基于多时间尺度复合子带噪声语音识别新方法

蒋文建,韦 岗
(华南理工大学无线电与自动控制研究所,广东广州 510640)

摘 要: 本文根据多时间尺度分析与子带方法,提出了一种多时间尺度复合子带的噪声环境下语音识别新方法.新方法在不同的时间尺度下分别进行子带特征提取和全带特征提取,并分别进行识别,然后在识别概率层相结合得到最终识别结果.本方法兼有多时间尺度方法和子带方法的抗噪性能.此外,进一步引入频谱差分方法提高语音特征的抗噪性能.对 E-SET 在 NoiseX92 下白噪声的识别实验表明,新方法具有良好的抗噪性能.

关键词: 语音识别; 噪声; 子带; 多时间尺度
中图分类号: TN912 文献标识码: A 文章编号: 0372-2112 (2001) 12A-1829-04

A New Method for Noisy Speech Recognition Based on Multiple Timescale and Hybrid Sub-Band

JIANG Wen-jian, WEI Gang
(*Institute of Electronic Engineering and control, South China University of Technology, Guangzhou, Guangdong* 510640, *China*)

Abstract: We present a new method for noisy speech recognition, which combines multiple timescale analysis with sub-band method. In the new method, sub-band features and full-band features are extracted in different timescales, then combined at probabilistic layer. The method has robustness of both multiple timescale method and sub-band method. Differential method is also introduced to further enhance the robustness of the features. Evaluated by a task on E-SET database under white noise from NoiseX92 noise database, the new method has higher recognition accuracy than conventional multiple timescale method on both noisy speech and clean speech.

Key words: speech recognition; noise; sub-band; multiple timescale

1 引言

随着隐马尔可夫模型在语音识别中的广泛应用,语音识别技术得到很大发展.但是在大多数实际应用环境中总是存在各种各样的噪声,由于训练环境与识别环境的不匹配,现有语音识别系统即使在低噪声情况下,系统的识别率也下降得十分严重[1].因此要使语音识别技术真正达到实用,噪声环境下语音识别的研究有着十分重要的意义.噪声语音识别是当前语音识别研究领域中非常活跃的研究课题之一[2].

通过对人耳识别语音的实验研究表明:人耳识别语音时,首先将语音信号分为不同的频率子带,各子带独立识别,然后再将这些识别结果结合成更上一级的音节、字、词等[3].基于该声学结论的子带识别方法近年来得到广泛的研究[4,5].子带识别方法分为特征结合方法和概率结合方法.其中特征结合方法是指在特征层上将不同子带的特征结合为一个统一的特征矢量用于训练语音模型并进行识别[6].概率结合方法指不同子带的特征分别训练语音模型,分别识别,在概率层上将不同子带的识别概率结合,得到最终的识别结果[7].子带方法具有良好的抗噪性能.结合不同的语音数据流,使得不同的语音数据流提供不同的语音信息,可以提高语音识别性能[8].例如人在识别语音时,通过视觉系统输入视觉信息进入大脑参与语音识别.在嘈杂的公共场所,人们可以通过判断对方嘴唇的变化来辅助识别语音,在语音识别中专门有唇语识别研究(Lip-Reading).文献[9]分别提取一倍帧长、三倍帧长和五倍帧长,再进行特征结合用于噪声下语音识别,结果表明多时间尺度方法可以提高语音识别抗噪性能.

本文将多时间尺度方法和子带识别方法相结合,提出了一种新的多时间尺度复合子带的噪声语音识别方法,以更好地利用子带方法和多时间尺度方法各自的抗噪特性.全文的结构如下:第二部分讨论子带方法,第三部分讨论多时间尺度差分复合子带方法,第四部分为实验和结果分析,最后是全文的总结.

2 子带特征

Mel 频率倒谱系数(MFCC)在现有语音识别系统中得到广泛的应用.如图 1 所示:首先对语音信号进行 FFT 变换到频域,通过 Mel 尺度的滤波器阵列后,将滤波器能量输出进行离

收稿日期: 2001-05-21; 修回日期: 2001-09-15
基金项目: 国家自然科学基金(No. 60072048)

散余弦变换(DCT). 即:

$$X = AE \tag{1}$$

其中: A 表示离散余弦变换, E 表示通过三角滤波器阵列后的各滤波器输出频谱能量和.

X 表示MFCC特征.

子带方法(Subband), 先将语音信号经过 FFT 变换到频域, 通过 Mel 尺度的滤波器阵列后, 将滤波器能量输出按照频率分为多个子带, 分别对每个子带进行离散余弦变换. 即:

$$[X_1, X_2] = [(A_1E_1), (A_2E_2)] \tag{2}$$

其中: A_j 表示离散余弦变换, E_j 表示三角滤波器阵列的各滤波器输出频谱能量和. X_j 表示子带 MFCC 特征.

子带特征结合如图 2 所示: 将各子带特征级联起来, 即将 X_j 级联起来作为一个统一的语音特征矢量, 用于训练语音模型和识别语音.

子带概率结合如图 2 所示: 各子带特征分别训练语音模型, 分别识别, 得到识别概率, 在概率层上将各子带结合起来:

$$p(X \mid M) = \prod_{j=1}^{J} p(X_j \mid M_j) \tag{3}$$

其中 $p(X_j \mid M_j)$ 表示各子带识别的似然概率. J 为划分的子带数目. 研究表明子带方法有良好的抗噪性能.

我们以往的研究结果表明, 由于噪声频谱的变化比语音频谱的变化小, 对频谱进行差分能有效地减少噪声的影响[10], 因此在进行频域变换后, 我们采用频谱差分方法进行去噪处理, 计算公式如下:

$$\frac{\partial \mid Y(w) \mid}{\partial t} = \frac{\sum_{t=-T}^{T} t \mid Y(w+t) \mid}{2\sum_{t=-T}^{T} t^2} \tag{4}$$

其中: $Y(w)$ 表示语音频谱, t 表示时间, T 表示参与差分的帧数.

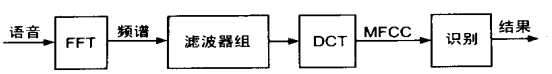

图 1 标准 MFCC 特征示意图

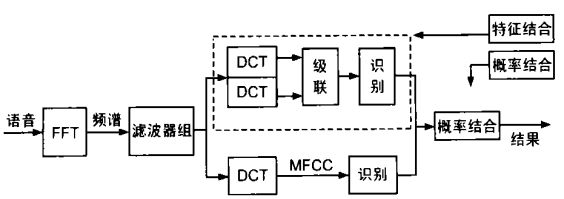

图 2 复合子带方法示意图

3 多时间尺度差分复合子带方法

在语音识别中, 通常使用单一的时间尺度, 即使用固定的帧长和固定的帧移. 例如通常使用 20ms 帧长和 10ms 帧移. 由于不同时间尺度的语音特征能描述不同的语音特性, 可以对语音提取不同时间尺度的语音特征. 对于噪声环境下的语音识别, 由于噪声频谱特性和语音频谱特性的区别, 长时间尺度的特征能较好地描述噪声特性, 而短时间尺度的特征能较好地描述语音的特性, 将两个时间尺度特征结合起来可以更好地区分噪声和语音特性, 从而提高语音识别系统的抗噪性能.

语音分析表明语音的有效信息可能存在于局部语音特性中[3], 子带特征能较好的描述语音信号的局部特性, 提高语音识别性能. 本文对语音进行多时间尺度分析, 提取不同帧长的语音特征. 即分别提取 N_1 点帧长和 N_2 点帧长的语音特征. 并且将不同时间尺度的语音全带特征、子带特征相结合. 使得语音识别系统能描述不同时间尺度的语音特性. 为了消减噪声的影响, 进一步引入频谱差分方法. 下面分别介绍各算法.

传统多时间尺度全带算法:

(1)对语音进行多时间尺度分帧处理, 分别取 N_1 点帧长和 N_2 点帧长. 对每帧语音加汉明窗.

(2)计算每帧语音的 DFT.

(3)对频域信号提取 MFCC 特征, 作为全带特征.

(4)各时间尺度的全带特征分别训练语音模型.

(5)识别时将多时间尺度全带特征的识别概率叠加, 得到识别结果.

本文提出了两种基于多时间尺度算法.

算法 **1** 多时间尺度子带算法:

(1)对语音进行多时间尺度分帧处理, 分别取 N_1 点帧长和 N_2 点帧长. 对每帧语音加汉明窗.

(2)计算每帧语音的 DFT.

(3)对频域信号按照频段分为两个子带, 各子带分别提取子带 MFCC 特征, 进行子带特征结合.

(4)各时间尺度的子带特征分别训练语音模型.

(5)识别时将各时间尺度子带特征的识别概率按照式(3)叠加进行概率结合, 得到最终识别结果.

算法 **2** 多时间尺度复合子带算法:

在算法 1 的基础上, 我们根据子带特征反映语音局部特性和全带特征反映语音整体特性的这一特点, 将多时间尺度子带特征识别和多时间尺度全带特征识别两者在识别概率层上结合起来. 即: 将各时间尺度的全带特征和子带特征的识别概率按照式(3)叠加进行概率结合, 作为最终识别结果.

此外, 为了更好地提高算法的抗噪性能, 我们使用频谱差分方法减少噪声干扰. 在进行 DFT 变换后按照式(4)对频谱进行差分计算, 减少噪声影响, 得到以上各方法的差分方法.

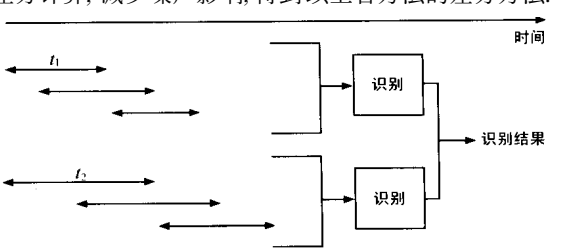

图 3 多时间尺度识别方法

4 实验

本实验对 E-SET 进行非特定人语音识别. E-SET 数据包包括{B、C、D、E、G、P、T、V、Z}九个音, 这九个音的元音部分相同, 只有辅音部分不同, 是易混淆集合. 在语音识别中, E-

SET 是较难识别的集合. 国际文献中采用 MFCC 特征通用方法的纯净语时的 E-SET 识别率为 84.0% 左右[5]. 由于连续语音识别一般采用音素作为识别单位, E-SET 的识别对于从孤立字识别向连续语音识别过渡有重要的意义. 该数据包训练集合包括八个男性录制的语音,其中每个人读所有的九个音,每个数字发音十遍. 测试集合包括另外八个男性录制的语音,每人读所有的九个音,每个数字读十六遍. 数据包的采样率为 12.5kHz,16bits 量化.

噪声使用 NoiseX92 噪声数据包的白噪声(White). 带噪语音用原始纯净语音根据不同的信噪比与白噪声叠加得到:

$$SNR = 10\log_{10}\frac{\sum_k |s(k)|^2}{\sum_k |n(k)|^2} \quad (5)$$

其中 $y(k)$ 表示带噪语音信号, $n(k)$ 表示噪声信号.

语音识别的前端特征提取时, N_1 取为 200, N_2 取为 400. 全部 MFCC 方法提取 12 维的倒谱系数;子带特征如上所述分别计算两个子带的倒谱系数. 所有的倒谱系数再经过差分得到一阶差分系数. 所得 12 维倒谱系数和 12 维倒谱差分系数组成 24 维的语音特征.

在本实验的子带特征中,我们将语音频谱分为两个子带,分别为 150Hz~2000Hz 和 1800Hz~6250Hz,两个子带分别提取 7 维和 5 维倒谱参数.

用隐马尔可夫模型(HMM)为每个子建模,其中每个 HMM 包含 5 个状态,用 3 个高斯拟合每个状态的概率分布.

$$b_j(o) = \sum_{k=1}^{M} c_{jk} N(o, \mu_{jk}, U_{jk}),\ M = 3 \quad (6)$$

式中 c_{jk} 表示各高斯拟合系数, $N(o, \mu_{jk}, U_{jk})$ 表示高斯密度函数.

表 1 给出了各种特征在不同信噪比(SNR)下的实验结果.

表 **1**(*a*)　多时间尺度全带方法、子带方法识别率(%)

	10dB	15dB	20dB	25dB	Clean
全带_单_200	31.42	44.79	58.59	68.49	84.46
全带_单_400	32.73	47.40	61.72	68.14	74.83
全带算法_多	34.20	49.22	63.11	70.40	76.04
子带_单_200	35.24	50.17	62.67	72.31	82.20
子带_单_400	36.63	48.09	62.93	70.31	75.35
算法 1_多	35.85	50.26	63.98	73.61	83.25

表 **1**(*b*)　多时间尺度复合子带方法识别率(%)

	10dB	15dB	20dB	25dB	Clean
复合子带_单_200	32.73	50.43	64.50	73.26	84.90
复合子带_单_400	34.46	51.04	65.97	72.05	76.04
算法 2_多	36.81	53.99	68.49	76.39	84.29

其中:全带_单_200:单时间尺度全带方法,即传统的 MFCC 方法,取帧长为 200 点. 全带_单_400:单时间尺度全带方法,帧长为 400 点. 全带算法_多:传统的多时间尺度全带方法,帧长分别为 200 点和 400 点. 子带_单_200:单时间尺度子带方法,帧长为 200 点. 子带_单_400:单时间尺度子带方法,帧长为 400 点.

算法 **1**_多　多时间尺度子带方法,帧长分别为 200 点和 400 点.

复合子带_单_200:单时间尺度复合子带方法,将 200 点帧长的全带和子带相结合.

复合子带_单_400:单时间尺度复合子带方法,将 200 点帧长的全带和子带相结合.

算法 **2**_多　多时间尺度复合子带方法,将 200 点帧长和 400 点帧长的全带和子带相结合.

表 2 为表 1 对应方法的差分方法.

各差分方法识别率:

表 **2**(*a*)　多时间尺度差分全带方法、差分子带方法识别率(%)

	10dB	15dB	20dB	25dB	Clean
全带_单_200_差分	44.70	57.73	66.49	71.70	81.51
全带_单_400_差分	46.44	61.46	69.10	73.52	75.00
全带_单_多_差分	56.25	67.88	73.18	76.82	82.20
子带_单_200_差分	44.53	58.51	64.76	69.97	82.29
子带_单_400_差分	44.79	60.42	67.71	72.92	76.13
算法 1_多_差分	55.12	65.80	71.44	75.26	82.90

表 **2**(*b*)　多时间尺度差分复合子带方法识别率(%)

	10dB	15dB	20dB	25dB	Clean
复合子带_单_200_差分	46.44	59.81	67.88	73.61	84.11
复合子带_单_400_差分	48.87	63.63	71.09	74.39	77.26
算法 2_多_差分	57.99	68.92	74.57	78.21	84.03

实验结果分析讨论:

(1)从表 1(*a*)、表 2(*a*)可以看出,不同时间尺度方法的识别率不同. 在无噪声情况,帧长 200 点的方法识别率比帧长 400 点的方法高大约 10 个百分点. 但帧长 400 点的方法有更好的抗噪性. 说明通常使用的 200 点帧长能较好的描述语音特性,而在噪声情况下,由于噪声的时间特性变化较慢,相对较长的帧长,400 点帧长更能描述噪声特性和语音特性的区别.

(2)从图 4(*a*)可以看出,传统的多时间尺度的全带方法可以提高噪声下语音识别性能,说明多时间尺度方法具有抗噪性. 但对于无噪声情况下,多时间尺度全带方法使识别率下降.

(3)从表 1(*a*)可以看出,本文提出的多时间尺度的子带方法不仅可以提高噪声下的语音识别性能,对于无噪声情况下,多时间尺度的子带方法能基本保持同样识别性能. 并且多时间尺度子带方法的识别性能比多时间尺度全带方法稍好.

(4)从图 4(*c*)、(*d*)可以看出,多时间尺度差分全带方法和多时间尺度差分子带方法在有噪声能提高系统识别率.

(5)在噪声情况下,差分方法性能比通常方法识别率高,因为经过差分消减了噪声的干扰,但在无噪声情况下,差分方法性能比通常方法识别率低,因为经过差分同样消减了部分语音信息.

(6)多时间尺度复合子带方法有比其他方法更好的识别性能. 说明结合了不同语音特征数据流,可以提高语音识别性能.

(7)从表 2(*b*)可以看出,本文的多时间尺度差分复合子带特征有最好的噪声识别性能. 在无噪声情况下,改进了传统多时间尺度全带方法识别性能下降的缺点.

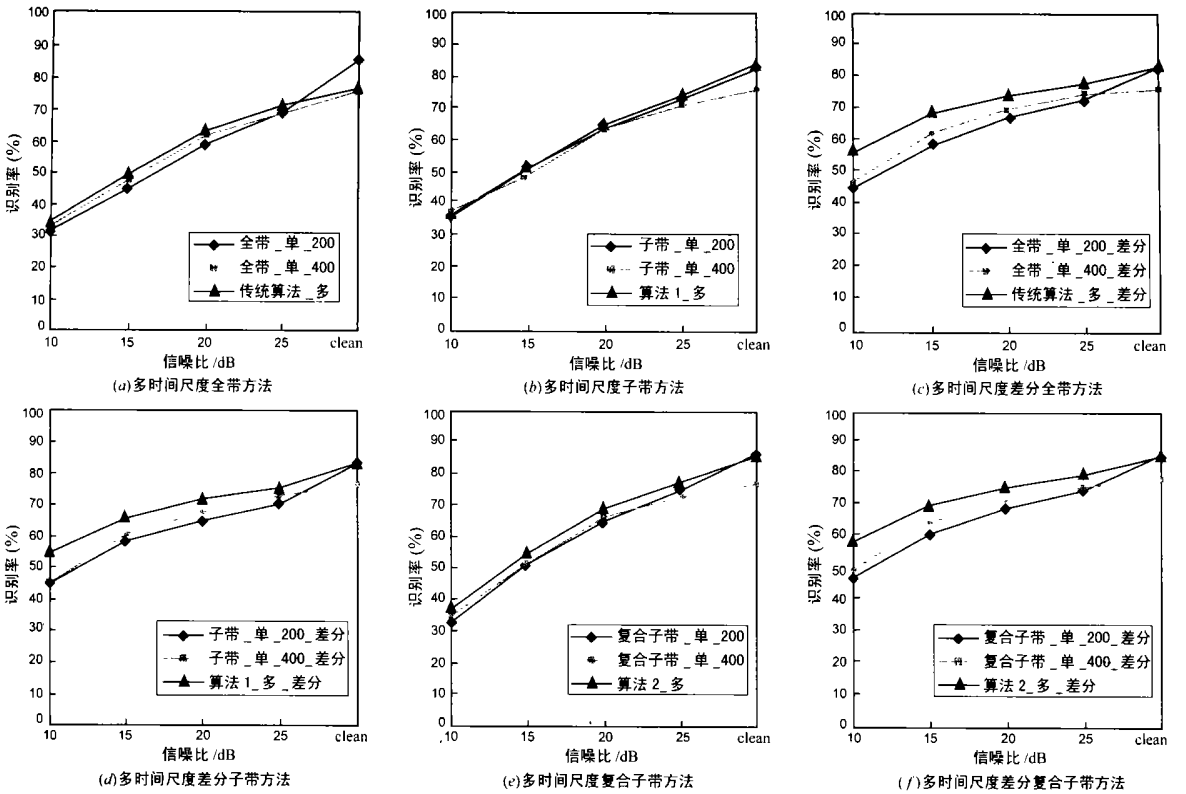

图 4

综上所述, 多时间尺度方法比单时间方法有更好的抗噪性能. 本文提出的多时间尺度子带方法进一步反映了人耳听觉特性, 描述了语音的局部特性, 比传统的多时间尺度全带方法有更好的抗噪性能, 并改进了全带方法在无噪情况下的识别率下降的缺点. 多时间尺度差分复合子带方法由于兼有子带方法、多时间尺度方法和差分方法的抗噪性能, 有最好的识别效果.

5 结论

本文提出一种多时间尺度复合子带语音识别新方法, 新方法将不同时间尺度的全带和子带特征结合起来. 将新方法应用于 E-SET 在 NoiseX92 数据包白噪声下的语音识别, 实验结果表明, 新方法有最好的抗噪性能. 说明不同的时间尺度特征和子带特征、全带特征提供了不同的语音识别特性, 将不同的语音特征数据流有效的结合可以提高语音系统的抗噪性.

参考文献:

[1] Y Gong. Speech recognition in noisy environment: a survey [J]. Speech Communication, 1995, 16(3): 261- 291.

[2] 刘加. 汉语大词汇量连续语音识别系统研究进展 [J]. 电子学报, 2000, 28(1): 85- 91.

[3] J Allen. How do humans process and recognize speech [J]. IEEE Trans. on SAP, 1994, 2(4): 567- 576.

[4] J Ming, F J Smith. A probabilistic union model for subr band based rr bust speech recognition [A]. Proc. of ICASSP2000 [C], 2000: 739- 742.

[5] H Christensen, Lindberg B, Anderson O. Employing heterogeneous ir formation in a multistream framework [A]. Proc. of ICASSP2000 [C], 2000: 1571- 1574.

[6] S Okawa, E Bocchieri, A Potamianos. Multr band speech recognition in noisy environments [A]. Proc. of ICASSP98 [C], 1998, 641- 644.

[7] S Tibrewala. H Hermansky. Subr band based recognition of noisy speech [A]. Proc. of ICASSP1997 [C], 1997: 1255- 1258.

[8] S Wu. Incorporating information from syllable-length time scales in au tomatic speech recognition [D]. Ph. D dissertation, Dept. of EECS, UC Berkeley, 1998.

[9] A Hagen, H Bourlard. Using multiple time scales in the framework of multi stream speech recognition [A]. Proc. of ICSLP 2000 [C], 2000.

[10] Jinfu Xu, Gang wei(韦岗). Noise_ robust speech recognition based on difference of power spectrum [J]. Electronics Letters, 2000, 36(14): 1247- 1248.