北京大学学报(自然科学版),第 37 卷,第 3 期,2001 年 5 月
Acta Scientiarum Naturalium
Universitatis Pekinensis, Vol. 37, No. 3 (May, 2001)

语音识别和说话人识别中各倒谱分量的相对重要性[1)]

甄 斌 吴玺宏 刘志敏 迟惠生
(北京大学信息科学中心,100871,北京)

摘 要 采用增减特征分量的方法研究了 MFCC 各维倒谱分量对说话人识别和语音识别的贡献。使用 DTW 测度,在标准英文数字语音库上的实验表明,最有用的语音信息包含在 MFCC 分量 C_1 到 C_{12} 之间,最有用的说话人信息包含在 MFCC 分量 C_2 到 C_{16} 之间。MFCC 分量 C_0 和 C_1 包含有负作用的说话人信息,将其作为特征会引起识别率的降低。低阶 MFCC 分量较高阶分量更容易受加性噪声和卷积噪声干扰。

关键词 MFCC; 说话人识别; 语音识别
中图分类号 TP 319; TP 39119

0 引 言

与人听觉系统非凡的感知能力比较,目前的语音识别和说话人识别等机器系统还存在许多问题,尤其是在不利的噪声环境下,系统性能急剧下降[1,2]。语音识别和说话人识别系统中特征提取过程就是抽取保持语音最重要特征,并消除与语音无关信号的干扰,其性能对识别系统的性能有直接影响[1]。寻找具有良好性能的特征及其提取算法是提高识别系统性能的根本途径之一。目前,常用的语音特征包括基于声道的 LPCC、基于临界带的 MFCC 以及基于临界带和等响度曲线的 PLP[1~3],考虑语音动态特性的一阶和二阶差分倒谱[4],考虑语音时域特性的 RASTA 滤波[5],还有其他基于听觉模型的特征[6]。

在语音识别和说话人识别的特征中,通常丢弃第零阶倒谱系数以归一化功率谱,为什么只丢弃第零阶倒谱系数[1,2]?还有没有其他阶倒谱系数需要丢弃?Juang 提出适合语音识别的升余弦(Raised2Sine)倒谱提升,甄斌等提出适合说话人识别的半升余弦(Half Raised2Sine)倒谱提升[7,8]。不同的倒谱窗口对识别率有较大的影响,它暗示不同倒谱系数项对识别的贡献是不一样的。由此引出这样一个重要问题,即提取的所有特征分量是否都对识别有贡献?它们对识别目标的重要程度是否相同?同时还有特征的抗噪声性问题。

本文以目前较为常用的 MFCC 特征为例,采用增减特征分量的方法评价 MFCC 各阶系数对语音识别和说话人识别性能的影响。在第二部分介绍了倒谱分量相对重要性评价方法,第三部分是使用的标准语音库,第四部分是 MFCC 各分量对语音识别和说话人识别的相对重要

1) 国家自然科学基金(69635050)、北京市自然科学基金(4002012)和高等学校骨干教师资助计划资助项目
收稿日期:2000204205

性,最后是讨论和结论。

1 倒谱分量相对重要性评价方法

评价特征对识别的贡献有两种方法,通过定义各分量的F比可以得到各特征分量的区分能力,还可直接进行识别,通过增减分量的方法考察每个特征分量的贡献[2,9,10]。本文采用第二种方法,每个MFCC倒谱分量的平均贡献R(i)由下式计算[10]:

$$R(i)=\frac{1}{n}\left(\sum_{j>i}(p(i,j)-p(i+1,j))+\sum_{j<i}(p(j,i)-p(j,i-1))\right)。\quad(1)$$

式中,n是倒谱阶数,p(i,j)是以i阶到j阶倒谱系数为特征的识别率。图1是式(1)的图示说明,比如,以第0至2阶倒谱系数为特征的识别率p(0,2)减去以第0至1阶倒谱系数为特征的识别率p(0,1),就得到在以C_0~C_1为特征时倒谱分量C_2对识别的贡献。将在所有可能的顺序组合的C_2贡献的平均就得到本文定义的C_2的平均贡献(相对重要性)。

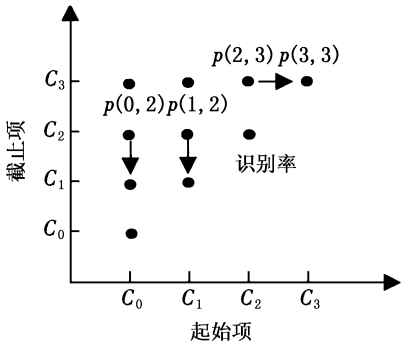

图1 MFCC分量平均贡献评价方法

Fig. 1 Grid for evaluating the average importance of MFCC components

正值的平均贡献R(i)表明由于添加该特征往往会使识别率增加(识别率增加的多少同所利用的其他倒谱特征分量有关。在某种特定的特征分量组合下,也有可能使得识别率降低。但从所有特征分量可能的顺序组合平均,包含该特征的识别率总体是增加的),负的平均贡献R(i)则相反。由于本文实验仅顺序添加或舍弃特征分量,因此平均贡献R(i)仅表示该分量的相对重要性,而不表示各分量之间的相互依赖关系,对语音识别和说话人识别都是如此。

2 语音数据库

识别数据库为TI46,数据库包含8男8女,样本内容为0~9共10个孤立的英文数字,包括训练集和测试集,训练集为一次录音,每个数字发10遍,测试集8次录音,每个数字每次发2遍,抽样频率12 500 Hz。

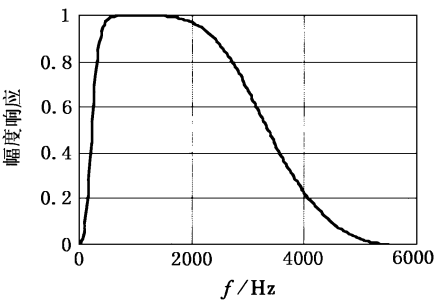

图2 卷积噪声滤波器频率响应

Fig. 2 Frequency response of channel distortion filter

识别时,从训练集中取的10遍录音作为训练样本,以测试集中16遍录音作为测试样本,测试中加入零均值的高斯加性白噪声和卷积噪声。卷积噪声指训练语音和识别语音通过不同的通信信道,如不同的麦克风或电话线路所引入的失真,这里将测试集语音通过4阶Butterworth带通滤波器(滤波器23 dB带宽为100~3 300 Hz,幅度频响如图2所示),近似模拟固定电话信道的频率响应。识别试验在各种噪声不同信噪比SNR进行,SNR定义为干净语音能量与加性噪声能量之比。

由于试验的目的是比较MFCC各分量对识别的

贡献, 而不是刻意追求最高的识别率, 本文采用简单而有效的 DTW 识别器[10]。MFCC 帧长 256 点, 约 20148 ms, 帧移 128 点, 计算 FFT 后由 Mel 滤波器规整。Mel 滤波器中心频率在 100~1 000 Hz间隔 100 Hz, 1 000 Hz 以上中心频率与带宽之比为 11149。

3　MFCC 分量相对重要性

3.1　干净语音

表 1 是干净语音条件下 MFCC 各分量顺序组合的说话人识别, 行方向为起始 MFCC 分量, 列方向为截止 MFCC 分量。由表 1 按式(1) 计算可得到说话人识别各倒谱分量的平均贡献, 如图 3 所示, 每个填充块的高低表示由于识别时包含该维 MFCC 特征系数而增加的平均识别率, 横坐标是倒谱系数序号。C_0 和 C_1 负的平均贡献表示使用包含该分量的 MFCC 特征往往引起识别率降低, 比如利用 MFCC 分量 C_0 到 C_{16}的识别率是 70162%, 利用 MFCC 分量 C_1 到 C_{16}的识别率增加到 8516%, 而利用 MFCC 分量 C_2 到 C_{16}的识别率又增加至 91138%。同时由图可知, 最有用的说话人信息包含在 MFCC 分量 C_2 到 C_{16}间, 其他倒谱系数项包含的有用信息较少, C_6 的平均贡献最大为 912%。其间 C_6 到 C_{13} 的贡献又较其他分量大, 其平均的平均贡献为 7132%。在大多数说话人识别中, MFCC 分量舍弃 C_0 是合理的, 但 C_1 往往被保留作为有用特征则是不应该的[11]。

表 1　干净语音条件下, MFCC 相邻分量顺序组合的说话人识别率

Table 1　The average speaker recognition ratio for clean speech

	C_0	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	C_{16}	C_{17}	C_{18}	C_{19}
C_0	19.61	27.66	38.89	47.81	53.43	58.00	62.01	63.63	65.87	66.65	67.24	67.20	68.15	69.21	70.07	70.30	70.62	70.62	70.34	70.70
C_1		28.58	46.17	60.63	67.74	72.22	75.89	77.90	79.94	81.27	82.53	83.40	84.50	85.01	85.17	85.28	85.60	85.80	86.07	86.34
C_2			39.35	64.22	71.89	78.37	82.03	85.09	86.35	87.96	89.30	90.00	90.28	90.79	91.26	91.14	91.38	91.42	91.34	91.66
C_3				45.18	61.11	71.36	79.74	82.72	85.59	87.40	89.22	89.96	90.79	91.46	91.38	91.78	91.81	90.91	91.06	91.38
C_4					36.37	57.04	73.80	80.48	83.04	85.87	87.13	88.86	89.49	89.60	90.11	90.35	90.75	90.58	90.51	90.07
C_5						41.63	69.04	78.25	82.54	86.54	87.72	89.10	90.24	91.34	91.50	91.54	91.93	91.50	91.73	90.91
C_6							49.66	70.73	78.09	83.80	86.62	87.88	89.30	90.20	90.72	90.40	90.52	90.36	90.40	90.28
C_7								43.96	65.99	77.26	80.77	83.79	85.95	88.00	88.75	88.98	89.10	88.98	88.83	88.55
C_8									45.99	65.14	74.82	80.64	82.57	85.12	86.46	86.74	87.25	87.29	86.98	86.11
C_9										43.67	62.74	73.29	78.16	80.64	83.08	84.62	85.10	84.73	84.93	84.10
C_{10}											43.07	65.41	72.14	76.04	80.91	82.02	82.93	83.04	83.52	82.92
C_{11}												46.13	63.37	71.68	75.68	78.76	79.50	80.33	79.86	80.28
C_{12}													41.03	60.97	68.89	72.56	76.45	76.84	77.82	77.66
C_{13}														44.44	58.89	66.33	70.94	73.14	74.95	75.57
C_{14}															40.41	54.85	66.57	68.15	70.31	71.17
C_{15}																38.21	57.87	62.99	67.63	68.65
C_{16}																	41.43	52.48	59.96	64.04
C_{17}																		34.59	49.42	57.31
C_{18}																			33.44	48.78
C_{19}																				34.76

同样, 由 MFCC 各分量顺序组合的语音识别可以得到各分量对语音识别的平均贡献, 如图 4 所示, 最有用的语音信息包含在 MFCC 分量 C_1 到 C_{12}之间, 其他倒谱系数包含的有用信息较少, C_6 的平均贡献最大为 7123%。其间又以 C_3 到 C_9 的平均贡献又较其他分量大, 其平均贡献为 6163%。

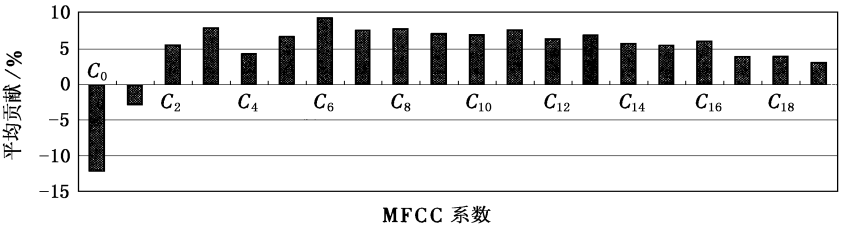

图3 干净语音条件下,说话人识别MFCC各分量的平均贡献

Fig. 3 Improvement of recognition accuracy by including each MFCC component in speaker identification with clean speech

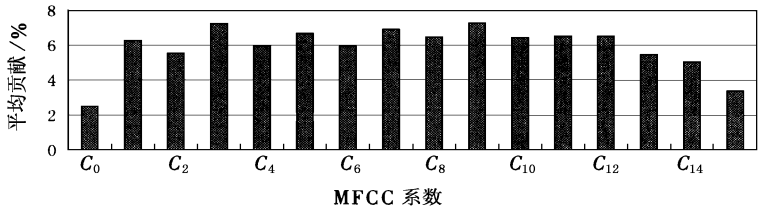

图4 干净语音条件下,语音识别MFCC各分量的平均贡献

Fig. 4 Improvement of recognition accuracy by including each MFCC component in speech recognition with clean speech

3.2 加性白噪声失真语音

图5是加性白噪声MFCC各分量说话人识别的归一化平均贡献(相对重要性),各图纵坐标均按最大值归一化,图5(a)是图3的归一化。受加性白噪声的干扰,低阶MFCC分量的相对重要性下降,高阶MFCC分量的相对重要性不变或略有增加,随着信噪比增加此规律更加明显。如,对于10 dB语音,MFCC分量C_3和C_4的平均贡献由正值变为负值,而C_5以上分量的平均贡献依旧是正值。因此,加性噪声主要影响MFCC低阶项,对高阶项影响则较小。这可能同白噪声的功率谱平坦有关,平坦功率谱的低阶MFCC数值较大,尤其是C_0,而高阶分量的值则较小。加性噪声的功率谱同语音的功率谱叠加,因而数值较大的噪声低阶MFCC分量对语音信号的低阶MFCC分量影响较大,而数值较小的噪声高阶分量对信号的高阶分量的影响较小。

图6是加性白噪声失真语音识别MFCC各分量的相对重要性,图6(a)是图4的归一化。各阶MFCC分量相对重要性随SNR的变化规律同说话人识别类似,低阶MFCC分量的相对贡献下降,高阶MFCC分量的相对贡献则上升。如,对于20 dB语音,MFCC分量C_0和C_1的平均贡献由正值变为负值,而C_2以上分量的平均贡献依旧是正值。

3.3 卷积噪声失真语音

图7是在卷积噪声(信道失真)条件下,语音识别和说话人识别MFCC各分量的归一化平均贡献。奇怪的是增加信道失真,使MFCC偶阶项分量贡献显著下降,而奇阶项分量贡献显著增加。如对于干净语音,MFCC分量C_6对语音识别和说话人识别的平均贡献分别为51.93%和91.27%,增加卷积噪声后分别减低至01.34%和11.42%。比较而言,对低阶MFCC分量的影响较

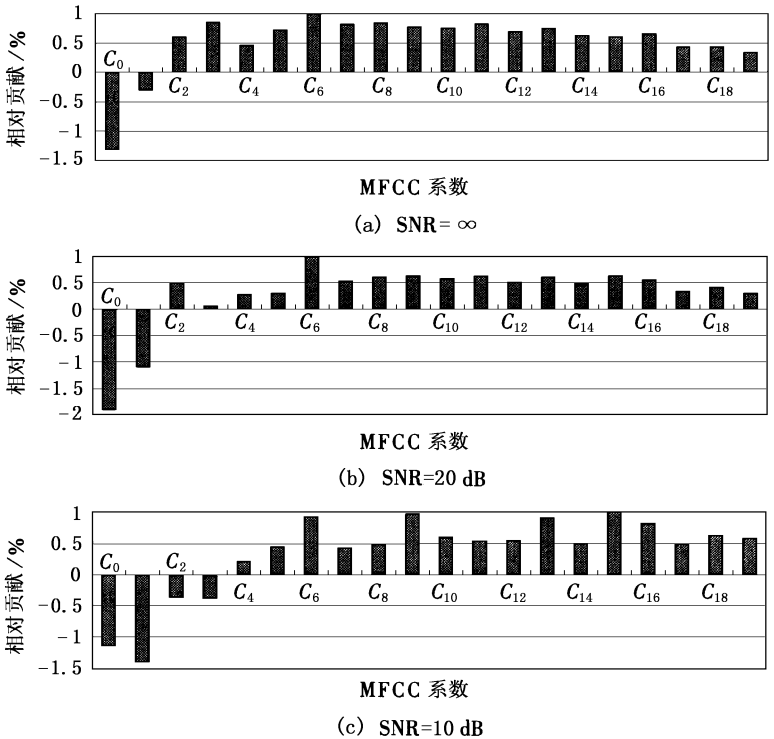

图 5　加性噪声说话人识别 MFCC 各分量的相对重要性

Fig. 5　Relative importance of components of MFCC for speaker recognition using additive noise degraded speech

高阶分量大。这可能同信道失真滤波器的幅度频率响应有关, 所加信道滤波器含有较少的快变化成分。

3.4　加性噪声与卷积噪声失真语音

在加性噪声与卷积噪声失真混合干扰条件下, 语音识别和说话人识别 MFCC 各分量的归一化平均贡献分别如图 8 所示。其影响基本上表现为信道失真和加性噪声影响之和。偶阶项的 MFCC 分量贡献明显下降, 而奇阶项分量贡献明显增加, 其中低阶 MFCC 分量的贡献下降影响较高阶分量显著。

4　讨论与结论

在 DTW 定义的欧氏距离测度下, 最有用的语音信息包含在 MFCC 分量 C_1 到 C_{12} 之间, 而最有用的说话人信息包含在 MFCC 分量 C_2 到 C_{16} 之间, 注意到对于说话人识别有用的 MFCC 起始和截止分量均大于语音识别。较高阶倒谱系数表示语音功率谱快变化的成分, 语音识别需要谱包络的慢变化信息, 而说话人识别则更多从谱的快变化中提取信息。C_0 和 C_1 负的贡献率表示说话人识别需要谱能量 C_0 和谱斜率 C_1 的归一化, 在通常的说话人识别系统着, C_0 被丢弃, 而 C_1 被保留作为有用信息[11]。

在 DTW 定义的欧氏距离测度下, MFCC 各分量对语音识别和说话人识别的相对平均贡献

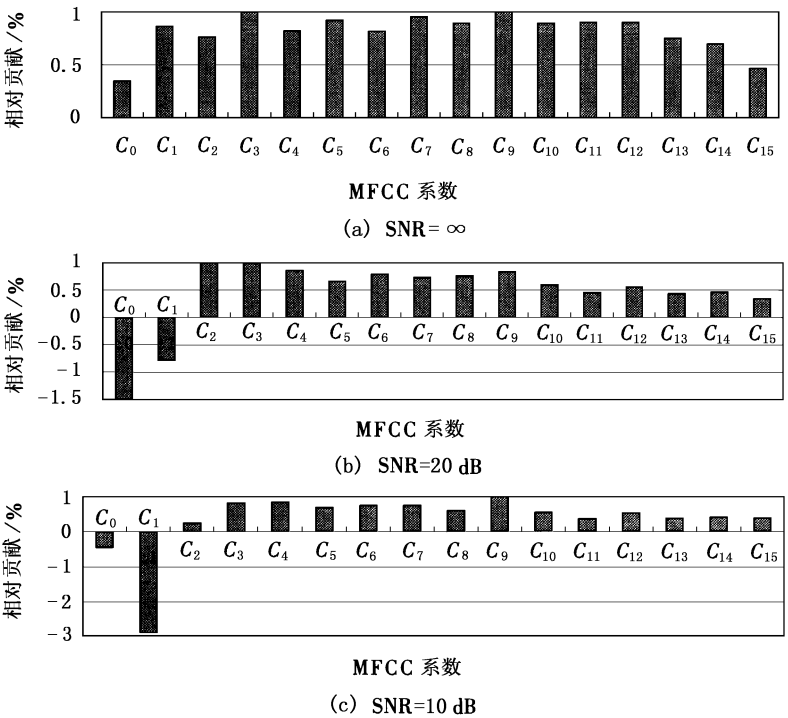

图6 加性噪声语音识别MFCC各分量的相对重要性

Fig. 6 Relative importance of components of MFCC for speech recognition using additive noise degraded speech

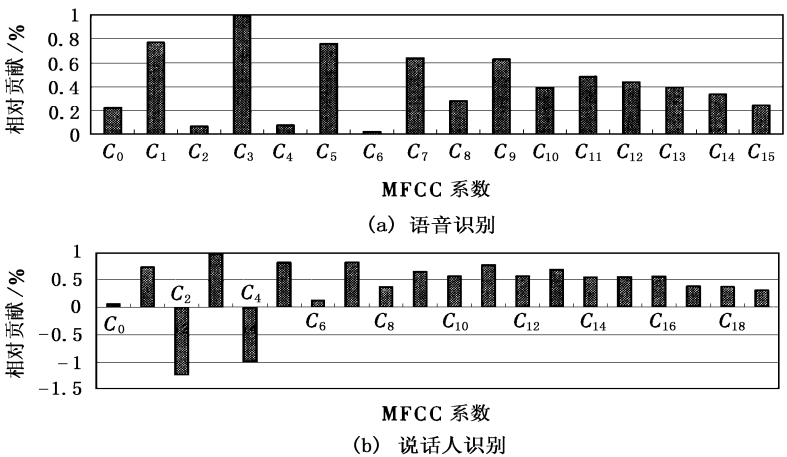

图7 卷积噪声MFCC各分量的相对重要性

Fig. 7 Relative importance of components of MFCC on additive noise degraded speech

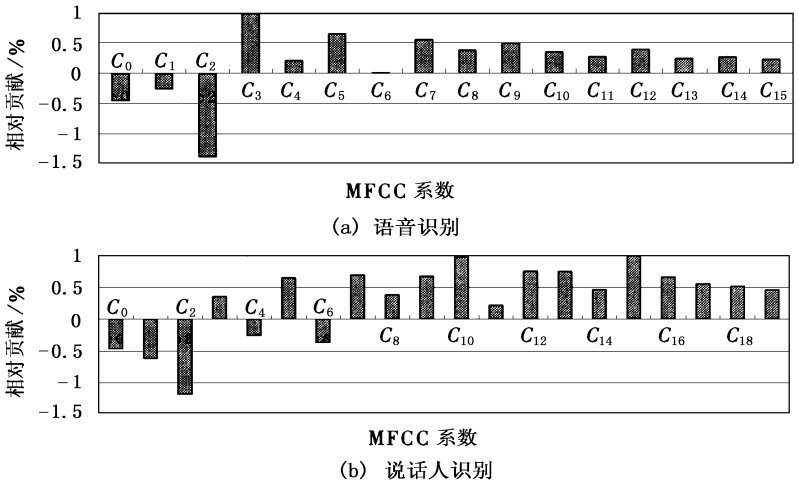

图 8　10 dB 加性噪声语音在信道失真条件下 MFCC 各分量的相对重要性

Fig. 8　Relative importance of components of MFCC on additive and convolution noise degraded speech

在加性噪声和卷积噪声干扰下的变化规律是相同的, 各阶分量的贡献变化可能同噪声的功率谱有关, 比较而言, 低阶 MFCC 分量较高阶 MFCC 分量易受噪声的干扰。卷积噪声引起偶阶项的 MFCC 分量贡献下降, 而奇阶项 MFCC 分量贡献增加, MFCC 奇阶项和偶阶项不同变化的原因有待进一步研究。Juang 在研究加性噪声和信道失真对 LPC 倒谱系数项的影响时也有相同的结论[7]。这也解释了倒谱提升能够改善低 SNR 时特征性能的原因, 中间阶数和高阶 MFCC 分量较低阶 MFCC 分量稳定, 增加其在识别中的权重显然有助于提高特征的抗噪声性。甄斌等提出的特别抑制低阶倒谱项的减半升正弦(MHRS)倒谱提升窗口较其他窗口低 SNR 性能好也说明低阶倒谱项易受加性和卷积噪声的干扰, 其增加高阶倒谱项权重的说话人提升窗口也说明短时谱的快变化中包含说话人信息[7]。

本文讨论的 MFCC 分量的相对重要性是通过对相邻若干分量的组合得到的, 不能推广到不相邻分量组合的情况, 对语音识别和说话人识别都是如此。对后一种情况, 由于不相邻分量的组合涉及天文数字的计算, 不适合本方法讨论, 抽样统计是可能的解决方法之一。

统计学中常用的 F 比方法, 通过计算每个特征分量的类内距离和类间距离的比值说明各特征分量的对识别目标的区分能力[1,2,9]; 本文则通过对特征的若干相邻分量的组合进行识别实验, 得出各特征分量的相对重要性。F 比方法的结果是在特征分量之间相互独立的意义下的得到的, 并且没有考虑到识别系统测度准则的影响; 本文方法考虑到特征的各分量之间具有一定的相互依赖关系, 得出在特定测度下各分量之间的相对重要性, 这一结果对于具体的识别系统更具有直接意义。此外, F 比方法在统计学上具有计算简单方便的优点, 而本文方法则需要较大的计算量。

虽然本文结果是由 MFCC 特征和 DTW 分类器得到, 但结果对语音特征提取和使用是有启发意义的, 即对给定特征, 各特征分量对识别结果的贡献是不同的, 而且最重要的是并非所有特征都是有用的, 某些特征可能会有负作用, 其原因可能同聚类准则和测度准则有关。

参考文献

1 Rabiner L, Juang B H. Fundamental of Speech Recognition. New York: Prentice Hall, 1993

2 杨行峻, 迟惠生. 数字语音信号处理. 北京: 电子工业出版社, 1995

3 Hermansky H. Perceptual Linear Predictive(PLP) Analysis for Speech. J Acoust Soc Am, 1990, 87: 1 738~ 1 752

4 Furui S. Speaker Independent Isolated Word Recognition Using Dynamic Feature of Speech Spectrum. IEEE Trans on Acoustics, Speech, Signal Processing, 1986, 34(1): 52~ 59

5 Hermansky H, Morgan N. RASTA Processing of Speech. IEEE Trans Speech and Audio Processing, 1994, 2(4): 578~ 589

6 Seneff S. A Joint Synchrony/mean2rate Model of Auditory Speech Processing. Journal of Phonetics, 1988, 16: 55~ 76

7 Juang B H, Rabiner L, Wilpon J G. On the Use of Bandpass Liftering in Speech Recognition. IEEE Tran on Acoustics, Speech, Signal Processing, 1987, 35(7): 947~ 953

8 Zhen B, Wu X H, Liu Z M, et al. On the Use of Bandpass Liftering in Speaker Recognition. In: Proceedings of ICSLP, 2000, Beijing, Ò: 933~ 936

9 Atal B S. Automatic Recognition of Speakers from Their Voices. Proceeding of IEEE, 1976, 64(4): 460~ 475

10 Kanedera N, Arai T, Hermansky H, et al. On the Importance of Various Modulation Frequencies for Speech Recognition. In: Proceedings of EUROSPEECH, 1997, Rodos, Greece

11 Reynolds D A. Experimental Evaluation of Features for Robust Speaker Identification. IEEE Trans on Speech and Audio Processing. 1994, 2(4): 639~ 643

On the Importance of Components of the MFCC in Speech and Speaker Recognition

ZHEN Bin WU Xihong LIU Zhimin CHI Huisheng

(Center for Information Science, Peking University, Beijing, 100871)

Abstract The analysis of the relative importance of components of MFCC for both speech recognition and speaker recognition using DTW recognizer in various noise environments are given. For English digit and under the Euclidean distance definition, the experiment results show cepstral components from C_2 to C_{16} contain the most useful speaker information, while C_0 and C_1 are usually harm to speaker recognition. Cepstral terms from C_1 to C_{12} are found to contain the most useful speech information. In both tasks, the additive noise decreases the relative importance of low MFCC terms faster than that of the middle and high MFCC terms, and the decrement depends on the speech SNR. The channel distortion will deteriorate low terms more than the middle and high MFCC terms in both tasks, also.

Key words MFCC; speech recognition; speaker recognition