第 59 卷 第 7 期/2022 年 4 月/激光与光电子学进展 研究论文

激光与光电子学进展

基于谱峰值点特征的汉语音节匹配算法

唐维康，邵玉斌*，龙华，杜庆治，彭艺，陈亮

昆明理工大学信息工程与自动化学院，云南 昆明 650500

摘要 为提升噪声环境中汉语语音音节的匹配效果，依据汉语语音的谱峰值点特征，提出了一种音节匹配算法。采用离散余弦变换提取语音信号包络语谱图，利用人耳掩蔽效应进行谱能量判决，获取每一帧谱能量的极大值点；接着在对数频率范围内作二值量化，将音节信号对应为二进制序列；然后根据二进制序列的模板对比，确定音节匹配结果。本算法对无噪汉语语音的音节匹配效果优于传统方法，且在低信噪比情况下仍具有较高的匹配准确率。

关键词 信号处理；音节匹配；极值点；对数频率域

中图分类号 TN912.34 **文献标志码** A **doi**：10.3788/LOP202259.0707001

Syllable Matching Algorithm with Spectral Peak Point Feature for Chinese Speech

Tang Weikang, Shao Yubin*, Long Hua, Du Qingzhi, Peng Yi, Chen Liang

School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, Yunnan 650500, *China*

Abstract Based on the spectral peak point characteristics of Chinese speech, this study proposes a syllable matching algorithm to improve the matching effect of Chinese speech syllables in noisy environments. First, a discrete cosine transform is used to extract the speech signal envelope spectrogram, and the human ear masking effect is used for spectral energy judgment to obtain the extreme value points of spectral energy in each frame. Then, the syllable signal is corresponded to a binary sequence by performing binary quantization in the logarithmic frequency range. Finally, the syllable matching result is determined based on the template comparison of the binary sequence. The results show that the proposed algorithm outperforms the conventional methods for matching syllables in the noiseless Chinese speech. Additionally, it has a high matching accuracy at low signal-to-noise ratios.

Key words signal processing; syllable matching; extreme value point; logarithmic frequency range

1 引 言

在语音识别领域，汉语语音音节的匹配步骤主要分为两步[1]：首先提取音节的特征参数[2]，如语音音节信号的基音周期、短时能量、包络、滑动差分倒谱[3](SDC)、感知线性预测系数[4](PLP)、美尔频率倒谱系数[5](MFCC)等；然后使用距离方法测量出这些特征的间隔，利用间隔的大小得出两音节的匹配度[6]。文献[7]指出：一方面，传统的特征参数的提取都以更准确地描述音节为目的，音节的很多细节信息都会映射到特征参数上，如性别信息，而对于音节的匹配，性别信息是需要抹除的；另一方面，如果两个音节含有噪声干扰，两个音节就会产生较大差异[8]，导致语音的质量和可懂度下降[9-10]，故以

收稿日期：2021-06-07；**修回日期**：2021-06-24；**录用日期**：2021-07-06

基金项目：国家自然科学基金(61761025)

通信作者：*shaoyubin@kust.edu.cn

传统的音节特征参数作为匹配特征会使两音节的匹配准确度降低，甚至发生误判。

近年来，在语音音节匹配方面的研究成果主要包括两类：一类是基于距离的匹配方法[11]，如，先提取 MFCC 特征参数，将其视为 Mahalanobis 距离的测量值，值与值之间的距离越大，表明匹配度越低；另一类是基于相关系数的匹配方法[12]，即利用余弦相似度进行匹配，所得的匹配范围是 0～1，0 代表两音节完全不相关，1 代表完全相同，其他值代表不同的匹配度。本文提出了一种基于谱峰值点特征的语音音节匹配方法，即：先在音节的语谱图上剔除与匹配不相关的部分信息，获取每一帧谱能量的极大值点，然后在对数频率范围内将其量化成二进制数据，从而将音节信号对应为二进制序列，之后根据二进制序列的排列组合确定是否为同一音节的发音。本文所提方法的匹配正确率优于传统的匹配方法，即使在加噪环境下，语音音节的匹配正确率也可以达到实际应用的要求。

2　谱峰值点的引入

在汉语语音音节匹配过程中，从时域方面提取到的特征不足以进行有效匹配，而相同音节的频域特征有一定的规律可循。图 1(a)为同一个人不同时刻不同语速下汉字“的”“高”“是”“他”“兴”的语谱图，每个汉字发音 4 次。图中越明亮的地方，代表此处的能量越大。同一音节的能量在某些频率范围内相对集中，因此可以将能量的分布特性作为音节匹配的特征输入。原始语谱图上保留着很多说话人信息，为了剔除不必要的说话人信息，本文提出了包络语谱图的计算方法，将原始的语谱图进行“模糊”处理，处理后形成的包络语谱图仍然保留音节的基本信息，如图 1(b)所示。在匹配过程中，选取具有代表性的音节信息不仅可以提高匹配准确率，还可以减少运算量。首先将包络语谱图进行谱能量判决，随后获取每一帧谱中能量的极大值点，如图 1(c)所示。根据谱能量极大值点的分布特性，本研究团队设计了基于谱峰值点特征的汉语音节匹配方法。

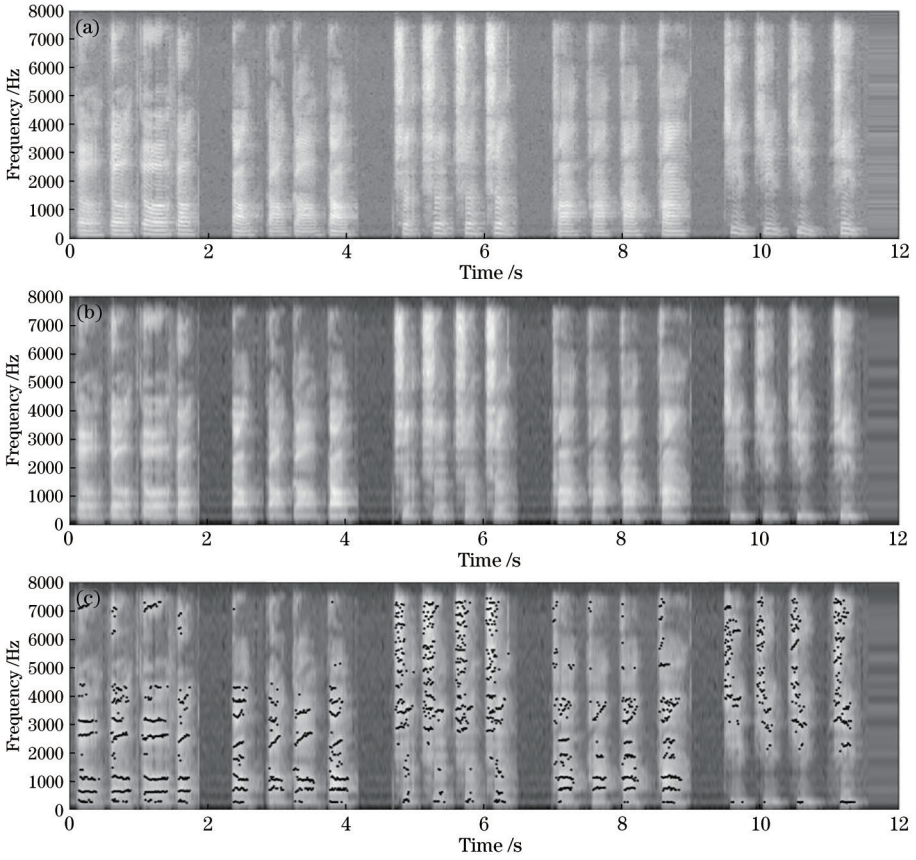

图 1　汉语音节的语谱及能量峰值点分布。(a)原始语谱图；(b)包络语谱图；(c)能量峰值点分布

Fig. 1　Speech spectrograms and energy peak point distribution of Chinese phonetic syllables. (a) Original speech spectrogram; (b) speech envelope spectrogram; (c) energy peak point distribution

3　基于谱峰值点特征的音节匹配算法

3.1　算法设计的原理与框图

汉语音节的匹配主要分为4部分：首先利用谱峰值点生成器提取有效的匹配特征；然后将这些谱峰值点特征与模板库中存放的音节谱峰值点特征进行比较，计算出匹配得分；再设立一个得分阈值，大于该阈值的认为是同一音节的发音。具体原理图如图2所示。

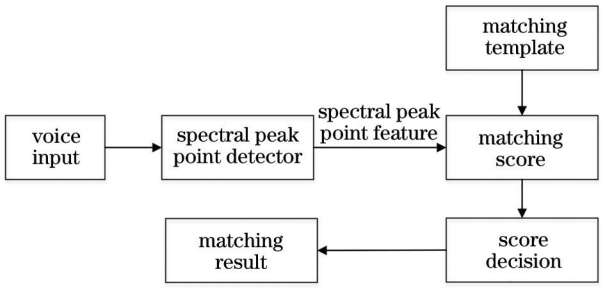

图2　音节匹配算法步骤

Fig. 2　Syllable matching algorithm steps

如何提取到有效的匹配特征成为重中之重。本文设计了谱峰值点生成器，用来提取匹配特征，如图3表示。首先求出语音信号的灰度语谱图，对其进行“模糊”处理后形成包络语谱图；在包络语谱图上进行谱能量判决，获取每一帧谱能量的极大值点；再将每个频率段量化成一个二进制数据，这一串二进制数据就是谱峰值点特征。

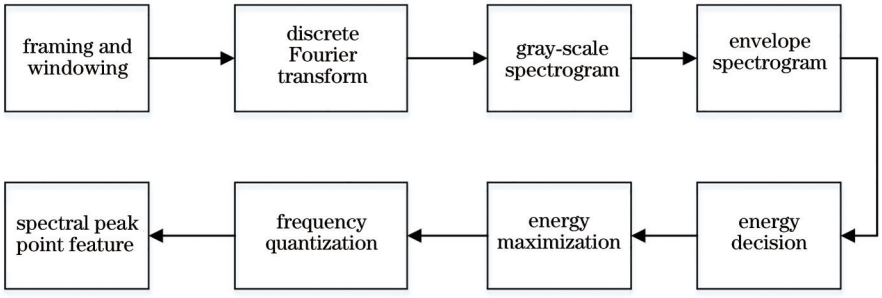

图3　谱峰值点特征的提取

Fig. 3　Extraction of spectral peak point feature

3.2　灰度语谱图

语谱图含有语音信号的时域和频域分布特征，展示的是语音频谱随时间变换的特性，其纵轴方向代表频率，横轴方向代表时间。灰度语谱图的提取步骤如下：

1）分帧加窗。对语音信号$x(n)$分帧加窗，第i帧信号为$x_i(n)$。

2）短时离散傅里叶变换。对第i帧信号$x_i(n)$进行短时离散傅里叶变换，变换公式为

$$X_i(k)=\sum_{n=1}^{N}x_i(n)\cdot\exp\left(-\mathrm{j}2\pi k/N\right),\tag{1}$$

式中：$0\leqslant k\leqslant N-1$；$N$为帧长；$X_i(k)$为信号$x_i(n)$的离散傅里叶变换。

3）计算$X_i(k)$的能量谱，计算公式为

$$P_i(k)=\left|X_i(k)\right|^2。\tag{2}$$

对$P_i(k)$进行对数化，即

$$P_i^{(\mathrm{dB})}(k)=10\lg P_i(k),\tag{3}$$

式中：$P_i^{(\mathrm{dB})}(k)$为对数化能量谱；$P_i(k)$为能量谱。

4）得到灰度语谱图。把对数化的时变能量谱表达为矩阵，可得二维图像，如图4所示。图4的纵坐标表示离散频率点k，横坐标表示帧序号i对应的时间。

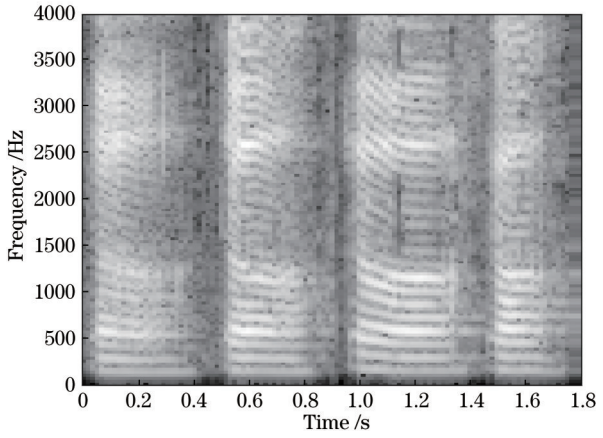

图4　语音信号的灰度语谱图

Fig. 4　Gray-scale spectrogram of speech signal

图4是同一个人在不同时刻、不同语速下“的”字发音4次的灰度语谱图，该谱图保存了说话人的原始信息，灰度值的大小表示对应频率成分在对应时刻能量的高低。灰度语谱图中的纹理结构反映了说话人的音律学、基音频率、共振峰等特征信息，

可被广泛应用于说话人识别领域。

3.3　包络语谱图

灰度语谱图中的纹理是区分不同说话人的重要信息，而在音节匹配中，不同说话人的信息需要被进一步消除，只需要考虑同一音节的相似特征，因此可以将灰度语谱图中的声纹进一步进行“模糊”处理。本文采用离散余弦变换[13](DCT)方法进行“模糊”处理，变换后的能量大部分集中于前面的最大特征量上。对数化能量谱$P_i^{(\mathrm{dB})}(k)$变换后的系数矩阵为

$$\boldsymbol{F}_i(v)=\frac{1}{\sqrt{N}}C(v)\sum_{k=0}^{N-1}P_i^{(\mathrm{dB})}(k)\cos\left[\frac{\pi(2k+1)v}{2N}\right],\tag{4}$$

式中：i代表帧序号；$v=0,1,\cdots,N-1$；$C(v)=\begin{cases}\dfrac{1}{\sqrt{2}} & v=0\\ 1 & v\neq 0\end{cases}$。

$\boldsymbol{F}_i(v)$的变换系数分为低频分量和高频分量，其中高频分量反映的是灰度语谱图中声纹间隔等细节信息，而音节匹配需要减少这部分细节信息，因此本文默认将每帧40点后的数据置0，即将许多系数值较小的高频分量量化成0。令置0后的系数矩阵为$\boldsymbol{Z}_i(v)$。离散余弦变换是可逆的，$\boldsymbol{Z}_i(v)$的逆变换为

$$\tilde{\boldsymbol{P}}_i^{(\mathrm{dB})}(k)=\frac{1}{\sqrt{N}}\sum_{v=0}^{N-1}C(v)\boldsymbol{Z}_i(v)\cos\left[\frac{\pi(2k+1)v}{2N}\right]。\tag{5}$$

恢复出的灰度语谱图上的声纹间隔和走势已经不明显，将其称为“包络语谱图”，如图5所示。

根据人耳的掩蔽效应[14]，人耳对对数能量谱上的峰值点刺激最敏感，反映在包络语谱图上就是人耳对能量较高且较集中的那部分刺激最敏感。首先获取谱能量的峰值点，随后设定能量阈值为T，小于T的峰值点舍弃，大于T的峰值点用黑色圆点标识，即保留对人耳刺激最敏感的谱峰值点，如图6所示。

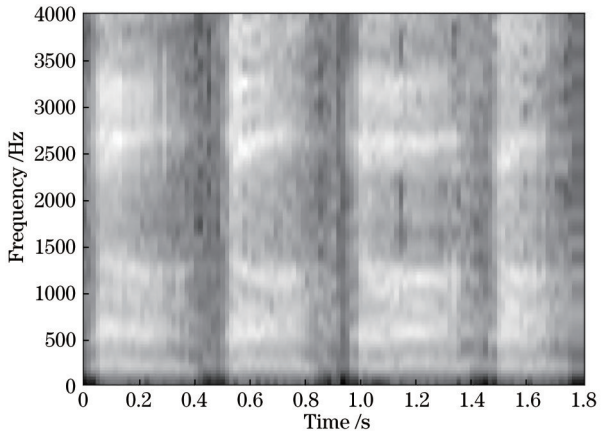

图5　语音信号的包络语谱图

Fig. 5　Envelope spectrogram of speech signal

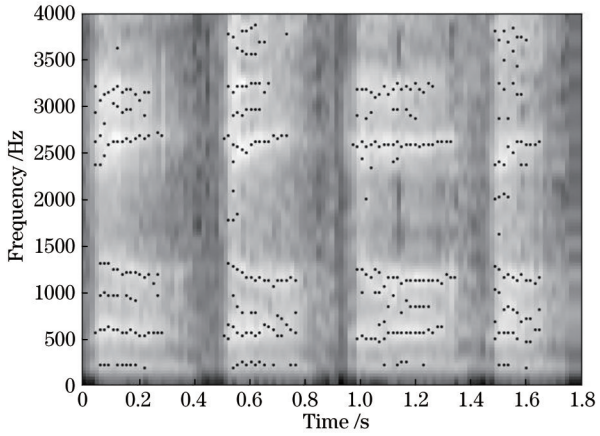

图6　阈值处理后的能量点

Fig. 6　Energy point after thresholding

由图6可以看出，同一音节的能量集中于特定频率区间，过滤后的能量也集中于特定频率段，可以进一步加以改进。

3.4　能量的极大值点

通常情况下，人说话的语音频率主要集中在中低频，大多数汉语语音的频率在300～4000 Hz，部分性别信息主要集中在300 Hz以下。因此，舍弃300 Hz以下的信息有利于减小性别对音节匹配的干扰，同时也有助于减少运算量，而4000 Hz以上的信息对于音节匹配的影响很小，也可以舍弃。由于实验中采用的语音信号为8 kHz的wav格式音频文件，故只舍弃了300 Hz以下的信息。图7(a)是常用音节“的”字发音的包络语谱图，该音节分为10帧。实验发现，对于音节的匹配分帧帧数过多或过少都不能达到匹配的最佳效果，在10帧左右的匹配效果最佳。图7(b)是舍弃了300 Hz以下信息的包络语谱图。

另外，掩蔽效应指出，一个强音会掩蔽周围较弱的音，反映在包络语谱图上就是在一个频率范围内，人耳基本只对那几个较亮能量点的刺激最敏感，越靠近这几个较亮点的能量值越容易被掩蔽。在该频率范围内求取分贝的极大值点，频率范围的划分方法最常见的是等间隔划分。图8(a)是频率范围内空间分布呈等间隔的能量点图像。在每个

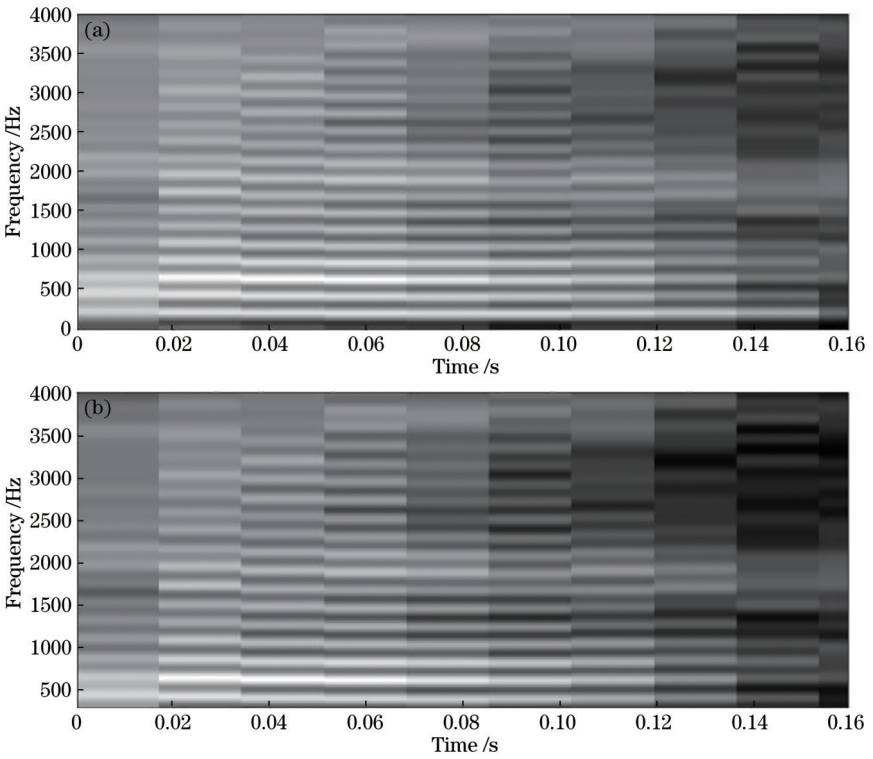

图 7　包络语谱图。(a)常用音节“的”字发音的包络语谱图;(b)舍弃 300 Hz 以下信息的包络语谱图

Fig. 7　Envelope spectrograms. (a) Envelope spectrogram of commonly used Chinese character pronunciation “de”; (b) envelope spectrogram after discarding partial information blow 300 Hz

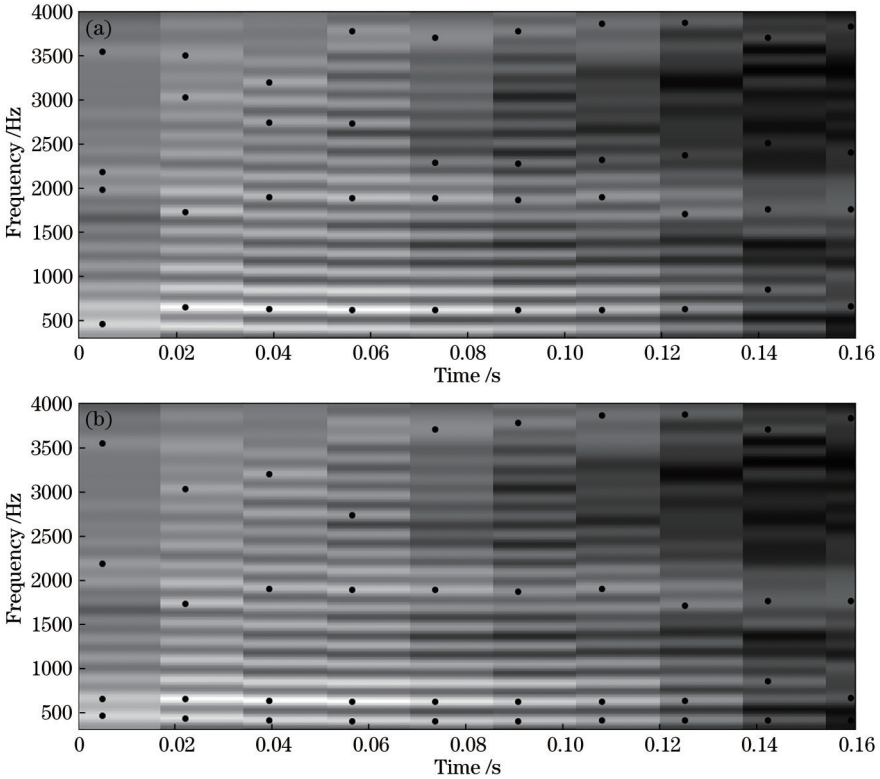

图 8　频率段不同划分方法下能量的极大值点分布情况。(a)频率段等间隔划分;(b)频率段对数形式划分

Fig. 8　Distribution of energy maximum points using different division methods in frequency bands. (a) Frequency band is equally spaced; (b) logarithmic division of frequency bands

频率范围找寻找能量的极大值点，如图 8(a)所示，每帧上找到 4 个特征值点。研究表明，人耳对音高及音强的响应分布呈对数特性[15]，人耳的构造决定了频率范围的分布是接近对数形式的，因此采用对数频率范围可以更好地反映耳蜗的听觉特性。同时，实验也表明，对数分布频率范围比等间隔分布的匹配效果更好。图 8(b)的频率范围呈对数分布，分别划分为 300～600 Hz、>600～1200 Hz、>1200～2400 Hz 和>2400～4000 Hz 频率范围带。

3.5　频率量化

特征点的排列组合形式越多，越可以体现音节的更多信息。特征点的排列组合可以根据二项式 $(a+b)^n$ 的展开式系数 $C_n^0, C_n^1, C_n^2, \cdots, C_n^n$ 获得，$C_n^0, C_n^1, C_n^2, \cdots, C_n^n$ 是离散组合数，将变换的上标看成自变量 r，特征点的排列组合数量看成 $f(r)$，$f(r)=C_n^r(r=0,1,2,\cdots,n)$，由杨辉三角特性[16]可知，$f(r)$要想取得最大值，可以分为两种情况讨论：1) 当 n 为偶数时，最大值在中间项，即 $f(r)=C_n^{\frac{n}{2}}$ 为最大值；2) 当 n 为奇数时，最大值在中间两项，即 $f(r)=C_n^{\frac{n-1}{2}}$ 或 $f(r)=C_n^{\frac{n+2}{2}}$ 为最大值。在本文算法中，经过多次实验发现，经对数分割后形成 4 个频率范围带的匹配效果最佳，即 $n=4$ 且 $r=2$ 时，特征点的排列组合数 $f(r)$ 是最多的。

每帧信号可以用 4 个特征值点表示，按照能量值的高低进行特征值点的排序，特征值点的排列组合状态数越多，其所能表示的音节信息也越多。在每帧信号中取出前两位能量值较大的特征值点，如图 9 所示。

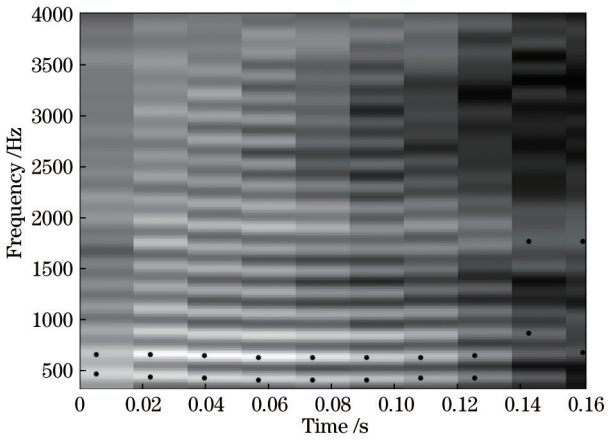

图 9　每帧信号最大的两个特征值点

Fig. 9　Illustration of two maximum feature points in each signal frame

将能量较大的两个特征值所在的频率范围量化为 1，剩下的两个频率范围量化为 0，即每帧信号根据特征值的分布可以用两个 1 和两个 0 代替，这样一个音节可用一个二进制矩阵表示为

$$\boldsymbol{B}=\begin{bmatrix}0&0&0&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&0&1&1\\1&1&1&1&1&1&1&1&1&1\\1&1&1&1&1&1&1&1&0&0\end{bmatrix}。\tag{6}$$

$\boldsymbol{B}$ 矩阵为 4 行 10 列矩阵，4 行代表有 4 个对数频率范围，10 列代表有 10 帧。将矩阵 $\boldsymbol{B}$ 转化为一维二进制数据 B_1，B_1=0b1100110011001100110011001100110001100110。同样，也可以将另一个音节表示为一个二进制数据 B_2。将两个二进制数据进行按位异或运算，即

$$C=B_1\hat{}B_2,\tag{7}$$

计算出 C 中 0 的个数并将其记为 s，s 代表两音节对应的二进制数据中相同位的总数。设定得分阈值为 s_t，当 $s\geqslant s_t$ 时，认为两个音节是同一个发音，可以用一个简单的循环确定 s_t 的最佳值。

4　数值实验

4.1　实验设计

本次实验采用的语料集通过语音合成软件合成，匹配的汉语语音数据集 E 由 I 个常用汉字的发音集组成，每个汉字的发音集有 J 条。本次实验中，$I=20$，$J=100$，匹配的汉语语音数据集共计 $I\times J$ 条语音，即 2000 条语音。模板库数据集 F 是这 I 个汉字的二进制数据，其中一个汉字由 R 个不同的人发音，随后采用本文算法将这 R 个发音生成一个 $R\times L$ 的二进制矩阵数据 $\boldsymbol{Y}$，$\boldsymbol{Y}$ 的每一行代表一个人的发音。设 $r\in(0,R/2)$，$l\in(0,L)$，若 $\boldsymbol{Y}_{rl}$ 为同一个数据 u，u 为 0 或者 1，那么模板库中该字的第 l 列数据也为 u，即模板库中每个汉字的二进制数据是通过多个发音人的平均数获得的。

汉语语音音节的匹配正确率定义为

$$\eta=\frac{\sum_{i=1}^{I}\sum_{j=1}^{J}M(\tilde{a}_i,a_{ij})}{IJ},\tag{8}$$

式中：M 代表本文的匹配算法；a_{ij} 表示匹配数据集 E 中第 i 个汉字发音集 a_i 下的第 j 条语音；$\tilde{a}_i$ 表示模板库数据集 F 中的第 i 个汉字发音集。若 a_{ij} 与 a_i 匹配，则 $M(\tilde{a}_i,a_{ij})=1$，否则，$M(\tilde{a}_i,a_{ij})=0$。$\sum_{i=1}^{I}\sum_{j=1}^{J}M(\tilde{a}_i,a_{ij})$ 表示所有匹配数据集下的 a_i 匹配到

模板库数据集下$\tilde{a}_i$的数量。

4.2　实验结果分析

共设计了6组实验，其中：第一组实验和第二组实验是在高斯白噪声环境下分别寻找不同信噪比下语音信号分为多少帧、对数频率带分为几个时可以达到最佳匹配效果；第三组实验和第四组实验是在无噪和加噪环境下，将传统算法与本文算法对同一个人语音音节的匹配正确率进行比较；第五组实验和第六组实验是在无噪和加噪环境下将传统算法与本文算法对不同说话人语音音节的匹配正确率进行比较。需要说明的是，第三组实验和第四组实验汉语语音数据集中的每个汉字发音集是由同一个说话人在不同时间按照不同语速发音组成的；其余组实验的语音数据集中每个汉字的发音集包含20个不同人的发音，其中20人中一半是男生，一半是女生，每个人按照3～7音节/s的语速发音，每个人按不同语速发音5次。

先进行第一组和第二组实验，即寻找语音信号分为多少帧和划分为多少对数频率带时的匹配效果最佳。实验结果如表1和表2所示。

表1　不同信噪比(SNR)下不同帧数N_f的匹配正确率

Table 1　Matching accuracy of different frame numbers under different signal-to-noise ratios

SNR /dB	Accuracy /%			
	N_f=5	N_f=10	N_f=15	N_f=20
30	61.2	75.6	70.7	65.6
25	59.0	73.4	68.7	62.4
20	57.3	71.8	66.4	60.1
15	55.6	70.1	62.5	57.8

表2　不同信噪比下不同对数频率带数N_b的匹配正确率

Table 2　Matching accuracy of different logarithmic frequency bands under different signal-to-noise ratios

SNR /dB	Accuracy /%			
	N_b=2	N_b=4	N_b=8	N_b=16
30	35.6	75.6	70.1	58.6
25	32.3	73.4	68.4	54.7
20	30.1	71.8	65.8	51.2
15	29.3	70.1	62.4	48.7

从表1和表2所示的实验结果可以看出，对于语音音节的匹配，一个音节的分帧数量过多或过少都达不到匹配的最佳效果。帧数越少，代表生成的音节的二进制数据少，说明语音音节信息被掩盖；帧数越多，则很多与匹配无关的信息就会表达出来。实验得出一个音节分为10帧左右时，匹配正确率最高。如果对数频率带只划分为两个，则匹配的准确率直线下降，两个对数频率带代表的音节信息是最少的；对数频率带越多，冗余信息也越多，同时计算量也加大。实验得出一帧信号分为4个对数频率带时，匹配效果最好。

为使实验结果更加有对比性，针对同一说话人生成的语音数据集：一方面，在无噪语音条件下，分别采用基于Mahalanobis距离的算法、基于余弦相似度的算法和本文算法进行匹配正确率的验证，结果如表3所示；另一方面，在语音信号中添加高斯白噪声，使信噪比分别为15 dB、10 dB、5 dB和0 dB，再分别采用多种匹配方法进行匹配正确率的验证，结果如表4所示。

表3　不同算法对于无噪环境下同一个人发音的匹配正确率

Table 3　Matching accuracy of different algorithms for the same person's pronunciation in a noise-free environment

Matching algorithm	Accuracy /%
Mahalanobis distance[11]	62.3
Cosine similarity[12]	71.6
Our algorithm	80.4

表4　不同算法针对加噪环境下同一个人发音的匹配正确率

Table 4　Matching accuracy of different algorithms for the same person's pronunciation in a noisy environment

Matching algorithm	Accuracy /%			
	SNR of 25 dB	SNR of 20 dB	SNR of 15 dB	SNR of 10 dB
Mahalanobis distance[11]	58.3	56.8	54.6	51.1
Cosine similarity[12]	68.4	66.7	64.0	61.2
Our algorithm	76.4	74.8	72.2	71.1

从表3可以看出，在无噪环境下，基于Mahalanobis距离的算法的匹配正确率在60%左右，而本文算法的语音音节匹配正确率超过了80%。从表4可以看出，在多种信噪比环境下，本文算法的语音音节匹配正确率均保持在70%以上，而基于Mahalanobis距离的算法的匹配正确率是最低的。这是因为该方法提取的音频特征向量是MFCC，MFCC提取了很多说话人本身的信息。在不同说话人生成的语音数据集下，多种匹配方法在无噪和加噪环境下的音节匹配正确率如表5表6所示。

由表5和表6可以看出：在不同说话人生成的语音音节数据集中，本文算法在无噪语音环境下的匹配正确率在70%以上，在语音信号信噪比较低的情

表 5 不同算法针对无噪环境下不同人发音的匹配正确率

Table 5 Matching accuracy of different algorithms for different people's pronunciation in a noise-free environment

Matching algorithm	Accuracy /%
Mahalanobis distance[11]	58.3
Cosine similarity[12]	65.5
Our algorithm	74.4

表 6 不同算法针对加噪环境下不同人发音的匹配正确率

Table 6 Matching accuracy of different algorithms for different people's pronunciation in a noisy environment

Matching algorithm	Accuracy /%			
	SNR of 25 dB	SNR of 20 dB	SNR of 15 dB	SNR of 10 dB
Mahalanobis distance[11]	56.3	54.8	53.6	52.1
Cosine similarity[12]	63.4	62.1	60.7	58.2
Our algorithm	70.8	68.9	67.1	64.5

况下，也取得了不错的语音匹配效果。此外，基于余弦相似度的匹配算法在无噪和加噪环境下也取得了不错的效果，这是因为余弦相似度减小了维数的影响，考虑了音节之间的相关性，但其计算更加复杂。

由以上 6 个实验结果可以看出，本文算法在低信噪比的噪声环境下进行音节匹配是有效的，匹配正确率可以满足工程应用。本文算法较其他算法在匹配准确率上有所提升，并且易于实现。

5 结 论

针对传统汉语语音匹配技术对实际说话环境中常用汉语语音音节匹配效果不佳的问题，本研究团队提出了一种音节匹配算法。该算法利用离散余弦变换法获取信号的包络语谱图，通过能量判决和掩蔽效应获取每一帧能量的极大值对应的频率点，并依据人耳听觉的对数特性，将频率范围量化成对数形式；然后在对数频率范围内将极大值点频率量化成二进制数据，根据相同音节的二进制数据排列组合大致一样的原则，确定出两个音节的匹配度。在无噪和加噪语音环境下的实验结果表明，所提算法的匹配效果能较好地满足工程应用，具有较高的匹配正确率。

参 考 文 献

[1] Zhang L, Zhou T, Du Q Z, et al. Audio comparison algorithm based on physical characteristics[J]. Video Engineering, 2017, 41(Z4): 110-114.
张琳, 周韬, 杜庆治, 等. 基于物理特征的音频相似度比对算法研究[J]. 电视技术, 2017, 41(Z4): 110-114.

[2] Guo X J, Fan B Q. Feature-based comparison of audio[J]. Journal of Henan Normal University (Natural Science), 2006, 34(2): 35-38.
郭兴吉, 范秉琪. 基于特征的音频比对技术[J]. 河南师范大学学报(自然科学版), 2006, 34(2): 35-38.

[3] Torres-carrasquillo P A, Singer E, Kohler M A, et al. Approaches to language identification using Gaussian mixture models[C]//7th International Conference on Spoken Language Processing, September 16-20, 2002, Denver, Colorado, USA. New York: ISCA, 2003.

[4] Chen S, Wang H C, Jia J, et al. Comparison of Mel frequency ceptrum coefficient and perceptual linear predictive in perceptual measurement of Chinese initials[J]. Applied Mechanics and Materials, 2013, 411/412/413/414: 291-297.

[5] Tanaka K, Ichikawa K, Kittiwattanawong K, et al. Automated classification of dugong calls and tonal noise by combining contour and MFCC features[J]. Acoustics Australia, 2021, 49(2): 385-394.

[6] Sonnleitner R, Widmer G. Robust quad-based audio fingerprinting[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2016, 24(3): 409-421.

[7] Gan T, He Y M, Huang X G, et al. A fast broadcast audio comparison method: CN104992713A[P]. 2015-10-21.
甘涛, 何艳敏, 黄晓革, 等. 一种快速广播音频比对方法: CN104992713A[P]. 2015-10-21.

[8] Ye Y Q, Zhang S R, Hong W J, et al. Audio similarity comparison system of English dubbing on android platform[C]//2017 IEEE International Conference on Information and Automation (ICIA), July 18-20, 2017, Macao, China. New York: IEEE Press, 2017: 692-697.

[9] Fang B, Chen J Y. Wavelet threshold denoising algorithm for removing impulse noise[J]. Laser & Optoelectronics Progress, 2021, 58(22): 2210016.
方斌, 陈家益. 去除脉冲噪声的小波阈值去噪算法[J]. 激光与光电子学进展, 2021, 58(22): 2210016.

[10] Hua C J, Ma J K, Chen Y. Improved non-local mean denoising algorithm based on difference hash algorithm [J]. Laser & Optoelectronics Progress, 2020, 57(14): 141007.
化春键, 马金科, 陈莹. 基于差异哈希算法的改进非局部均值去噪算法[J]. 激光与光电子学进展, 2020,

57(14): 141007.

[11] Lee C Y. A study on the optimal mahalanobis distance for speech recognition[J]. Speech Sciences, 2006, 13(4): 177-186.

[12] Ai J Q, Zuo Y, Liu J X, et al. A hierarchical clustering approach for speech feature extraction based on cosine similarity[J]. Application Research of Computers, 2020, 37(S2): 147-149.
艾佳琪, 左毅, 刘君霞, 等. 基于余弦相似度的动态语音特征提取算法[J]. 计算机应用研究, 2020, 37(S2): 147-149.

[13] Brahimi N, Bouden T, Brahimi T, et al. A novel and efficient 8-point DCT approximation for image compression[J]. Multimedia Tools and Applications, 2020, 79(11/12): 7615-7631.

[14] Tejani V D, Brown C J. Speech masking release in hybrid cochlear implant users: roles of spectral and temporal cues in electric-acoustic hearing[J]. The Journal of the Acoustical Society of America, 2020, 147(5): 3667-3683.

[15] Ben Messaoud M A, Bouzid A. Pitch estimation of speech and music sound based on multi-scale product with auditory feature extraction[J]. International Journal of Speech Technology, 2016, 19(1): 65-73.

[16] Hung T L, Kien P T. On the order of approximation in limit theorems for negative-binomial sums of strictly stationary m-dependent random variables[J]. Acta Mathematica Vietnamica, 2021, 46(1): 203-224.