

基于支持向量机的复杂场景中多人 对话语音智能识别方法研究

刘子寒[†], 沈力, 奚梦婷, 陆佳鑫, 朱佳佳, 查俊杰

(国网江苏省电力有限公司信息通信分公司, 江苏 南京 210000)

摘 要: 面对多人对话语音单一特征表征性别组合信息不足, 导致语音识别结果不精准的问题, 提出了基于支持向量机的复杂场景中多人对话语音智能识别方法。使用距离度量方法, 检测复杂场景多人对话变化点。计算任意两个数据集的对数似然概率值, 构建得分集。结合 T-Test 相似性度量方法, 判断两个数据集显著差异性。构造支持向量机判别函数, 利用支持向量机的映射逻辑实现相似语音的分离。使用支持向量机的二元分类超线性分类器构建最优判别函数, 结合男性、女性基音频率、信号非谐振频率特征, 实现多人对话语音智能识别。由实验结果可知, 所研究方法对于基音频率识别结果, 男性、女性幅度波动范围分别为 $-0.5 \sim 0.5$ 、 $-0.7 \sim 0.7$, 与实验数据一致; 对于信号非谐振频率识别结果, 男性、女性频率波动范围分别为 $-600 \sim 600$ Hz、 $-360 \sim 405$ Hz, 男性频率波动范围与实验数据仅存在 50 Hz 的误差, 女性频率波动范围与实验数据一致。

关键词: 支持向量机; 复杂场景; 多人对话; 语音智能识别

中图分类号: TP274

文献标识码: A

Research on Intelligent Speech Recognition Method for Multi Person Conversation in Complex Scenes Based on Support Vector Machine

LIU Zihan[†], SHEN Li, XI Mengting, LU Jiaxin, ZHU Jiajia, ZHA Junjie

(Information and Telecommunication Branch, State Grid Jiangsu Electric Power Co., Ltd., Nanjing, Jiangsu 210000, China)

Abstract: In the face of the problem of insufficient gender combination information represented by a single feature in multi-person conversation speech, resulting in inaccurate speech recognition results, a support vector machine based intelligent recognition method for multi-person conversation speech in complex scenes is proposed. Using distance measurement methods to detect changes in multi-person conversations in complex scenes. Calculate the logarithmic likelihood probability values of any two datasets and construct a diversity set. Using the T-Test similarity measurement method, determine the significant differences between the two datasets. Construct a support vector machine discriminant function and use the mapping logic of the support vector machine to achieve the separation of similar voices. The binary classification super linear classifier of support vector machine is used to construct the optimal discriminant function, and combined with male and female pitch frequency and signal non resonant frequency characteristics, the intelligent recognition of multi person conversation speech is realized. From the experimental results, it can be seen that the range of amplitude fluctuations for pitch frequency recognition in the research method is $-0.5 \sim 0.5$ for males and $-0.7 \sim 0.7$ for females, which is consistent with the experimental data; For the non resonant frequency identification results of the signal, the frequency fluctuation ranges for males and females are $-600 \sim 600$ Hz and $-360 \sim 405$ Hz, respectively. There is only a 50 Hz error between the male frequency fluctuation range and the experimental data, while the female frequency fluctuation range is consistent with the experimental data.

Key words: support vector machine; complex scenarios; multi person dialogue; speech intelligent recognition

伴随着移动网络时代的来临,信息的处理与储存能力得到了极大提升,人类日常生活中存在大量的、各种各样的声音,其中包含了电视和电台声音、电话声音、会议声音、有声邮箱等等。面对如此庞大的声音资料,如何利用电脑来进行有效的搜寻和检索,以获得所需的资讯,已是一个热门课题。通常来说,目前大多数以统计模式识别技术为基础的语音处理技术都是以单个人为对象的。面对复杂场景中的多人对话语音识别,上述方法还不能很好满足实际应用需求。文献[1]提出了基于深度学习的检测方法,该方法通过使用深度学习方式,从一个单一通道声音水平上,利用一个相对低的频率来进行端到端的学习。在学习样本中采取消除静音方式,随机抽取比重较多的类,并使用维特比算法检测对话重叠语音片段。由于大多数情况下,多人对话中语音是同时存在的,使得精准识别对话语音变得十分困难,容易产生误判;文献[2]提出了基于多操作网络的识别方法,该方法利用多领域情感数据库和层次化操作,将原始数据映射到不同的转换空间,建立多操作的语情学网络模型。对左、右两种不同特征进行多层次融合,从而达到对语音进行智能识别的目的。然而,该方法是直接在匹配模型上进行语音识别与分离的,但是实际语音分离时混合语音音色是未知的,误识别的概率较大。针对复杂场景的多人对话语音智能识别,提出了一种基于支持向量机的识别方法。该方法通过构建支持向量机判别函数,实现多人对话语音智能识别。

1 基于距离度量的复杂场景多人对话变化点检测

复杂场景下的多人对话变化点检测,表示在输入的声音中发现说话人的身份已经变化的点,并基于这些变化点,将整个语音切成较小碎片^[3]。通常,被切分每一个片段仅包含一个说话人声音。一般情况下,尽量减少对说话人变化点的漏检,从而使得在每个语音切片中,说话人变化点错误检测概率更高。因此,采用了距离度量检测方法,该方法将检测问题视为一个假定检验问题^[4]。采用一种基于极大似然法的语音特征序列检测方法,提取语音相关信号。

为了检测复杂场景下多人对话变化点,现作出如下假设:

原假设 F1:假设从语音信号中提取了相同说

话人,不存在变化点的语音特征序列,在距离度量步长上,所有序列都服从相同概率分布。

备假设 F2:假设从语音信号中提取了不同说话人,距离度量本征变化点与距离测度步长之间不符合不同概率分布^[5]。

采用距离度量方法比较两个假设,距离度量数值较大的一方,则认为该假设成立,用两个假设之间的距离度量差值作为比较依据,公式为:

$$\Delta I = I(F1) - I(F2) = \log P(A | \alpha_A) + \log P(B | \alpha_B) - \frac{1}{2}\beta(n_A + n_B)\log(N_C) - \log P(C | \alpha_C) \quad (1)$$

公式(1)中, β 表示调整系数; A, B 表示两个说话者的语音序列; C 表示 A, B 两个序列合并后的序列; $\alpha_A, \alpha_B, \alpha_C$ 表示 A, B, C 序列对应的模型参数; P 表示概率; n_A, n_B 表示两个序列模型参数对应的个数; N_C 表示合并后序列的样本数^[6,7]。

针对多人会话中的变化点检测问题,在距离度量中的样本数目很小情况下,假定多人对话的特征序列服从单一正态分布。对于距离度量步长内是否存在变化点进行判断,可用如下公式表示:

$$\max_k \Delta I(k) \begin{cases} > 0 & \text{存在多个} \\ = 0 & \text{存在一个} \\ < 0 & \text{不存在} \end{cases} \quad (2)$$

公式(2)中, k 表示变化点^[8]。充分考虑变化效率,以0.1距离度量步长为标准进行遍历,将最可能变化的点作为检测到的变化点,用如下公式表示:

$$a = \arg \max_k \Delta I(k) \quad (3)$$

基于上述假设,可确定复杂场景多人对话变化点。

2 基于支持向量机的多人对话语音智能识别

通过确定复杂场景多人对话变化点,使用 T-Test 相似性度量方法检测数据集间的差异。使用支持向量机映射逻辑,分离经过相似性判断的语音。结合支持向量机二元分类超线性分类器,构建智能识别框图,以此识别多人对话语音。

2.1 基于多人对话变化点的多人对话相似性度量

在确定多人对话变化点的前提下,结合 T-

Test 相似性度量方法,假设选取任意两个数据集 $H_1 = \{h_1^1, h_2^1, \dots, h_k^1\}$ 、 $H_2 = \{h_1^2, h_2^2, \dots, h_k^2\}$, 这两个数据集所属于的类分别为 C_1 、 C_2 , 通过计算对数似然概率构建得分集:

$$\begin{aligned} A_{11} &= \{\log P(h | C_1) - \log P(h | C_k) \mid \forall h \in X_1\} \\ A_{12} &= \{\log P(h | C_2) - \log P(h | C_k) \mid \forall h \in X_2\} \\ A_{21} &= \{\log P(h | C_1) - \log P(h | C_k) \mid \forall h \in X_2\} \\ A_{22} &= \{\log P(h | C_2) - \log P(h | C_k) \mid \forall h \in X_1\} \end{aligned} \quad (4)$$

公式(4)中, P 表示对数似然概率值^[9]。将前两个得分集和后两个得分集分别合并,可得到如下两个分布:

$$\begin{aligned} A_1 &= A_{11} \cup A_{12} \\ A_2 &= A_{21} \cup A_{22} \end{aligned} \quad (5)$$

结合这两个分布,使用如下公式度量两个类别相似度:

$$d(A_1, A_2) = \frac{v_1 - v_2}{\sqrt{\frac{\sigma_1^2}{N_1}} + \sqrt{\frac{\sigma_2^2}{N_2}}} \quad (6)$$

公式(6)中, v_1 、 v_2 分别表示两组数据集均值; σ_1 、 σ_2 分别表示两组数据集标准差; N_1 、 N_2 分别表示两组数据集样本数^[10]。如果两种语言类型是同一个人的,则 A_1 、 A_2 分布是比较靠近的,这两个数据集相似性较高;反之,如果两个类别属于不同说话者,那么 A_1 、 A_2 分布相对较远,这两个数据集相似性较小^[11]。

2.2 支持向量机映射逻辑的相似语音分离

在实际语音分离过程中,复杂场景多人对话者性别组合信息是未知的。如果采用一般的分离方法,则不能得到理想的分离结果。为此,采用支持向量机算法对各信道的声频信号进行有效的分离。

充分考虑两个时间域求和的方法,研究一种基于单信道的语音分离方法:

$$f(t) = g_A(t) + g_B(t) \quad (7)$$

公式(7)中, $g_A(t)$ 、 $g_B(t)$ 分别表示 A 、 B 两个说话者的语音信号; t 表示时间^[12]。

在获取单通道语音分离结果后,利用支持向量机映射逻辑,构造一个支持向量机的决策函数,如图 1 所示,将公式(7)转化到频域展开分析。

在决策过程中,支持向量机的输入是复杂场景下多人对话语音的幅度谱,目标语音信号的理想比率掩码可用如下公式表示:

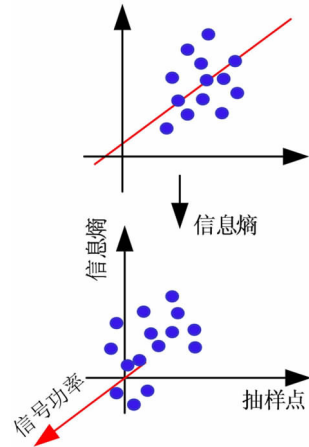


图 1 支持向量机映射逻辑示意图

$$M_k(x, y) = \sqrt{\frac{g_A^2(x, y)}{g_B^2(x, y) + Z}} \quad (8)$$

公式(8)中, x 、 y 表示频域横纵坐标; Z 表示极小整数,避免分母为 0^[13]。将该计算结果与估计比率掩码进行差值分析,结果如下所示:

$$J = \frac{1}{2n} \sum_{k=1}^n \| M_k(x, y) - \hat{M}_k(x, y) \|_2^2 \quad (9)$$

公式(9)中, n 表示语音分帧总数; $\hat{M}_k(x, y)$ 表示估计比率掩码^[14]。该公式用于衡量估计比率掩码是否在合理范围内,如果 J 计算数值小于等于设定的阈值,那么可直接使用估计比率掩码。将其和混合信号振幅谱的特征矩阵作哈达玛乘积,可以得出:

$$E = \hat{M}_k(x, y) \odot H \quad (10)$$

公式(10)中, H 表示输入语音信号幅度谱的特征矩阵; $\odot$ 表示哈达玛乘积算子^[15]。

为了解决语音频域信号分离过拟合问题,引入支持向量机 ROC 边界的音频域信号分离方法,可表示为:

$$\begin{cases} \min_{\omega} = \frac{\|\omega\|^2}{2} + V \cdot \left(\sum_{k=1}^j \tau_k \right) \\ \text{s. t. } g(t) \geq 1 - \tau_i \end{cases} \quad (11)$$

公式(11)中, ω 表示加权向量; V 表示惩罚系数; j 表示遍历次数; $g(t)$ 表示分离决策函数; τ_k 表示松弛变量。将满足公式(11)的作为多人对话语音,包括男男语音混合、男女语音混合、女女语音混合,将不满足公式(11)的作为单人语音。

2.3 基于语音分离特征的智能识别

根据语音分离特征,构建智能识别框图如图 2 所示。

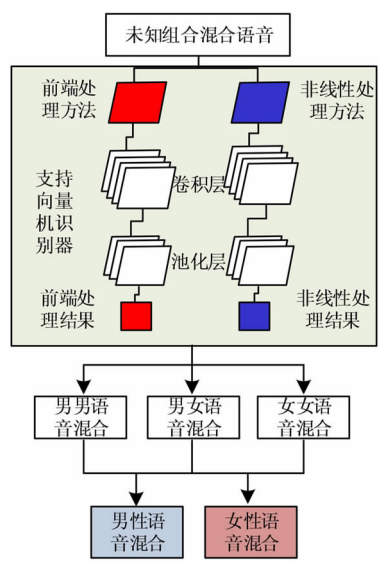


图 2 基于语音分离特征的智能识别框图

结合框图，设计一个超平面：

$$L = [v_i u_i o(x, y)]^T + b \tag{12}$$

公式(12)中， v_i 表示第 i 个拉格朗日系数； u_i 表示输出标签； $o(x, y)$ 表示核函数； T 表示分离周期； b 表示偏置。

利用多种核函数可以实现多种类型的超平面映射，利用其与数据之间的非线性关联，将数据与数据之间的关联，建立对应的最佳超平面，达到对数据进行有效识别的目的。采用多个识别特征时，支持向量机方法有着很好的性能，当样本为 (x_i, y_i) 时，最优判别函数可表示为：

$$g(x) = \text{sgn} \left(\sum_{i=1}^j L \right) \tag{13}$$

人在发声时，主要依赖于喉部声带的振动来发出基本的声音，并且还会经过口腔、咽腔、鼻腔等器官区域来对声腔辐射共振性质控制，进而产生了变化的音色。因此，结合公式(13)识别复杂场景中多人对话语音，结合男性基音频率 η_{male} 小于女性基音频率 η_{female} ，男性语音信号非谐振频率 $\mathfrak{S}_{\text{male}}$ 大于女性语音信号非谐振频率 $\mathfrak{S}_{\text{female}}$ 的特点，设定如下识别依据：

$$\begin{cases} \eta_{\text{male}} < \eta_{\text{female}} \\ \mathfrak{S}_{\text{male}} > \mathfrak{S}_{\text{female}} \end{cases} \tag{14}$$

在该识别依据条件下，能够精准智能识别对话语音。

3 实验

3.1 实验数据集和样机

基于支持向量机的复杂场景中多人对话语音

智能识别方法实验验证，它的使用方法是基于 IF-FLY-DIARIZATION-I 数据库，数据为中文，来自几个访谈节目，一共有 1050 个语音，其中说话人的数量为 3~10，时间由 15~40 min。讲话时，说话人语言比较随意，在不同人之间跳跃很大，其中还包括了现场欢呼声、鼓掌声、笑声、音乐等噪声。实验参数设置如表 1。

表 1 实验参数设置

序号	参数名称	参数值	取值范围
1	c_1	1.6	1~2
2	c_2	1.8	1~2
3	$\max gen$	100	0~100
4	K	0.6	0.1~1

所有测试集合、培训集合和发展集合都是从 IFLY-DIARIZATION-I 的资料库中提取的，其中，测试集含有 300 条语音，语音中说话人的数量在 3~10 个之间；训练集含有 500 条语音，总时间长度为 180 h；开发集总共含有 80 条语音，总时间长度为 15 h。

音频采集与播放实验设备如图 3 所示。



(a) 音频采集装置



(b) 音频播放装置

图 3 实验设备

为了确保所获取的语音为远场信号，需要将各个话筒探测信号之间的时延显示，话筒阵列中的相邻话筒间隔为 2.5 cm，将扬声器放在其顶部，并与话筒指向呈 90°夹角，以避免由于音箱与话筒之间距离太小而导致的谐振啸叫。

3.2 实验数据分析

任意选择 1 条语音，分别分析基音频率、信号非谐振频率，以此为依据进行实验对比分析，如图 4、图 5 所示。

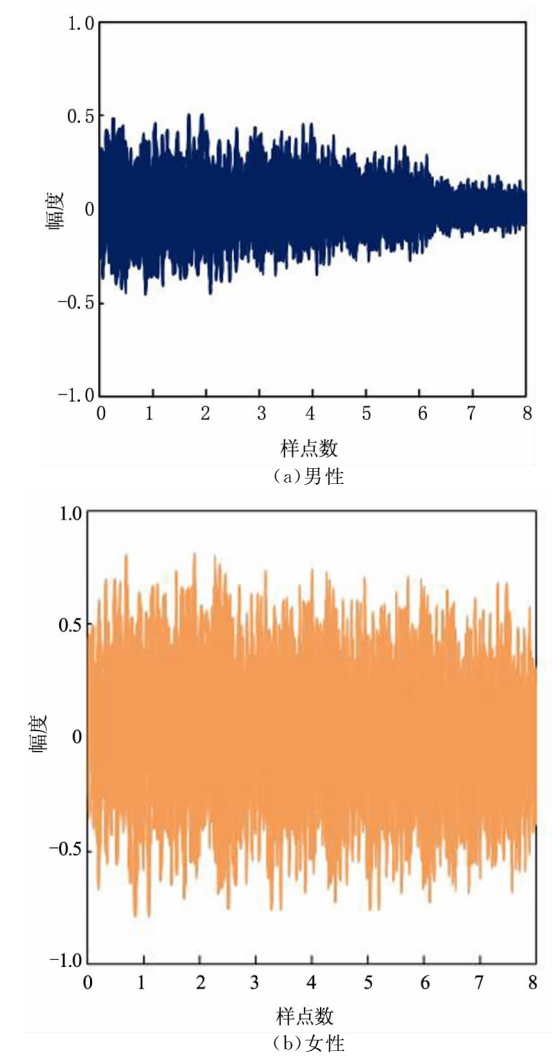


图 4 基音频率实验数值

由图 4 可知,男性基音频率的幅度波动范围为 $-0.5\sim0.5$,女性基音频率的幅度波动范围为 $-0.7\sim0.7$,男性基音频率小于女性基音频率。

由图 5 可知,男性声音信号非谐振频率的波动范围为 $-600\sim650$ Hz,女性声音信号非谐振频率的波动范围为 $-360\sim405$ Hz,男性声音信号非谐振频率大于女性声音信号非谐振频率。

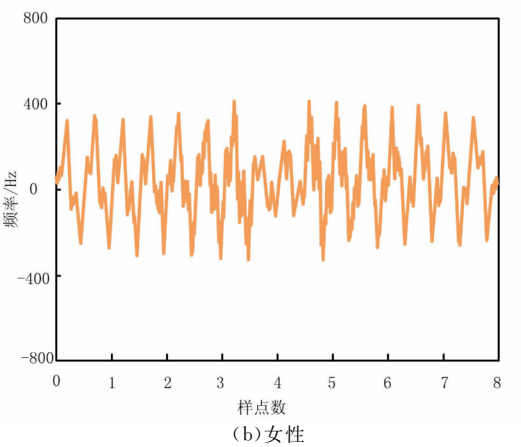
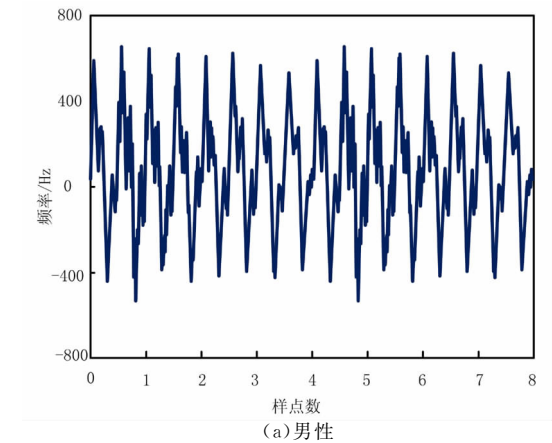


图 5 信号非谐振频率实验数值

将上述实验数据作为语音智能识别结果,越接近实验数值越说明识别结果精准。

3.3 实验结果与分析

使用基于情境分析的检测方法、基于 LM-BP 神经网络的识别方法和基于支持向量机的识别方法对比分析基音频率,对比结果如图 6 所示。

由图 6(a)可知,男性基音频率的幅度波动范围为 $-0.7\sim0.6$,女性基音频率的幅度波动范围为 $-0.6\sim0.55$,与图 4(a)、(b)实验数值不一致。

由图 6(b)可知,男性基音频率的幅度波动范围为 $-0.65\sim0.5$,女性基音频率的幅度波动范围为 $-0.5\sim0.5$,男性基音频率大于女性基音频率,与图 4(a)、(b)实验数值不一致。

由图 6(c)可知,对于男性基音频率的识别,只有在 6~8 的样点数下出现了与图 4(a)波动幅度不完全一致的情况,其余均一致,男性基音频率的幅度波动范围为 $-0.5\sim0.5$,女性基音频率的幅度波动范围为 $-0.7\sim0.7$,与图 4(b)波动范围一致。

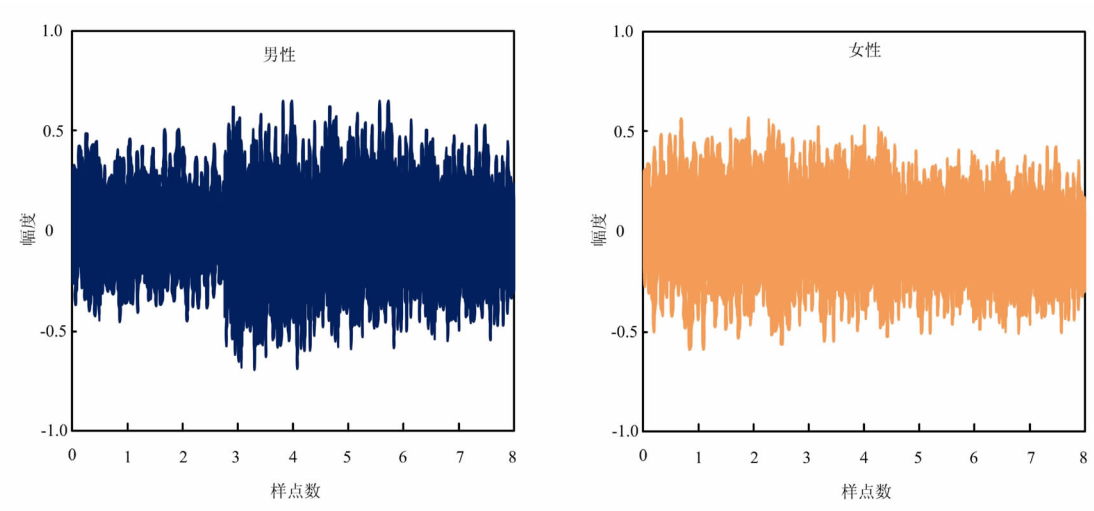
使用三种方法对比分析信号非谐振频率,对比结果如图 7 所示。

由图 7(a)可知,基于情景分析的检测方法识别的男性声音信号非谐振频率的波动范围为 $-500\sim400$ Hz,女性声音信号非谐振频率的波动范围为 $-390\sim390$ Hz,与实验数据不一致。

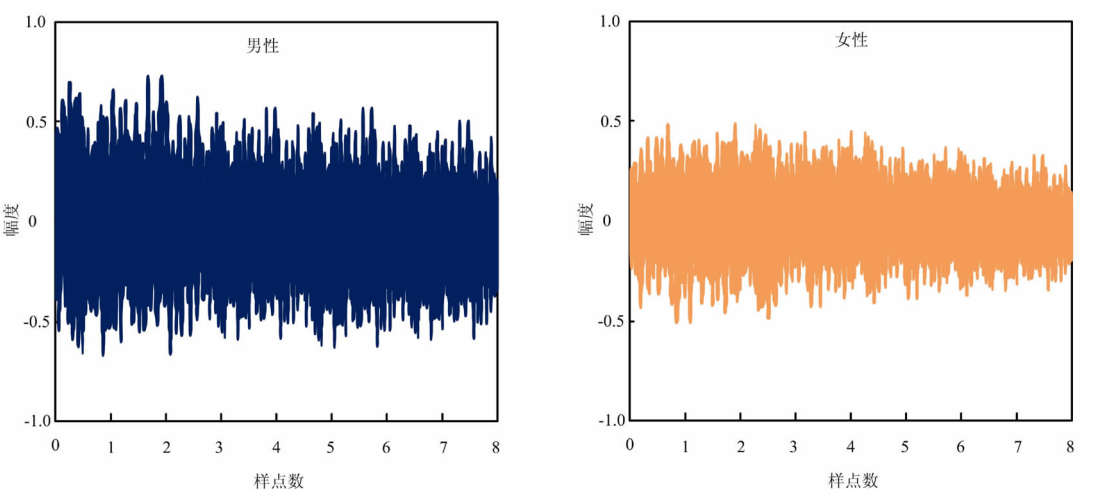
由图 7(b)可知,基于 LM-BP 神经网络的识别方法识别的男性声音信号非谐振频率的波动范围为 $-350\sim370$ Hz,女性声音信号非谐振频率的波动范围为 $-450\sim500$ Hz,与实验数据不一致。

由图 7(c)可知,使用所研究方法男性声音信号非谐振频率的波动范围为 $-600\sim600$ Hz,女性声音信号非谐振频率的波动范围为 $-360\sim405$ Hz,男性波动范围与实验数据存在最大为 50 Hz

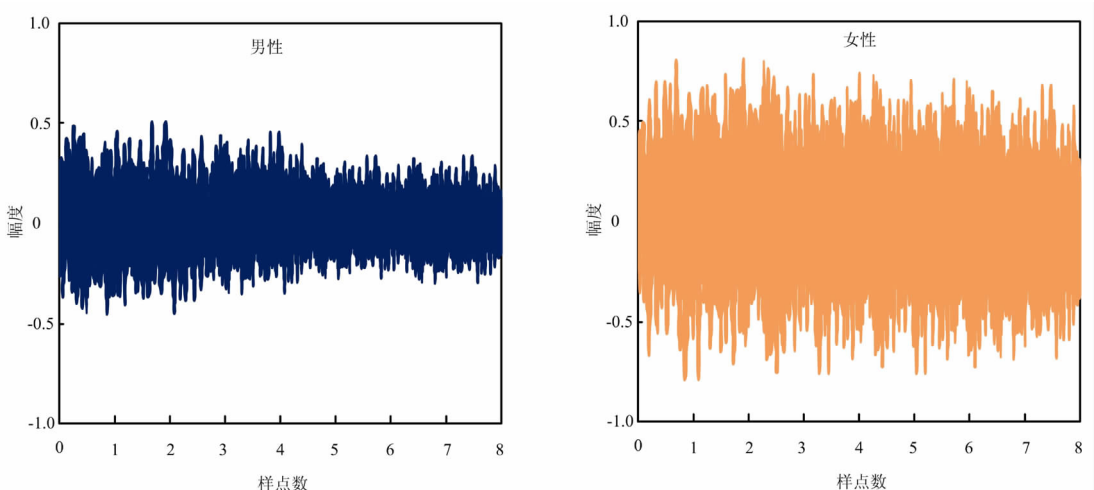
的误差,女性波动范围与实验数据一致。



(a) 基于情境分析的检测方法



(b) 基于 LM-BP 神经网络的识别方法



(c) 基于支持向量机的识别方法

图 6 不同方法基音频率识别结果分析

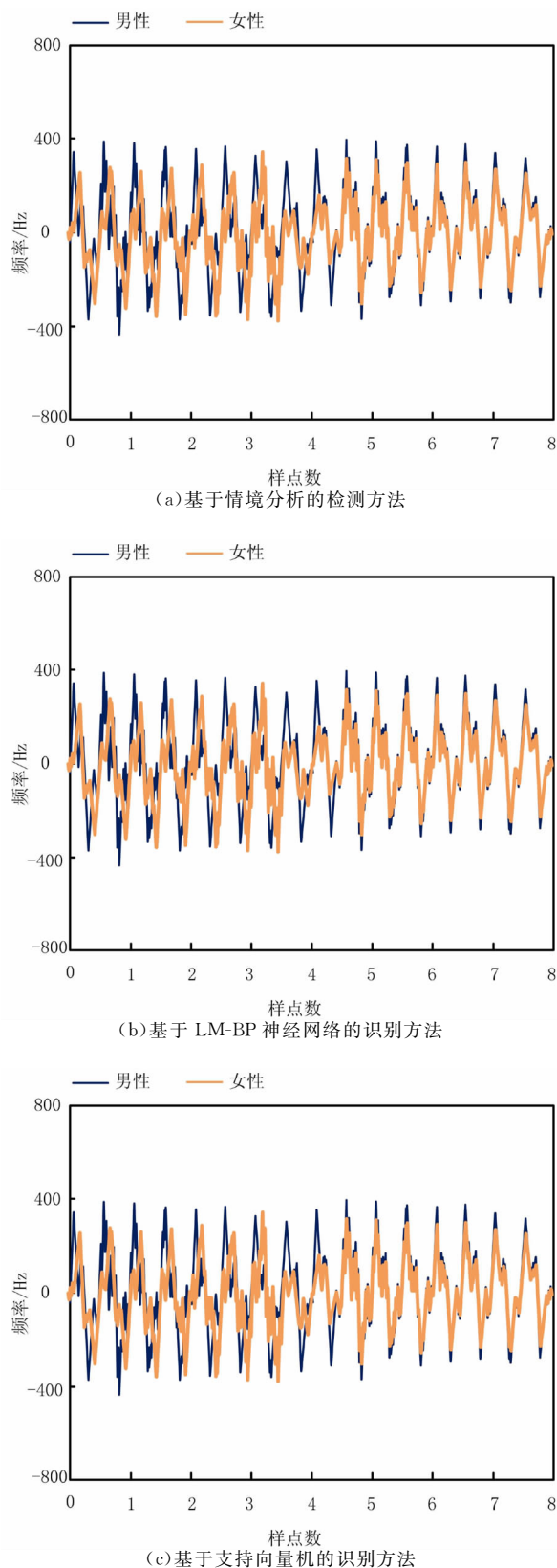


图 7 不同方法信号非谐振频率识别结果分析

解决了传统复杂场景中多人对话语音识别不精准的问题。通过使用距离度量方法,能够获取精准多人对话变化点;通过分离相似语音,能够为语音识别提供有效音频段,方便后续识别;经过对男性、女性基音频率、信号非谐振频率特征分析,能够获取精准识别结果。

参考文献

- [1] 魏金太,高穹. 基于深度学习的对话重叠语音片段检测[J]. 中北大学学报(自然科学版),2021,42(1):34—39.
- [2] 张会云,黄鹤鸣. 基于多操作网络的图式多域语音情感识别研究[J]. 计算机工程,2022,48(7):59—65.
- [3] 张家扬,童峰,陈东升,等. 面向多说话人分离的深度学习麦克风阵列语音增强[J]. 电子技术应用,2022,48(5):31—36.
- [4] 黄江岚,卿粼波,姜雪,等. 融合场景及交互性特征的多人行为识别[J]. 四川大学学报(自然科学版),2022,59(6):77—88.
- [5] 张开生,赵小芬. 复杂环境下基于自适应深度神经网络的鲁棒语音识别[J]. 计算机工程与科学,2022,44(6):1105—1113.
- [6] ZHOU D, YANG Y, ZHU J, et al. Intelligent reading recognition method of a pointer meter based on deep learning in a real environment[J]. Measurement Science and Technology, 2022, 33(5):055021.
- [7] 江官星,付悦. 基于多任务训练的用户登入语音识别模型仿真[J]. 计算机仿真,2022,39(9):190—194.
- [8] 覃京燕,王赞壘. 高低语境下对话式语音交互系统研究[J]. 包装工程,2021,42(10):85—91.
- [9] 王怡,王黎明,柴玉梅. 融合多特征的语音情感识别方法[J]. 小型微型计算机系统,2022,43(6):1232—1239.
- [10] 姜特,陈志刚,万永菁. 基于注意力机制的多任务 3D CNN—BLSTM 情感语音识别[J]. 华东理工大学学报(自然科学版),2022,48(4):534—542.
- [11] 张晓宇,张华熊,高强. 基于深度学习的多模式权重网络语音情感识别[J]. 大连理工大学学报,2022,62(5):526—534.
- [12] LIN P C, YANKSON B, CHAUHAN V, et al. Building a speech recognition system with privacy identification information based on Google voice for social robots[J]. The Journal of Supercomputing, 2022, 78(13):15060—15088.
- [13] 胡文轩,王秋林,李松,等. 基于端到端的多语种语音识别研究[J]. 信号处理,2021,37(10):1816—1824.
- [14] 张禹尧,蒋玉茹,张仰森. 基于多尺度自注意力增强的多方对话角色识别方法[J]. 中文信息学报,2021,35(5):101—109.
- [15] 薛艳飞,毛启容,张建明. 基于多任务学习的多语言语音情感识别方法[J]. 计算机应用研究,2021,38(4):1069—1073.

4 结 论

提出的基于支持向量机的智能识别方法,有效