2—5—2　自動音声認識の特徴量抽出への変調 Wavelet 変換の応用*

◎岡田 賢治　荒井 隆行 (上智大・理工)
金寺 登 (石川高専)　百村 裕智　村原 雄二 (上智大・理工)

1. はじめに

近年、音声認識の技術が進んできているが、音声認識の特徴量抽出において、雑音環境下を含むあらゆる環境において有効な、特徴量抽出の方法が必要である。Arai[1]らは音節明瞭度の知覚実験により、1-16Hz の変調周波数バンドが重要であるということをあきらかにした。さらに Kanedera et al.[2]は、自動音声認識において 1-16Hz、特に 2-10Hz が重要であるということを明らかにした。

Kanedera et al.[3]は、特徴量抽出において、PLP 係数の時間軌跡に対して周波数解像度の高い FFT と低い FFT を用いて、解像度の異なる 2 種類の FFT 係数を求め、2.5, 5, 7.5Hz 付近の変調周波数バンドに対応する係数を取り出すことで、認識率が向上すると報告している。疑似的に異なる解像度の変調周波数帯を複数抽出する際、低い変調周波数に対しては帯域幅を狭く、高い変調周波数に対しては帯域幅を広くすることが効果を生んでいると考えられる。この手法を変調フーリエ変換 (modulation FT) と呼ぶ。

Wavelet 変換もまた、高い周波数成分では時間分解能を高く、低い周波数成分では時間分解能が低いという特徴がある。本研究ではこの性質を利用し、従来変調フーリエ変換で行なっていた方法を、効率的に行なうことができると考えられる。この手法を変調 Wavelet 変換 (modulation wavelet transform) と呼ぶ。

2. 変調 Wavelet 変換を用いた単語音声認識実験

従来法と今回提案した新しい方法を比較するため、単語音声認識実験を行なった。実験環境は表 1の通りである。

表 1. 実験環境

タスク	Bellcore digit(0-9, zero, oh, yes, no の 13 種類 200 人発話の 2600 個)
標本化周波数	8 kHz
シフト	10 ms
フレーム	25 ms
学習	150 人話者 (男性 75 人 女性 75 人)
評価	50 人話者 (男性 25 人 女性 25 人)

Wavelet 変換を PLP 係数の時間軌跡に対して施した。これは PLP の時間軌跡に対する処理が、MFCC の時間軌跡に対するよりも認識率が向上するからである[4]。変調 Wavelet 変換により、音声認識に重要である変調周波数帯域 1-10 Hz を分割する。利用したスケールと変調周波数帯域の分割数との関係は表 2の通りである。

表 2. 分割数とスケールの関係

帯域分割数	スケール
2	16 8
3	32 16 8
4	64 32 16 8
5	128 64 32 16 8

認識・学習には HMM ToolKit(HTK[6]) を利用し、単語毎に状態数 6、混合数 2 の HMM を用いた。

雑音は、NOISEX-92 database[5]を利用し、その中の babble, buccaneer1, buccaneer2, destroyerengine, destroyerops, f16, factory1, factory2, hfcannel, leopard, m109, machinegun, pink, volvo, white の雑音を利用した。雑音は、SNR 比が 10dB になるように混ぜ合わせている。

3. 認識実験とその結果

3.1 変調 Wavelet 変換と一般的な手法との比較

まず clean 環境に対してと雑音環境として babble 雑音を選び実験した。従来法の MFCC + delta, PLP + delta, PLP + Modulation FT を行なった。そして meyer 型を利用して変調 Wavelet の帯域 2,3,4,5 等分割の実験を行なった。

従来法と 2-5 分割の変調 Wavelet による結果を表 3に示す。babble 雑音環境下で、3 等分割の変調 Wavelet が従来の MFCC + delta や PLP + delta や modulation FT よりも良い結果となった。modulation FT は 2.5, 5, 7.5 Hz 近辺の変調周波数バンドであるのに対して、変調 Wavelet の 3 等分割は (meyer)、2, 4, 8Hz 近辺の分割を行なっている。この認識率の差は、変調 Wavelet の方法とこの中心となる周波数の関係があるのではないかと考えられる。

3.2 mother wavelet の種類による比較

次に、'meyer 型' で clean, babble 雑音環境とも認識率が良かった帯域 2 分割と帯域 3 分割で、さまざまな mother wavelet を利用した。利用した mother wavelet は 'meyer', 'mexican hat', 'haar' である。

---

* Applying of the modulation wavelet transform on feature extraction in automatic speech recognition

By Kenji Okada, Takayuki Arai (Sophia University), Noboru Kanedera, (Ishikawa National College of Technology), Yasunori Momomura, Yuji Murahara(Sophia University)

表 3. 従来法と 2～5 等分割の結果 (単語誤り率 [%])

	clean	babble
MFCC + delta	1.65	21.5
PLP + delta	1.42	27.7
modulation FT	1.61	18.6
変調 Wavelet (2 band)	4.61	21.7
変調 Wavelet (3 band)	3.62	17.9
変調 Wavelet (4 band)	4.96	28.1
変調 Wavelet (5 band)	7.26	36.3

表 4. mother wavelet の比較 (単語誤り率 [%])

	2 等分割		3 等分割	
	clean	babble	clean	babble
meyer	4.61	21.7	3.62	17.9
mexican hat	3.1	22.8	5.03	33.5
haar	2.4	23.2	1.96	25.0

表 5. 変調 Wavelet 変換 ('meyer') と MFCC の、さまざまな雑音下における認識率の比較 (単語誤り率 [%])

雑音	meyer	MFCC
babble	17.9	21.5
buccaneer1	19.8	21.7
buccaneer2	19.0	21.8
destroyerengine	16.9	19.0
destroyerops	16.8	16.9
f16	17.0	21.5
factory1	18.6	20.9
factory2	13.1	16.0
hfchannel	15.8	23.1
leopard	13.4	15.5
m109	14.6	15.8
machinegun	41.5	50.2
pink	16.2	19.0
volvo	9.5	7.0
white	17.3	19.6
平均	17.8	20.6

clean 環境下と babble 雑音環境下で、さまざまな mother wavelet の認識実験の結果を表 4に示す。clean 環境で、meyer 型よりも低い誤り率となるものもあった。しかし、雑音環境下では、meyer 型よりも低い誤り率となるものがなかった。

3.3 'meyer' 型による全種類の雑音雑音による比較

認識率の良かった 'meyer' 型に対して全種類の雑音環境での認識実験を行ない、従来法の MFCC との比較を表5に示した。従来用いられていた MFCC の手法に比べ、全体でおよそ 3%の認識率向上が見られた。

4. まとめ

音声認識の特徴量抽出において、雑音に対して頑強な特徴量抽出の方法を調べた。従来用いられていた MFCC を用いた方法と、新しく提案した Wavelet 変換を用いた方法の比較を行なった。Wavelet 変換を用いた新しい方法は、従来の方法よりも高い認識率を出した。これから、雑音と用いる mother Wavelet の種類の検討により、有効に利用できる可能性があることがわかった。

5. 謝辞

本研究の一部は、平成１２年度 科学技術振興事業団 地域研究開発促進拠点事業の一環により行われた。

参考文献

[1] T. Arai, M. Pavel, H. Hermansky, C. Avendano, "'Syllable intelligibility for temporally filtered LPC cepstral trajectories." *J. Acoust. Soc. Am.*, Vol. **105**, No.5, pp 2738 - 2791, 1999.

[2] N. Kanedera, T. Arai, H. Hermansky and M. Pavel, "On the importance of various modulation frequencies for speech recognition." *Proc. of Eurospeech*, pp 1079–1082, 1997.

[3] N. Kanedera, T. Arai, H. Hermansky and M. Pavel, "On the relative importance of various components of the modulation spectrum for automatic speech recognition." *Speech Communication* **28**, pp. 43–55, 1999.

[4] N. Kanedera, , H. Hermansky and T. Arai, "On properties of modulation spectrum for robust automatic speech recognition," *Proc. IEEE ICASSP*, pp. II-613 - II-616, 1998.

[5] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, Vol. 12, No. 3, pp. 247 - 251, 1993.

[6] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, P. Woodland . "The HTK Book," Ver. 2.2, Entropic, 1999