宽窄带语谱图融合分带投影的特定人汉语词汇识别

魏 莹[1] 王双维[1] 潘 迪[1] 张 玲[2] 许廷发[3] 梁士利[1]
(东北师范大学物理学院 长春 130024)[1] (长春理工大学理学院 长春 130022)[2]
(北京理工大学光电成像技术与系统教育部重点实验室 北京 100081)[3]

摘 要 提出一种基于宽窄带语谱图融合分带投影的方法对特定人二字汉语词汇进行识别。该方法将图像处理技术应用到语音识别领域,在图像特征提取过程中,首先对窄带语谱图进行等宽度分带行投影和二进宽度分带行投影,并将其分别作为窄带语谱图的第 1 个特征集合和第 2 个特征集合,同时将窄带语谱图进行再次图像傅里叶变换之后进行等宽度行投影,作为第 3 个特征集合。然后对宽带语谱图进行等宽度分带列投影,作为第 4 个特征集合。将上述特征集合作为识别的特征向量,以支持向量机为分类器进行特定人二字汉语词汇整体识别。采用 1000 个语音样本进行仿真实验,结果表明,采用前 3 个特征集合的特征向量对特定人二字汉语词汇识别的正确识别率可达 92.4%,采用第 4 个特征集合的特征值对特定人二字词汇识别的正确识别率可达 80%,而采用上述 4 个特征集合的特征值融合对特定人二字汉语词汇识别的正确识别率可达 95.4%。该特征融合的方法为汉语词汇的识别提供了新的思路。

关键词 语音识别,语谱图,特征融合,行投影,列投影,支持向量机(SVM)
中图法分类号 TP391.42 **文献标识码** A

Specific Two Words Chinese Lexical Recognition Based on Broadband and Narrowband Spectrogram Feature Fusion with Zoning Projection

WEI Ying[1] WANG Shuang-wei[1] PAN Di[1] ZHANG Ling[2] XU Ting-fa[3] LIANG Shi-li[1]
(Department of Physics, Northeast Normal University, Changchun 130024, China)[1]
(College of Science, Changchun University of Science and Technology, Changchun 130022, China)[2]
(Key Laboratory of Photo Electronic Imaging Technology and System, Ministry of Education of China, Beijing Institute of Technology, Beijing 100081, China)[3]

Abstract A method based on broadband and narrowband spectrogram fusion with zoning projection of specific two words Chinese lexical recognition was presented. In the process of image feature extraction, the image processing technique is applied to the speech recognition field. Firstly, equal width zoning line projection and binary width zoning line projection are carried out to the narrowband spectrogram, and they are set respectively as the narrowband spectrogram of the first characteristic set and the second characteristic set. Meanwhile, equal width zoning line projection is carried out again to the narrowband spectrogram after Fourier transform, treating it as the third feature set. Then, equal width column projection is carried out to the broadband spectrogram, regarding it as the fourth feature set. The above three feature sets are used as feature vectors to support vector machine(SVM) as a classifier for the overall recognition of specific two words Chinese vocabulary. 1000 voice samples are used in simulation experiment. The results show that the correct recognition rate of the two words Chinese word recognition by the first three feature sets is 92.4%. The correct recognition rate of two words vocabulary recognition using fourth feature sets is 80%. The correct recognition rate of the two words Chinese word recognition by using the feature value fusion of the above four features can reach 95.4%. This method of feature fusion provides a new way of thinking of Chinese vocabulary overall recognition.

Keywords Speech recognition, Spectrogram, Feature fusion, Line projection, Column projection, Support vector machine (SVM)

1 前言

语音信号是一种非平稳随机信号,考虑到人类发音器官在发声过程中的变化速度具有一定的限度,可以认为语音信号是短时平稳的,即在 10～30ms 的时间段内可以应用平稳随机过程的分析方法。然而,汉语语音清辅音多,开口呼出的音节占全部音节的一半以上,大多是弱轻音,且汉语的信息为声母、韵母及声调的整体表现。若采用常规的语音端点检测和语音参数分析,如平均能量、共振峰、倒谱等,会漏掉汉语语音中的重要信息。因此,寻找能够体现汉语语音整体化特征

本文受国家自然科学基金项目(61471111)资助。
魏 莹(1991－),女,硕士生,主要研究方向为信号处理,E-mail:lsl@nenu.edu.cn;王双维(1957－),男,硕士,教授,主要研究方向为声信号处理;潘 迪(1991－),女,硕士生,主要研究方向为信号处理;张 玲(1969－),女,硕士,副教授,主要研究方向为专业光学;许廷发(1968－),男,博士,教授,主要研究方向为光电子;梁士利(1968－),男,博士,副教授,主要研究方向为信号处理。

的处理方法对汉语语音识别显得尤为重要。

20 世纪 50 年代，贝尔实验室的 Davis 等研究人员发明了语音信号的语谱图，它是频谱能量随着时间变化的二维图，从而使我们看到了“语音”[1]。1995 年，潘凌云等人将语谱图应用到语音识别中的语音音素分割中[2]，该方法作为语音识别系统的一部分，已经在一个语音分析系统中使用。Victor W. Zue 和 Ronald A. Cole 做了若干关于语谱图阅读的实验来尝试用语谱图进行语音识别[3]。而 Denntis H. Klatt 和 Kenneth N. Stevens 则通过可视语谱图检验和机器帮助下的词汇搜索对一组未知句子进行识别[4]。Brian E. D Kingsbury 等人将调制语谱图应用到自动语音识别系统中[5]，该方法能够提高 ASR 系统的鲁棒性。Higuchi[6] 提出将两幅语谱图进行图像融合处理并提取出特征量进行语音分离的方法。近年来，应用语谱图模式匹配的“唱歌声音识别”[7] 方法，采用融合增强光谱图频道的监控语音识别方法[8]，应用语气核模型来进行汉语词汇语气的识别[9]，采用角色标注的方法进行中文人名自动识别[10]，采用构建支持向量机（SSVM）的方法进行中等大词汇量语音识别[11]，这些方法为语音识别提供了一个全新的视角。基于段式情感识别的汉语普通话连续语音，Neammalai P 等人提出将语音信号变换成语谱图图像并对语谱图图像进行二维傅里叶变换，计算出的能量作为语音和音乐分类的特征向量[12]，使得语谱图与傅里叶变换更好地结合。Prabhakar Chandrasekaran 等研究人员从语音信号的可视化特征——语谱图中提取熵序列作为语音识别系统的特征参数，开辟了语音信号处理与图像信号处理相结合的新领域[13]。与此同时，Awais[14] 将快速傅里叶变换语谱图的音素分割算法应用到连续阿拉伯语言中，并取得了显著的成果。Asahi Keepsake[15] 将带噪的语音信号通过对其相应的语谱图进行图像处理，提高了语音识别的鲁棒性。Ajmera 和 PK[16] 提出了一种对语音语谱图进行氡转换（RT）和离散余弦变换（DCT）的新的说话人识别的特征提取技术。Kekre H. B 使用光谱图的行平均向量进行说话人识别[18]。Steinberg R 等研究员提出数学形态学操作和分水岭变换的算法对宽带语谱图进行分割，并用于自动语音识别[20]。将语谱图用于低信噪比环境中的“语音端点检测”[19] 的方法；将谱图的“纹理图像”作为特征的语音情感识别[21] 的方法；王双维课题组依据语谱图纹理方位的数学形态学特征进行了汉语韵母声调识别研究[19]；Dutta Tridibesh[22] 研究使用频率变化和振幅随时间变化的复杂模式，通过谱图分割和模板匹配方式对一个给定的单词进行识别，这些方法使我们对语谱图的广泛应用有了更加深刻的认识。语谱图作为语音分析和语音学的有利工具，可将密切相关的时域与频域特征及其相互关系同时展现出来。因此，语谱图更加有利于表征语音信号的整体性。随着图像处理技术的发展，利用语谱图作为提取语音识别的特征参数实现语音识别，已经取得了一定效果。

由于窄带语谱图有较高的频域分辨率而宽带语谱图有较高的时间分辨率，这二者的结合能够更好地反映出语音信号的整体性。因此本文提出以窄带语谱图行投影值和宽带语谱图列投影值的融合作为特征集合，采用支持向量机作为分类器，实现特定人二字汉语词汇的识别。仿真实验表明，选取窄带语谱图分带行投影特征值作为识别特征量对特定人二字词汇的识别率可达 92.4%；将选取宽带语谱图分带列投影得到的特征值作为识别特征量的识别率可达 80%；而采用宽、窄带语谱图特征融合得到的特征值对特定人二字汉语词汇的识别率可达 95.4%。这为解决汉语词汇整体语音识别提供了一种新的思路。

2 语谱图

语谱图（Spectrogram）[23] 是表示语音频谱随时间变化的图形，其纵轴为频率，横轴为时间，任一给定频率成分在给定时刻的强弱用相应点的灰度或色调的浓淡来表示。语谱图在语音分析中具有重要的价值，被视为可视语言。不同语音的语谱图其形成纹络均有很大区别。语谱图中显示了大量的与语音的特性有关的信息，它综合了频谱图和时域波形的特性，明显地显示出了语音频谱随时间的变化情况。因此，语谱图所承载的信息量远远大于单纯时域和单纯频域承载信息量的总和[24]。语谱图通常分为宽带语谱图、窄带语谱图、断面语谱图以及等强度线语谱图 4 种。本文尝试采用宽带、窄带两种语谱图信息融合方法进行汉语二字词汇语音识别。

信息融合[25] 是在多级别、多方面对单源和多源的数据和信息进行自动检测、关联、相关、估计和组合的过程。它是依据输入信息的抽象层次将信息融合分为 3 级，包括数据级（或称像素级）融合、特征级融合以及决策级融合[26]。

本文采用的是特征级融合。它是指在各个传感器提供的原始信息中，首先提取一组特征信息，形成特征矢量，并在目标进行分类或其他处理前对各组信息进行融合，一般称其为中级融合。其优点是实现了可观的信息压缩，有利于实时处理，所提取的特征直接与决策分析有关。但是由于宽窄带语谱图都是图像，它反映的是同一词汇的不同信息，所以宽窄带语谱图特征级融合也相当于图像融合[27]。

窄带语谱图有着较高的频率分辨率，在谱图上能显示出两个纯音，但其时间分辨率较差，看不出两个纯音所产生的拍音[28]；而宽带语谱图有较高的时间分辨率，便于观察频谱包络，以便确定共振峰，同时也可以给出精确的时间结构。所以应用宽、窄带语谱图融合可以更好地反映语音信号的整体特性。

因为现代汉语词典[29] 中汉字大约有 10 万条，其中单字词约有 5000 条，二字词约有 6 万条，三字词和四字词约有 1 万条，因此二字词汇占汉字词汇的 60%左右，这说明二字汉语词汇在权威字典中占的比例非常大，这也是我们采用二字汉语词汇来进行识别的原因。根据现代汉语常用词汇表中使用频率较高的汉语普通话常用词语 56008 个，其中单音节词 3181 个，双音节词语 40351 个，三音节词语 6459 个，四音节词语 5855 个，五音节和五音节以上词语 162 个，由此可见双音节词语占所有常用词语比例的 72%，在常用词语中起到一个不可估量的作用。

因此本文选用 10 个二字汉语词汇进行语音识别，每个词汇分别用词汇编码代替，汉语词汇表如表 1 所列。其中第 1 行为词汇编码，第 2 行为汉语词汇发音，第 3 行为中文含义。

表 1 汉语词汇表

词汇编码	词 1	词 2	词 3	词 4	词 5	词 6	词 7	词 8	词 9	词 10
词汇发音	[meiou]	[tsitθi]	[tsuŋkuo]	[k′γji]	[w∂nt′i]	[kuŋtsuo]	[s′∂ŋxuo]	[tsγaŋ]	[itθiŋ]	[tsγθiε]
词汇含义	没有	自己	中国	可以	问题	工作	生活	这样	已经	这些

图 1 为某发音人词汇内容为“中国”的时域波形图，图 2 为相应的窄带语谱图(带宽为 43Hz)，图 3 为相应的宽带语谱图(带宽为 344Hz)。为重点突出，图中显示 4kHz 以下部分。

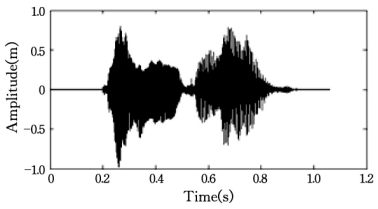

图 1　词汇“中国”的时域波形图

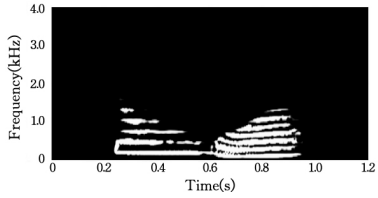

图 2　词汇“中国”的窄带语谱图

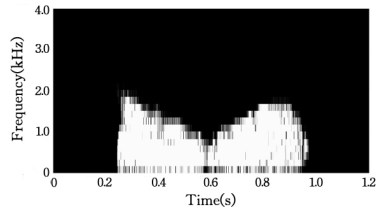

图 3　词汇“中国”的宽带语谱图

3　宽窄带语谱图矩阵的特征提取

3.1　语谱图样本构成

用 Cool Edit Pro 2.0 软件进行语音录制，采样率为 44.1 kHz，使得语谱图频域表达范围约为 0～22kHz，单声道，16bit 量化。实验共选取 10 人(男、女各 5 人)10 个词汇的读音样本，10 个词汇均为二字词汇，重复 10 遍，即每个词汇有 10 个样本，一个词汇的语音时长约为 1.2s，将每个词汇的 10 个样本截取为 10 个语音帧长分别为 1.2s 的音频文件，10 个人的 10 个词汇共为 1000 个语音样本文件，将所有语音样本文件转化为 MATLAB 数据文件，即语音样本序列。

窄带语谱图的构成：对每个样本序列进行分帧，帧长 1024 点，为保持其连续性，采用重叠率为 25%的帧移量，窗函数采用汉明窗(Hamming)，其公式为

$$w(n)=0.54-0.46\cos\left(\frac{2\pi n}{N-1}\right),0\leqslant n\leqslant N-1 \tag{1}$$

每个样本分为 68 帧，构造出 1024 行 68 列时域分帧矩阵。对时域分帧矩阵做 FFT，生成 1024 行 68 列时频分析矩阵，频域分辨率为 43Hz。时频分析矩阵的模矩阵即为样本所对应的语谱图矩阵。由于傅里叶变换具有对称性，取该矩阵的上半部或下半部作为语谱图即可，因此，每一幅语谱图的矩阵为 512 行 68 列，共 1000 幅灰度图像。

宽带语谱图的构成：对每个样本序列进行分帧，帧长 128 点，同样采用重叠率为 25% 的帧移量，窗函数采用汉明窗(Hamming)，每个样本分为 540 帧，构造出 128 行 540 列时域分帧矩阵。对时域分帧矩阵做 FFT，生成 128 行 540 列时频分析矩阵，频域分辨率为 344Hz。时频分析矩阵的模矩阵即为样本所对应的语谱图矩阵。由于傅里叶变换具有对称性，取该矩阵的上半部或下半部作为语谱图即可，因此，每一幅语谱图的矩阵为 64 行 540 列，共 1000 幅灰度图像。

以上过程形成了参数可调的 Matlab 语谱图生成程序，以备随时调用。为消除由于音量不同造成的各个样本幅度差异，对每个图像矩阵均进行归一化处理。

3.2　窄带语谱图矩阵特征值提取

3.2.1　等宽度分带行投影

窄带语谱图矩阵的每一行代表某一频率通道幅度特性随时间的变化，行投影则反映了某频率通道在整个语音时长过程中的总体特征。如果简单地对语谱图矩阵进行行投影，可以得到 512 个频率通道的投影值。但这种频域上过于细化的投影方式不仅对语义识别没有益处，反而会降低识别系统的容错能力。因此，我们采取了等宽度行投影方式，由于 512 行分成 10 带不是整数，还余 2 行，且高频部分信息量很少可忽略，所以从第 3 行开始划分，即语谱图矩阵每 51 行作为一个带进行行投影，每个矩阵可以得到 10 个带的投影值，构成一个 10 行列向量。因此可以构造每个词汇 10 行 10 列等宽度分带投影矩阵，对每个词汇投影矩阵值的各个行求平均值和方差。对 10 个词汇对应带投影值的两两 U 检验发现，只有分带投影矩阵的第 7 行投影值具有显著性差异，可以作为特征数据集合。U 检验公式如(2)式所示。

$$u=\frac{\overline{X_1}-\overline{X_2}}{\sqrt{a+b}} \tag{2}$$

其中，$a=\frac{S_1^2}{N_1}$，$b=\frac{S_2^2}{N_2}$。

U 检验结果(部分数据)如表 2 所列。为了清楚地观察到数据之间的识别差异，在此规定当 $U\geqslant 1.96$ 时，取值为 1；$U<1.96$ 时，取值为 0，得出 U 检验真值表如表 3 所列。由表 3 可以看出，第 7 行投影值能够作为识别词汇特征量的比例为 90%，由于篇幅所限，其他结果不再详述。

表 2　U 检验结果表(部分数据，第 7 行投影值)

词	没有	自己	中国	可以	问题	工作	生活	这样	已经	这些
没有	0	5.2862	10.8767	8.0706	16.4753	14.4032	1.3704	8.3833	2.2353	1.7512
自己	5.2862	0	8.7069	7.9601	9.7214	9.2471	5.4561	7.8522	3.5273	3.8115
中国	10.8767	8.7069	0	2.1092	3.5421	1.6447	4.3536	3.0806	7.9826	7.5688
可以	8.0706	7.9601	2.1092	0	5.6614	3.9286	3.1287	0.6371	6.6843	6.2610
问题	16.4753	9.7214	3.5421	5.6614	0	2.5396	6.0837	7.6860	9.8764	9.4752
工作	14.4032	9.2471	1.6447	3.9286	2.5396	0	5.2593	5.5364	9.0366	8.6246
生活	1.3704	5.4561	4.3536	3.1287	6.0837	5.2593	0	2.8921	2.7604	2.3902
这样	8.3833	7.8522	3.0806	0.6371	7.6860	5.5364	2.8921	0	6.5651	6.1288
已经	2.2353	3.5273	7.9826	6.6843	9.8764	9.0366	2.7604	6.5651	0	0.3902
这些	1.7512	3.8115	7.5688	6.2610	9.4752	8.6246	2.3902	6.1288	0.3902	0

表 3　U 检验真值表(第 7 行投影值)

词	没有	自己	中国	可以	问题	工作	生活	这样	已经	这些
没有	0	1	1	1	1	1	0	1	1	0
自己	1	0	1	1	1	1	1	1	1	1
中国	1	1	0	1	1	0	1	1	1	1
可以	1	1	1	0	1	1	1	0	1	1
问题	1	1	1	1	0	1	1	1	1	1
工作	1	1	0	1	1	0	1	1	1	1
生活	0	1	1	1	1	1	0	1	1	1
这样	1	1	1	0	1	1	1	0	1	1
已经	1	1	1	1	1	1	1	1	0	0
这些	0	1	1	1	1	1	1	1	0	0

3.2.2　二进宽度分带行投影

从窄带语谱图灰度图像中可以发现,图像的大量信息集中分布在图像的中下部分,这一点符合人类语言信息主要分布在低频段的特征。为了使特定人的二字词汇的语义识别更加准确,同时又能将灰度图像的中下部分的信息更清楚地显示,我们采取了二进宽度分带投影的方法,从第 1 行开始二进分,即将每个语谱图矩阵分为 7 个带,分别为 1－256 行(带宽 256 行)、257－384 行(带宽 128 行)、385－448 行(带宽 64 行)、449－480 行(带宽 32 行)、481－496 行(带宽 16 行)、497－504行(带宽 8 行)、505－512 行(带宽 8 行),不将最后 8 行再分带是因为最后一个带的频率范围在 0～200Hz,而人类所能听到的频率在 100Hz 以上,所以最后 8 行相当于只有 4 行是有效的,将这 7 个带构造成 7 行 10 列二进宽度分带投影矩阵。通过对 10 个词汇之间对应带投影值进行两两 U 检验发现,第 4 行到第 6 行 3 个带投影值有显著性差异,可以作为特征数据集合。

3.2.3　语谱图傅里叶变换矩阵的等宽度分带行投影

窄带语谱图图像中像素的灰度值代表了信号在相应频率、相应时刻的幅度比重。考虑到语谱图虽然反映了语音的时频特征,但其可视化形式为图像,属于二维空域信号,基于图像处理思路,仍可以对其进行图像频谱分析。为此,本文对语谱图图像进行再次傅里叶变换,傅里叶变换后语谱图的图像低频部分被分到了图像的中间部分,将变换后的语谱图图像频域矩阵进行等宽度行投影,构造每个词汇 10 行 10 列等宽度分带投影矩阵。通过 U 检验发现,每个矩阵的第 6 行到第 9 行 4 个带投影值适合作为特征数据集合。

3.3　宽带语谱图矩阵特征值提取

经大量实验表明,上述 8 个特征值构造出的 8 维特征向量用于 10 人的二字汉语词汇的识别率为 91.4%,效果并不十分理想。

考虑到宽带语谱图矩阵的每列代表着某一时刻的频谱结构,列投影则反映了语音各时域帧长的能量分布。同样地,如果对语谱图矩阵进行列投影,这种细化的投影方式也会降低识别系统的容错能力。通过观察语音波形发现,语音帧长为 120ms,而且所有语音波形都分布在中间部分,若将 540 列分成 10 列,则相当于只有中间 8 列有效,而这 8 列基本上对应二字词汇的声母、韵母,因此,我们采取了等宽度列投影方式,将语谱图矩阵每 54 列作为一个带进行列投影,每个矩阵可以得到 10 个带的投影值,构成一个 10 行列向量。将这 10 个带的投影值构成 10 维特征向量并用于 10 人的二字汉语词汇识别,识别率达到 80%,效果也不理想。由于窄带语谱图有较高的频域分辨率而宽带语谱图有较高的时间分辨率,这二者的结合能够更好地反映出语音信号的整体性,因此将宽窄带语谱图矩阵特征值融合到一起来进行语音识别。

同样对 10 个带投影值进行两两 U 检验,可以发现,只有分带投影矩阵的第 3 行到第 7 行投影值具有显著性差异,因此将分带投影矩阵的第 3 行到第 7 行 5 个带投影值作为特征数据集合。

加上前 8 个特征量,特征量变为 13 个,将这 13 个特征量构成 13 维特征向量,用来进行特定人的二字词汇的语义识别,会使系统的识别率大大提高。

4　实验仿真结果与分析

4.1　系统设置

本次语音样本采用 10 个人对 10 个二字词汇进行录制,采样频率为 44.1kHz,单声道,16bit 量化,其中每个词汇 10 段重复录音,一共是 1000 个语音数据样本。为了采样数据的准确性和可说服性,将每个人的 10 个二字词汇的每前 5 遍作为训练集,后 5 遍作为测试集,即前 500 个语音数据作为训练集,后 500 个语音数据作为测试集。为了后面的数据处理的方便和保证程序运行时收敛加快,防止出现奇异样本数据(是指相对于其他输入样本特别大或特别小的样本矢量),对 1000 个语音数据进行归一化的预处理,使所有数据得到统一。

支持向量机(Supported Vector Machine,SVM)[30] 方法是建立在统计学习理论的 VC 维理论和结构风险最小原理基础上的,根据有限的样本信息在模型的复杂性(即对特定训练样本的学习精度)和学习能力(即无错误地识别任意样本的能力)之间寻求最佳折衷,以期获得最好的推广能力。

本文将处理后的特征数据送到支持向量机工具 LIBSVM[31] 中去训练和检测,同时采用线性核函数。在训练阶段,将准备好的前 500 个语音训练样本的特征数据存入数据库,作为支持向量机(SVM)的训练模板,对其进行训练。在检测阶段,将后 500 个语音检测样本提取的特征数据放入到训练好的网络中,对相应的特定人的二字词汇进行语义识别。

图 4 所示为本文所使用的支持向量机结构示意图:当 $N=9$ 时,为窄带语谱图特征提取的特征值集合对特定人二字词汇识别的支持向量机结构示意图;当 $N=10$ 时,为宽带语谱图特征提取的特征值集合对特定人二字词汇识别的支持向量机结构示意图;当 $N=13$ 时,为宽窄带语谱图融合的特征值集合对特定人二字词汇识别的支持向量机结构示意图。它包括输入层、中间层和输出层。首先通过非线性变换将输入空间变换到一个高维空间,然后在这个新空间中求取最优线性分类面。在本文中,这种非线性变换是通过中间层定义的线性核函数来实现的,其输出层是若干中间层节点的线性组合,而每一个中间层节点对应于输入样本与一个支持向量的内积。本文对 10 人的 10 个词汇的语音进行识别,所以采用基数词 1－10 的编码方式,即输出维度为 1。

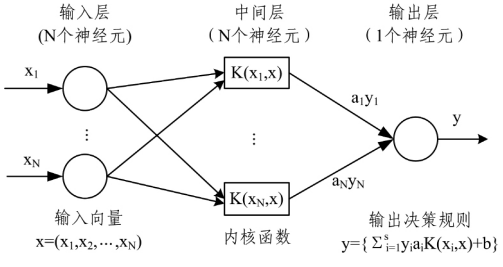

图 4　支持向量机结构示意图

4.2 仿真结果分析

4.2.1 窄带语谱图矩阵提取出的特征值作为特定人二字汉语词汇识别的仿真结果

对窄带语谱图矩阵等宽度分带行投影和二进宽度分带行投影，以及窄带语谱图傅里叶变换后矩阵的分带行投影，得出1000组数据，每组数据由9维特征向量组成，分别相应地将10个人的相同词汇前5遍作为训练样本数据，后5遍作为检测样本数据，即500组数据作为训练样本，500组数据作为检测样本。

通过前500组数据对多分类支持向量机进行反复训练，得到最佳识别系统。将后500组数据放入训练好的系统中进行检测，得出对10人的二字汉语词汇的正确率识别可达92.4%。

4.2.2 宽带语谱图矩阵提取出的特征值作为特定人二字汉语词汇识别的仿真结果

对宽带语谱图矩阵进行等宽度分带列投影，得出1000组数据，每组数据由10维特征向量组成，分别相应地将每个人的10个词汇前5遍作为训练样本数据，后5遍作为检测样本数据，即500组数据作为训练样本，500组数据作为检测样本。将后500组数据放入训练好的系统中进行检测，得出对特定人的二字词汇的语义识别正确率可达80%。

4.2.3 宽窄带语谱图矩阵提取出的特征值作为特定人二字汉语词汇识别的仿真结果

对窄带语谱图矩阵等宽度行投影和二进宽度行投影，语谱图傅里叶变换后矩阵的行投影，以及宽带语谱图矩阵的等宽度列投影，即宽窄带语谱图特征值融合构成特征向量，得出1000组数据，每组数据由13维特征向量组成，分别相应地将每个人的10个词汇前5遍作为训练样本数据，后5遍作为检测样本数据，即500组数据作为训练样本，500组数据作为检测样本。由于数据样本量太大，本文只各选择一组训练样本数据和检测样本数据给予显示，训练样本数据如表4所列，检测样本数据如表5所列。

表4 一组训练样本数据特征值

词 / 特征量	没有	自己	中国	可以	问题	工作	生活	这样	已经	这些
1	0.4442	0.5553	0.0502	0.1293	0.0641	0.0351	0.1244	0.1256	0.0701	0.0692
2	0.7658	0.1061	0.5558	0.5553	0.1289	0.5737	0.4288	0.9415	0.0849	0.4855
3	0.4977	0.4766	0.7337	0.3890	0.2569	0.5556	0.5288	0.4902	0.3678	0.5056
4	0.1390	0.2465	0.0734	0.1169	0.1009	0.0602	0.3270	0.0998	0.4258	0.7539
5	0.4403	0.4859	0.4562	0.5460	0.7447	0.5024	0.4552	0.4986	0.4891	0.4794
6	0.5271	0.5145	0.5493	0.6342	0.7716	0.6521	0.4891	0.5936	0.5429	0.5768
7	0.4587	0.5066	0.5311	0.5200	0.8136	0.5363	0.5318	0.4899	0.5111	0.4650
8	0.6771	0.5930	0.6237	0.8407	0.8473	0.7444	0.5044	0.7027	0.5935	0.6417
9	0.8757	0.7384	0.5918	0.0323	0.2204	0.7569	0.6416	0.7207	0.4598	0.6634
10	0.7999	0.7313	0.5587	0.4230	0.5903	0.3584	0.2084	0.5783	0.1534	0.2089
11	0.3850	0.1150	0.2317	0.7680	0.1926	0.2356	0.1001	0.2512	0.0930	0.1994
12	0.6565	0.1933	0.3726	0.3830	0.0539	0.0741	0.1560	0.7509	0.6187	0.2846
13	0.4248	0.2235	0.6452	0.1741	0.0992	0.5791	0.4264	0.3496	1.0000	0.5843

表5 一组检测样本数据特征值

词 / 特征量	没有	自己	中国	可以	问题	工作	生活	这样	已经	这些
1	0.5249	0.6212	0.0566	0.1526	0.0701	0.0240	0.1920	0.1299	0.0253	0.0501
2	0.7197	0.0926	0.6002	0.4397	0.0803	0.5384	0.6358	1.0000	0.0637	0.4286
3	0.5733	0.5453	0.7534	0.4949	0.2372	0.6069	0.6024	0.4761	0.2853	0.4820
4	0.1722	0.1924	0.1089	0.1442	0.0680	0.0560	0.2861	0.0975	0.1241	0.8552
5	0.4295	0.4362	0.4707	0.5619	0.9137	0.5402	0.4413	0.5071	0.6680	0.4958
6	0.4905	0.5001	0.5508	0.6257	0.9177	0.7161	0.4891	0.5385	0.7232	0.5323
7	0.4397	0.4743	0.5391	0.5705	0.8179	0.5891	0.4679	0.4922	0.6049	0.4760
8	0.5743	0.5591	0.6591	0.8359	0.8500	0.8720	0.5431	0.6243	0.8620	0.6975
9	0.8409	0.5336	0.7090	0.4259	0.3573	0.5779	0.6197	0.6849	0.1220	0.7929
10	0.7453	0.8584	0.3163	0.8228	0.2906	0.2437	0.3187	0.2516	0.1083	0.1833
11	0.4969	0.2362	0.1783	0.7782	0.0869	0.0354	0.1436	0.7000	0.0284	0.2814
12	0.7961	0.1477	0.0918	0.2762	0.0397	0.4256	0.0499	0.6265	0.2933	0.2760
13	0.4387	0.1858	0.5533	0.0680	0.0756	0.6401	0.3659	0.0621	0.8884	0.7604

将后500组数据放入训练好的系统中进行检测，对10人的二字汉语词汇的正确率识别可达95.4%。

结束语 本文提出了基于语谱图宽窄带融合分带投影的特定人二字汉语词汇识别的方法。由于窄带语谱图有较高的频域分辨率而宽带语谱图有较高的时间分辨率，这二者的结合能够更好地反映语音信号的整体性，从而更加精确地识别特定人不同语音的语义。因此将窄带语谱图矩阵进行等宽度分带行投影和二进宽度分带行投影，及其再次傅里叶变换矩阵的行分带投影，同时对宽带语谱图矩阵进行等宽度分带列投影，将投影值作为特征集合，以支持向量机为分类器，对特定人的二字汉语词汇进行识别。实验结果表明，采用宽、窄带语谱图特征融合得到的特征值对特定人二字汉语词汇的识别率可达95.4%。后续可以通过扩大语音数据库样本容量来提高二字词汇语义识别的精度和准确性。另外，该方法仅实现了二字词汇的语音识别，对于多字词汇量语音的识别还有待进一步研究。

参 考 文 献

[1] 蔡莲红，黄德智，蔡锐. 现代语音技术基础与应用[M]. 北京：清华大学出版社，2003

（下转第232页）

[6] LIU Xue. The color image registration technique based on feature [D]. Changsha: National University of Defense Technology, 2010

[7] 李婷婷,汤进,江波. 迭代的图变换匹配算法[J]. 中国图象图形学报,2014,19(5):723-729

[8] 李敏强. 遗传算法的基本理论与应用[M]. 科学出版社,2002

[9] 练仕榴,郑刚,牟善玲. 用于心电波形分析的相似性度量策略[J]. 计算机技术与发展,2011,37(9):263-265

[10] 刘小丹,吴笑嫦,曹耐. 基于颜色不变量和仿射不变性的彩色图像配准[J]. 计算机工程与设计,2014,35(12):4243-4248

(上接第 219 页)

[2] 潘凌云,孙达传,吴美朝. 语音识别中基于语谱图的语音音素分割方法[J]. 杭州大学学报(自然科学版),1995,22(1):42-46

[3] Zue V W, Lamel L F. An Expert Spectrogram Reader: A Knowledge—Based Approach to Speech Recognition[C]// IEEE International Conference on Acoustics, Speech, and Signal Processing. 1986:1197-1200

[4] Klatt D H, Stevens K N. On the Automatic Recognition of Continuous Speech: Implications from a Spectrogram—Reading Experiment[J]. IEEE Transactions on Audio and Electroacoustics, 1973,21(3):210-217

[5] Riley M D. Schematizing Spectrograms for Speech Recognition [J]. J. Acoust. Soc. Am. ,1983,73(1):36-46

[6] Kingsbury B E D, Morgan N, Greenberg S. Robust speech recognition using the modulation spectrogram[J]. Speech Commcination, 1998,25(1-3):117-132

[7] Hiroaki H, Kensaku A, Yuji S, et al. Sound Source Separation with Two Spectrograms by Image Processing[J]. IEEJ Transactions on Electronics, Information and Systems, 2005, 124(12): 2439-2445

[8] Khunarsal P, Lursinsap C, Raicharoen T P. Singing Voice Recognition Based on Mat-Chin of Spectrogram Pattern [C]// Proceedings of International Joint Conference on Neural Networks. 2009:1595-1599

[9] Shirin B, Richard R. A performance monitoring approach to fusing enhanced spectrogram channels in robust speech recognition [C]// Proceedings of the Annual Conference of the International Speech Communication Association. 2011:477-480

[10] Zhang Jin-song, Keikichi H. Tone nucleus modeling for Chinese lexical tone recognition[J]. Speech Communication, 2004, 42(3/4):447-466

[11] Zhang Hua-ping, Liu Qun. Automatic recognition of Chinese personal name based on role tagging[J]. Chinese Journal of Computers, 2004, 27(1):85-91

[12] Zhang S X, Gales M J F. Structured SVMs for automatic speech recognition[J]. IEEE Transactions on Audio, Speech and Language Processing, 2013, 21(5):544-555

[13] Neammalai P, Phimoltares S, Lursinsap C. Speech and Music Classification using Hybrid Form of Spectrogram and Fourier Transformation[C]// 2014 Annual Summit and Conference on Asia-Pacific Signal and Information Processing Association(APSIPA). 2014:1-6

[14] 马义德,袁敏,齐春亮,等. 基于 PCNN 的语谱图特征提取在说话人识别中的应用[J]. 计算机工程与应用,2005,41(20):81-84

[15] Awais M M, Waqas A, Masud S, et al. Continuous Arabic Speech Segmentation using FFT Spectrogram[C]// Innovations in Information Technology. 2006

[16] Kensaku A, Akira O. Reduction of Noise in Speech Signals through Image Processing Using the Spectrogram[J]. IEEJ Transactions on Electronics, Information and Systems, 2006, 126(12):1483-1489

[17] Ajmera P K, Djmera, Jadhav D V, et al. Text-independent Speaker Identification Using Radon and Discrete Cosine Transforms based Features from Speech Spectrogram[J]. Pattern Recognition, 2011, 44(10/11):2749-2759

[18] Kekre H B, Athawale A, Desai M. International Conference and Workshop on Emerging Trends in Technology[C]// ICWET Conference Proceedings. 2011:171-174

[19] Steinberg R, O'Shaughnessy D. Segmentation of a speech spectrogram using mathematical morphology. ICASSP[C]// IEEE International Conference on Acoustics, Speech and Signal Processing Proceedings. 2008:1637-1640

[20] Wu Di, Zhao He-ming, Huang Cheng-wei, et al. Speech Endpoint Detection in Low-SNRs Environment Based on Perception Spectrogram Structure Boundary Parameter[J]. Chinese Journal of Acoustical, 2014, 39(4):428-440

[21] Wang K C. The Feature Extraction Based on Texture Image Information for Emotion Sensing in Speech[J]. Journal Citation Reports, 2014, 14(9):16692-16714

[22] Xu Sen, Zhao Xu, Duan Cheng-hua, et al. A Mathematical Morphological Processing of Spectrograms for the Tone of Chinese Vowels Recognition[J]. Applied Mechanics and Materials, 2014 (571/572):665-671

[23] Dutta, Tridibesh. Dynamic time warping based approach to text-dependent speaker identification using spectrograms[C]// Proceedings-1st International Congress on Image and Signal Processing, CISP. 2008:354-360

[24] 赵力. 语音信号处理[M]. 北京:机械工业出版社,2009:29-30

[25] Zhang Yue. The Research on Spectrogram of a particular group of Small-Vocabulary Recognition Algorithm[D]. Changchun: Northeast Normal University, 2013

[26] Hll D L, Llinas J. Handbook of Multisensor Data Fusion[M]// The Electrical Engineering and Applied Signal Processing Series, CRC Press, 2001

[27] Liu Tong-ming, Xia Zu-xun, Xie Hong-cheng. Data Fusion Technology with Application[M]. Beijing: National Defence Industry Press, 1998

[28] Blum R S, Xue Z, Zhang Z. Multisensor Image Fusion and its Applications[M]. Boca Raton, CRC Press, 2005

[29] 张家騄. 汉语人机语言通讯基础[M]. 上海科学技术出版社,2010:328-332

[30] 李明宇. 现代汉语常用词表[M]. 北京:商务印书馆出版社,2008

[31] 邓乃扬,田英杰. 数据挖掘中的新方法:支持向量机[M]. 科学出版社,2009

[32] Chang C C, Lin C J. A Library for Support Vector Machines [M]. National Taiwan University, 2001