◎模式识别与人工智能◎

结合听觉模型的腭裂语音高鼻音等级自动识别

付方玲[1],何 飞[1],付 佳[1],尹 恒[2],黄 华[1],何 凌[1]
1.四川大学 电气信息学院,成都 610065
2.四川大学 华西口腔医院,成都 610041

摘 要:腭裂语音高鼻音等级的自动识别能为临床腭咽功能评估提供有效、客观、无创的辅助依据。对腭裂语音高鼻音等级自动分类系统进行了研究,利用听觉模型提取语音信号的听觉内部表达,并结合同步检测器提取软限制比(Soft Limited Ratio,SLR)谱特征作为特征参数,利用一对一支持向量机(1-v-1 Support Vector Machine,1-v-1 SVM)实现腭裂语音高鼻音四类等级(正常、轻度、中度和重度)的自动划分。实验采用56名儿童的共3 086个语音样本,并对比了使用不同基底膜滤波器种类和个数,使用同步检测器和侧抑制网络对识别效果的影响。实验结果表明,使用基于等效矩阵带宽(Equivalent Rectangular Bandwidth,ERB)尺度的Gammatone滤波器的识别效果优于基于Bark尺度的小波包滤波器;54个通道的滤波器能有效权衡算法时间成本和识别正确率;使用同步检测器提取SLR谱特征的识别效果优于侧抑制网络提取的LIN(Lateral Inhibition Network)谱特征。腭裂语音高鼻音四类等级自动识别系统最高分类正确率达91.50%。

关键词:腭裂语音;高鼻音;听觉模型;同步检测器
文献标志码:A **中图分类号:**TP391 doi:10.3778/j.issn.1002-8331.1803-0060

付方玲,何飞,付佳,等.结合听觉模型的腭裂语音高鼻音等级自动识别.计算机工程与应用,2019,55(10):127-134.
FU Fangling, HE Fei, FU Jia, et al. Automatic detection of hypernasality degrees in cleft palate speech based on human auditory model. Computer Engineering and Applications, 2019, 55(10):127-134.

Automatic Detection of Hypernasality Degrees in Cleft Palate Speech Based on Human Auditory Model

FU Fangling[1], HE Fei[1], FU Jia[1], YIN Heng[2], HUANG Hua[1], HE Ling[1]
1.College of Electrical Engineering and Information Technology, Sichuan University, Chengdu 610065, China
2.West China Hospital of Stomatology, Sichuan University, Chengdu 610041, China

Abstract: The automatic detection of hypernasality degrees in cleft palate speech can provide effective, objective and non-invasive basis for the assessment of velopharyngeal function in clinical. In this work, an automatic detection system of hypernasality degrees in cleft palate has been researched. The human auditory model is applied to extract the inner presentation of speech signal as the front-end processing, and the SLR(Soft-Limited Ratio) spectral features extracted from the synchronous detector is used as the acoustic characteristic parameters. The 1-v-1 SVM (1-v-1 Support Vector Machine) is utilized to automatically detect the hypernasality degrees (normal, mild, moderate and severe hypernasality). Experimental data include total 3 086 speeches from 56 kids, the comparisons of filter bank's kind and number, synchronous detector and lateral inhibitory network are discussed. And the results show that the Gammatone filter based on ERB

基金项目:国家自然科学基金青年科学基金项目(No.61503264)。
作者简介:付方玲(1996—),女,硕士研究生,研究领域为语音信号处理,E-mail:18384127060@163.com;何飞(1998—),女,硕士研究生,研究领域为语音信号处理;付佳(1998—),女,硕士研究生,研究领域为语音信号处理;尹恒(1971—),女,副主任护师,研究领域为腭裂语音评估;黄华(1961—),男,博士后,教授,博士生导师,研究领域为医学电子学;何凌(1981—),通讯作者,女,博士,副教授,研究领域为语音信号处理。
收稿日期:2018-03-05 **修回日期:**2018-04-20 **文章编号:**1002-8331(2019)10-0127-08
CNKI **网络出版:**2018-08-30, http://kns.cnki.net/kcms/detail/11.2127.TP.20180829.0834.002.html

(Equivalent Rectangular Bandwidth) scale performs better than the wavelet-packet filter based on Bark scale, and the filter bank with 54 channels can effectively weigh the time cost and recognition accuracy of our algorithm, and SLR spectral features extracted from the synchronous detector has better recognition than LIN spectral features extracted from the lateral inhibition network. The highest accuracy of the automatic detection of four-hypernasality degree is 91.50%.

Key words: cleft palate speech; hypernasality; auditory model; synchronous detector

1 引言

腭裂是世界上最常见的先天缺陷。腭裂会导致一系列问题,如进食问题、语言障碍、面部外观缺陷、耳部感染和心理障碍等。唇腭裂患者存在不同程度的腭咽闭合不全和代偿性发音问题,这将导致腭裂病理性语音[1]。腭咽功能不全是指口咽和鼻咽间的腭咽瓣不完全闭合而导致的语言及吞咽障碍。正常的腭咽闭合功能是人们获得正常语音的必要条件,正常情况下说话人发元音以及除鼻辅音外的大部分辅音均需要完整的腭咽闭合。完整的腭咽闭合在发音过程中为口腔气流提供足够的气流压力环境,使得元音和非鼻辅音能正确发音。而腭裂患者由于腭咽瓣的不完全闭合使得口鼻腔相通,在发音时不能形成完全的口鼻腔分离状态,导致气流总会从口腔泄漏到鼻腔,引起鼻腔内产生一个额外的共鸣腔,出现高鼻音临床特征[2]。

目前,治疗唇腭裂最主要的方法是唇腭裂序列治疗,这是一个漫长的过程,初期唇腭裂修复术后仍有20%~30%的患者存在不同程度的鼻腔共鸣障碍,这种情况下则需要进一步的二期唇腭裂修复术。在临床环境中,专业医生认为患者的高鼻音程度与其腭咽开口大小有关,因此对腭咽开口大小的评估是手术效果的评定指标,更是后续手术的关键参考指标。

在临床应用上,入侵式器械检查和主观言语治疗师评定常用于腭咽功能评估。常用的入侵式检查技术有鼻内窥镜(Nasendoscopy)和影像透视法(Videofluoroscopy),这类技术直接检查腭咽开口程度,常为患者带来不适,特别对于儿童腭裂患者体验极差;言语治疗师能无创地进行腭裂评估,但其评估结果也会受到言语治疗师的主观判断影响,中国还存在专业言语治疗师稀缺的现状。

而近年来,无创、客观的语音信号处理技术被广泛应用于病理语音的研究。该类技术主要是针对腭裂语音的高鼻音临床特征进行识别。目前,高鼻音检测技术主要分为三大类:(1)基于声学参数,国内外众多学者通过分析声学参数来检测高鼻音的存在。国外学者Castellanos、Bocklet和Maier[3-6]使用MFCC(Mel Frequency Cepstrum Coefficient)作为特征值;Castellanos[3]和Orozco-Arroyave[7]使用pitch、jitter、shimmer等参数;Lee等[8]使用低频和高频带的能量比(Voice Low Tone to High Tone Ratio, VLHR)作为特征值;Cruz等[9]使用EMD(Empirical Mode Decomposition)分解后的每个部分的Teager能量算子作为特征值。实验结果表明基于声学参数方法的分类正确率在63.2%~98.0%不等,部分声学参数对高鼻音有无的判别有不错的效果。(2)基于声道模型,腭裂语音不同于正常语音,现有研究表明腭裂语音的声道传输函数在频谱上存在零点,传统的AR(Autoregressive)模型并不适合腭裂语音。Akafi等[10]提出了适用于腭裂语音的ARMA(Autoregressive Moving Average)模型,并使用AR模型和ARMA模型的倒谱系数间的距离作为特征值;Rah等[11]使用高阶线性预测模型,用多个极点去近似表示一个零点,并计算低阶和高阶线性预测倒谱序列间的距离作为特征值。(3)基于鼻共振峰,国外学者Vijayalakshmi等[12-13]利用群时延可分离两相邻共振峰的特性,发现了腭裂语音在低频部分会出现额外的共振峰,元音/a/、/i/、/u/的额外共振峰依次出现在250 Hz、1 000 Hz、800 Hz左右。学者们将这个额外共振峰称为鼻共振峰。Dubey等[14]使用ZTW(Zero Time Windowing)技术来提高频谱分辨率,也能区别出第一共振峰和鼻共振峰。还有学者Cairns和Hansen等[15-16]利用Teager能量算子对复信号的敏感性,发现腭裂语音经过低通和带通滤波后的Teager能量包络有很大区别,而正常语音的能量包络则无区别,学者们利用该特征去检测高鼻音的有无。

目前,高鼻音检测的研究局限于对高鼻音有无的判别。而临床实践中,高鼻音等级(正常、轻度、中度、重度)的识别对唇腭裂序列治疗更具有临床参考价值。本文利用人耳听觉模型的高鲁棒性和良好的信号处理能力,将它作为腭裂语音的前端处理器,提取语音的听觉内部表达,结合同步检测器提取SLR谱特征作为特征参数,并利用1-v-1 SVM分类器,实现对高鼻音四个等级的自动识别。

2 腭裂语音高鼻音等级自动识别系统

本文提出基于听觉模型的腭裂语音高鼻音等级自动识别系统,系统流程如图1所示。

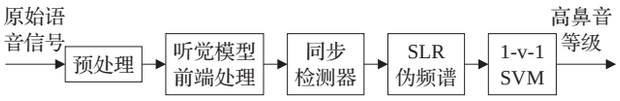

图1 腭裂语音高鼻音等级自动识别系统流程图

腭裂语音高鼻音等级自动识别系统主要分为四部分:(1)语音预处理,主要有幅值归一化和预加重滤波等

常规处理，重要的是利用过零率、短时能量结合修正算法对原始语音数据的词语进行切分，提取第一个字，并利用声韵母切分算法[17]提取第一个字中的韵母部分，后续的处理都是在韵母信号上进行；(2)听觉模型前端处理，利用人耳听觉感知的时频域分析特性，提取语音信号的听觉内部表达；(3)特征提取，对听觉模型每个通道的输出信号分别进行同步检测，把信号谱作为腭裂语音的特征参数；(4)等级分类，结合特征参数和标准标签，使用1-v-1 SVM分类器对腭裂语音等级(正常、轻度、中度、重度高鼻音)进行自动识别，并使用十折交叉验证方法测试算法的准确性。

2.1　听觉模型前端处理

人耳听觉系统在信号识别、分析和处理上都有重要作用。由于人耳能稳定适应噪声和不同的说话者，听觉系统作为前端处理器被广泛地用于自动语音识别(Automatic Speech Recognition，ASR)领域。Ghitza[18]使用基于听觉感知的后听觉神经模型为不同信号条件和音素变化提供感知不变性，提升语音识别效果。在预测语音质量领域，听觉模型也有重要意义。国外学者Karmakar[19]和Huber[20]等利用听觉模型具有内部表达特性，将语音信号转换为声音的激励模式。结果表明，这种感知转换对于语音质量和客观参数间的相关性至关重要。这种听觉内部表达特性在语音编码上也有重要作用。Plasverg[21]基于内部表达计算出的失真参数为语音编码提供了小误差的感知失真度量。在一些其他领域的应用中，基于听觉模型的听觉频谱在语音场景分类上有高鲁棒性表现[22]。

本文在国外学者Jepsen等[23]听觉感知模型上做了些许变化。整个听觉模型分为外、中耳模型和内耳模型，如图2所示。外、中耳模型通过两个线性相位的无限脉冲响应滤波器来模拟声音信号在外耳和中耳的传播过程；内耳模型依次模拟了基底膜、毛细胞、神经纤维的功能和机制。

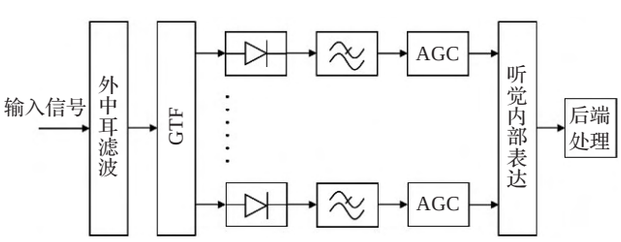

图2　人耳听觉模型框图

2.1.1　基底膜时频域分析模型

基底膜是耳蜗中重要的功能部分，它具有频率选择性，其功能相当于一个滤波器组对输入的语音信号进行滤波，并且输出信号在基底膜上呈现非线性分布。而这种滤波方式不同于普通的数字滤波，在听觉生理学上，用临界频带(Critical Band，CB)来表征基底膜的频率选择方式。最常用的临界频带的带宽有Bark尺度和ERB尺度。ERB尺度是一种矩形带通滤波器的带宽，比Bark尺度具有更窄的临界带宽，能够更加细致地反映信号的感知，因此本文使用ERB尺度和Gammatone滤波器组模拟基底膜的时频域分析特性。

Gammatone滤波器组是一种在语音识别和语音分析领域都有广泛应用的时频转换方法，由于其很好地模拟了人耳听觉特性，可以作为听觉滤波器组的一种。使用包含54个滤波器的Gammatone滤波器组，对同一输入信号进行分通道滤波，输出54个同步的信号。

Gammatone滤波器是一种线性滤波器，其时域脉冲响应如下：

$$h(t)=At^{(n-1)}\mathrm{e}^{-2\pi B_w t}\cos(2\pi f_c t+\varphi) \tag{1}$$

其中，A是调节比例的常数；n是滤波器阶数；f_c是中心频率；φ是相位，由于人耳对相位不敏感，一般该值取0；B_w是滤波器的衰减因子，决定了脉冲响应的衰减速度，并与滤波器带宽有关，其与ERB尺度的关系为：

$$B_w=1.019ERB(f_c) \tag{2}$$

$$ERB(f_c)=24.7\times(0.004\,37f_c+1) \tag{3}$$

2.1.2　耳蜗毛细胞模型

语音信号经过基底膜滤波后，带动耳蜗内的毛细胞运动。经过Gammatone滤波后的每个通道内的信号再用毛细胞模型进行处理。耳蜗毛细胞通过顶部纤毛的运动将声信号转换为电信号，通过蜗核等听觉中枢传导至听觉皮质而产生听觉。毛细胞位于螺旋器边缘，当其与盖膜在方向和时间上不一致时就会产生一种剪力，使得毛细胞上的纤毛倾斜。当基底膜移向蜗管时，剪力的方向从蜗轴向外拉，纤毛向外倾斜；当基底膜移向鼓阶时，剪力和纤毛方向向内。纤毛的倾斜使毛细胞受到刺激，产生去极化和超极化，这就是内耳毛细胞的半波整流效应[24]。

半波整流模型模仿了内耳毛细胞仅在一个方向发放电势的特性：

$$x_C=G(\arctan A[\mathrm{e}^{Bx}-1]+\arctan A) \tag{4}$$

其中，x是输入信号，即Gammatone滤波后每个通道的输出信号；A和B是常数，一般取$A=10$，$B=65$；G是增益。由于arctan函数的特性，保证了输出是恒为正值的。之后，再对半波整流的输出信号进行截止频率为1 kHz的一阶低通滤波，主要是为了消除理想半波整流后的高频分量。

2.1.3　耳蜗神经纤维模型

听觉神经纤维模型主要是模拟耳蜗神经纤维的非线性压缩特性。听觉系统接收的输入信号的动态范围接近10^{12}，而听觉神经上发放的脉冲速率的动态范围仅有10^2。在大部分听觉模型中，自动增益控制(Automatic Gain Control，AGC)耦合压缩网络可用来模拟这一非线性压缩特性。AGC网络用于动态范围的压缩以适应不同幅度的输入：

$$x_N[n]=\frac{x_C[n]}{1+K_{AGC}\{x_C[n]\}} \tag{5}$$

其中，K_{AGC} 是常数，{}代表对输入信号的低通滤波输出进行求期望操作。

2.2　同步检测器谱信号特性提取

在2.1节中，原始输入信号经过听觉模型的前端处理得到了听觉内部表达。经过听觉模型的最后一步AGC自动增益控制处理后，分别从54个通道得到了54个并行的时域输出信号。然后提取每个通道的输出信号的谱特征作为腭裂语音的特征参数。其中一种方法是基于人类感知系统的侧抑制效应。由2.1.1小节介绍的基底膜频率分解特性，分解后的信号传入中枢神经系统后，不同频段的信号之间会产生相互抑制和竞争，这种抑制和竞争使得能量占优势的频率分量得以加强。简单的听觉模型侧抑制神经网络（Lateral Inhibitory Network，LIN）采用拓扑结构，每个神经元的输出只与其相邻的神经元形成抑制连接[25]。侧抑制机制的主要功能有突出波峰，提高频谱分辨率，增强输入对比度。

本文使用同步检测器提取腭裂语音信号谱特征的特征参数。同步检测器具有突出共振峰谐振处的峰值，提高频谱分辨率，减少与声门激励相关的谱图特征，归一化幅值等优点。同步检测器的功能主要是比较两个输入信号，当两信号有相似波形和相同时序时，则会产生强烈的响应。它通过计算自相关类似的输出来检测在时间响应上的周期性。同步检测器计算每个通道的输出及其延时信号的和与差之间的软限制比（Soft-Limited Ratio，SLR），每个通道信号的延时与该通道滤波器的中心频率有关，SLR的计算公式为：

$$SLR_i(y)=A_s\arctan\frac{1}{A_s}\left[\frac{<|x_N[n]+x_N[n+n_i]|>-\delta}{<|x_N[n]-\beta^{n_i}x_N[n-n_i]|>}\right] \tag{6}$$

其中，$x_N[n]$ 是每个同步检测器的输入，也是每个AGC通道的输出信号；SLR_i 是第 i 个通道的同步检测输出；n_i 是每个同步检测通道信号的时延，$n_i=\frac{f_s}{f_i}$，f_s 是采样率，f_i 是第 i 个滤波通道的中心频率；<>代表包络检测处理，采用简单的50 Hz截止频率的低通滤波器进行包络处理；A_s、β 和 δ 都是常数。为了将分母的零点稍微设置在单位圆内，常数 β 一般取值略小于1.0；较小的阈值 δ 可以抑制对振幅较小信号的响应；A_s 用于控制输入信号的线性范围。

图3展示了同步检测器计算SLR的算法，如果特定频率 f 的信号出现突出的峰值，则其AGC输出信号会表现出周期性。这样，中心频率与特定频率 f 最为接近的同步检测通道就能通过该通道与相邻通道间明显的响应差异来检测AGC信号的周期性。使用同步检测器来提取腭裂语音的特征参数有如下优点：(1)由于同步检测器是针对听觉模型输出信号周期性的检测，而不是频率信息，这样就避免了对一个强峰的二次谐波进行同步检测，检测信号的周期性也能更适应噪声；(2)计算“和波形”与“差波形”的比值并进行能量归一化，可以减少由声门激励包络引起的响应上的时间波动[26]。每个听觉通道经过同步检测器的处理都能计算其相应的SLR，所有通道的SLR能大致表达信号在频率轴上的周期性的分布，有着伪频谱的类似效果。以整个SLR伪频谱作为语音的特征参数，可用来表征正常语音与腭裂语音的差异。

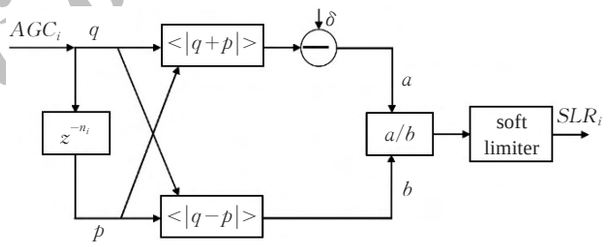

图3　同步检测器提取SLR算法

2.3　SVM分类器

SVM是一种典型的两类分类器，只能判断属于正类或是负类的问题。而腭裂语音高鼻音的正常、轻度、中度和重度等级的自动识别，属于一个四分类问题。为了处理多类问题的自动分类，国内外学者们一般采用间接法构造多类分类器，即通过组合多个二分类器来实现多分类器的构造。常用的构造方法有一对多法（one-versus-rest，1-v-r SVM）和一对一法（one-versus-one，1-v-1 SVM）。一对多法是依次将某个类别的样本作为正类，其余类别的样本都作为负类，分类时将未知样本判为具有最大分类函数值的那类。基于这种构造方法 K 个类别分类问题一共需要构造 K 个二类SVM分类器[27]。一对一法是在任意两类样本间设计一个二类SVM分类器，则 K 个类别分类问题需要 $\frac{K(K-1)}{2}$ 个二类SVM分类器，分类时将未知样本判为得票最多的那个类别。此外，应用于多分类的SVM还有层次支持向量机（H-SVM）、有向无环图多类SVM分类法（DAG-SVM）、纠错编码支持向量机（ECC-SVM）[28-30]。

本实验采用一对一法SVM分类器进行腭裂语音高鼻音等级的自动分类。对高鼻音等级的四分类需要设计六个SVM二分类器，分别为正常-轻度、正常-中度、正常-重度、轻度-中度、轻度-重度、中度-重度分类器。通过投票方式判定测试样本的所属样本，每个SVM分类器的超平面为：

$$\sum_{m=1}^{p}w_m x+b_m=0 \tag{7}$$

其中，w 表示SVM支持向量；p 表示支持向量个数；b 表示超平面偏置。

对一个全新的测试样本：

(1)计算测试样本与超平面的关系：

$$g(X)=\sum_{m=1}^{p}w_m X+b_m \tag{8}$$

(2)根据判别公式判定测试样本在当前分类器中的所属类别:

$$\delta(x)=\begin{cases}1, g(X)>0\\0, g(X)\leqslant 0\end{cases} \tag{9}$$

(3)统计六个SVM二分类器的投票结果:

$$Num(n)=\sum_{m=1}^{p}g(X_n) \tag{10}$$

(4)得票最多的类别则为该测试样本的所属类别:

$$n=\max(Num(n)) \tag{11}$$

3　实验结果与分析

3.1　实验数据

在临床应用上,医生希望得到腭裂语音高鼻音的等级(正常、轻度、中度、重度)。而目前的国内外研究主要局限于对高鼻音有无的判别,其一个原因是缺少标准的腭裂语音数据库。本实验采用的语音数据来自于四川大学华西口腔医院唇腭裂外科,数据按照"四川大学华西口腔医院语音矫治室普通话构音测量表"进行录制。该表充分考虑了普通话构音结构和腭裂语音的特性,分别有含元音/a/、/i/、/u/的词语各21个共63个,例如"爸爸""鼻子""布鞋"等词。语音数据由专业的言语治疗师进行判听,并为其人工划分高鼻音等级(四类)作为"金标准"。本实验一共采用56名儿童的语音数据,正常、轻度高鼻音、中度高鼻音、重度高鼻音儿童各14位,总共3 086个语音数据,具体包含的词语类别及高鼻音程度对应的个数如表1。

表1　样本种类及对应个数

含元音的词语	正常	轻度	中度	重度	总计
/a/	291	249	268	294	1 102
/i/	280	254	255	281	1 070
/u/	280	160	216	258	914
总计	851	663	739	833	3 086

3.2　腭裂语音高鼻音等级分类结果分析

本文对腭裂语音高鼻音等级(正常、轻度、中度、重度)自动分类的正确率达91.50%。分别测试了听觉滤波器种类和个数、侧抑制网络与同步检测器对分类结果的影响。分类器采用1-v-1 SVM分类器,使用十折交叉验证法测试系统识别准确率。

3.2.1　听觉滤波器种类的影响

基底膜是耳蜗的重要组织结构,从蜗底到蜗尖的基底膜可感受不同频率的音频。耳蜗可被简单认为是一个空间机械式频率分析器,可使输入信号在基底膜上呈现非线性分布。由此可将基底膜模拟成一个滤波器组,在听觉生理学中,临界频带描述了该"听觉滤波器"的中心频率和带宽。常用的临界频带有Bark尺度和ERB尺度,这两种尺度都能模拟人耳听觉的非线性特性,Bark尺度在掩蔽效应上有很大作用[31],而ERB尺度则更与心理声学模型中的等响度激励模型相关[32],两种尺度对应的中心频率和带宽也不相同。表2是Bark尺度和ERB尺度的频率群表,该频率群的划分是人耳听觉掩蔽效应的物理表现。本实验旨在对比分别使用两种临界频带尺度模拟基底膜非线性特性的性能,以及对腭裂语音高鼻音等级自动识别正确率的影响。分别测试了基于ERB尺度的Gammatone滤波器组(简称ERB滤波器组)[33]和基于Bark尺度的小波包滤波器组[34](简称Bark滤波器组)的识别正确率。使用Bark尺度时,各频段的中心频率是固定的,频带个数由信号采样率决定,而使用ERB尺度时,各频段的中心频率可根据频带个数计算。原始语音信号的采样率为22 050 Hz,信号截止频率为11 025 Hz,对应的1 Bark尺度滤波器共23个。ERB滤波器组同样使用23个滤波器。两种滤波器对应的中心频率(f_{ic})和临界带宽(B_{iw})如表2所示。

表2　Bark尺度和ERB尺度对应的中心频率和临界带宽

f_1~f_{23}	Bark尺度		ERB尺度	
	f_{ic}/Hz	B_{iw}/Hz	f_{ic}/Hz	B_{iw}/Hz
1	50	80	20	27
2	150	100	65	32
3	250	100	118	37
4	350	100	180	44
5	450	110	253	52
6	570	120	340	61
7	700	140	443	72
8	840	150	564	85
9	1 000	160	707	101
10	1 170	190	876	120
11	1 370	210	1 075	140
12	1 600	240	1 310	166
13	1 850	280	1 588	196
14	2 150	320	1 916	231
15	2 500	380	2 303	273
16	2 900	450	2 756	322
17	3 400	550	3 298	380
18	4 000	700	3 933	450
19	4 800	900	4 684	530
20	5 800	1 100	5 570	625
21	7 000	1 300	6 615	738
22	8 500	1 800	7 850	872
23	10 500	2 500	9 306	1 030

对比ERB滤波器组和Bark滤波器组的识别效果,只更改滤波器种类,采用相同的听觉模型、同步检测器和SVM分类器。表3和表4是两种滤波器的识别正确率。

表3　基于ERB尺度的23个GT滤波器的识别正确率　%

等级	正常	轻度	中度	重度
正常	85.18	5.00	1.88	2.86
轻度	7.87	89.09	6.58	5.24
中度	4.17	1.82	84.50	2.86
重度	2.78	4.09	7.04	89.04

表4　基于Bark尺度的23个小波包滤波器的识别正确率　%

等级	正常	轻度	中度	重度
正常	55.55	16.36	11.27	10.47
轻度	24.53	66.82	7.04	19.05
中度	7.42	4.09	70.89	8.58
重度	12.50	12.73	10.80	61.90

表3、表4中横排的等级为样本语音的正确等级，由四川大学华西口腔医院的专业语音师标注，列排为样本被分类器自动识别划分的等级。例如表3中第一列，一共有正常等级语音样本851个，其中被自动划分为正常等级的正确率为85.18%，分别有7.87%、4.17%、2.78%的样本被划分为轻度、中度和重度。从表3的基于ERB尺度的23个Gammatone滤波器算法的识别结果可以算出，四类等级自动划分的正确率为86.96%；从表4的基于Bark尺度的23个小波包滤波器算法的识别结果可以算出，四类等级自动划分的正确率为63.79%。明显得出ERB滤波器组比Bark滤波器组有更好的识别效果。由表2可以看出ERB尺度比Bark尺度具有更窄的临界带宽，能够更加细致地反映信号的感知，因此ERB滤波器组比Bark滤波器组能更好地模拟人耳基底膜时频域分析特性。

3.2.2　听觉滤波器个数的影响

基底膜在空间轴上被模拟为一组频率响应重叠的并联带通滤波器，原始语音信号经基底膜处理被分解为在不同位置上具有不同频率特性的并行输出时域信号。Gammatone(GT)滤波器的通道个数对应信号分频段分解的精度，通道个数越多，听觉谱的分辨率越高，提供的听觉参数特征越细致。同时，GT滤波器的通道个数也影响着ERB尺度的频率群的中心频率和带宽。国内外学者根据实际需求使用不同通道个数的GT滤波器，赵红等[35]使用21通道的GT滤波器修正多级线性预测以达到去混响目的；胡峰松等[33]、戴明扬等[36]和Hu等[37]使用64通道的GT滤波器实现听觉特征参数的提取；王迪等[38]使用128通道的GT滤波器实现了在听觉谱域上计算谐噪比。然而，由于相邻两个滤波器的功率谱响应曲线有一定的重叠部分，过于细致的临界频带划分会导致相邻滤波器的频窗重叠过多，使得两个相邻通道的输出信号也会出现“听觉掩蔽效应”[39]。同时，过多通道数的GT滤波器会大幅增加算法的时间成本。

本节实验目的在于权衡算法时间成本和识别正确率，获得本算法最适合的GT滤波器通道数。分别使用12、23、27、32、36、42、48、54、60、64、68、72通道的GT滤波器做基底膜滤波器组，其余听觉模型、同步检测处理、SVM分类器保持一致。表5是使用12种不同通道数的滤波器组对应的高鼻音四类等级识别正确率。其中，使用12通道GT滤波器时，识别正确率最低有80.56%；使用68通道GT滤波器时，识别正确率最高达91.50%。

表5　12种滤波器组对应的识别正确率

滤波器个数	识别正确率/%	滤波器个数	识别正确率/%
12	80.56	48	90.45
23	86.85	54	91.27
27	88.47	60	91.04
32	88.47	64	90.80
36	88.59	68	91.50
42	90.22	72	90.69

图4展示了它们对应的高鼻音四类等级识别的正确率趋势。从图4明显可以看出，随着GT滤波器的个数提高，高鼻音四类等级的识别率也在提高，GT滤波器增加56个通道，识别正确率提高了10.94%。当滤波器个数在48个及以上时，高鼻音识别正确率的提升不大，而增加滤波器个数会增加算法时间成本，因此权衡了识别正确率和算法时间成本，本文选择54通道的GT滤波器组作为基底膜模型。

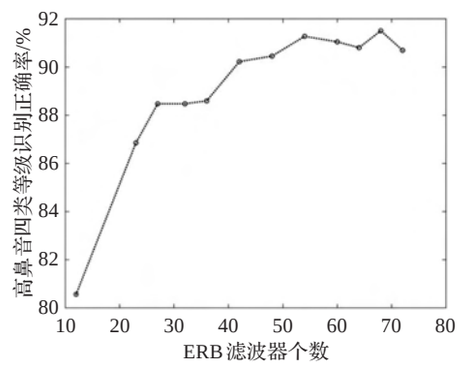

图4　使用不同通道的ERB滤波器的识别正确率

图5、图6、图7分别为对同一语音样本使用12、23、68通道GT滤波器的听觉伪频谱，可以明显看出使用68通道GT滤波器对应的听觉伪频谱具有更高的分辨率，提供了更多的信号谱特征参数的细节。

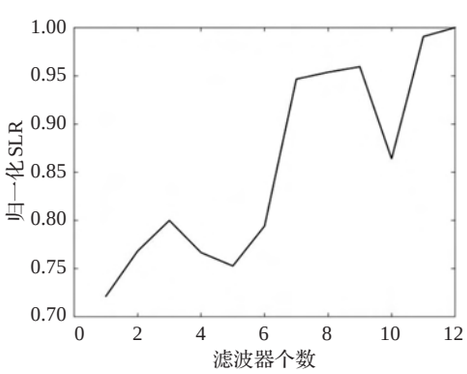

图5　12通道GT滤波器的听觉伪频谱

3.2.3　同步检测器和侧抑制网络的比较

听觉模型作为前端处理器，依次模拟了外中耳、基底膜、毛细胞和神经纤维的功能和机制，从多个并行通道上提取出听觉内部表达。听觉内部表达中包含着声激励的各种信息，正常语音和腭裂语音的频谱结构由于共振峰的差异而不尽相同，使用同步检测器和侧抑制网

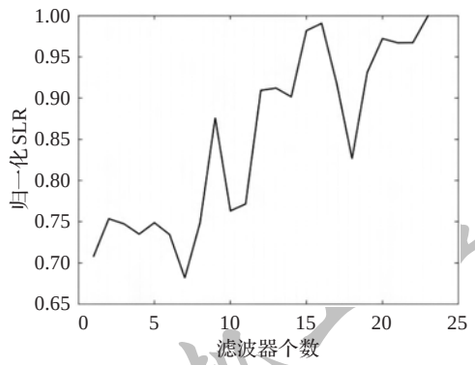

图6 23通道GT滤波器的听觉伪频谱

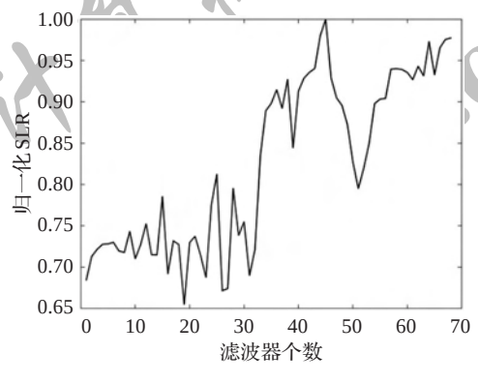

图7 68通道GT滤波器的听觉伪频谱

络等后处理都能从听觉内部表达中提取出听觉谱特征。这两种后处理的输入都是听觉神经纤维活动的时频域模式,两者都能锐化共振峰边缘,使谱结构轮廓更加明显,并提高频谱分辨率。侧抑制是一种非线性的神经抑制现象,侧抑制网络模拟相邻频段间的相互抑制和竞争关系,其每个通道的特征参数都会受到相邻通道的影响,其输出与相邻通道形成抑制连接[40];而同步检测器可加强共振峰结构和检测信号周期性[26],其特征参数只与当前通道有关,计算的是该通道的输入信号与其延时信号的比值。本小节实验对比了同步检测器和侧抑制网络提取的谱特征参数的识别正确率。使用相同的听觉模型、SVM分类器,其中基底膜模型使用54个基于ERB尺度的Gammatone滤波器。由表6可得,同步检测器提取的谱特征SLR作为特征参数进行高鼻音等级识别的正确率为91.27%;由表7可得,侧抑制网络提取的谱特征LIN作为特征参数的识别正确率为82.58%。

表6 同步检测器的SLR特征的识别正确率 %

等级	正常	轻度	中度	重度
正常	87.96	6.36	1.88	1.90
轻度	7.41	91.82	5.63	3.33
中度	2.31	0.91	91.08	0.48
重度	2.31	0.91	1.41	94.29

同步检测器是针对听觉模型输出信号周期性的检测,而不是频率信息,能更加适应噪声,其计算和波形与差波形的比值并进行能量归一化,可以减少响应中由声门激励包络引起的时间波动。

表7 侧抑制网络的LIN特征的识别正确率 %

等级	正常	轻度	中度	重度
正常	68.06	7.73	3.73	3.81
轻度	14.81	86.36	7.04	5.24
中度	6.94	1.36	86.38	1.43
重度	10.19	4.55	2.82	89.52

4 结束语

本文使用听觉模型和同步检测器提取了语音的伪频谱特征,并利用1-v-1 SVM分类器对腭裂语音高鼻音等级进行自动识别。对56位儿童共3 086个语音样本进行特征提取和模式识别,高鼻音等级自动识别率最高达91.50%。并对比了听觉模型中使用不同滤波器种类和个数,后处理中使用同步检测器和侧抑制网络提取谱特征对识别效果的影响。实验结果表明,使用基于ERB尺度的Gammatone滤波器组比基于Bark尺度的小波包滤波器组更能模拟人耳基底膜的时频域分析特性;使用54个ERB滤波器更能权衡算法时间成本和识别准确率;使用同步检测器提取的伪频谱参数有更高的识别率。充分说明,本文提出的基于听觉模型、同步检测器和SVM分类器的系统对腭裂语音高鼻音四类等级的自动识别具有有效性和可行性。高鼻音的等级评估可表征腭咽开口大小,在整个唇腭裂序列治疗中,可辅助医生诊断,客观、无创地为前期手术效果提供评定指标,并为后续手术提供参考指标,具有一定临床应用价值。

参考文献:

[1] 王光和,马莲.唇腭裂的序列治疗[J].医学研究杂志,2001,30(6):22-23.

[2] 周莹,王爱红.腭裂语音的研究进展[J].医学信息,2014(14):629-630.

[3] Castellanos G,Daza G,Sanchez L,et al.Acoustic speech analysis for hypernasality detection in children[C]//2006 International Conference of the IEEE Engineering in Medicine and Biology Society,2006:5507-5510.

[4] Bocklet T,Riedhammer K,Eysholdt U,et al.Automatic phoneme analysis in children with cleft lip and palate[C]//IEEE International Conference on Acoustics,Speech and Signal Processing,2013:7572-7576.

[5] Maier A,Hacker C,Schuster M.Analysis of hypernasal speech in children with cleft lip and palate[C]//International Conference on Text,Speech and Dialogue,2008:389-396.

[6] Nieto R G,Marín-Hurtado J I,Capacho-Valbuena L M,et al.Pattern recognition of hypernasality in voice of patients with cleft and lip palate[C]//2014 Symposium on Image,Signal Processing and Artificial Vision,2014:1-5.

[7] Orozco-Arroyave J,Belalcazar-Bolanos E,Arias-Londono J,et al.Characterization methods for the detection of multiple voice disorders:neurological,functional,and organic diseases[J].IEEE Journal of Biomedical & Health Infor-

matics,2015,19(6):1820-1828.
[8] Lee G S,Wang C P,Yang C C,et al.Voice low tone to high tone ratio:a potential quantitative index for vowel [a:] and its nasalization[J].IEEE Transactions on Biomedical Engineering,2006,53(7):1437-1439.
[9] Cruz C D L,Santhanam B.A joint EMD and Teager-Kaiser energy approach towards normal and nasal speech analysis[C]//50th Asilomar Conference on Signals,Systems and Computers,2016.
[10] Akafi E,Vali M,Moradi N.Detection of hypernasal speech in children with cleft palate[C]//19th Iranian Conference of Biomedical Engineering,2013:237-241.
[11] Rah D K,Ko Y L,Lee C,et al.A noninvasive estimation of hypernasality using a linear predictive model[J].Annals of Biomedical Engineering,2001,29(7):587-594.
[12] Vijayalakshmi P,Reddy M R,O'Shaughnessy D.Acoustic analysis and detection of hypernasality using a group delay function[J].IEEE Transactions on Biomedical Engineering,2007,54(4):621-629.
[13] Vijayalakshmi P,Nagarajan T,Rav J.Selective pole modification-based technique for the analysis and detection of hypernasality[C]//TENCON 2009-2009 IEEE Region 10 Conference,2009:1-5.
[14] Dubey A K,Prasanna S R M,Dandapat S.Zero time windowing analysis of hypernasality in speech of cleft lip and palate children[C]//22nd National Conference on Communication,2016:1-6.
[15] Cairns D A,Hansen J H L,Kaiser J F.Recent advances in hypernasal speech detection using the nonlinear teager energy operator[C]//4th International Conference on Spoken Language Processing,1996:780-783.
[16] Cairns D A,Hansen J H,Riski J E.A noninvasive technique for detecting hypernasal speech using a nonlinear operator[J].IEEE Transactions on Biomedical Engineering,1996,43(1):35.
[17] 黄生,何强,张有为.一种基于小波变换的声韵分割方法[C]//全国信号处理学术年会,1999.
[18] Ghitza O.Auditory models and human performance in tasks related to speech coding and speech recognition[J].IEEE Transactions on Speech & Audio Processing,1994,2(1):115-132.
[19] Karmakar A,Kumar A,Patney R K.A multiresolution model of auditory excitation pattern and its application to objective evaluation of perceived speech quality[J].IEEE Transactions on Audio Speech & Language Processing,2006,14(6):1912-1923.
[20] Huber R,Kollmeier B.PEMO-Q—a new method for objective audio quality assessment using a model of auditory perception[M].Piscataway:IEEE Press,2006.
[21] Plasberg J H,Kleijn W B.The sensitivity matrix:using advanced auditory models in speech and audio processing[J].IEEE Transactions on Audio Speech & Language Processing,2006,15(1):310-319.
[22] Chu W,Champagne B.A simplified early auditory model with application in audio classification[C]//2006 Canadian Conference on Electrical and Computer Engineering,2007:775-778.
[23] Jepsen M L,Ewert S D,Dau T.A computational model of human auditory signal processing and perception[J].Journal of the Acoustical Society of America,2008,124(1):422-438.
[24] 顾春.基于人耳感知模型的厅堂音质分析的研究[D].上海:同济大学,2005.
[25] 张焱.基于一种听觉模型的特征提取及语音识别[J].南京理工大学学报,1998,22(2):113-116.
[26] Seneff S.Pitch and spectral analysis of speech based on an auditory synchrony model[J].Journal of Hepatology,1985,32(S):2080-2082.
[27] 董荣胜,赵岭忠,蔡国永,等.基于对象的分布式实时系统调度模型研究[J].计算机研究与发展,2002,39(11):1464-1470.
[28] Azimi-Sadjadi M R,Zekavat S A.Cloud classification using support vector machines[C]//IEEE 2000 International Geoscience and Remote Sensing Symposium,2000:669-671.
[29] Platt J C.Large margin DAGs for multiclass classification[J].Advances in Neural Information Processing Systems,2000,12(3):547-553.
[30] Kindermann J,Leopold E,Paass G.Multi-class classification with error correcting codes[Z].2000.
[31] 高印寒,谢军,梁杰,等.基于小波分析的听觉滤波器组模型[J].吉林大学学报(工学版),2008,38(S1):179-183.
[32] 张伟豪,许枫.基于ERB尺度的心理声学模型及其数值计算[J].声学技术,2011,30(2):161-166.
[33] 胡峰松,曹孝玉.基于Gammatone滤波器组的听觉特征提取[J].计算机工程,2012,38(21):168-170.
[34] 王晓华,屈雷,张超,等.基于Fisher比的Bark小波包变换的语音特征提取算法[J].西安工程大学学报,2016,30(4):452-457.
[35] 赵红,李双田.Gammatone滤波器修正的多级线性预测去混响[J].信号处理,2014(9):1019-1024.
[36] 戴明扬,徐柏龄.基于听觉模型的话者特征参数提取及其在噪声背景下的话者辨识[J].应用声学,2001,20(6):6-12.
[37] Hu Z,Yue C,Luo Y,et al.Research of the auditory feature extraction algorithm based on Gammatone filter bank[C]//2017 IEEE 3rd Information Technology and Mechatronics Engineering Conference,2017:444-449.
[38] 王迪,付强,杨琳,等.基于人耳听觉模型的自动嗓音评估方法[J].物理学报,2008,57(7):4244-4250.
[39] 詹海峰,田红心,牛博,等.基于多分辨率高斯滤波器组的时频分析方法[J].中国电子科学研究院学报,2017(6):654-661.
[40] Ali A M A,Van der Spiegel J,Mueller P.Robust auditory-based speech processing using the average localized synchrony detection[J].IEEE Transactions on Speech & Audio Processing,2002,10(5):279-292.