1—Q—13 変調処理による自動音声認識のための頑強な特徴量抽出法
—変調複素 Wavelet 変換を用いて—

◎百村裕智 荒井隆行 岡田賢治（上智大・理工）
金寺登（石川高専） 村原雄二（上智大・理工）

1. はじめに

近年、自動音声認識の技術は雑音が少ない状態では非常に高精度の認識結果が得られているが、雑音環境下では認識率は著しく低下する。よって、本研究では様々な雑音環境下において頑強な特徴量抽出の手法を検討する。

頑強な自動音声認識のシステム構築において、特徴量の動的特徴を考慮するために、デルタ処理 [1], RASTA 処理 [2] などのような手法が提案されてきた。そして、このような時間方向への処理は近年、変調フィルタリングとして系統化されている。

Kanedera et al.[3] は変調 Fourier 変換 (modulation Fourier transform) を提案し、認識率が向上することを報告している。また、振幅に併せて位相成分も特徴量抽出において重要であることも報告している。

Okada et al.[4] は低い周波数成分では周波数分解能が高く、高い周波数成分では周波数分解能が低いという特徴を持つ Wavelet 変換を用い、変調 Fourier 変換で行なっていた方法を Wavelet 変換で行なう変調 Wavelet 変換 (modulation Wavelet transform) を提案した。

しかし、変調 Wavelet 変換において使用した母関数は全て実関数であり、これらの関数は位相成分を考慮していなかった。よって、本研究では PLP 係数の時間軌跡に対して、複素関数である Gabor 関数を母関数とする変調複素 Wavelet 変換 (modulation complex Wavelet transform) を提案し、実験を行なった。

2. 実験

2.1 Gabor 関数

本研究では Gabor 関数を変調複素 Wavelet 変換の母関数とした。Gabor 関数は次式のように表すことができる。

$$\psi(t) = \frac{1}{2\sqrt{\pi}\sigma} e^{-(\frac{t}{\sigma})^2} e^{-jt}. \qquad (1)$$

2.2 実験環境

Arai et al.[5] 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ことを明らかにした。さらに Kanedera et al.[6] は自動音声認識においては特に 2~10Hz が重要であることを報告している。そこで、変調複素 Wavelet 変換により、PLP 係数の時間軌跡に対して、自動音声認識において特に重要な 2~10Hz の変調周波数帯を対象とし、先行研究である変調 Fourier 変換、変調 Wavelet 変換と比較するために、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 計 2600 個)
標本化周波数	8kHz
フレーム周期	10ms
窓長	25ms
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)

使用した *scale* パラメータと変調周波数帯域の関係は表 2の通りである。Kanedera et al. による先行研究より、*scale* パラメータは 3 分割時に中心周波数が 2.5, 5.0, 7.5Hz になるように手動で与えた。他の分割数はこの 3 分割を基準に変調周波数バンドの幅が対数的に等分割になるように、それぞれ手動で与えた。

表 2. 帯域分割数と Scale パラメータの関係

帯域分割数	*Scales*
2	19 8
3	26.5 12.5 8
4	26.5 17 11 8
5	26.5 19 15 11 8

認識、学習には HMM ToolKit (HTK) 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 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 が 10dB になるように混ぜ合わせている。

---

* Using the Modulation Complex Wavelet Transform for Robust Feature Extraction in Automatic Speech Recognition

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

3. 認識実験とその結果

3.1 複素関数の有効性

先行研究と同様に clean 環境に対して雑音環境として babble noise を選び実験を行なった。そして、複素成分の有効性を調べるために Gabor 関数の実部のみを用いた場合との比較実験を行なった。*scale* パラメータは表2の3分割のパラメータを使用し、予備実験より σ の値が4以下では変調周波数での分割が不十分であったため、σ は 4, 8, 16 の値を使用した。

実験結果を表3に示す。babble noise 環境下では、実部のみを用いた時よりも虚部も考慮した複素関数のほうが認識率が良い結果となった。この結果より、位相成分は変調処理での自動音声認識において重要であることがわかる。

表 3. 実部のみと複素関数を用いた場合の比較実験 (%).

σ		clean	babble noise
4	real	98.5	80.6
	complex	98.7	84.8
8	real	96.6	77.8
	complex	96.5	79.3
16	real	93.7	73.0
	complex	94.1	75.8

3.2 分割数と σ の値による比較

次に、2~5分割の変調複素 Wavelet 変換による結果を表4に示す。全体的に帯域3分割時が他の分割数よりも認識率が高かった。表4より、σ=4、帯域3分割の時に最も良い認識率が得られた。

表 4. 分割数と σ の値による認識実験 (%)

帯域分割数		σ=4	σ=8	σ=16
2	clean	98.3	95.2	91.2
	babble noise	81.3	77.7	74.2
3	clean	98.7	96.5	94.1
	babble noise	84.8	79.3	75.8
4	clean	98.7	96.6	94.8
	babble noise	84.0	80.8	78.7
5	clean	98.5	96.5	94.4
	babble noise	83.2XS	80.3	78.1

3.3 先行研究との比較

認識率が最も良い結果となった帯域3分割、σ=4 に対して NOISEX-92 database の babble noise 環境下における先行研究との比較結果を表5に示す。

同様に NOISEX-92 database の全 noise 環境での実験を行ない、先行研究と比較した。この帯域3分割、σ=4 の時の認識率と先行研究との比較を表6に示す。

表 5. babble noise 環境下における先行研究との比較結果 (%)

手法	clean	babble noise
変調フーリエ変換	98.3	82.1
変調 Wavelet 変換		
Meyer, 3 分割	96.4	83.1
変調複素 Wavelet 変換		
Gabor, 3 分割	98.7	84.8
Standard approach		
MFCC + delta	98.3	78.5
PLP + delta	98.6	72.3

表 6. 全 noise 環境下における先行研究との比較結果 (%)

手法	clean	all noise
変調フーリエ変換	98.3	87.5
Meyer, 3 分割	96.4	82.2
Gabor, 3 分割	98.7	88.2

4. まとめ

本論文では、自動音声認識のための頑強な特徴量の抽出法を提案した。特徴量の時間方向への処理として、PLP 係数の時間軌跡に対して Gabor 関数を母関数とした変調複素 Wavelet 変換を適用した。

変調 Fourie 変換と変調 Wavelet 変換による手法との比較した結果において認識率が向上した。よって、Gabor 関数を用いる変調複素 Wavelet 変換の有効性が確認された。

参考文献

[1] S. Furui, "Speaker-independent isolated word recognition using dynamic feature of speech spectrum," *IEEE Trans. on Acoust.*, Speech Signal Processing, ASSP-34, 1, pp.52–59, 1986.

[2] H. Hermansky, N. Morgan, "RASTA Processing of speech," *IEEE Trans. on Speech and Audio Processing*, Vol.2, No.4, pp.578–589, 1994.

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

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

[5] T. Arai, M. Pavel, H. Hermansky and C. Avendano, "Syllable intelligibility for temporally filtered LPC cepstral trajectories," *J. Acoust. Soc. Am.*, Vol.105, No.5, pp.2738–2791, 1999.

[6] 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.