理, 对转换后的梅尔倒谱进行滤波, 有效去除转换过程中产生的梅尔倒谱“杂波”, 通过这种全局方差一致性滤波, 有效提升了转换的语音质量。

全局方差一致性滤波的公式及步骤如下:

1)计算目标语句的每一维梅尔谱的均值方差, 公式如(1)所示

$$\begin{cases}\overline{x}_{i_{\text{tar}}}=1/(N\times M)\sum_{n=1}^{N}\sum_{m=1}^{M}x_i\\ \sigma^2{}_{i_{\text{tar}}}=1/(N\times M)\sum_{n=1}^{N}\sum_{m=1}^{M}\left(x_i-\overline{x}_{i_{\text{tar}}}\right)^2\end{cases}(i=1,2,3,\cdots,T), \tag{4}$$

其中, N 表示训练阶段目标语句的数量, M 表示每个语句包含的帧数, T 表示梅尔谱的维度, i 表示第 i 维

梅尔谱, 求出所有训练语句所有帧的各维度梅尔谱均值 $\overline{x}_{i_{\text{tar}}}$ 和方差 $\sigma^2{}_{i_{\text{tar}}}$。tar 表示 target 目标语句。

2) 利用公式(4)计算训练阶段, 所有转换语句所有帧的各维度梅尔谱均值 $\overline{x}_{i_{\text{con}}}$ 和方差 $\sigma^2{}_{i_{\text{con}}}$, 以及待转换语句所有帧的各维度梅尔谱均值所构成的向量 $\overline{y}$ 和方差所构成的向量 $\sigma_y{}^2$, 其中 con 表示转换语句。

3) 构造全局均方差一致滤波器, 如公式(5)所示, 得到初级滤波后数据

$$\hat{y}=\sqrt{\left(\overline{x_{\text{tar}}}/\overline{x_{\text{con}}}\right)}\times\left(y-\overline{y}\right)+\overline{y}, \tag{5}$$

公式(5)中 $\overline{x_{\text{tar}}}$ 表示由目标语句各维度梅尔谱均值构成的向量 $\overline{x_{\text{con}}}$ 表示由转换语句各维度梅尔谱均值构成的向量, y 表示待转换语句的梅尔谱向量。

4) 根据公式(6)设置参数 α, 调整得到最后的滤波数据

$$\hat{y}=\alpha\times\hat{y}+\left(1\text{-}\alpha\right)\times y \tag{6}$$

$\hat{y}$ 是最后经过滤波后, 最终得到的待转语音梅尔谱。

4 实验及分析

4.1 评价指标

转换语音的评价指标主要有客观和主观两方面。客观评价指标主要为: MSE、谱失真(Spectral Distortion, SD)和梅尔倒谱失真(Mel Cepstral Distortion, MCD), MSE、SD 和 MCD 的值越小, 说明失真越小, 转换精度越高。

主观评价是以人为主体, 通过人的主观感受来对语音进行测试。由于语音信号最终是用来给人聆听的, 因而人对语音转换效果好坏的感受是最为重要的评价结果。相对于客观评价来说, 主观评价结果更具有可信度。主观评价一般从转换语音质量和转换语音与目标语音相似度两个方面进行衡量, 通常采用的方法主要是平均意见分(Mean Opinion Score, MOS)和 ABX 测试。

本实验主要目的是测试转换语音对目前声纹认证系统的欺骗效果, 所以在客观指标上以欺骗不同声纹认证系统的成功次数和认证系统给出的相似得分来衡量转换语音的质量, 主观指标方面则主要选取了 MOS 指标来主观评价转换语音质量。MOS 测试的主要原理是让测评人根据 5 个等级划分对测试语音的主观感受进行打分, 它既可以用于对语音质量进行主观评价, 也可以用于对说话人相似度的评价。MOS 分是对所有测试语句和所有测评人的综合平均结果, 得分越高说明转换质量越好, 与目标语音相似度越高。其具体评测标准可参考文献[20]。

4.2 实验设置及结果

为了更好的衡量出本文所提方法的优越性, 实验中还选取了两种不同的对比方法。一种是语音转换挑战赛 2018 年中作为排名第二的 GMM 语音转换方法[21], 一种是传统的 BLSTM 网络实现的语音转换方法[22]。实验数据集选取的是公开的 CMU ARCTIC 数据集[23], 并选取了跨性别男到女(Male to Female, M-F)之间的转换为例进行了欺骗攻击实验。

实验中所选取的认证系统主要是 i-vector 和 x-vector 的认证系统。测试实验主要是用不同转换系统所得到的转换语音作为输入, 目标语音作为注册语音, 分别观察转换语音能够成功欺骗认证系统。如果认证系统能够有效区分出转换语音和目标语音, 则证明转换语音和目标语音虽然听起来相似但是还不足以欺骗认证系统。如果认证系统不能区别转换语音和目标语音, 一方面说明转换语音与目标语音的相似度极高, 另一方面也说明认证系统需要进一步改进。此时以各认证系统判定的综合得分来作为评价转换语音质量好坏的客观指标。

表 1 是随机选取 15 句不同转换方法的转换语音用于欺骗两个不同认证系统的结果统计。通过统计观察可以看出, 除了 BLSTM 系统的转换语音可以被低概率的识别出来外, 其他方法的转换语音均以 100%的欺骗成功率欺骗了两个认证系统。说明目前的认证系统还需要不断完善以应对来自语音转换的潜在威胁攻击。其中源—目标表示源语音和目标语音对。

表 1 不同转换语音对认证系统的欺骗成功概率统计
Table 1 Statistics of success probability of deception of different converted voice to authentication system

转换方法	认证系统	
M-F(*jmk-slt*)	x-vector	i-vector
源-目标	0%	0%
GMM	100%	100%
BLSTM	100%	70%
本文方法	100%	100%

同时为了更加客观反映不同转换方法所转换语音的具体性能, 实验中还统计了两个认证系统对转换语音的评价得分, 综合得分越高说明转换语音和目标语音越相似, 认证系统越难以区分, 具体得分如表 2 所示。由表 2 横向对比可以看出声纹认证方法越先进、越不容易被欺骗; x-vector 与 i-vector 相比多数相同情况下, x-vector 得分更低, 鉴别能力更强

一些。同时纵向对比可以看出语音转换方法越先进, 越容易欺骗别人, 如本文方法比 GMM 和 BLSTM 更容易欺骗过两个声纹认证系统。

结合表 1 和表 2 可以看出本文所提方法能够以 100%的欺骗成功率攻击两个不同的认证系统, 即使与 GMM 转换方法具有同样欺骗成功的概率时, 所表现的综合得分也更高, 说明所提方法转换的语音相较于其他对比方法具有更高的相似性。

表 2 认证系统对不同转换语音的平均得分

Table 2 The average score of authentication system for different converted voice

转换方法	认证系统	
M-F(*jmk-slt*)	x-vector	i-vector
源-目标	–49.89	–138.61
GMM	24.34	56.18
BLSTM	22.14	13.86
本文方法	**26.81**	**62.26**

同时本文也对不同语音转换方法进行了主观 MOS 测试, 其统计结果如图 5 所示。

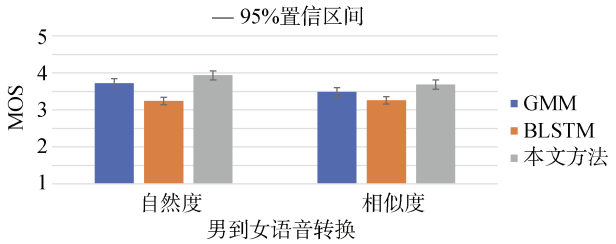

图 5 转换语音置信区间为 95%的自然度和相似度 MOS 得分

Figure 5 MOS score of naturalness and similarity of converted voice with 95% confidence interval

从图 5 可以看出, 本文所提方法转换语音无论是在相似度还是自然度的主观感知上均优于其他两种对比算法。

5 结论

本文提出一种采用改进深度残差的双向长短时记忆网络对联合动态特征的梅尔倒谱进行映射的语音转换方法, 并通过全局方差一致性滤波对转换过程中产生的杂波进行了有效滤波, 提升了语音转换的整体质量。同时通过实验重点研究了转换语音对两个目前先进的声纹认证系统进行欺骗的效果, 实验结果客观反映了本文所提出方法生成的转换语音的相似度和自然度较高, 与其他基线方法对比, 本文所提出方法生成的转换语音质量无论是客观指标还是主观指标均有了明显的提升。同时根据对声纹认证系统的欺骗攻击测试也反映了目前声纹认证系统还存在一定不足, 需要进一步改善。

未来我们将针对语音攻防等问题展开更加深入的研究, 将声纹认证系统与语音转换系统融合到一起, 一方面进一步提升转换语音的质量, 另一方面也提升声纹认证系统的准确性, 使其可以有效抵抗转换语音的欺骗攻击。

参考文献

[1] Yuan M, Duan Z. Spoofing Speaker Verification Systems with Deep Multi-speaker Text-to-speech Synthesis [EB/OL]. 2019: arXiv: 1910.13054.

[2] 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.

[3] Snyder D, Garcia-Romero D, Povey D, et al. Deep Neural Network Embeddings for Text-Independent Speaker Verification[C]. *INTERSPEECH* 2017, 2017: 999-1003.

[4] Snyder D, Garcia-Romero D, Sell G, Povey D, Khudanpur S. X-vectors: Robust DNN embeddings for speaker recognition[C]. *ICASSP 2018*, 2018:5329–5333.

[5] Li C, Ma X, Jiang B, et al. Deep speaker: an end-to-end neural speaker embedding system [EB/OL]. 2017: arXiv: 1705.02304.

[6] Wan L, Wang Q, Papir A, Moreno I L. Generalized end-to-end loss for speaker verification[C]. *ICASSP 2018*, 2018: 4879–4883.

[7] Toda T, Black A W, Tokuda K. Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory[J]. *IEEE Transactions on Audio, Speech and Language Processing*, 2007, 15(8): 2222-2235.

[8] Helander E, Virtanen T, Nurminen J, et al. Voice Conversion Using Partial Least Squares Regression[J]. *IEEE Transactions on Audio, Speech, and Language Processing*, 2010, 18(5): 912-921.

[9] Erro D, Alonso A, Serrano L, et al. Towards Physically Interpretable Parametric Voice Conversion Functions[M]. Advances in Nonlinear Speech Processing. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013: 75-82.

[10] Tian X, Wu Z, Lee S W, Hy N Q, Chng E S, Dong M. Sparse representation for frequency warping-based voice conversion[C]. *ICASSP 2015*, 2015: 4235–4239.

[11] Kawahara H, Masuda-Katsuse I, de Cheveigné A. Restructuring Speech Representations Using a Pitch-adaptive Time–Frequency Smoothing and an Instantaneous-frequency-based F0 Extraction: Possible Role of a Repetitive Structure in Sounds[J]. *Speech*

Communication, 1999, 27(3/4): 187-207.

[12] Takashima R, Takiguchi T, Ariki Y. Exemplar-Based Voice Conversion Using Sparse Representation in Noisy Environments[J]. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, 2013, E96.A(10): 1946-1953.

[13] Wu Z Z, Virtanen T, Chng E S, et al. Exemplar-Based Sparse Representation with Residual Compensation for Voice Conversion[J]. *ACM Transactions on Audio, Speech, and Language Processing*, 2014, 22(10): 1506-1521.

[14] Hansen J H L, Hasan T. Speaker Recognition by Machines and Humans: A Tutorial Review[J]. *IEEE Signal Processing Magazine*, 2015, 32(6): 74-99.

[15] 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*, 2011, 19(4): 788-798.

[16] Li J, Zhang X, Sun M, et al. Attention-Based LSTM Algorithm for Audio Replay Detection in Noisy Environments[J]. *Applied Sciences*, 2019, 9(8): 1539-1550.

[17] Pawar R V, Jalnekar R M, Chitode J S. Review of Various Stages in Speaker Recognition System, Performance Measures and Recognition Toolkits[J]. *Analog Integrated Circuits and Signal Processing*, 2018, 94(2): 247-257.

[18] Martin W, Angeliki M, Nassos K, Björn S, Shrikanth N. Analyzing the memory of BLSTM neural networks for enhanced emotion classification in dyadic spoken interactions[C]. *ICASSP 2012*, 2012: 4157-4160.

[19] He K M, Zhang X Y, Ren S Q, et al. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification[EB/OL]. 2015: arXiv:1502.01852.

[20] Shuang Z W, Bakis R, Qin Y. IBM Voice Conversion Systems for 2007 TC-STAR Evaluation[J]. *Tsinghua Science and Technology*, 2008, 13(4): 510-514.

[21] Sun L, Kang S, Li K, Meng H. Voice conversion using deep Bidirectional Long Short-Term Memory based Recurrent Neural Networks[C]. *ICASSP 2015*, 2015:4869–4873.

[22] Lorenzo-Trueba J, Yamagishi J, Toda T, et al. The Voice Conversion Challenge 2018: Promoting Development of Parallel and Nonparallel Methods[C]. *2018 The Speaker and Language Recognition Workshop*, 2018: 26-29.

[23] Kominek J, Black A W. The CMU Arctic speech databases[C]. *Fifth ISCA workshop on speech synthesis*. 2004: 223-224.

苗晓孔 于 2017 年在军械工程学院通信工程专业获硕士学位。现在陆军工程大学网络空间安全专业攻读博士学位。研究领域为智能信息处理。研究兴趣包括: 语音转换。Email: miao_xk@163.com

孙蒙 于 2012 年在比利时鲁汶大学电子系获得博士学位。现任陆军工程大学智能信息处理实验室副教授。研究领域为智能语音处理、机器学习。研究兴趣包括: 语音转换, 语音识别。Email: sunmengccjs@163.com

张雄伟 现任陆军工程大学智能信息处理实验室教授。研究领域为语音与图像处理、智能信息处理。研究兴趣包括: 语音增强, 语音转换。Email: xwzhang9898@163.com

李嘉康 于2017年在陆军工程大学基础部数学专业获得理学硕士学位。现在陆军工程大学网络空间安全专业攻读博士学位。研究领域为智能信息处理。研究兴趣包括: 声纹识别。Email: jkangli@163.com

张星昱 于 2019 年在陆军工程大学电子信息工程专业获工学硕士学位, 现在陆军工程大学网络空间安全专业攻读博士学位。研究领域为信息内容安全, 研究兴趣包括: 说话人识别, 对抗样本。Email: zxynbnb@126.com