基于MLP的伪装语音说话人性别鉴定

张晓, 管林玉

引用本文

张晓, 管林玉. 基于MLP的伪装语音说话人性别鉴定[J]. 计算机科学, 2024, 51(11A): 240400021-4.

ZHANG Xiao, GUAN Linyu. Gender Recognition of Electronic Disguised Voices Based on MLP[J]. Computer Science, 2024, 51(11A): 240400021-4.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240400021

基于 MLP 的伪装语音说话人性别鉴定

张 晓 管林玉
公安部第三研究所 上海 201204

摘 要 文中提出了一种基于神经网络的伪装语音说话人识别模型,用以实现从共振峰的中心频率、带宽、音强等参数识别伪装语音说话人的性别。该模型以多层感知机(Multi-Layer Perceptron,MLP)为框架,经全连接的非线性堆叠计算获取识别结果,并在模型的训练阶段采用 L-BFGS 进行优化参数的求解。实验中采用 SoundTouch 对男性和女性的自然语音进行伪装,探讨了网络结构与激活函数对该模型的影响,以及该识别模型对不同电子伪装手段的适应能力。实验结果表明,基于 MLP 的识别模型能高效区分采用不同电子伪装手段伪装后的语音对应的说话人的性别。

关键词:多层感知机;电子伪装语音;性别鉴定;共振峰;说话人
中图分类号 TP391

Gender Recognition of Electronic Disguised Voices Based on MLP

ZHANG Xiao and GUAN Linyu
The Third Research Institute of Public Security,Shanghai 201204,China

Abstract A neural-network-based disguised voices recognition model is proposed to realize the gender identification of the disguised speech speaker from the parameters such as the formant center frequency,bandwidth and intensity of sound. The model uses multi-layer perceptron(MLP) as the framework to obtain the gender recognition results through the fully connected nonlinear stacking calculation,and uses L-BFGS to solve the parameters optimization in training. This paper uses SoundTouch to disguise the original voices of the male and the female respectively,and then linear predictive coding(LPC) extracts various parameters such as the center frequency,bandwidth and sound intensity of the formant,and eliminates the outliers. Then experiment is carried out to explore the influences of network structure and activation function on the model as well as the adaptability of this recognition model to different electronic disguised methods. The experimental results show that the MLP-based recognition model can effectively distinguish the gender of the speaker corresponding to the voice disguised by different methods. This laid the foundation for electronic disguised voice speaker recognition.

Keywords Multi-layer perceptron(MLP),Electronic disguised voice,Gender recognition,Formant,Speaker

1 引言

语音识别是司法鉴定中的一个重要领域。语音信号是非平稳随机过程,其特性是随着时间变化的,且存在诸多干扰因素对语音的说话人身份判断的准确性造成影响。Kinnunen等[1]在 2010 年的研究表明,通信场景与传输信道会影响语音信号,而语音特征也会随着说话者的情绪、年龄、身体状况等发生变化。而现有的语音自动识别系统虽然已经能够处理声道变化所带来的影响,但是对于伪装语音的说话者识别仍存在缺陷。

自 20 世纪 70 年代起,已涌现出大量有关伪装语音的研究。伪装语音是一种经人为扭曲、模糊处理后的语音,伪装方式可以分为物理伪装与电子伪装[2-3]。相较于物理伪装后的语音,电子伪装语音与原始语音的相似性更小。研究表明,经伪装的语音的频谱与其对应的原始频谱有着较大的差别,频谱识别的错误率大大增加[4]。通过机器识别伪装语音说话人的准确率近乎随机选择的概率。而在语音自动识别系统中,电子伪装语音的说话人识别错误率高于 40%。随着声音转换技术的普及与发展,电子伪装语音一旦被不法分子所利用,后果将十分严重。电子伪装语音的说话人身份鉴别已经成为了当前语音识别的关键问题,而随着人工智能的不断发展,近年来涌现出诸多采用机器学习与深度学习等方法来提取电子伪装语音的特征以及判断语音是否经过电子伪装。然而,对于电子伪装语音的说话人的身份识别仍然没有显著突破。为此,本文提出了一种多层感知机的识别模型来判断电子伪装语音说话人的性别,并以音调(Pitch)、节拍(Tempo)和速度(Rate) 3 种不同伪装方式下生成的电子伪装语音为例,进行了说话人性别识别的实验。实验结果表明,该模型不仅能在较低的时间复杂度和空间复杂度下实现十分精确的电子伪装语音说话人性别识别,AUC 最高可达 97.89%。此外,经过

基金项目:国家重点研发计划(2021YFC3320105);教育部人文社会科学研究项目(23YJA820015)
This work was supported by the National Key Research and Development Program of China(2021YFC3320105) and Program for the Humanities and Social Science of Ministry of Education of China(23YJA820015).
通信作者:张晓(526993512@qq.com)

检验，该模型还适用于多种不同的电子伪装方式所伪装的语音的说话人性别识别，这为电子伪装语音识别奠定了基础。

2 电子伪装语音性别识别

伪装后的语音会失真且会改变音调、共振峰中心频率、带宽等参数数值。然而，不论语音是否经过电子伪装，不同说话人的语音的声学参数均具有较大的差异。Hautamaki 等发现，男女性的语音依旧存在明显差异；Zhang 等[5]的研究也证实了伪装后男女的声纹鉴定有较大的偏差，但两性的语音参数确实存在差异。因此，根据声学参数进行电子伪装语音的说话人性别区分是可行的。

相较于传统方法（如 i-vector），Larcher 等的研究证明了机器学习的方法在无伪装的语音识别上更有效。而在区别电子伪装语音与未经伪装的语音的研究中，机器学习与深度学习也发挥了重要作用，例如：采用广义矩估计（Generalized Method of Moments，GMM）进行鉴别的准确率约为 80%；Support Vector Machine（SVM）进一步将误差率降低至 5%～10%；而卷积神经网络（Convolutional Neural Networks，CNN）的引入，使得这一误差率被控制在 5%以内，甚至可达 0.1%。而当前针对电子伪装语音说话人性别的研究大多仍局限于传统的统计方法，并不能实现高效的性别区分。作为模式识别领域中标准的监督学习算法，多层感知机（Multi-Layer Perceptron，MLP）已在语音识别、图像识别等领域有着广泛的应用。通过 MLP 对复杂的电子伪装语音声学参数测量值进行分类，可以有效地改进之前的研究结果。

3 基于 MLP 的识别模型

3.1 MLP 的模型结构

多层感知机（MLP）是一种前馈的人工神经网络，由输入层、隐藏层和输出层构成，隐藏层位于输入层与输出层之间，层数可以不止一层。MLP 的层与层之间是全连接的，即上一层的任何一个神经元与下一层的所有神经元都有连接关系，如图 1 所示。

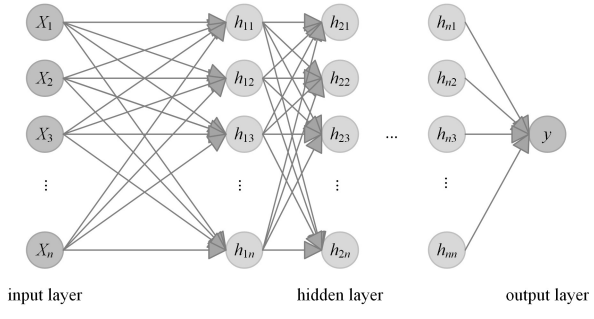

图 1 MLP 结构示意图

Fig. 1 Schematic diagram of MLP structure

假设输入层输入的共振峰参数用向量 $\boldsymbol{x}=(x_1,x_2,\cdots,x_n)^T$ 表示，则隐藏层的输出为：

$$h_i=\begin{cases}f(W_i\boldsymbol{x}+b_i), & i=1\\ f(W_ih_{i-1}+b_i), & i>1\end{cases}$$

其中，W_i 是权重，即连接系数；b_i 是偏置系数，函数 f 是激活函数。当接收到上一神经元传输而来的数据 $\boldsymbol{x}$ 时，该神经元将会对其进行处理，如 $W_i\boldsymbol{x}+b_i$ 的加权运算，生成一个初始的输出数据。而后激活函数 f 将对该初始数据进行转换，生成 $f(W_ix+b_i)$ 或 $f(W_ih_{i-1}+b_i)$ 作为该神经元的最终输出。

3.2 激活函数

在语音信号的建模中，非线性模型的效果比一般的线性模型效果更好，为了使神经网络能更好地解决语音问题，利用激活函数将非线性的因素引入识别模型，使其具备分层的非线性映射学习能力。Bengio 等将激活函数定义为一个处处可微的映射，如式(1)所示：

$$f:R\to R' \tag{1}$$

常见的激活函数有 Sigmoid，tanh 和 ReLU 等，其函数形式如式(2)、式(3)和式(4)所示：

$$\mathrm{Sigmoid}(x)=\frac{1}{1+e^{-x}} \tag{2}$$

$$\tanh(x)=\frac{\sinh(x)}{\cosh(x)}=\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \tag{3}$$

$$\mathrm{ReLU}(x)=\max(0,x) \tag{4}$$

3.3 L-BFGS 优化求解

机器学习中经常利用梯度下降法求最优解问题，通过大量的迭代来得到最优解，但是对于高维度的数据，除了占用大量的内存还会很耗时。在本文的 MLP 识别模型中，采取 L-BFGS 算法作为求取参数的优化算法，该算法优化计算效率高且能适应小样本的数据预测处理。

L-BFGS 算法是在 Newton method 基础上提出的一种求解函数根的算法。在 MLP 识别模型中，需要优化的目标函数如式(5)和式(6)所示：

$$s_k=x_k-x_{k-1} \tag{5}$$

$$t_k=\nabla f(x_k)-\nabla f(x_{k-1}) \tag{6}$$

其中，k 表示迭代次数，x_k 与 x_{k-1} 表示第 k 次和第 $k-1$ 次迭代时的点，s_k 为步长，t_k 为 x_k 与 x_{k-1} 处目标函数的梯度差值。记 ρ_k 与 V_k 如式(7)和式(8)所示：

$$\rho_k=\frac{1}{t_k^{\mathrm{T}}s_k} \tag{7}$$

$$V_k=I-\rho_kt_ks_k^{\mathrm{T}} \tag{8}$$

对于非二次函数，牛顿法并不能保证经过有限次迭代就可以求得最优解。此外，牛顿法需要计算二阶导，即 Hesse 矩阵，因此采用牛顿法求最优解时，不仅计算量巨大，且需要较大的存储空间，而且目标函数的 Hesse 矩阵可能非正定。故而，在 L-BFGS 算法中，用不含二阶导数的矩阵 $\boldsymbol{B}$ 近似替代 Hesse 矩阵，且只保存最近的 m 次迭代信息。假设当前迭代次数为 i，迭代 m 次时，可得 B_i，如式(9)所示：

$$\begin{aligned}B_i=&(V_{i-1}^{\mathrm{T}}\cdots V_{i-m}^{\mathrm{T}})B_i^{0}(V_{i-m}\cdots V_{i-1})+\rho_{i-m}(V_{i-1}^{\mathrm{T}}\cdots\\&V_{i-m+1}^{\mathrm{T}})s_{i-m}s_{i-m}^{\mathrm{T}}(V_{i-m+1}\cdots V_{i-1})+\rho_{i-m+1}(V_{i-1}^{\mathrm{T}}\cdots\\&V_{i-m+2}^{\mathrm{T}})s_{i-m+1}s_{i-m+1}^{\mathrm{T}}(V_{i-m+2}\cdots V_{i-1})+\cdots+\rho_{i-1}s_{i-1}\\&s_{i-1}^{\mathrm{T}}\end{aligned} \tag{9}$$

则第 k 次迭代的可行最优化方向 r_k 为：

$$r_k=-B_k\nabla f(x_k) \tag{10}$$

综上所述，L-BFGS 算法具备牛顿法收敛速度快的特点，但不需要像牛顿法那样存储 Hesse 矩阵，因此节省了大量的空间以及计算资源。经 L-BFGS 算法优化求解可得各层的连接系数 W_i 与偏置系数 b_i。

3.4 识别结果的输出

隐藏层到输出层需对数据进行多分类输出。在 MLP 识

别模型中采用 Softmax 函数对运算后的数据进行离散化分类。Softmax 函数能将一个 N 维的任意实数向量映射为一个各个元素的取值都在(0,1)中的 N 维向量，如式(11)所示：

$$\begin{pmatrix} P(k=1|h) \\ P(k=2|h) \\ \vdots \\ P(k=n|h) \end{pmatrix} = \begin{pmatrix} \zeta(h)_1 \\ \zeta(h)_2 \\ \vdots \\ \zeta(h)_n \end{pmatrix} = \frac{1}{\sum_{k=1}^{n} e^{x_i}} \begin{pmatrix} e^{x_1} \\ e^{x_2} \\ \vdots \\ e^{x_n} \end{pmatrix} \tag{11}$$

所以，输出层的输出结果如式(12)所示：

$$y=\mathrm{Softmax}(f(\boldsymbol{W}_n\boldsymbol{x}+b_n)) \tag{12}$$

4　实验分析

本文采用了 SoundTouch 3 个基本变声功能音调(Pitch)、节拍(Tempo)和速度(Rate)对男女声的自然语音进行了电子伪装。将音频转换为采样频率为 8000 Hz、16 位、单声道的统一格式后，使用 Linear Predictive Coding(LPC)提取各电子伪装语音的前 4 个共振峰 F1－F4 的中心频率、带宽与声强等各项参数指标。提取参数时，限定只过滤提取中心频率在 200 Hz 以上、最大带宽为 2000 Hz，最小频率间隔为 80 Hz 的共振峰。通过预先限定共振峰值的范围，可以消除离群值，使共振峰参数估计更为精确[6]。音频采集的参数配置如表 1 所列。

表 1　音频采集参数配置

Table 1　Parameter configuration of audio acquisition

参数名称	参数配置
加窗方式	Hann
阶数	8
FFT 点数	512
高频提升系数	0.97

4.1　神经网络层的影响

对于神经网络而言，网络结构会对其生成的结果产生重要影响。过深的网络结构不仅会造成较大的时间开销，也容易产生过拟合等现象。因此，本文通过实验就网络层数对 MLP 识别模型的影响进行了测试。实验测试数据如表 2 所列。

表 2　不同网络层数对性别识别结果的影响

Table 2　Impact of different network layers on the rsults of gender recongnition

Model	Hid_layer	Max_iter	Alpha	*AUC*
MLP_1	1	1000	0.0001	0.9306
MLP_2	2	1000	0.0001	0.9628
MLP_3	3	1000	0.0001	0.9717
MLP_4	4	1000	0.0001	0.9789

表 2 列出了在激活函数为 ReLU 时，不同网络层数下的 MLP 识别模型的 AUC($AUC<1$)，AUC 越大，模型识别的准确率越高。其中，Hid_layer 表示隐藏层层数，Max_iter 表示最大的迭代次数，$Alpha$ 表示正则化项参数。实验结果表明，随着 MLP 隐藏层数量的增加，电子伪装语音的说话人性别识别的测试集准确率会有所提升，在隐藏层为 4 层时，AUC 最高可达 0.9789%。但隐藏层层数越多，AUC 的增加减缓，隐藏层由 1 层变为 2 层时，AUC 提升约 3%；而当隐藏层由 2 层变为 3 层时，AUC 提升不到 1%，增长幅度不明显。相对应地，随着隐藏层层数的增加，识别模型的训练时间呈指数倍增加。因此，选取隐藏层为 2 层时可以取得较好的电子伪装语音说话人性别识别的效果，且不会耗费过长的时间。

4.2　激活函数的影响

前文中已经讨论过激活函数在语音信号处理中的重要性，本节将探讨激活函数对 MLP 识别模型结果的具体影响，并研究迭代次数对不同激活函数的影响。实验测试数据如表 3 所列。

表 3　不同激活函数对性别识别结果的影响

Table 3　AUC for gender recognition with different activation functions in test set

Act_func	*Max_iter*			
	500	1000	1500	2000
Sigmoid	0.9241	0.9467	0.9696	0.9696
tanh	0.9220	0.9303	0.9303	0.9303
ReLU	0.9622	0.9628	0.9673	0.9673

表 3 列出了 Sigmoid，tanh 和 ReLU 这 3 种激活函数在两层隐藏层(Alpha=0.0001)情况下的模型总体识别情况。总体而言，这 3 种激活函数均可使测试集的 AUC 达到 92%以上，采用 Sigmoid 时识别模型可以获得最优结果，AUC 达 96.96%。但采用 ReLU 激活函数时，识别模型的总体分类识别效果最好，在低迭代水平下 AUC 即可达到 96%以上，获得高准确率所需的模型训练时间更短。显然，不论采用何种激活函数，随着迭代次数的增加，识别模型的准确率都呈现出先增长后稳定不变的趋势：ReLU 与 Sigmoid 在迭代次数为 1500 之后，AUC 不再增加；而 tanh 的 AUC 拐点出现在迭代次数为 1000 时。

4.3　识别模型的敏感性与稳定性

进一步研究发现，MLP 识别模型对于不同的电子伪装手段的敏感度不同。如表 4 所列，采用 ReLU 的两层隐藏层(Alpha=0.0001)MLP 时，分别分析了不同迭代次数对不同电子伪装方式的说话人性别分析的影响。可以看出，在同一迭代次数下，经不同伪装手段(Pitch，Rate and Tempo)伪装后的语音的性别识别 AUC 并不相同。相较而言，MLP 识别模型对 Tempo 这一伪装手段更为敏感，AUC 最高达 0.9937，基本实现无误差；对 Pitch 的敏感性次之；最低为 Rate，AUC 也可达 0.9330。实验测试数据如表 4 所列。

表 4　不同电子伪装手段的性别识别结果

Table 4　AUC for gender recognition of different electronic disguise methods in test set

Ele_dis	Max_iter			
	250	500	750	1000
Pitch	0.9196	0.9464	0.9463	0.9530
Rate	0.9140	0.9330	0.9330	0.9330
Tempo	0.9222	0.9861	0.9867	0.9937

这一实验结果也验证了 MLP 识别模型具有良好的稳定性，能适用于多种不同的电子伪装手段的说话人性别识别，并能获得良好的识别分类结果。此外，由于不同伪装手段的说话人性别识别率最高时对应的迭代次数不尽相同，因此通过调整迭代率，可以使识别模型适应于不同的电子伪装手段的说话人性别识别[7]。

结束语　本文采用 MLP 框架下的深度学习识别模型，

实现了对不同伪装方式下的电子伪装语音说话人性别识别。该模型通过 L-BFGS 实现模型的参数求解优化，提升了模型的训练效率。此外，还针对 MLP 的网络结构、激活函数以及其他的超参对电子伪装语音说话人性别识别的影响进行了分析，并测试了该模型对不同的伪装手段敏感性。实验表明，MLP 识别模型可以高效实现电子伪装语音说话人性别识别，并且对于不同的伪装手段具有良好的普适性。如何改进该模型，使其能更广泛地应用于电子伪装语音说话人身份识别，将是未来的工作方向。

参 考 文 献

[1] ZHANG G Q, JIN Y Z, LIU H W, et al. Study on changing rules of electronic camouflage audio [J]. Evidence Science, 2010, 18(4): 503-509.

[2] ENDRES W, BAMBACH W, FLOSSER G. Voice spectrograms as a function of age, voice disguise and voice imitation [J]. J. Acoust. Soc. Am., 1971, (49): 1842-1848.

[3] HANSEN J H, HASAN T. Speaker recognition by machines and humans: a tutorial review [J]. IEEE Signal Process Magazine, 2015, 32(6): 74-99.

[4] ZHANG C. Acoustic Analysis of Disguised Voices with Raised and Lowered Pitch [C]// IEEE. ISCSLP, 2012: 353-357.

[5] RODMAN R. Computer Recognition of Speakers who Disguise Their Voice [C] // Proceedings of the International Conference on Signal Processing Applications & Technology. USA: Texas, 2000.

[6] ZHAO L. Speech signal processing [M]// Beijing: Machinery Industry Press, 2009: 11.

[7] Gender Recognition of Electronic Disguised Voices: Chinese [P]. Patent ZL 2019 1 0959040. [2020-10-23].

ZHANG Xiao, born in 1987, master, associate professor, is a member of CCF (No. 37630M). Her main research interests include network information security, electronic data and audio-visual information, and computer juridical expertise.