音声のスパース表現のためのフィルタバンクの検討と聴覚特性との比較*

◎小野順貴, 嵯峨山茂樹 (東大院・情報理工)

1 はじめに

計算論的聴覚情景解析 (Computational Auditory Scene Analysis; CASA) においても，ブラインド音源分離 (Blind Sound Separation; BSS) においても，時間周波数分解はその基本的な信号処理過程の一つである。CASA においは聴覚フィルタを模擬するために Gammatone フィルタバンクなどが，BSS においては計算の容易さから短時間 FFT などが用いられるが，様々な音が混合した時間軸上の入力信号を，より分離しやすい基本的な成分に分解するという役割は共通している。

一方近年，信号分離においてはスパース性という概念が重要であることが指摘されている [1]。スパース性とは，信号のエネルギーがまばらにしか存在しない性質のことであり，信号がスパースに分解・表現されているならば，複数の信号が混合していてもその重なりは少なくなるため，分離はより容易になる [2]。

ここで重要になるのはスパースな表現を与えるような信号の分解形式であり，これは対象とする信号の性質に依存する。我々は前報告において，音声を対象とし，定帯域幅から定 Q をカバーするようなクラスのフィルタバンクで統計的な検証を行なった結果，短時間 Fourier 変換にほぼ等価な定帯域幅よりも，定 Q，さらには定帯域幅と定 Q の中間のようなフィルタバンクの方が，よりスパースな信号表現を与えるという結果を得た [3]。このような構造は調波成分抽出の枠組みでも議論されているが [4]，聴覚フィルタと類似点をもち，聴覚フィルタによる音声信号処理の信号処理的合理性になんらかの知見を与える可能性があると考えられる。

以上のような動機から，本研究では音声をスパースに表現するフィルタバンクを前報告より広いクラスから求め，またその特性を聴覚フィルタと比較することにより，音声のスパース表現と聴覚フィルタの関係について得られた結果を報告する。

2 時間周波数領域でのスパース性

2.1 フィルタバンクによる時間周波数表現

フィルタバンク分析により時間周波数領域に展開された信号 $s(t, f_c)$ は一般に

$$s(t, f_c) = \int_{-\infty}^{\infty} F(f) H(f; f_c) e^{j2\pi ft} df \qquad (1)$$

のように表わせる。ただし $F(f)$ は入力信号 $f(t)$ の Fourier 変換，$H(f; f_c)$ は中心周波数 f_c のフィルタの周波数特性を表わす。ここで本研究では，後述の解析のための規格化条件として，フィルタバンクは以下の 2 つの条件を満たすものとする [3]。

$$H(f; f_c) = 0 \quad (f < 0) \qquad (2)$$

$$\int_{-\infty}^{\infty} |H(f; f_c)|^2 df_c = 1 \qquad (3)$$

式 (2) は解析表現 (複素数表現) を得るため，式 (3) は信号エネルギーを時間周波数領域での二乗積分により評価可能にするための条件であり，ともにフィルタバンク特性を本質的に制限する条件ではない。

2.2 スパース性の指標

スパース性を定量的に評価する指標として，ここでは計算の容易さと分離性能との関連から，規格化された L_1 ノルム

$$P = E[a]^2 / E[a^2] \qquad (4)$$

を用いる [3]。ここで $E[\ ]$ は期待値，a は $s(t, f_c) = a(t, f_c) e^{j\phi(t, f_c)}$ のように，時間周波数領域で複素数で表現された信号 $s(t, f_c)$ の振幅 (絶対値) である。P は $0 < P \leq 1$ のように無次元化された指標であり，P が小さいほど信号エネルギーがまばらであり，スパース性が大きことを表す。

3 実験条件と実験方法

3.1 フィルタ形状と帯域幅

本研究では聴覚特性との比較を念頭に，フィルタ形状としては

$$H_b(f; f_c) = \frac{1}{C_b(f)} \exp\left(-\frac{(f - f_c)^2}{2B(f_c)^2}\right) \qquad (5)$$

$$H_m(f; f_c) = \frac{1}{C_m(f)} \left\{ \frac{1}{(1.019B(f_c) + j(f - f_c))^4} + \frac{1}{(1.019B(f_c) + j(f + f_c))^4} \right\} \qquad (6)$$

の 2 つを選んだ。H_b は信号処理によく用いられる Gabor フィルタ，H_m は聴覚フィルタの形状をよく近似するといわれる，次数 $n = 4$ の Gammatone フィルタの周波数特性 [5]，また $C_b(f)$，$C_m(f)$ は式 (3) を満足するための規格化関数である。$B(f_c)$ は帯域幅を決める関数であり，ここでの目的は，式 (4) を最小にするような $B(f_c)$ を求めることである。ただしこの最適化を容易にし，またデータに対する過学習を防ぐために，$B(f_c)$ を以下のような多項式でパラメトリックに表現し，式 (4) を最小にするような係数 b_k を求めることとした。

$$B(f_c)/[\mathrm{Hz}] = \sum_{k=0}^{K-1} b_k (f_c/[\mathrm{kHz}])^k \qquad (7)$$

3.2 実験データと実験方法

実験データとしては，日本音響学会研究用連続音声データベースから，男性 25 人女性 25 人計 50 名の話者が発話した音素バランス文 100 文を用いた。式 (4) を求めるための時間周波数領域での振幅分布の k 乗 ($k = 1, 2$) の期待値は，

$$E[a^k] = \frac{1}{N} \sum_i \sum_n \sum_m |f_i(nT, m\Delta f_c)|^k \qquad (8)$$

*Investigation of Filterbank for Sparse Representation of Speech and its Comarison with Auditory Characteristics by ONO, Nobutaka and SAGAYAMA, Shigeki (The University of Tokyo)

により求めた。ただし，i はデータのインデックス，$T = 1/16000$[s] はサンプリング周期，$\Delta f_c/2\pi = 8000/512$[Hz] はフィルタバンクの中心周波数間隔，$N$ は全サンプル数をそれぞれ表わす。すなわち本実験では，中心周波数はナイキスト周波数までの帯域を等間隔に 512 分割して固定した。

以上の条件の下，2 種類のフィルタ形状と帯域幅を表す多項式のパラメータ数 K を変え，スパース性の指標である式 (4) を最小にするような係数 b_k を数値的に求めた。パラメータ数 K は 1 から 4 まで変化させた他，帯域幅を b_1 の項のみで表した定 Q フィルタ条件 (この条件もパラメータ数 1) でも求めている。

4 実験結果

4.1 帯域幅の多項式近似の次数の評価

Table 1 に結果を示す。ただし Gabor フィルタと Gammatone フィルタでは $B(f_c)$ の意味が異なるため，これら 2 つのフィルタ形状間での係数の値自体の比較は意味がないことに注意する。Fig. 1 には，パラメータ数と規格化 L_1 ノルムの関係をグラフとして示した。Fig. 1 より，1) 音声のスパース表現のためには，Gabor フィルタの方が若干 Gammatone フィルタよりよい，2) どちらの場合もパラメータ数を 1 から 2 に増やした場合にはスパース性の指標である規格化 L_1 ノルムは減少するが，パラメータ数を 2 以上に増やしても値はほとんど変化しない，の 2 点が確認できる。特に 2) は，音声のスパース表現のためには，帯域幅は $B(f_c) = b_0 + b_1 f_c$ のような 1 次式の形でほぼ最適化されることを表している。

4.2 聴覚特性との比較

前節での議論より，パラメータ数 2 の場合についてフィルタバンク間での帯域幅を比較するために，帯域幅を次式で定義される等価矩形帯域幅 (Equivalent Rectangular Bandwidth; ERB) に換算した。

$$B_{ERB}(f_c) = \frac{\int_0^\infty |H(f;f_c)|^2 df}{(\max_f |H(f;f_c)|)^2} \qquad (9)$$

結果を Fig. 2 に示す。Gabor フィルタも Gammatone フィルタも帯域幅を ERB に換算すると，スパース性に関して最適化されたフィルタバンクはほとんど等しいことがわかる。また，聴覚フィルタの特性として，Zwicker らによって提案された臨界帯域 (CB)[6]，Glasberg and Moore によって求められた聴覚フィルタの等価矩形帯域幅 (ERB)[7] の関数表現

$$\mathrm{CB} = 25 + 75(1 + 1.4(f/[\mathrm{kHz}])^2)^{0.69} \qquad (10)$$

$$\mathrm{ERB} = 24.7(4.37(f/[\mathrm{kHz}]) + 1) \qquad (11)$$

を同図に示した。いずれと比較しても，今回得られたフィルタバンクの帯域幅に対して聴覚フィルタは 5 ～ 10 倍以上広く，聴覚フィルタと音声を最適にスパース表現するフィルタバンクとは直接には一致しないことがわかる。しかし特に臨界帯域 (CB) と比較すると，低周波数帯域では定帯域幅，高周波数帯域では定 Q に近いという相対的な帯域幅構造はよく類似しており，ほぼ 6 倍程度の関係にある。聴覚とは独立に音声という信号から導出したフィルタバンクと聴覚フィルタがこのような類似点をもっているのは興味深く，今後は聴覚系の動的な帯域幅制御機能も考慮して考察をすすめていく予定である。

Table 1 各条件で最小化された規格化 L_1 ノルム P とそれを与える帯域幅の係数：b, m はそれぞれ Gabor フィルタ，Gammatone フィルタを表す。一番下は比較のため，聴覚フィルタの ERB を用いたときの値。

type	b_0	b_1	b_2	b_3	L1 norm
b		20.6			0.157757
b	12.0				0.156662
b	7.50	6.33			0.152784
b	7.50	7.27	-0.459		0.152714
b	7.50	7.50	-0.791	0.0574	0.152704
m		33.7			0.160713
m	19.3				0.160872
m	11.3	11.7			0.156398
m	10.0	15.9	-1.38		0.156261
m	10.0	18.0	-3.67	0.370	0.156187
m	24.7	108			0.196416

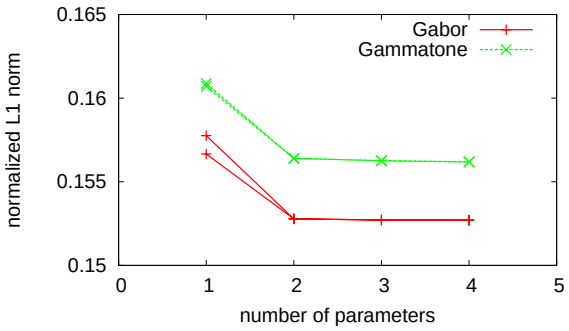

Fig. 1 パラメータ数と規格化 L_1 ノルム (スパース性指標) の関係

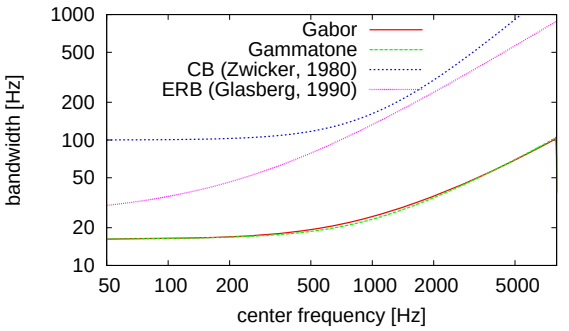

Fig. 2 中心周波数と帯域幅 (ERB 換算値) の関係

謝辞 本研究の一部は科学研究費補助金・若手研究 (B)(課題番号 18760303) の補助を受けて行なわれたので，ここに謝意を表する。

参考文献

[1] P. D. O'Grady, et al., Int. J. Imaging Systems and Technology, vol. 15, no. 1, 2005.
[2] Ö. Yilmaz et al., IEEE Trans. on SP, vol. 52, no. 7, pp. 1830–1847, 2004.
[3] 小野他, 音講論 (春), 3 月, 2006.
[4] 西他, 電気学会論文誌, vol. 122-E, no. 6, pp. 338-344, 2002.
[5] D. Dimitriadis, et al., Proc. Interpeech 2005, pp. 3013-3016, Sep. 2005.
[6] E. Zwicker, et al., JASA, vol. 68, no. 5, pp. 1523-1525, Nov. 1980.
[7] B. R. Glasberg, et al., Hearing Research, vol. 47, pp. 103-138, 1990.