融合CFCC和Teager能量算子倒谱参数的语音识别

史燕燕 白 静
(太原理工大学信息与计算机学院 太原 030024)

摘 要 针对现有表征语音特性的特征提取不完善的问题,提出了一种耳蜗滤波倒谱系数(Cochlear Filter Cepstral Coefficients,CFCC)和Teager能量算子倒谱参数(Teager Energy Operators Cepstral Coefficients,TEOCC)相互融合的方法。该方法将表征人耳听觉特性的CFCC和体现非线性能量特性的TEOCC的融合特征应用到语音识别系统中,并联合主成分分析(Principal Components Analysis,PCA)对该融合特征进行特征选择和优化,最后通过支持向量机进行语音识别。实验结果表明:该融合特征与单一特征相比具有更佳的语音识别性能,结合PCA后其语音识别的准确率平均提高了3.7%。

关键词 耳蜗滤波倒谱系数,Teager能量算子倒谱参数,主成分分析,语音识别
中图法分类号 TN912.34 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.05.044

Speech Recognition Combining CFCC and Teager Energy Operators Cepstral Coefficients

SHI Yan-yan BAI Jing
(College of Information and Computer,Taiyuan University of Technology,Taiyuan 030024,China)

Abstract In view of the imperfection of the existing features which represent the speech characteristics,this paper proposed a mutual integration method based on Cochlear Filter Cepstral Coefficients and Teager Energy Operators Cepstral Coefficients. First,the fusion feature of CFCC that reflects human auditory characteristics and TEOCC that embodies nonlinear energy characteristics is applied to speech recognition system. Then principal component analysis is applied to the selection and optimization of fusion features. Finally,support vector machine is used for speech recognition. The results show that the proposed fusion features can achieve better speech recognition performance than single feature,and after combining PCA,the accuracy rate of speech recognition is increased by 3.7% on average.

Keywords CFCC,TEOCC,PCA,Speech recognition

1 引言

语音识别包括两大部分:特征提取和模式识别。特征提取作为语音识别中的重要环节,对识别系统的性能有直接且显著的影响。因此,如何从语音信号中提取能够充分表征其语义信息的最优特征参数,从而进一步提高识别率是语音识别面临的最大问题之一。目前最普遍且有效的语音特征是基于人耳听觉特性提出的,其源于人耳具有良好的抗噪能力。越来越多的研究者致力于研究人耳听觉特征,并建立了更符合人耳听觉特性的语音特征参数模型[1]。其中,最主流的特征参数是梅尔频率倒谱系数(Mel Frequency Cepstrum Coefficent,MFCC),但是MFCC的性能随着信噪比的降低会大幅下降[2],从而导致语音识别系统的稳定性较差。2009年,贝尔实验室的Peter Li博士首次提出听觉变换[3]的概念。Li等[4]基于听觉变换,提取了耳蜗滤波倒谱系数特征,并将其应用于不匹配条件下的鲁棒说话人辨识系统。另外,针对单一特征不足以表征语音信号的完整特性,李作强等[5]将相位信息和CFCC特征融合应用到说话人辨识系统。Patel等[6-7]将CFCC及其衍生特征与MFCC特征进行多特征融合来检测自然语音和合成语音,但他们并未考虑能量信息。Bandela等[8]和Sreeraj等[9]将Teager能量算子和MFCC进行特征融合并分别应用于应激语音情感识别和自动方言识别,以此来验证能量特征的有效性。本研究将结合人耳听觉特性和非线性能量特性,提出一种融合特征提取算法,将CFCC特征及其一阶差分和TEOCC特征相互融合,并验证该融合特征对于单一特征的优化作用。此外,本研究还联合PCA对所提出的融合特征进行特征选择和优化,最后采用支持向量机进行识别。实验结果验证了该方法的有效性。

2 特征参数的提取

2.1 CFCC特征参数的提取

Peter Li将耳蜗滤波函数作为一种新的小波基函数,运用小波变换实现滤波过程,此过程被称为听觉变换(Auditory Transform,AT)。听觉变换利用人耳听觉机理,首先定义一

到稿日期:2018-03-23 返修日期:2018-06-12 本文受山西省青年科技研究基金,山西省科技攻关(社会发展)项目资助。
史燕燕(1994-),女,硕士生,主要研究方向为语音信号处理,E-mail:690742874@qq.com;白 静(1965-),女,博士,教授,硕士生导师,主要研究方向为语音信号处理,E-mail:bj613@126.com(通信作者)。

个耳蜗滤波函数 $\varphi(t)\in L^2(R)$，要求 $\varphi(t)$ 满足式(1)－式(3)[10]。

$$\int_{-\infty}^{+\infty}\varphi(t)\mathrm{d}t=0 \tag{1}$$

$$\int_{-\infty}^{+\infty}|\varphi(t)|^2\mathrm{d}t<\infty \tag{2}$$

$$\int_{-\infty}^{+\infty}\frac{|\psi(\omega)|^2}{\omega}\mathrm{d}\omega=C \tag{3}$$

其中，$0<C<\infty$，并且

$$\psi(\omega)=\int_{-\infty}^{+\infty}\varphi(t)\mathrm{e}^{-j\omega t}\mathrm{d}\omega \tag{4}$$

假设 $f(t)$ 为任意一个平方可积的函数，则 $f(t)$ 的听觉变换输出为：

$$T(a,b)=\int_{-\infty}^{+\infty}f(t)\ \varphi_{a,b}(t)\mathrm{d}t \tag{5}$$

其中，$\varphi_{a,b}(t)$ 是耳蜗滤波函数，其表达式为：

$$\varphi_{a,b}(t)=\frac{1}{\sqrt{a}}\frac{(t-b)^a}{a}\exp[-2\pi f_L\beta(\frac{t-b}{a})]\times[\cos 2\pi f_L(\frac{t-b}{a})+\theta]u(t-b) \tag{6}$$

其中，$\alpha>0$，$\beta>0$。α 和 β 的取值决定了耳蜗滤波函数的频域形状和宽度，一般取经验值 $\alpha=3$，$\beta=0.2$。$u(t)$ 为单位步进函数，b 为随时间可变的实数，a 为尺度变量，θ 为初始相位。一般情况下，a 可由滤波器组的中心频率 f_c 和最低中心频率 f_L 决定：

$$a=f_L/f_c \tag{7}$$

将式(6)代入式(5)可得 $f(t)$ 经过 AT 变换的输出 $T(a,b)$。

经过时频转换的语音信号会被人耳耳蜗的内毛细胞转变成人脑分析的电信号，这一过程可用式(8)－式(10)来模拟。

$$h(a,b)=[T(a,b)]^2 \tag{8}$$

$$S(i,j)=\frac{1}{d}\sum_{b=1}^{l+d-1}h(i,b),\forall i,j \tag{9}$$

其中，$l=1,L,2L,\cdots$；$d=\max\{3.5\tau_i,20\ \mathrm{ms}\}$，$d$ 是第 i 频带毛细胞函数的窗长，τ_i 是第 i 个滤波器中心频带中心频率的时间长度，$\tau_i=1/f_c$；L 为帧移，一般情况下取 $L=d/2$；j 是窗的个数。

将式(9)的输出 $S(i,j)$ 进一步应用于响度函数的尺度变换，如非线性对数变换或立方根响度变换。按照文献[4]中的 CFCC 特征的提取方法，本文使用非线性对数变换，其过程用式(10)来模拟。最后，采用离散余弦变换进行去相关，得到 CFCC 特征参数。

$$y(i,j)=\log[S(i,j)] \tag{10}$$

原始的 CFCC 参数反映的是语音信号的静态特性，因为人耳对动态参数更加敏感[11]，加入一阶差分后能更完整地表征语音信号的动态特征，从而提高识别系统的性能。

2.2　TEOCC 参数的提取

Teager 能量算子[12]是由 Kaiser 提出的一种非线性差分算子，具有跟踪信号非线性能量的特性，能够合理地呈现信号能量的变换。

对于一离散时间信号 $x(n)$，TEO 定义[13]为：

$$\Psi[x(n)]=x(n)^2-x(n+1)x(n-1) \tag{11}$$

其中，$\Psi[x(n)]$ 是 TEO 的输出，$x(n)$ 是离散信号在 n 点时的采样值。

含有加性噪声的语音信号 $x(n)$ 可以表示为纯净语音信号 $s(n)$ 与零均值加性噪声 $\omega(n)$ 之和，即：

$$x(n)=s(n)+\omega(n) \tag{12}$$

$x(n)$ 的 TEO 可以表示为：

$$\Psi[x(n)]=\Psi[s(n)]+\Psi[\omega(n)]+2\ \tilde{\Psi}[s(n),\omega(n)] \tag{13}$$

其中，$\tilde{\Psi}[s(n),\omega(n)]$ 是 $s(n)$ 与 $\omega(n)$ 的互 Teager 能量，且

$$\tilde{\Psi}[s(n),\omega(n)]=s(n)\omega(n)-0.5s(n-1)\omega(n+1)-0.5s(n+1)\omega(n-1) \tag{14}$$

由于 $s(n)$ 和 $\omega(n)$ 均为零均值且两者相互独立，因此有：

$$E\{\tilde{\Psi}[s(n),\omega(n)]\}=0 \tag{15}$$

推导出：

$$E\{\Psi[x(n)]\}=E\{\Psi[s(n)]\}+E\{\Psi[\omega(n)]\} \tag{16}$$

一般情况下，与纯净语音信号的 TEO 能量相比较，噪声的 TEO 能量几乎可以忽略不计，因此可以得到：

$$E\{\Psi[x(n)]\}\approx E\{\Psi[s(n)]\} \tag{17}$$

由此可见，Teager 能量算子能够消除零均值噪声的影响，达到增强语音的目的[14]。因此，将 Teager 能量算子用于特征提取，不仅能够更好地反映语音信号的能量变化，而且能够抑制噪声，增强语音信号，从而获得良好的语音识别效果。

经过预处理的语音信号根据 2.2 节的式(11)求出每帧语音信号的平均 TEO 能量，进行归一化处理并取对数得到：

$$\hat{\Psi}[x(n)]=\log\{\Psi[x(n)]/\max(\Psi[x(n)])\} \tag{18}$$

然后进行 DCT 变换得到一维的 TEOCC。

3　融合特征的提取和优化

为了构造更有效的语音特征子集，本研究将听觉特征和非线性能量特征进行融合，在提取的 CFCC 和其一阶差分的基础上，加入反映信号能量变化的 TEOCC，得到的融合特征既表征了人耳听觉感知特性，又结合了语音瞬时能量的特性，还在一定程度上抑制了零均值噪声对语音信号的影响，因此更能完整地描述语音的特性。

为减少特征数据的存储量，进一步获得最优特征集合，将融合特征进行主成分分析，达到降维并缩短识别时间的目的，以进一步提高识别系统的性能。在目前的研究中，PCA[15]是一类基于 K-L 变换的统计分析方法，该方法利用协方差矩阵将原来维数较高的具有相关性的数据线性组合成维数较少且互不相关的数据。

首先对提取到的融合特征矩阵进行数据归一化，将每列数据样本设为 x_i，归一化后的数据样本为：

$$x_i^*=\frac{x_i-x_{\min}}{x_{\max}-x_{\min}} \tag{19}$$

其中，x_i 是原始数据，x_i^* 是归一化后的数据，$x_{\min}$ 和 $x_{\max}$ 分别代表 x_i 的极小值和极大值。

然后进行主成分分析，计算相关系数矩阵及其特征值和特征向量，基于该特征向量组成一个新的矩阵，使该新矩阵的所有列根据贡献率降序排列。主成分的累积贡献率为：

$$\alpha_p=\frac{\sum_{i=1}^{p}\lambda_i}{\sum_{i=1}^{n}\lambda_i} \tag{20}$$

其中，λ 表示每一维的特征值，p 表示前 p 个主成分，α_p 表示

累积贡献率。本研究选取累积贡献率为 97%。

由于特征矢量各维数对语音识别的贡献率不同，本研究采用增减分量法[16]对特征矢量 CFCC 和 ΔCFCC 的每一维分量进行识别率比较，各去除两维贡献率最低的分量，再将 TEOCC 特征加入最后一维，按照不同的组合方式组成特征向量 ECFCC，CFCC＋ΔCFCC 和本研究提出的融合特征。采用 PCA 对该融合特征进行优化，将优化后的特征集合定义为 PCA-Features，并将其作为支持向量机的输入。具体算法流程如图 1 所示。

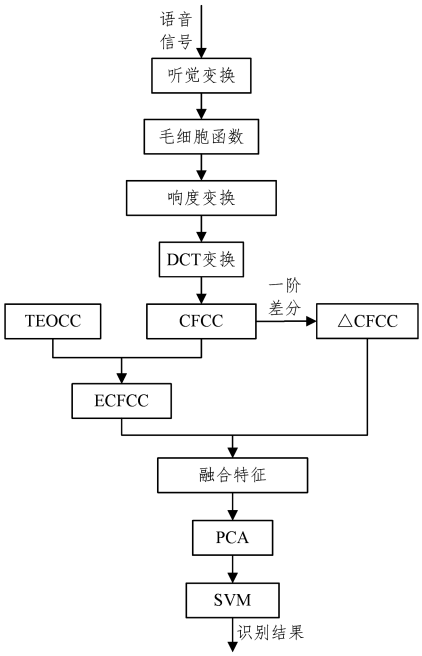


图 1 本文算法的流程

Fig. 1 Flow of the proposed algorithm

4 实验结果与分析

4.1 实验准备

本研究的实验环境是基于 MATLAB R2011a 平台实现的语音识别系统，采用的是针对非特定人的孤立词语音库。实验将 16 个人在不同信噪比(0 dB，5 dB，10 dB，15 dB，20 dB)下对 10 词、20 词的发音作为语音数据，每人每词发音 3 次，其中 9 人的发音作为训练数据，7 人的发音作为测试数据。

4.2 实验结果分析

本实验采用支持向量机作为语音识别模型，针对非特定人的孤立词的语音库提取 CFCC，ΔCFCC 和 TEOCC 特征，按照不同的特征组合方式进行语音识别，在不同信噪比环境下设计 4 组实验来验证提出的融合特征在语音识别中的优越性能。实验结果如表 1 和表 2 所列。

(1)从实验一和实验二的识别率可知：加入反映语音信号能量变化的 TEOCC 特征之后，不同信噪比下的识别率都有了一定的提升，尤其是在 10 词情况下 0 dB 时提升最为明显，增加了 4.48%，证明 Teager 能量算子倒谱参数中包含语音信号的语义信息，可以作为辅助特征参数来提高语音识别系统的性能。

(2)从实验一和实验三的识别结果可知：在 CFCC 参数提取过程中加入一阶差分系数，即用语音信号静态特征的一阶差分谱来描述其动态特性，在 5 种信噪比环境下 CFCC＋ΔCFCC 的识别率均比 CFCC 参数的识别率高，在 10 词情况下平均高出 2.75%，在 20 词情况下平均高出 4.91%，说明动、静态特征的结合能更有效地表征语音信号的完整特性。

(3)对比实验一、实验三和实验四可知：本文提出的融合特征与 CFCC 以及 CFCC＋ΔCFCC 的组合特征参数相比识别率最高，在 10 词情况下的信噪比为 20 dB 时识别率高达 94.29%，说明结合 TEOCC 参数能反映语音信号非线性能量特性并消除语音信号噪声，使得语音识别系统表现出较好的分类性能，进一步证明此融合特征的有效性。

表 1 10 词情况下不同特征组合方式的识别率

Table 1 Comparison of recognition rates of different features combinations in 10 words

(单位：%)

实验	特征参数	SNR/dB					平均识别率
		0	5	10	15	20	
实验一	CFCC	73.14	79.84	82.81	87.14	86.19	81.82
实验二	ECFCC	77.62	82.38	86.67	87.62	87.62	84.38
实验三	CFCC＋ΔCFCC	81.43	81.43	86.19	87.73	87.14	84.78
实验四	ECFCC＋ΔCFCC	85.24	87.62	91.90	93.81	94.29	90.57

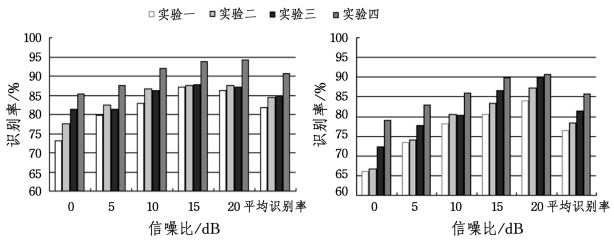
表 2 20 词情况下不同特征组合方式的识别率

Table 2 Comparison of recognition rates of different features combinations in 20 words

(单位：%)

实验	特征参数	SNR/dB					平均识别率
		0	5	10	15	20	
实验一	CFCC	66.10	73.42	78.07	80.45	83.99	76.41
实验二	ECFCC	66.67	73.98	80.48	83.43	87.34	78.38
实验三	CFCC＋ΔCFCC	72.25	77.70	80.30	86.59	89.76	81.32
实验四	ECFCC＋ΔCFCC	78.96	82.90	85.87	89.94	90.69	85.67

图 2 更加直观地描述了 10 词、20 词时 4 组实验在不同信噪比环境下的语音识别结果。从图 2 可以看出，10 词情况下在 CFCC 基础上分别加入 ΔCFCC 和 TEOCC 特征的识别效果几乎相当，而本文提出的融合特征的识别率均高于前 3 种特征。这说明能量特征 TEOCC 可以对人耳听觉倒谱特征起到特征补偿的作用，应用于语音识别可取得良好的效果。从整体上看，不管是 10 词还是 20 词的情况，本文提出的融合特征参数的识别率最高，因此结合听觉感知特征和能量特征的方法可以构造更优的语音特征集合。



(a)10 词情况下的语音识别结果 (b)20 词情况下的语音识别结果

图 2 4 组实验的识别率比较

Fig. 2 Comparison of recognition rates of four experiments

为了进一步验证优化特征集的有效性，将融合特征集与

优化特征集进行识别率比较实验，实验结果如表 3 所列。

表 3　基于 ECFCC＋△CFCC 和 PCA-Features 的识别率对比

Table 3　Comparison of recognition rates based on ECFCC＋ΔCFCC and PCA-Features

（单位：%）

词汇量	特征参数	SNR/dB					平均识别率
		0	5	10	15	20	
10 词	ECFCC＋ΔCFCC	85.24	87.62	91.90	93.81	94.29	90.57
	PCA-Features	89.05	89.52	92.38	94.29	94.76	92.00
20 词	ECFCC＋ΔCFCC	78.96	82.90	85.87	89.94	90.69	85.67
	PCA-Features	85.85	89.59	93.87	94.23	94.23	91.55

从表 3 可以看出，融合特征经 PCA 优化后，识别率均得到提升。这是因为 PCA 分析能够减小特征参量之间的相关性，保留特征参数中重要的成分，去除冗余特征，突出特征参数之间的差异性，从而使语音识别系统的性能得到进一步提高。

结束语　本文采用非特定人的孤立词的语音库，使用语音信号的能量信息特征参数对听觉倒谱特征 CFCC 进行特征补偿，提出新的融合特征集合，并对其进行优化，最后建立支持向量机的语音识别模型，并在语音识别系统上取得了良好的效果。首先在语音库的不同词汇量和不同信噪比环境下，采用单一类型特征和融合特征进行实验，验证了融合特征对于单一特征的优化作用；在此基础上，结合 PCA 对该融合特征进行特征选择，得到优化后的特征集合，使得该集合在降低识别时间的同时得到了更高的识别准确率。实验结果表明，本文提出的融合特征在低信噪比(0 dB)10 词条件下，识别率可以达到 85.24%，结合 PCA 后的识别率达到 89.05%，进一步证明了本文提出的最优特征参数集合的有效性和可行性。

参 考 文 献

[1] GAO Y. Cochlear Filter Cepstral Feature in Speech recognition [D]. Taiyuan: Taiyuan University of Technology, 2011. (in Chinese)
高扬. 耳蜗滤波器倒谱特征在语音识别中的应用[D]. 太原：太原理工大学，2011.

[2] WANG L, MINAMI K, YAMAMOTO K, et al. Speaker Recognition by Combining MFCC and Phase Information in Noisy Conditions[J]. IEICE Transactions on Information & Systems, 2010, 93-D(9): 2397-2406.

[3] LI Q. An auditory-based transform for audio signal processing [C]// IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2009 (WASPAA'09). IEEE, 2009: 181-184.

[4] LI Q, HUANG Y. An Auditory-Based Feature Extraction Algorithm for Robust Speaker Identification Under Mismatched Conditions[J]. IEEE Transactions on Audio Speech & Language Processing, 2011, 19(6): 1791-1801.

[5] LI Z Q, GAO Y. Robust speaker identification based on CFCC and phase information[J]. Computer Engineering and Applications, 2015, 51(17): 228-232. (in Chinese)
李作强，高勇. 基于 CFCC 和相位信息的鲁棒性说话人辨识[J]. 计算机工程与应用，2015，51(17)：228-232.

[6] PATEL T B, PATIL H. Combining Evidences from Mel Cepstral, Cochlear Filter Cepstral and Instantaneous Frequency Features for Detection of Natural vs. Spoofed Speech[C]// The Conference of International Speech Communication Association. 2015.

[7] PATEL T B, PATIl H A. Cochlear Filter and Instantaneous Frequency Based Features for Spoofed Speech Detection[J]. IEEE Journal of Selected Topics in Signal Processing, 2017, 11(4): 618-631.

[8] BANDELA S R, KUMAR T K. Stressed speech emotion recognition using feature fusion of teager energy operator and MFCC [C]// International Conference on Computing, Communication and Networking Technologies. IEEE Computer Society, 2017: 1-5.

[9] SREERAJ V V, RAJAN R. Automatic dialect recognition using feature fusion[C]// International Conference on Trends in Electronics and Informatics. 2017: 435-439.

[10] LI J J, AN D, YANG D, et al. TEO-CFCC Characteristic Parameter Extraction Method for Speaker Recognition in Noisy Environments[J]. Computer Science, 2012, 39(12): 195-197. (in Chinese)
李晶皎，安冬，杨丹，等. 噪声环境下说话人识别的 TEO-CFCC 特征参数提取方法[J]. 计算机科学，2012，39(12)：195-197.

[11] WU D, CAO J, WANG J H. Speaker recognition based on adapted Gaussian mixture model and static and dynamic auditory feature fusion[J]. Optics and Precision Engineering, 2013, 21(6): 1598-1604. (in Chinese)
吴迪，曹洁，王进花. 基于自适应高斯混合模型与静动态听觉特征融合的说话人识别[J]. 光学精密工程，2013，21(6)：1598-1604.

[12] KAISER J F. On a simple algorithm to calculate the 'energy' of a signal[C]// International Conference on Acoustics, Speech, and Signal Processing. IEEE, 2002: 381-384.

[13] WANG M R, ZHOU P, JING X X. Mixed Peramaters of Mel Frequency Cepstral and Short-time TEO Energy in Speaker Recognition[J]. Microelectronics & Computer, 2016, 33(1): 144-148. (in Chinese)
王茂蓉，周萍，景新幸. MFCC 和短时 TEO 能量的混合参数应用于说话人识别[J]. 微电子学与计算机，2016，33(1)：144-148.

[14] LI J, ZHOU P, DU Z R. Application of short-time TEO energy in noisy speech endpoint detection[J]. Computer Engineering and Applications, 2013, 49(12): 144-147. (in Chinese)
李杰，周萍，杜志然. 短时 TEO 能量在带噪语音端点检测中的应用[J]. 计算机工程与应用，2013，49(12)：144-147.

[15] JIANG H H, HU B. Speech Emotion Recognition in Mandarin based on PCA and SVM[J]. Computer Science, 2015, 42(11): 270-273. (in Chinese)
蒋海华，胡斌. 基于 PCA 和 SVM 的普通话语音情感识别[J]. 计算机科学，2015，42(11)：270-273.

[16] YUE Q Q, ZHOU P, JING X X. The Auditory Feature Extraction Algorithm Based on Power-law Nonlinearity Function[J]. Microelectronics and Computers, 2015(6): 163-166. (in Chinese)
岳倩倩，周萍，景新幸. 基于非线性幂函数的听觉特征提取算法研究[J]. 微电子学与计算机，2015(6)：163-166.