Computer Science and Application 计算机科学与应用, 2016, 6(6), 384-392
Published Online June 2016 in Hans. http://www.hanspub.org/journal/csa
http://dx.doi.org/10.12677/csa.2016.66047

Hans汉斯

The Research of Text-Independent Feature Extraction Based on Single Training Sample

Jianmin Guo

School of Physics and Information Technology, Shaanxi Normal University, Xi'an Shaanxi
Email: 1060397306@qq.com

Received: Jun. 7th, 2016; accepted: Jun. 26th, 2016; published: Jun. 30th, 2016

Abstract

The existing speaker identification are based on Linear Predictive Coding Cepstral (LPCC) coefficients, Mel-Frequency Cepstral Coefficients (MFCC), local normalized cepstral coefficients (LNCC) and wavelet packet transform (WPT) method; these features are sensitive to noisy and environmental sounds. This paper describes a novel robust text-independent feature extraction method using single training sample. In the proposed method, the features can reflect a person's basic phonation characteristic and distinguish different speakers. This paper introduces the four methods in single training sample and compares them with the proposed method. Experimental results on speech databases in English and Chinese demonstrate that the proposed approach can implement feature extraction in speaker identification based on single training sample, and yields a better performance in single training sample.

Keywords

Feature Extraction, Linear Predictive Coding Cepstral, Mel-Frequency Cepstral Coefficients, LNCC, WPT

与文本无关的单训练样本特征点提取研究

郭建敏

陕西师范大学物理学与信息技术学院，陕西 西安
Email: 1060397306@qq.com

文章引用: 郭建敏. 与文本无关的单训练样本特征点提取研究[J]. 计算机科学与应用, 2016, 6(6): 384-392.
http://dx.doi.org/10.12677/csa.2016.66047

收稿日期：2016年6月7日；录用日期：2016年6月26日；发布日期：2016年6月30日

摘　要

现有的说话人识别是基于语音的线性预测编码(LPCC)、Mel频率倒谱系数(MFCC)、局部归一化倒谱系数和小波包变换等特征，这些特征对环境噪声都比较敏感。针对上述问题，本文提出了一种与文本无关的单训练样本的特征提取方法。该方法提取的语音特征能够充分反映说话人的基本发声特性，可以很好的将不同的说话者区分开。本文列出了以上四种特征提取方法在但语音训练样本上对于不同说话者的识别效果，也将其与本文的方法进行了比较。对英文与汉语语音数据库的仿真实验表明，该特征提取方法可以实现单训练样本下的说话人识别中对于特征的提取，而且在单样本识别中会有相对好的效果。

关键词

特征提取，线性预测编码，Mel频率倒谱系数，局部归一化倒谱系数，小波包变换

1. 引言

现在与文本无关的说话者身份认证方法[1]越来越多，这种方法不用规定说话人的说话内容，可以在被识别者无感知的情况下进行识别，这样应用起来就会更加的方便。要想识别出测试语音所对应的说话者，就必须能够提取出有区别性的语音特征。一般情况下，我们都会提取出语音信号的短时频谱进行研究，由于短时频谱会随着时间的变化而变化，而这种变化又能反映人的发音习惯，也就可以用来识别不同的说话者。除了人的发音习惯以外，我们还可以通过研究人耳的听觉特性来利用美尔倒谱系数、感知线性预测参数来模拟人耳对声音频率的感知特性。

在说话者识别中，所提取出的语音特征的好坏在很大程度上会影响分类器的性能，也会影响到识别模型的训练和以及对特征参数的确定。因此，在语音识别中，能够提取出辨别效果好的特征很重要，这样我们就可以使用同样的分类器而获得很好的识别率。

在以往的研究中，大多数研究都是用同一个人的多句语音作为训练样本来提取其语音特征，并利用这些语音特征来建模，在多训练样本中，所使用的特征提取方法有线性预测倒谱系数、美尔频率倒谱系数、局部归一化倒谱系数和小波包变换等方法。这些方法在多句训练样本的语音识别中的识别效果比较好，但是对于单句的训练样本来说识别效果就不是很好。而我们在实际应用中训练样本并不是永远充足的，这样对于单训练样本的语音识别研究就显得尤为重要。

为了能够在语音训练样本单一的时候仍然能够区分或确定出不同的说话人，本文提出了一种新的语音信号的特征提取方法，该方法是在一个短时的语音信号上提取对应说话人的有区别性的特征向量。

通过将本实验所提出的语音信号特征提取方法在多样本说话人识别中的识别效果和多训练样本的语音识别中经常使用的 LPCC、MFCC、局部归一化倒谱系数

LNCC(Locally-Normalized Cepstral Coefficients)和 WPT 的方法进行比较，发现当训练样本是一段短时语音信号时，这种新的特征提取方法的识别效果相对较好。

2. 本文方法

2.1. 语音信号的特征提取

在实际应用中，输入的语音信号并不稳定，而且还会受到周围环境噪声的干扰，这样就会使得在频

率位上的一些主频率成分有一点帧移。这样通过频率位的值来判断，就很难准确的确定每一个局部特征。但是每个人的语音都各有特点，不同的人在说话的时候声音信号的频率高低在不同的时刻不同，我们可以利用这个不同的特征来区分不同的说话者。

受此启发，本文在进行特征提取时所用到的方法是：1、先将语音信号进行分帧处理再得到该训练语音信号的语谱图；2、在该语谱图中按帧从左到右，一帧一帧的处理。对于每一帧从低频率位到高频率位，依次以 8 个频率为一组，寻找最大值作为提取出的特征点，并求出这 8 个频率位中的最小值。以此类推，得到对应训练样本的特征点所对应的矩阵，用每个特征值减去对应的最小值，再进行之后的处理。

2.2. 实验所用数据库

在本实验中用到的数据库是标准的 TIMIT 英文数据库和自己建立的中文数据库。在每个数据库中用到 9 个人，每个人说 10 句话，用每个人的其中一句话最为训练样本，剩下的 9 句话作为测试样本。我们根据语音频率能量的分布对于不同的人分布不同来区别不同的说话者。

3. 对比实验

3.1. 线性预测倒谱系数

线性预测倒谱系数 LPCC 是线性预测系数 LPC 在倒谱域中的表示。得到 LPCC 系数的方法，不是求原始语音信号语谱图的反傅里叶变换，而是用递归算法通过 LPC 得到。由于在 LPCC 中不需要求语音信号的傅里叶变换，所以计算量会少一些。LPCC 继承了 LPC 的优点，在进行 LPC 参数的求解过程中，最主要的思想是模拟人说话时声道的工作原理[2]，在对比实验中 LPC 预测采用的是预测的阶数 等于 14 的自相关分析方法。

通过对语音信号进行线性预测(LPC)分析，可以假设声道模型如下：

$$H(z)=\frac{G}{1-\sum_{i=1}^{p}a_i z^{-i}} \tag{1}$$

其中 G 为增益常数，可令其为 1。 p 为预测系数 a_i 的个数。

由自相关法得到的 LPC 系数既可以使系统具有稳定性，也可以使上式所对应的人说话的声道模型的传输函数的相位最小。其中语音信号的倒谱参数 $c(n)$ 和 LPC 系数之间的递推的关系为：

$$c(1)=a_1 \tag{2}$$

$$c(n)=a_n+\sum_{k=1}^{n-1}\left(1-\frac{k}{n}\right)a_k c(n-k) \quad 1<n\le p \tag{3}$$

$$c(n)=\sum_{k=1}^{p}\left(1-\frac{k}{n}\right)a_k c(n-k) \quad n>p \tag{4}$$

或者可以通过 LPC 得到

$$C_{\mathrm{LPCC}}(n)=C_{\mathrm{LPC}}(n)+\sum_{k=1}^{n-1}\frac{n-k}{n}C_{\mathrm{LPCC}}(n-k)C_{\mathrm{LPC}}(k) \tag{5}$$

根据同态处理的方法和产生声音信号的模型可以得出，用激励信号的倒谱加上声道传输函数的倒谱可以来表示语音信号的倒谱 $c(n)$ [3] [4]。经过分析声道传输函数的零极点的分布情况和激励信号的语音特点，可得出产生激励的信号的倒谱分布的范围很宽，该倒谱从低时域一直分布到高时域，但是声道传输函数的倒谱大部分分布在低时域里。由于声道传输函数上携带者一些重要信息，而这些信息对于不同

搞得说话者来说是不同的，所以我们可以模拟人说话时产生声音的声道，进而去识别不同的说话者。LPC倒谱的特征来自于语音信号的倒谱所对应的低时域部分用 c 来表示：

$$c=\left[c(1),c(2),\cdots,c(q)\right] \quad 10\le q\le 16 \tag{6}$$

式中， q 为 LPC 倒谱特征的阶数。

利用 LPCC 方法提取语音信号的特征向量的优点是：1、该系数可以很好的描述元音。2、用这种方法进行特征提取的时候计算量比较小，也很容易编程。LPCC 来提取语音信号的缺点是：第一、不容易识别辅音信号，容易受噪声性的干扰；第二、实际原理是不符合人的发声原理的，因为人所发出的声音信号不是随着时间成线性变化的。第三、得到的特征向量中也括一些高频的噪声信号。针对以上问题，最早由 Steven B. Davis 提出了 MFCC 系数，现在也有关于端点检测的研究使用 MFCC(美尔频率倒谱系数)[5]，MFCC 的特征提取方法在设计的时候考虑了把人耳的实际听音原理，所以 MFCC 方法把信号的频谱转化到了基于 Mel 频标的非线性频谱之上，然后再转换到倒谱域上，之后才提取语音信号的特征向量[6]。

3.2. 美尔频率倒谱系数(MFCC)

美尔频率倒谱系数 MFCC 是基于滤波器组进行特征提取。语音信号首先要经过一个滤波器进行预加重处理，来加强高频成分。加窗处理时必须注意两个问题：第一是窗宽的选择，第二是帧移大小的选择。MFCC 是把频谱转化到了 Mel 频谱上了，主要是因为在 Mel 频谱上进行处理要比一般的线性倒谱更接近人耳的听觉原理。这种方法主要是先用 FFT 把短时处理后的信号的 $x(n)$ 转化为频域信号的 $X(m)$，并计算其短时能量谱 $P(f)$。再将 $P(f)$ 的频率轴转换到 Mel 刻度上[7]，其转换公式如下：

$$Mel(f)=2595\log_{10}\left(1+\frac{f}{700}\right) \tag{7}$$

接下来的处理过程为：

1、计算信号在美尔域中的能量谱。

2、将该能量谱通过三角滤波器组进行滤波。

3、再在 Mel 域中求其倒谱系数。

如果第 k 个三角滤波器的能量输出为 $\theta(M_k)$，则在美尔刻度谱上可以采用修改的离散余弦反变换(IDCT)来求得美尔频率倒谱 $C_{mel}(n)$：

$$C_{mel}(n)=\sum_{k=1}^{K}\theta(M_k)\cos\left(n(k-0.5)\frac{\pi}{K}\right),\ n=1,2,\cdots,p \tag{8}$$

式中，K 表示滤波器的个数， p 为 MFCC 参数的阶数[8] [9]。实验表明，在多样本语音识别中，用美尔频率倒谱系数提取的特征也具有较好的识别效果和噪声鲁棒性。但识别过程的计算量和计算精确度要求比较高，它也会受到滤波器组中滤波器个数、形状、分布及能量谱等个因素的影响。所以 Victor Poblete 和 Felipe Espic，在美尔频率倒谱系数的基础上进行改进，提出了局部归一化倒谱系数。

3.3. 局部归一化倒谱系数 LNCC

局部归一化倒谱系数 LNCC 方法中所使用过的滤波器组如下：

$$\text{Numerator}_{\text{LNCC}}=\begin{cases}-\dfrac{2}{B}\left|f-f_i^c\right|+1 & \left|f-f_i^c\right|\le\dfrac{B}{2}\\ 0 & \text{otherwise}\end{cases} \tag{9}$$

$$\text{Denominator}_{\text{LNCC}} = \begin{cases} \frac{2}{B}(1-d_{\min})\left|f - f_i^c\right| + d_{\min} & \left|f - f_i^c\right| \le \frac{B}{2} \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

在上式中，Numerator 是局部归一化特征提取方法中分子滤波器，Denominator 是分母滤波器，B 指的是在某个特定频率周围所取得带宽，也就是在这个带宽上加上所设计的滤波器。实验结果表明，分子滤波器在中心频率上有很好的频率响应，将量个滤波器进行组合我们可以将所处理的信号进行的局部归一化，之后的特征提取过程和之前提到的美尔频率倒谱系数的特征提取过程一样，也就是说局部归一化倒谱系数的特征提取方法和美尔频率倒谱系数的特征提取方法的区别在于所用滤波器不同，在美尔频率倒谱系数的特征提取中使用的是三角滤波器，而在该方法中使用的是分子滤波器与分母滤波器的比值的滤波器。

经过以上分析可以得到局部归一化倒谱系数的特征提取方法如下：

1) 为了提高信号的高频部分，先对输入的语音信号预加重(信号通过预加重滤波器)，这样也可以将语音信号的频谱变得平滑。

2) 再将预加重后的语音信号进行加窗处理，通过窗函数在语音信号上平滑移动，将语音信号分成若干帧，也就将随时间变化的语音信号变成了短时稳定的信号。

3) 之后再对每一个窗所对应的短时信号进行傅里叶变换，求得其频域所对应的值 $X(m)$，再计算其短时能量谱。

4) 再将该短时能量谱在频率轴上的频谱转化成在 Bark 坐标上。

5) 在 Bark 频域内，将分子分母滤波器加在特定频率周围上。所使用的分子分母滤波器共有 28 对，带宽 B 为 3.5 Bark。

6) 将分子滤波器和分母滤波器的输出求比值，再求其对数，然后将输出用离散余弦变换进行降低最终特征的维数，最后即可得到所需要的特征系数[10]。

3.4. 小波包变换

在对比实验中所使用的小波包变换是在 MFCC 的基础上进行改进的，在 MFCC 中首先用到了傅里叶变换 FFT，我们知道，用傅里叶变换的结果会使得信号的低频分量很大，这样就会使得其他分量不够明显，所以我们对信号进行傅里叶变换之后，又取其对数，进而获得其幅值谱。

对语音信号 x 进行 N 点的离散傅里叶变换，结果为

$$\widehat{x_k} = \sum_{n=0}^{N-1} x_n \exp\left(-j2\pi nk/N\right), \quad k \in \{0,1,\cdots,N-1\} \tag{11}$$

第 i 个滤波器的对数能量输出为

$$x_k = \log_{10}\left(\sum_{k=0}^{N-1}\left|\widehat{x_i}\right|\widehat{h_k(i)}\right) \tag{12}$$

在 MFCC 中，其美尔频率倒谱系数的计算方法如下

$$c_t = \sum_{k=1}^{M} x_k \cos\left(t\left(k-0.5\right)\frac{\pi}{M}\right), \quad t \in \{1,2,\cdots,T\} \tag{13}$$

其中，T 是倒谱系数的个数，也是 MFCC 参数的阶数，而 M 是三角滤波器的个数。而在小波包变换 WPT 的方法中，从细节分量 $d_x(j,i,k)$(此细节分量表示的是在尺度为 j 时，第 i 个空间里的第 k 个系数所对应的细节分量)中得到的细节信号为：

$$D_j(t)=\left(\sum_{i\in Ij}\sum_{k=0}^{n_j-1} d_x(j,i,k)_{j,k}(t)\right) \tag{14}$$

用 $D_j(t)$ 替换滤波器的能量输出 x_k，可得到用小波包变换得到的能量输出 Z_{kj}，则

$$Z_{kj}=\log_{10}\left(\sum_{i=0}^{n-1}\left|D_j(t)\right|\widehat{h_k(i)}\right),\quad j\in\{1,2,\cdots,L\} \tag{15}$$

L 为用小波包变换提取语音信号的特征时所用的总的尺度数。最后将美尔频率倒谱系数求解公式中的 x_k 用上式求出的 Z_{jk} 替换可得到用小波包变换 WPT 提取的语音信号的特征 W_t，其中 W_t 的计算公式如下：

$$W_t=\sum_{k=1}^{M} Z_{kj}\cos\left(t(k-0.5)\frac{\pi}{M}\right),\quad t\in\{1,2,\cdots,T\} \tag{16}$$

在上式中，相当于将 MFCC 中的短时傅里叶变换变换成了小波包变换。该小波包变换在小波 $\{db1,db2,\cdots,db10\}$ 上重复进行，所使用的 WPT 的尺度共有 6 个，滤波器的个数 M = 20，预加重时，用到的系数为 a = 0.95，加窗处理时，所选的窗长为 320，重叠 160 个采样点[11]。

4. 实验过程及实验结果

将本文所提出的特征提取方法和 LPCC、MFCC、LNCC 和 WPT 这四种特征提取方法进行对比。用到的数据库都是标准的 TIMIT 数据库和自己建立的中文数据库。所用的语音训练样本是单样本集。在每个数据库的实验都分为两部分，第一部分是在不加噪声时看其识别率的高低，第二部分是看被噪声感染了的待测试语音在信噪比 dB 分别为 0、10、15、20、25、30、35 时的识别结果。其中每次实验中所加的噪声都是与验证本文提出的特征提取方法的识别率时所使用的相同的白噪声，相同的粉红噪声和相同的工业噪声。将本文所提出的特征提取的方法和 LPCC、MFCC、LNCC 和 WPT 的特征提取方法在单训练样本上进行相比较。实验结果如下：

实验 1：在没有噪声的环境下，将本文所提出的方法和其他对比方法进行验证，实验结果如下表 1。

实验 2：对实验 1 的每个测试样本分别添加白噪声与粉红噪声，感染信噪比分别为 0、10、15、20、25、30、35 dB 组成测试样本，其识别结果如图 1、图 2、图 3、图 4 所示，可以看出，本文方法对白噪

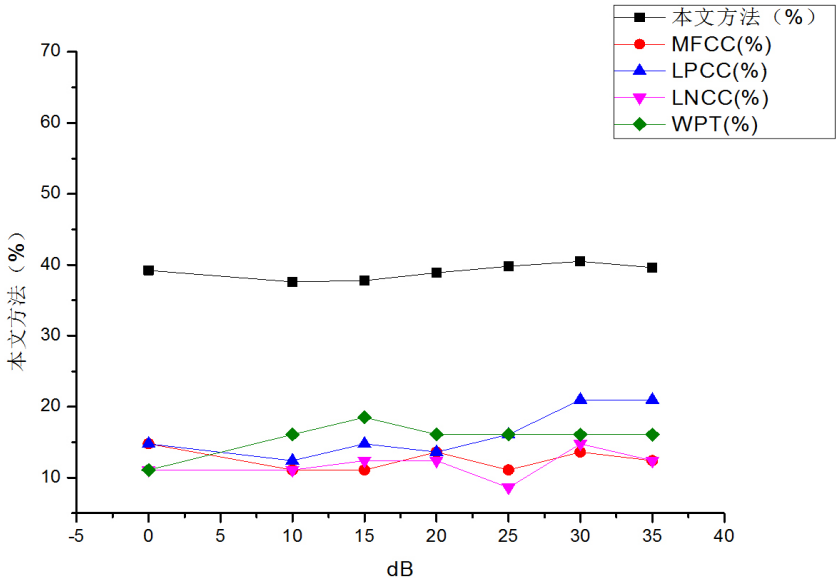

Figure 1. Recognition rate comparisons under white noise in TIMIT database
图 1. TIMIT 数据库加白噪声时的识别率

Table 1. Recognition rate comparison between different algorithms in clean environment
表 1. 在没有噪声的环境下，不同算法识别率的比较

特征提取的方法 \ 语音数据库	TIMIT 数据库	自己建立的中文数据库
LPCC	16.05%	25.93%
MFCC	18.52%	11.11%
LNCC	20.99%	12.35%
WPT	17.28%	13.58%
本文的方法	40.6%	39.5%

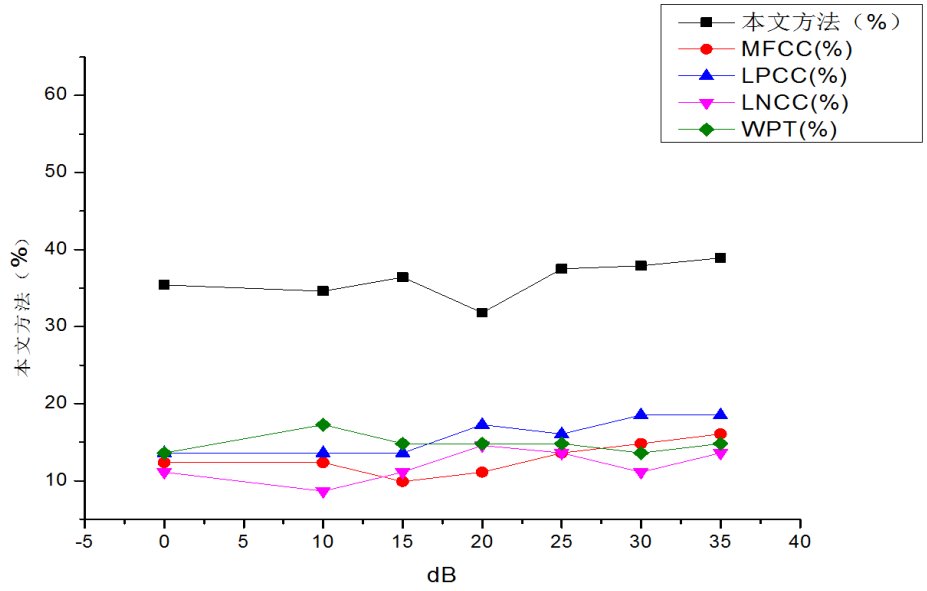

Figure 2. Recognition rate comparisons under pink noise in TIMIT database
图 2. TIMIT 数据库加粉红噪声时的识别率

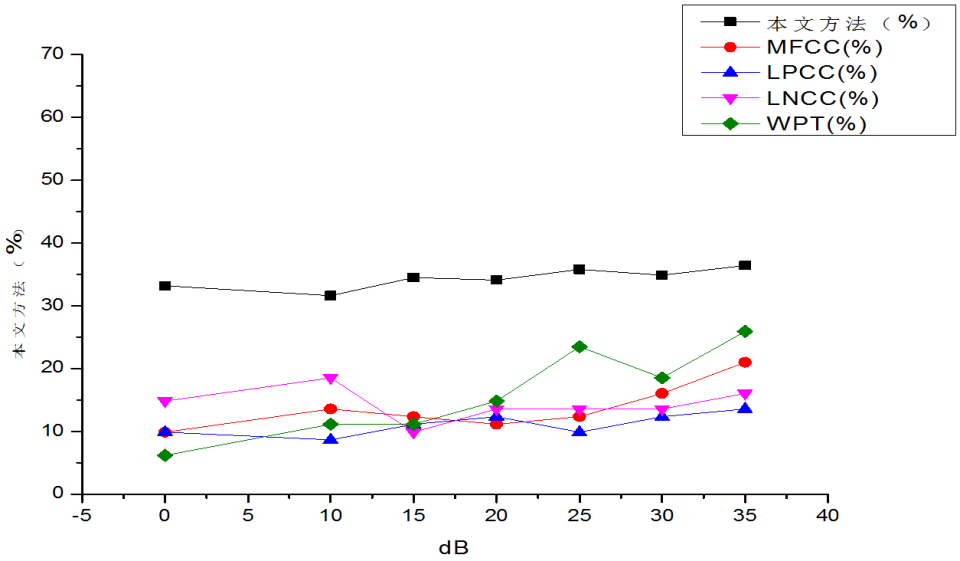

Figure 3. Recognition rate comparisons under white noise in Chinese database
图 3. 自建中文数据库加白噪声时的识别率

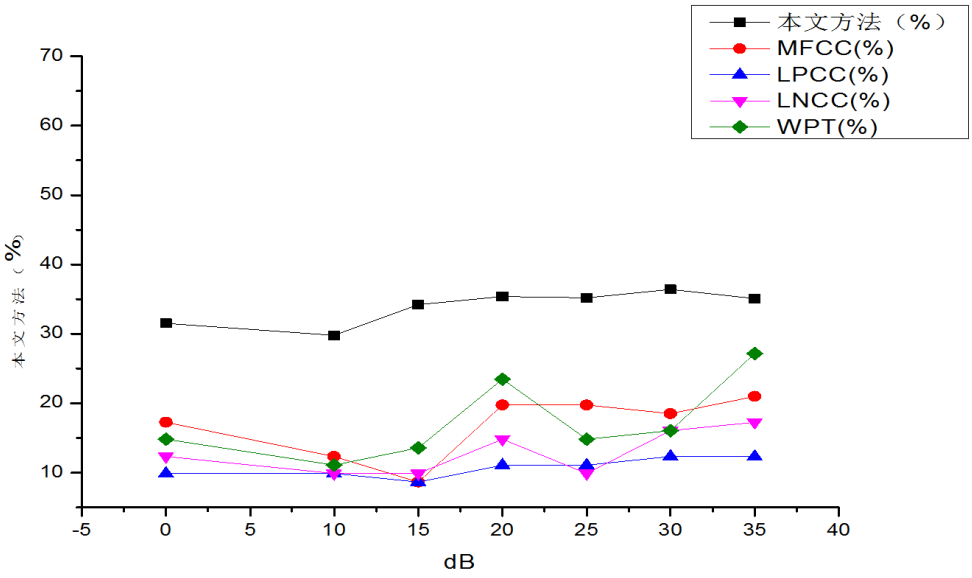

Figure 4. Recognition rate comparisons under pink noise in Chinese database
图 4. 自建中文数据库加粉红噪声时的识别率

声与粉红噪声具有较高的鲁棒性。

5. 结论

本文提出了一种新的简单的语音信号的特征提取方法，该方法所提取的特征能够充分反映说话人的基本发声特性，而且对语音常见的白噪声、粉红噪声具有很好的鲁棒性。该方法在单样本的说话者识别中识别效果要比 MFCC、LPCC、LNCC 和 WPT 方法好。在以后的工作中将进一步验证与提升其在语音样本比较少的情况下的识别性能，也会尝试将几种特征提取方法相结合进而寻求更高的识别效果。

致　谢

在此非常感谢我的导师对我论文的指导和帮助，感谢提供参考文献的研究人员，也感谢编辑部对我论文审核，感谢所有帮助我的人。

参考文献 (References)

[1] Pohjalainen, J. and Räsänen, O. (2015) Feature Selection Methods and Their Combinations in High-Dimensional Classification of Speaker Likability, Intelligibility and Personality Traits. *Computer Speech and Language*, **29**, 145-171.

[2] Kinnunen, T. and Li. H.Z. (2010) An Overview of Text-Independent Speaker Recognition: From Features to Supervectors. *Speech Communication*, **52**, 12-40. http://dx.doi.org/10.1016/j.specom.2009.08.009

[3] Vijayasenan, D. and Valente, F. (2012) Multistream Speaker Diarization of Meetings Recordings beyond MFCC and TDOA Features. *Speech Communication*, **54**, 55-67. http://dx.doi.org/10.1016/j.specom.2011.07.001

[4] 王彪. 基于 LPCC 参数的语音识别系统[J]. 电子设计工程, 2012, 20(7).

[5] 许昊, 张二华. 基于改进 C0 复杂度和 MFCC 相似度的端点检测[J]. 现代电子技术, 2015, 38(10).

[6] Madikeri, S. (2012) Effect of Feature Warping and Decorrelation on Mel Filter bank Slope for Speaker Recognition, IEEE, 978-1-4673.

[7] R. Shantha Selva Kumari, S. Selva Nidhyananthan and Anand. G. (2012) Fused Mel Feature Sets Based Text-Independent Speaker Identification Using Gaussian Mixture Model. *Procedia Engineering*, **30**, 319-326. http://dx.doi.org/10.1016/j.proeng.2012.01.867

[8] Ai, O.C. and Hariharan, M. (2012) Classification of Speech Dysfluencies with MFCC and LPCC Features. *Expert Systems with Applications*, **39**, 2157-2165. http://dx.doi.org/10.1016/j.eswa.2011.07.065

[9] El-Henawy, I.M. and Khedr, W.I. (2014) Recognition of Phonetic Arabic Figures via Wavelet Based Mel Frequency Cepstrum Using HMMs. *HBRC Journal*, **10**, 49-54.

[10] Poblete, V. and Espic, F. (2015) A Perceptually-Motivated Low-Complexity Instantaneous Linear Channel Normalization Technique Applied to Speaker Verification. *Computer Speech and Language*, **31**, 1-27. http://dx.doi.org/10.1016/j.csl.2014.10.006

[11] Turner, C. and Joseph, A. (2015) A Wavelet Packet and Mel-Frequency Cepstral Coefficients-Based Feature Extraction Method for Speaker Identification. *Procedia Computer Science*, **61**, 416-421. http://dx.doi.org/10.1016/j.procs.2015.09.177