自動音声認識における特徴量としての変調 Wavelet 変換の検討:変調帯域の分割法が認識率に与える影響について *

◎浅井健司　△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

より良い音声認識システムを実現するには、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 の変調周波数帯が重要であるということを明らかにした。さらに Kanedera et al.[2] は、自動音声認識において 1-16Hz、特に 1-10Hz 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、解像度の異なる 2 種類の FFT 係数を求め、2.5, 5.0, 7.5Hz 付近の変調周波数帯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) と呼んだ。

Okada et al.[4] は、特定の変調周波数帯域成分を抽出する際、従来変調フーリエ変換で行っていた方法を効率的に行う為に、高い周波数成分では周波数分解能を低く、低い周波数成分では周波数分解能が高いという特徴をもつ Wavelet 変換を利用し、認識率が向上することを確認している。この手法を変調 Wavelet 変換 (modulation wavelet transform) と呼んだ。

本実験では、変調周波数帯域成分の分割法と認識率の関係について調べる。

2. 帯域分割による単語音声認識実験

変調周波数帯域の分割を行う際、その分割方法が認識率に与える影響について調査するために孤立単語音声認識実験を行った。実験環境は表 1の通りである。

手法としては、PLP 係数の時間軌跡に対して Wavelet 変換を施した。Wavelet 変換では帯域分割を行う際の周波数帯を変えることができる。帯域分割する際、分割する帯域の最高変調周波数を変化させ、認識率に与える影響を調査した。帯域分割数は、文献 [4] において最も認識率が良かった 3 帯域で行った (図 1)。さらに 3 帯域のうち、中央の帯域の中心変調周波数を変化させたときの認識率への影響について調査した。用いた変調 Wavelet 変換の Mother Wavelet は Okada ら [4] による実験で最もよい認識率を与えた Meyer 型で、特徴量は 9 次の PLP 係数を 3 帯域に分けた、計 27 次の特徴量である。

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)

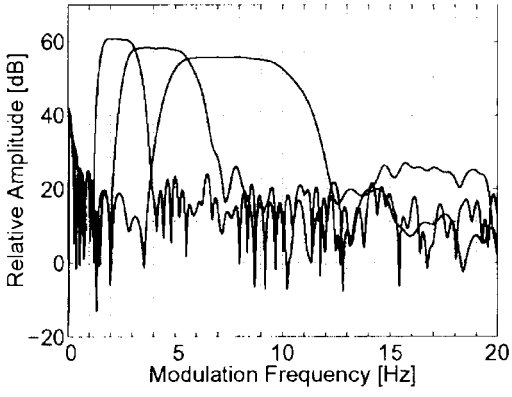

図 1. 変調 Wavelet 変換により抽出される帯域

認識・学習には HMM ToolKit(HTK[5]) を利用し、単語毎に状態数 6、混合数 2 の HMM を用いた。学習・評価データは Jack-knife 方式で 4 組用意した。

雑音は、NOISEX-92 database[6] を利用し、その中の babble, buccaneer1, buccaneer2, destroyerengine, destroyerops, f16, factory1, factory2, hfchannel, leopard, m109, machinegun, pink, volvo, white の 15 種類の雑音を利用した。雑音は、SNR が 10dB になるように波形上で混ぜ合わせている。

3. 認識実験とその結果

* Modulation wavelet transform as a feature for automatic speech recognition : the effect of division in the modulation domain

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

表 2. 最高変調周波数と認識率の関係

(単語誤り率 [%])

最高変調周波数 [Hz]	clean	babble
6.7	4.7	20.2
7.2	4.3	19.8
7.7	4.7	19.4
8.3	4.3	18.3
9.1	3.9	17.5
10.0	3.7	17.1
11.1	3.7	17.3
12.4	3.6	17.9

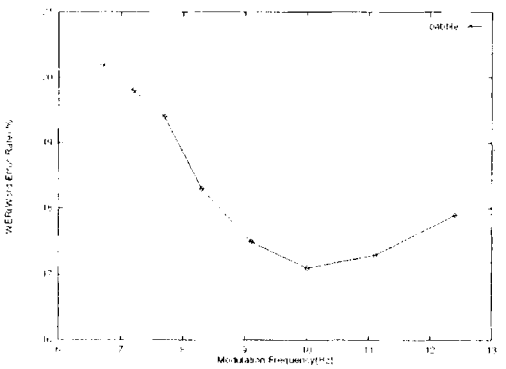

図 2. 最高変調周波数と認識率の関係 (単語誤り率 [%])

3.1 最高変調周波数と認識率

3 帯域の抽出を行う際、最高変調周波数を変えていったときの認識率に与える影響について調査した。最高変調周波数と認識率の関係を表 2に示す。抽出する最高変調周波数を 10Hz とした時、一番良い結果 (17.1%) を出しており (図 2)、同じく最高変調周波数が 10Hz のとき、clean でも Word Error Rate が 3.7%と比較的良い結果が観測された。

3.2 帯域等分割と認識率

抽出する 3 帯域のうち、中央の帯域の中心変調周波数を変化させ、認識率に与える影響について調査した。その際、第 **3.1** 節の結果を考慮して、最高変調周波数 10Hz まで抽出するために、高い帯域の中心変調周波数を 7.50Hz, 低い帯域の中心変調周波数を 2.35Hz に固定した。結果を表 3に示す。中央の帯域の中心変調周波数が 4.25Hz と 4.19Hz の時に最高の認識率である 16.9%が得られた。

ところで、両側の帯域の中心変調周波数 2.36Hz と 7.50Hz の相乗平均を求めたところ、4.21Hz となった。この値は最高の認識率を出した中心変調周波数とほぼ一致していることから、変調周波数を対数的に等分割することより、より良い認識率が得られることが分かった。

4. まとめ

変調 Wavelet 変換における変調周波数帯域を操作することによって認識率の変化を調査した。帯域分割

表 3. 中央帯域の中心変調周波数と認識率の関係

(単語誤り率 [%])

中心周波数 [Hz]	clean	babble
4.08	4.0	17.4
4.10	3.8	17.6
4.12	3.8	17.5
4.14	4.0	17.2
4.16	4.0	17.0
4.19	4.1	16.9
4.23	3.8	17.4
4.25	4.0	16.9
4.30	4.0	17.3
4.31	4.0	17.1
4.33	3.4	17.0
4.36	3.5	17.1
4.38	3.9	17.0

する際、最高周波数を 10Hz にすることが有効であると確認された。さらに変調周波数 1-10Hz を 3 帯域に対数的等分割することにより、より良い認識率を得ることができた。

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] K. Okada, T. Arai, N. Kanedera, Y. Momomura and Y. Murahara, "Using the mudulation wavelet transform for feature extraction in automatic speech recognition," *Proc.of ICSLP* , Vol.I, pp. 337-340, 2000.

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

[6] 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.

[7] 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.