文章编号:1001 - 9081(2005)06 - 1342 - 03

基于小波调制尺度的语音特征参数提取方法

马 昕,杜利民
(中国科学院 声学研究所, 北京 100080)
(max@ iis. ac. cn)

摘 要:在时频分析的理论基础上,提出了一种基于小波调制尺度特征的参数提取方法。根据人对调制谱信息的感知特性及干扰在调制谱中的特点,采用小波分析技术及归一化处理求得归一化的小波调制尺度特征参数,并以此作为语音的动态特征应用于语音识别系统。通过与 MFCC 一阶、二阶系数对比的汉语音节识别实验表明,该方法在抗噪声干扰和说话速率变化等方面比 MFCC 的一阶、二阶系数的性能优越,为提高语音识别鲁棒性提供了一种新途径。

关键词:语音识别; 小波调制尺度; 语音特征
中图分类号: TP18 **文献标识码**:A

Speech features extraction based on wavelet modulation scale

MA Xin, DI Li-min
(*Institute of Acoustics, Chinese Academy of Sciences, Beijing* 100080, *China*)

Abstract: Based on time-frequency analysis, the theory of estimating a modulation scale representation was discussed, and a new method of features extraction for speech recognition was proposed. Considering specialty of human auditory perception and disturbances, wavelet analysis was used instead of Fourier analysis for modulation frequency transform, and wavelet modulation scales was acquired as speech features for recognition. For further attenuating the effects of disturbances, subband normalization was introduced with the wavelet modulation scales. Experiments for the Chinese syllables recognition show extracting the wavelet modulation scales as the dynamic features outperform the frequency differences both in noise environments and in time misalignment cases.

Key words: speech recognition; wavelet modulation scale; speech features

0 引言

提取语音特征参数是自动语音识别技术中的重要步骤。语音的最小单位是音素,一个音素(phoneme)的发音不仅与其语音特征有关,还受前后音素的发音影响[1],这就要求短时谱特征除了要反映该时间段的静态特征外,还应反映一定时间段的动态特征。当前的识别算法通常用对频谱作差分的方法来弥补其动态特性的不足,对提高识别率及抗噪性能起到一定作用[2],但差分系数尚不能很好地反映频谱的变化规律。此外,采用统计方法涵盖可能出现的语音特征也是解决此问题的方法之一,但需要建立庞大的语音模型库,并且在识别时需要较长的搜索运算过程。

在语音识别系统中,面临的另一问题是识别环境与训练环境不匹配,从而使得提取的语音特征常常含有不利于识别目的的干扰因素。常见的因素有噪声干扰和信道畸变干扰以及话音速率变化等。如果这些干扰不能在语音特征中得到有效的抑制,将会降低语音识别的准确率。在识别系统前端进行语音增强是有效抑制干扰的方法之一[3],但却往往需要预先获得现场环境的有关信息,如噪声特性、信噪比等,但这在实际应用中往往是难以解决的。

因此,需要设计一种既能较好地反映语音信号动态信息,又具有较高抗干扰能力的语音特征参数。调制谱是反映语音谱分量随时间变化规律的特征量,能很好地反映语音信号的动态特征。在调制谱中,加性噪声和通道畸变产生的干扰成分易于与语音信号的有用成分分离并易于抑制[4]。根据当前语音心理学的研究表明,人耳对于语音调制频率的感知整个频带上具有恒 Q 特性[5]。为有效模拟这种特性,本文采用小波变换提取被污染语音信号的小波调制尺度特征,通过适当的归一化处理以降低噪声、通道畸变及说话速率变化对语音识别的不利影响,并用这种小波调制尺度特征单独作为识别特征参数,将其与 MFCC 参数联合作为识别参数分别进行了语音识别实验。

1 调制谱与小波调制尺度

1.1 调制谱原理简介

语音信号的调制谱特征是基于语谱图的特征表示。信号 $x(t)$ 的语谱图 S_x 定义为短时傅立叶变换的幅值,即:

$$|S_x^{(\gamma)}(t,\omega) = STFT_x^{(\gamma)}(t,\omega)|^2 \tag{1}$$

式中,上标 γ 表示以 $\gamma(t)$ 作为分析窗。语谱图满足二次迭加原理,即当 $x(t) = c_1x_1(t) + c_2x_2(t)$ 时:

$$S_x(t,\omega) = |c_1|^2 S_{x1}(t,\omega) + |c_2|^2 S_{x2}(t,\omega) + c_1c_2S_{x1,x2}(t,\omega) + c_1c_2S_{x2,x1}(t,\omega) \tag{2}$$

由(2)式可以看出,语谱图有较明显的相干项,其时频集聚(或分辨率)也比较差[6]。

语音信号的调制谱 $M(\omega,\eta)$ 可由谱图经(3)式变换得

收稿日期:2004 - 11 - 19;**修订日期**:2005 - 01 - 28 **基金项目**:国家 973 规划项目(G1998030505)

作者简介:马昕(1969 -) 男,山东泰安人,博士研究生,主要研究方向:语音信号处理、DSP 应用; 杜利民(1957 -)男,四川人,研究员,博士生导师,主要研究方向:信号处理与语音交互技术.

到,具体变换过程如图 1 所示。

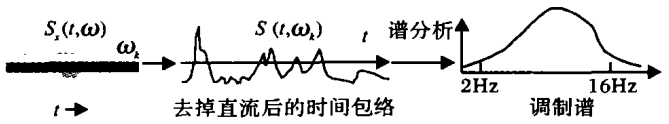


图 1　由谱图变换为调制谱的过程

$$M_x(\omega,\eta) = \int_{-\infty}^{+\infty} S_x(t,\omega)e^{-j\eta t}\mathrm{d}t \tag{3}$$

式中 ω 和 η 分别为语音频率和调制频率;$M_x(\omega,\eta)$ 既也可看作是语音信号 $x(t)$ 的自相关函数的二维变换,也可看作 $X(\omega)$ 的自相关函数[7]。然而,$M_x(\omega,\eta)$ 也会出现相干项。对 $M_x(\omega,\eta)$ 进行平滑处理可有效地消除部分相干项[8],以 $M_x^{sp}(\omega,\eta)$ 表示平滑后的结果,则有:

$$M_x^{sp}(\omega,\eta) = M_W(\eta,\omega) *_{\omega} M_x(\eta,\omega) \tag{4}$$

式中,$M_W(\eta,\omega)$ 是所加时间窗函数 $W(t)$ 的调制谱表示。由上式看出,$M_x^{sp}(\omega,\eta)$ 是 $M_x(\eta,\omega)$ 与 $M_W(\eta,\omega)$ 在 ω 上卷积的结果,根据文献[8],平滑后的调制谱特征对相干项有明显抑制作用,这是将调制谱分析方法用于抗干扰技术的理论依据。

1.2　小波调制尺度特征表示

语音心理学的研究表明[5],对数尺度在整个频带内具有恒 Q 特性,该特性可较好地模仿人对调制频率的感知特性[12]。小波变换是对信号进行时间—尺度分析的一种方法,通过对基本小波的伸缩和平移可以对被分析信号作多尺度分析,其作用也相当于用一组恒 Q 的带通滤波器对信号作多分辨率分析。这也是作者采用小波变换来模拟这种感知特性的主要依据,其具体计算步骤如下:

首先计算语音信号 $x(t)$ 的谱图,即:

$$S_x(t,\omega) = \frac{1}{2\pi}\left|\int x(u)w^*(u-t)e^{-j\omega u}\right|^2 \tag{5}$$

式中,$w^*(t)$ 为短时窗函数。

对于连续尺度参数 s,小波函数 $\psi(t)$,沿 $S_x(t,\omega)$ 的时间方向作小波变换,即:

$$M_x(s,\zeta,\omega)\ \frac{1}{s}\int S_x(t,\omega)\psi\left(\frac{1-\zeta}{s}\right)\mathrm{d}t \tag{6}$$

式中 ζ 是连续时间平移参数,$M_x(s,\zeta,\omega)$ 即为 $x(t)$ 的小波调制尺度特征。

1.3　对干扰的抑制

设含噪语音信号 $y(t) = [x(t) + d(t)]*h(t)$,其中 $x(t)$ 为纯净语音信号,$d(t)$ 为加性噪声。为便于表达,令 $s(t) = x(t) + d(t)$,$h(t)$ 为产生卷积干扰的通道冲激相应。$y(t)$ 的谱图可表示为:

$$S_y(t,\omega) = S_s(t,\omega)S_h(t,\omega) \tag{7}$$

式中,$S_s(t,\omega)$ 与 $S_h(t,\omega)$ 分别为 $s(t)$ 与 $h(t)$ 的谱图表示。对 $S_y(t,\omega)$ 加窗,并沿时间方向对谱图作小波变换,即可得到 $y(t)$ 的小波调制尺度表示,即:

$$M_y(s,\zeta,\omega)\ \frac{1}{s}\int S_y(t,\omega)W_L(t-B)\psi\left(\frac{1-\zeta}{s}\right)\mathrm{d}t \tag{8}$$

式中 $W_L(t)$ 为平滑窗函数。在此加窗的意义不仅可避免分帧时的截短效应,同时也可减弱由于加性干扰在调制信息中产生的相干项[7]。假设在平滑窗内,产生卷积干扰的通道特性近似为线性时不变的,则 $y(t)$ 小波调制尺度可表示为:

$$M_y(s,\zeta,\omega) \approx M_s(s,\zeta,\omega)M_h(\omega) \tag{9}$$

$M_s(s,\zeta,\omega)$ 为仅含加性噪声的信号 $s(t)$ 的小波调制尺度表示,$M_h(\omega)$ 为卷积干扰的小波调制尺度表示。对 $M_y(s,\zeta,\omega)$ 归一化可得:

$$M_{y,norm}(s,\zeta,\omega) = \frac{M_y(s,\zeta,\omega)}{\int M_y(s,\zeta,\omega)\mathrm{d}s} = \frac{M_s(s,\zeta,\omega)M_h(\omega)}{\int M_s(s,\zeta,\omega)M_h(\omega)\mathrm{d}s} = \frac{M_s(s,\zeta,\omega)}{\int M_s(s,\zeta,\omega)\mathrm{d}s} = M_{s,norm}(s,\zeta,\omega) \tag{10}$$

此外,在语音系统中,另一种信道干扰是时间尺度畸变干扰,这种干扰可能源于说话人发音速度的变化,也可能由声音通道延迟造成。假设尺度变化由时间 t 变为 at,a 为时间尺度因子,速率变化前的谱图为 $S_x(t,\omega)$,则变化后的谱图可表示为:

$$S_y(t,\omega) = S_x(at,\omega) \tag{11}$$

因子 a 对小波调制尺度的影响相当于调制尺度参数线性变化为 as,调制尺度幅度则变为原来的 a^{-1},即:

$$M_y(s,\zeta,\omega) = a^{-1}M_x(as,\zeta,\omega) \tag{12}$$

当 a 接近于 1 时,可作如下近似:

$$M_x(as,\zeta,\omega) \approx M_x(s,\zeta,\omega) \tag{13}$$

即:

$$M_y(s,\zeta,\omega) \approx a^{-1}M_x(s,\zeta,\omega) \tag{14}$$

由(14)式可以看出,这时的小波调制尺度调制值仅与 a 有关,按式(10)作归一化处理,则 a 的影响可被削弱。

在实际应用公式(8)计算小波调制尺度时,要将尺度参数 s 离散化为 s_d,时间平移参数 ζ 离散化为 ζ_n,则离散化的归一小波调制尺度为:

$$M_{y,norm}(s_d,\zeta_n,\omega) = \frac{M_s(s_d,\zeta_n,\omega)}{\sum_{s_d} M_s(s_d,\zeta_n,\omega)} \tag{15}$$

对调制谱的研究表明,干扰与语音信号的主要成分在调制谱中分布在不同的区域[3]。文献[9]针对日语音节做了调制谱的研究,结果表明调制谱成分中最能反映语音特性的信息在2Hz ~ 16Hz的范围,因此,通过合理选择 s_d 的范围,亦可达到对加性噪声的抑制。(15)式可进一步近似为:

$$M_{y,norm}(s_d,\zeta_n,\omega) = M_{x,norm}(s_d,\zeta_n,\omega) \tag{16}$$

式中,$M_{x,norm}(s_d,\zeta_n,\omega)$ 为纯净语音信号 $x(t)$ 的归一化小波调制尺度特征表示。通过上述处理,加性干扰、卷积干扰以及语音速率变化在归一化小波调制尺度上的影响被明显削弱。

2　模拟实验与结果分析

实验采用非特定人汉语音节识别的系统。实验采用的语音数据是 863 汉语音节库,该音节库涵盖了汉语中可能出现的所有音节。无噪语音信号 $s(k)$ 由 10 个男生、10 个女生在比较纯净的环境下录制而成。将这些无噪信号提取的特征参数序列存入模板库作为各人语音的参考模板。数据为 16kHz 采样,16bit 量化。加性噪声使用 NoiseX92 数据包中的白噪声(White)。按下式将无噪语音和噪声进行叠加,形成不同信噪比的实验测试信号:

$$SNR = 10\log\left(\sum_k |s(k)|^2 / \sum_k |n(k)|^2\right) \tag{17}$$

信道畸变采用 Cooledit2.0 中的 Dynamic EQ 调整,发音速度的变化由 cooledit2.0 中的 time/stretch 调整。语音信号按每帧 25ms(400 点)分帧并加汉明窗,帧移选择 125 点,这样相当于调制谱计算的采样率为 128Hz。由于一般意义的语谱图得到的是均匀的频率子带,而 Mel 子带更能反映人耳的听觉特性,所以在按公式(5)得到谱图表示 $S_x(t,\omega)$ 后,进一步转换得到 Mel 刻度下的功率谱表示 $S_x(t,\omega_k)$。在此,将频带划分为 26 个 Mel 子带,即 k 等于 26,将每一个子带沿时间方向

分帧并加汉明窗。为了获得足够的分辨率,帧应有足够的长度,选择 1s(128 点)作为帧长,这样每一长帧包含 128 个短帧子带能量值 $E_n(0 \leqslant n \leqslant 128)$。所选小波函数应具有较好的紧支性,本文实验采用 bior5.5 双正交小波函数对每一长帧作 8 尺度二进小波变换,小波平移参数 ζ_n 的步长选为 1。由于 bior5.5 小波滤波器的长度是 9 和 11,所以在帧与帧间重叠 11 点。变换后,将前 11 点对应的变换系数去除,类似于重叠保留法滤波[11]。计算得到的调制尺度参数按前文所述作归一化及滤波处理。根据小波分析理论[10]及关于人对语音调制谱的感知理论[1],第 3,4,5 层的小波系数最能反映调制信息中的有用成分。因此,本文仅保留这三层小波系数作为调制尺度参数,这样每一长帧得到 3×117 的小波系数矩阵,该矩阵的每一列都是相同平移因子 $\zeta_n(0 \leqslant \zeta_n < 117)$ 的调制尺度系数,组合在一起作为此长帧中短时帧号为 ζ_n 的语音特征参数。

各种特征在不同的噪声、不同的语音速率下的识别率分别如表 1、表 2 所示。

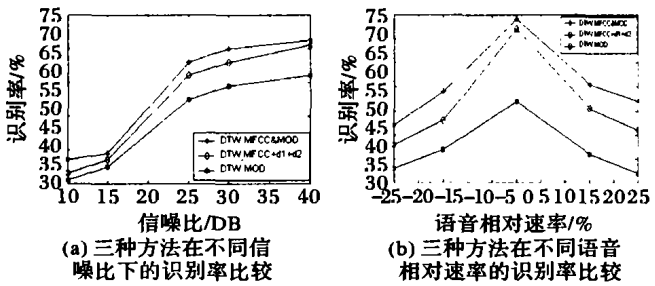
表 1 汉语音节在不同的信噪比、卷积畸变下的识别率(%)

噪声类型	白噪声 10dB	白噪声 15dB	白噪声 25dB	白噪声 30dB	白噪声 40dB	卷积噪声
DTW_MFCC + d1 + d2	33	37	62	66	71	65
DTW_MFCC&MOD	37	39	66	70	72.5	68
DTW_MOD	31	35	55	59	62	57

表 2 汉语音节在不同语音速率下的识别率(%)

语速变化率	-25%	-10%	不变	10%	25%
DTW_ MFCC + d1 + d2	48	53	72	55.5	51
DTW_ MFCC&MOD	52	59	74	60.5	57
DTW_MOD	43	47	57	46	42

表 1 给出了各种参数在不同的信噪比及卷积畸变的识别率。其中,DTW_MFCC + d1 + d2 :表示用传统的 MFCC 参数(含一阶、二阶差分)作为语音特征参数进行 DTW 运算;DTW_MFCC&MOD:表示将每帧语音的 MFCC 参数与按前述方法得到的小波调制尺度参数进行组合作为每帧语音的识别参数进行 DTW 运算;DTW_MOD:表示单独使用上述小波调制尺度参数进行 DTW 运算。



(a) 三种方法在不同信噪比下的识别率比较　(b) 三种方法在不同语音相对速率的识别率比较

图 2 三种方法的识别率比较

上述的三种方法在不同信噪比和不同语音相对速率下的识别率分别如图 2(a)、(b)所示。可以看出,采用了小波调制尺度结合 MFCC 参数的识别方式对在噪声环境中的识别性能有明显改善,同时对因语音速率变化而对识别造成的干扰也有明显的抑制。

3 结语

调制谱是语音信号动态特征的另一种有效表示方法。本文基于时频分布理论和语音感知的心理实验结论,提出了以小波调制尺度特征作为动态特征参数的语音参数提取新方法,即在语音感知敏感的调制谱区域,提取语音的动态特征。通过进行归一化处理,进一步提高了特征的抗干扰能力。初步的汉语音节实验证实这种新的特征参数对于提高语音识别鲁棒性是有效的,进一步的研究将探讨这种方法对于连续语音识别鲁棒性的贡献方式和程度。

参考文献:

[1] HERMANSKY H. Human Speech Perception: Some Lesson From Automatic Speech Recognition[A]. in Proceedings of TSD'01, Zelezna Ruda, Czech Republic[C], September 2001.

[2] RABINER LR, JUANG BH. Fundementals of Speech Recognition [M]. Prentice Hall, Englewood Cliffs, NJ, USA, 1993. 194 - 200.

[3] BOLL S. Suppression. of acoustic noise in speech using spectral subtraction[J]. IEEE and Signal Processing, April 19791: 13 - 120.

[4] HERMANSKY H. The Modulation Spectrum in Automatic Recognition of Speech[J]. IEEE Workshop on Automatic Speech Recognition and Understanding, 1997: 140 - 147.

[5] KANDEL ER, SCHWARTZ JH, JESSELL TM, *et al.* Principles of Neural Science[M]. Third Edition, Chapter32 Hearing, Elsevier Science Publishing Co, Inc, 1991. 481 - 498.

[6] 张贤达. 现代信号处理[M]. 北京:清华大学出版社, 1995.

[7] SUKITTANON S, ATLAS LE. Channel Compensation of Modulation Spectral Features[A]. in Proceedings of the 2003 IEEE ISCAS[C], 2003.

[8] SUKITTANON S, ATLAS LE. Modulation Frequency Features for Audio Fingerprinting[A]. Proc of ICASSP'2002[C], 2002. 1173 - 76.

[9] ARAI T, PAVEL M, HERMANSKY H, *et al*. Intelligibility of speech with filtered time trajectories of spectral envelopes[A]. Proc ICSLP-96[C], Philadelphia, October 1996. 2490 - 2493.

[10] MALLAT SG. A Wavelet Tour of Signal Processing[M]. Academic Press, Second Edition, 1999.

[11] OPPENHEIM AV, Schafer R W. Digital signal Processing[M]. Prentice Hall, Inc, 1975. 85 - 86.

[12] EWERT S, DAU T. Characterizing frequency selectivity for envelope fluctuations [J]. JASA, 2000, 108: 1181 - 96.

(上接第 1341 页)

[3] SUN GZ, GILES CL, CHEN HH, *et al.* The neural network pushdown automata: model, stack and learning simulations[R]. Report UMIACS-TR-93-77 and CS-TR-3118, University of Maryland, 1993.

[4] 孙增圻. 智能控制理论与技术[M]. 北京: 清华大学出版社, 1992. 133 - 140.

[5] HAMMERTON JA. A *hybrid connectionist shift reduce parser, Final* year project report[R]. Department of artificial intelligence and computer science, University of Edinburgh. 1994.

[6] TOMITA M. Efficient Parsing for Natural Language[M]. Boston: Kluwer academic Publishers, 1986.

[7] SAMPSON G. English for the Computer [M]. Oxford University Press, 1995.