「聴覚脳プロジェクト」における ウェーブレットの応用と展開

河原英紀　（和歌山大学／ ATR ／ CREST）

kawahara@sys.wakayama-u.ac.jp

1　はじめに

「聴覚脳プロジェクト」は，科学技術振興事業団の戦略的基礎研究推進事業 (CREST) の「脳を創る」領域において進められている「聴覚の情景分析に基づく音響・音声処理システム」の略称である．人間の聴覚が行っている処理と同型の処理を工学的に実現することを目標に，研究を進めている．ただし，生理学的知見に忠実なモデルを目指すのではなく，本質的な機能のレベルにおいて同型の処理を行うことを目指しているのである．

内耳のフィルタ特性がウェーブレットに類似していることは，早くから指摘されており，同等な処理は広く用いられていた．聴覚とウェーブレットのこのようなつながりは，偶然のものではなく必然的な背景をもっている．以下では，本プロジェクトのキーワードである『聴覚の情景分析』[1] を軸として，本プロジェクトにおけるウェーブレットの役割に関する話題を紹介し，その背景に触れてみたい．

2　聴覚の情景分析

我々が日常的に接する音の中には，様々なものが含まれている．ある程度の長さの時間にわたって一つの音だけが聴こえていることは稀である．普通は，同時に様々なものが音を出していて，しかも，それぞれの音は，音源から耳に届くまでに様々な経路を伝わって変形されたものが混合されている．我々は二つの耳に到達するこれらの音の混合物の中から，自分がいる空間の広さやどのような音源がどのような位置にあるかを無意識のうちに把握することができる.『聴覚の情景分析』は，このような聴覚の機能を表わす概念である．

聴覚の情景分析の研究において大きな関心を持たれている問題の一つに,「どのような音が一つのものとして抽出されまとまったものとして知覚されるか？」というものがある．この問題に対する簡単な答は,「近いものはまとまる」である．この系として「音の変化は，継続している音に新しい音が付け加わって生ずると知覚される」という経験的法則がある．音の知覚的な『近さ』に基づいて，音の混合音の中から「継続している音」が取り出されるのである．

ただし『近さ』には，空間的な近さ，時間的な近さ，音の大きさや，高さや，音色の近さ等の様々な側面がある．本プロジェクトでは，その中で重要でかつ検討の遅れている音の高さや音色に関連した情報の表現と抽出方法を明らかにし，知覚的に意味のある領域で自由に変形して再合成することのできるシステムを開発することを具体的な目標としている [6, 5, 4].

3　似ている音

スケール独立な性質：　子供の話す言葉と大人の言葉は，文字に表わした場合には同じでも物理的には大きな違いがある．しかし，大人の「ア」も子供の「ア」も同じように「ア」と聞こえる．バイオリンとチェロは，寸法が大きく違い，出てくる音の物理的性質も違っているのに似た音に聞こえる．やや乱暴に言えば，形状の相似なものは，寸法が異なっていても似た音を出す．つまり，聴覚はスケールに依存せず形状だけに関連するような音の性質を取り出しているのである．

本プロジェクトメンバーの入野らは，内耳での周波数分析特性を良く近似できるとして彼等が提案した gammachirp 特性がスケール-時間領域で最小の不確定性を有するものであることを明らかにしている [2, 3]．入野によれば，聴覚は wavelet-Mellin 変換を行うシステムであるということになる．本プロジェクトとは独立に得られた内耳の応答に関する生理学的データは，この仮説が予測した瞬時周波数の遷移と同じ傾向を示している．

音の高さ　聴覚的な『近さ』の様々な側面が同等に重要な訳ではない．音の高さとして知覚される属性は，その中でも特に重要な役割を果たしている．音の高さには，二つの側面がある．一つは，響きとしての高さであり，もう一つは音楽で用いる音名で呼ぶことのできる高さである．前者は，周波数領域でのエネルギー分布の平均値に関連し，後者は，時間領域でのエネルギー分布の周期的構造に関連している．前の節で触れたように音から形状情報を抽出するには，そもそも音が発されなければならない．周期構造に関連した音の高さは，音を生み出すための駆動源についての情報でもある．

時間構造に関連した音の高さを求める問題は，近似的には，周期信号の基本周波数を求める問題となる．本プロジェクトで実現を狙うような高品質な分析変換合成を実現するためには，途中の処理の過程での高い時間分解能と周波数分解能とが必要となる．しかも，音声のように基本周波数が常に変化し続けるような信号に対しては，周期信号の定義を当てはめるのは不適切だという問題がある．これらが，瞬時周波数に基づいた sinusoidal mode あるいは，イベントに基づいた source filter model が必

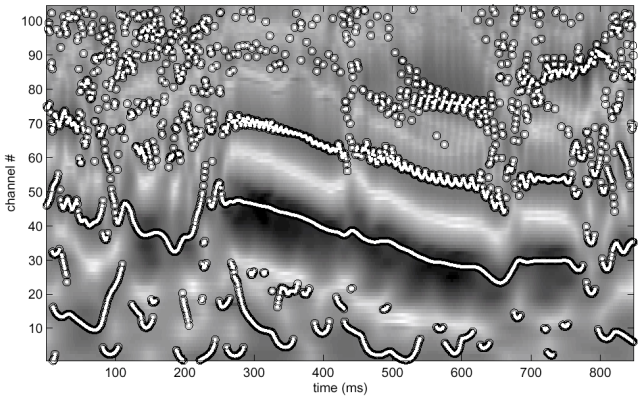

図 1: 男性が発声した連続数字音声「ひゃくにじゅうご」(125) の基本波成分の抽出

要となる理由である．次に紹介するようにウェーブレットは，それぞれのモデルにおいて重要な役割を果たしている．

4　ウェーブレットの応用

図 1 は，男声の発声した連続数字音声からの基本周波数の推定にウェーブレットを応用した例である．任意の基本周波数の調波複合音の基本波成分だけを取り出すように設計した帯域フィルタのインパルス応答を用いたウェーブレット変換では，応答のキャリア周波数から瞬時周波数への写像の不動点が基本周波数を与えるという性質がある．この性質を利用し，さらに，フィルタとして Gabor 関数と 2 次の cardinal B-spline を畳込んだものを用いることで，不動点近傍での時間周波数領域での写像の形状から精度よく瞬時周波数に含まれる雑音の影響を評価する方法を提案した [5]．図の縦軸はスケール，横軸は時間を表わし，散在する白点は不動点を表わす．この例では，40 Hz から 800 Hz までのスケールについて分析している．背景の濃淡図形の明度は，推定した瞬時周波数に含まれるノイズを表わす．基本周波数成分は，最もノイズの影響の少ない不動点として選択される．

図 2 は，同じ音声について，エネルギーの時間的局在として定義したイベントを手掛りに，振幅スペクトルから求められる最小位相システムの群遅延を補償することで音の原因となった駆動情報を抽出した例を示す．上の図は，縦軸がスケール，横軸が時間を表わす．図中の点がそれぞれのスケールにおける駆動の位置である．ここでは，分析に用いた時間窓の重心から窓内のエネルギーの重心への写像の不動点としてイベント時刻が求められる [4]．下の図は対応する音声波形である．この処理は，多重解像度表現でのエッジ検出に音響信号のような時系列に特有の因果律による影響の補償を加えたものと見ることができる．

ここで紹介した処理は，音声の高品質な変換を実現するために発明されたものであり，聴覚の神経生理学的知見に基づくものではない．しかし，そうして出来上がった処理では，内耳のフィルタに類似したウェーブレット

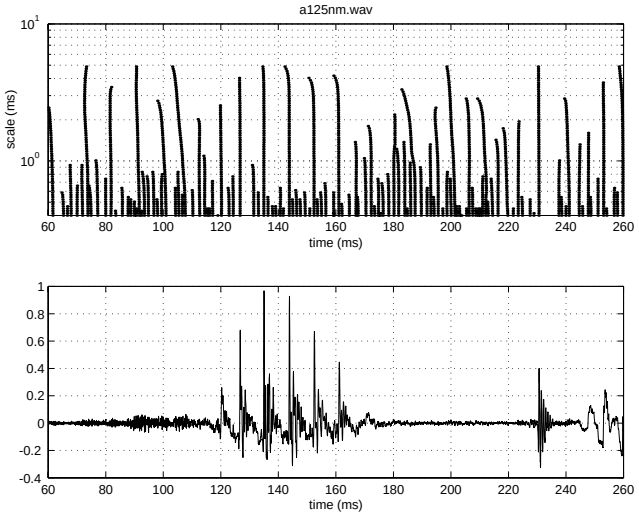

図 2: 同じ音声の最初の部分におけるイベントの抽出

変換が基本周波数の抽出に適した性質を有していることが利用され，聴覚神経に見られる同期発火の意味についても新しい解釈を示唆していることは大変興味深い．

5　まとめ

「聴覚脳プロジェクト」において，ウェーブレットがどのように応用されいかに重要な役割を果たしているかを紹介した．聴覚においては，wavelet-Mellin 変換のスケール不変な性質とともに，ここでは紹介しきれなかった over complete なシステムとしての側面が重要であることが，今後ますます明らかになって来るものと思われる．

参考文献

[1] A. S. Bregman. *Auditory Scene Analysis*. MIT Press, Cambridge, MA, 1990.

[2] Toshio Irino and Roy D. Patterson. A time-domain, level-dependent auditory filter: the gammachirp. *J. Acoust. Soc. Am.*, Vol. 101, No. 1, pp. 412–419, 1997.

[3] Toshio Irino and Masashi Unoki. An analysis/synthesis auditory filterbank based on an iir implementation of the gammachirp. *J. Acoust. Soc. Jpn.(E)*, Vol. 20, No. 6, pp. 397–406, 1999.

[4] Hideki Kawahara, Yoshinori Atake, and Parham Zolfaghari. Accurate vocal event detection method based on a fixed-point analysis of mapping from time to weighted average group delay. In *Proc. ICSLP'2000*, Beijin, 2000. [to appear].

[5] Hideki Kawahara, Haruhiro Katayose, Alain de Cheveigné, and Roy D. Patterson. Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of f0 and periodicity. *Proc. Eurospeech'99*, Vol. 6, pp. 2781–2784, 1999.

[6] Hideki Kawahara, Ikuyo Masuda-Katsuse, and Alain de Cheveigné. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction. *Speech Communication*, Vol. 27, No. 3-4, pp. 187–207, 1999.