音声のスパース性を最大化するフィルタバンクの検討*

◎小野順貴, 和泉洋介, 嵯峨山茂樹 (東大院・情報理工)

1 はじめに

近年，ブラインド信号分離の分野においては，音声のスパース性を利用した手法が提案され，観測信号数より音源信号数が多くても分離可能な方法論として注目されている [1][2][3][4]。音声のスパース性とは，音声のエネルギーが時間周波数領域においてまばらにしか存在せず，混合された複数の音声信号も時間周波数領域上ではあまり重ならない性質のことであり，ブラインド信号分離のみならず，他のアレイ信号処理 [5][6] にも関連する重要な概念である。

スパース性を前提とした信号分離の最初の処理は，信号がよりスパースに表現される基底への分解であり，その良し悪しは後段の特徴抽出や分離自体の性能に直結する。信号を最もスパースに表現する分解形式は対象とする信号に依存するため，先行研究においては音声信号を対象とし，短時間 FFT を前提に窓関数やその窓幅を変化させてスパース性が最大となる条件が検討されている [1][2]。

一方，短時間 FFT をフィルタバンク分析ととらえると，これは定帯域幅フィルタバンクに相当するが，音声に対してこのような帯域幅の選び方が最適かどうかは必ずしも自明ではない。例えば定 Q フィルタバンクのように，周波数帯域ごとに帯域幅を変化させるような分解を行なった方が，より高いスパース性が得られる可能性もある。

そこで我々は今回，フィルタバンク分析により音声のスパース性の変化について検討を行なった。本研究では，まず信号のスパース性の指標として時間周波数領域での振幅分布を統計的に扱うことにより得られる指標を導入し，これに基づきフィルタバンク形状と音声スパース性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確率的モデリングに基づくスパース性の評価

2.1 時間周波数分解

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 $f(t,\omega_c)$ は一般に

$$f(t,\omega_c)=\frac{1}{2\pi}\int_{-\infty}^{\infty}F(\omega)H(\omega;\omega_c)e^{j\omega t}d\omega \quad (1)$$

のように表わせる。ただし $F(\omega)$ は入力信号 $f(t)$ の Fourier 変換，$H(\omega;\omega_c)$ は中心周波数 ω_c のフィルタの周波数特性を表わす。

本研究ではフィルタバンクとして，以下の 2 つじの条件を満たすクラスを扱う。1 つ目の条件は，後の議論で用いる解析信号表現を得るために必要であり，

$$H(\omega;\omega_c)=0 \quad (\omega<0) \quad (2)$$

で与えられる。2 つ目の条件は，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に必要であり，任意の入力信号 $f(t)$ に対し，

$$\int_{-\infty}^{\infty}|f(t)|^2dt=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}|f(t,\omega_c)|^2dtd\omega_c \quad (3)$$

が成り立つことと表わされる。これを同値なフィルタバンクの条件として表わすと (導出は略),

$$\int_{-\infty}^{\infty}|H(\omega;\omega_c)|^2d\omega_c=1 \quad (4)$$

を得る。

2.2 スパース性の指標

Yilmaz らは信号のスパース性として，2 つの短時間 Fourier 変換分布 $s_1(t,\omega)$，$s_2(t,\omega)$ が時間周波数領域 (t,ω) において重ならず,

$$|s_1(t,\omega)|\cdot|s_2(t,\omega)|=0,\ {}^{\forall}t,\omega \quad (5)$$

が成り立つ性質を W-disjoint orthogonality と呼んでいる。また，この性質がどの程度満たされているかをはかるため，混合信号に対しある binary mask が 1) 目的信号をどの程度保存しているか，2) 妨害信号をどの程度除去しているか，の 2 つの量により決まる 0 ～ 1 に規格化された WDO という指標を導入している [2]。しかし，WDO のように複数信号の重なり具合を直接用いる指標は信号がどのような時間関係で混合されたかに依存することに注意が必要である。。

ここでは，あるクラスの信号の統計的なエネルギー分布の性質に基づき，信号間の時間関係などに依らないスパース性の指標を得ることを考える。そのためにまず，時間周波数分布 $f(t,\omega_c)=a(t,\omega_c)e^{j\phi(t,\omega_c)}$ の振幅 a，位相 ϕ を時間周波数領域 (t,ω) 上の確率過程として扱い (式 (2) より $f(t,\omega_c)$ は複素数になることに注意)，式 (5) のような 2 信号の重なり具合を積の期待値として考える。$E[a_1a_2]$ はそのままでは信号のエネルギーに依存するが，Schwartz の不等式：$E[a_1a_2]^2\le E[a_1^2]E[a_2^2]$ に注意して

$$0\le\frac{E[a_1a_2]^2}{E[a_1^2]E[a_2^2]}\le 1 \quad (6)$$

のように規格化すれば，0 ～ 1 の無次元量を得る。さらに a_1 と a_2 の独立性を仮定すれば,

$$\frac{E[a_1a_2]^2}{E[a_1^2]E[a_2^2]}=\frac{E[a_1]^2}{E[a_1^2]}\cdot\frac{E[a_2]^2}{E[a_2^2]} \quad (7)$$

のように積の形に分解できるので，あるクラスの信号 f_i の統計的なスパース性の尺度として,

$$P_i=\frac{E[a_i]^2}{E[a_i^2]} \quad (8)$$

を定義することができる。P_i は以下の性質をもつ。

*Investigation of Filterbank to Maxmize Sparseness of Speech by ONO, Nobutaka, IZUMI, Yosuke, and SAGAYAMA Shigeki (The University of Tokyo)

- $0 < P_i \leq 1$ であり，$a_i(t,\omega) = C$(定数) のときに限り $P_i = 1$ となる。すなわち P_i は，時間周波数領域上での振幅 a_i の分布の平坦さのようなものを表わしており，P_i が小さいほどスパース性が大きいといえる。
- 式 (3) の条件下では，同じクラスの混合された 2 信号に対し，振幅依存の連続値マスク $M(t,\omega_c) = a_1(t,\omega_c)^2/(a_1(t,\omega_c)^2 + a_2(t,\omega_c)^2)$ を用いた際の SIR の期待値の上限が $2/P$ で与えられる。

3 実験

3.1 フィルタバンク形状

先行研究によれば，短時間 Fourier 変換の場合には，スパース性には窓関数の形状はあまり関係がなく，窓関数の長さに強く依存することが示されている [1][2]。よって本研究では，フィルタバンクの周波数領域での形状はガウシアンに限定し，

$$H(\omega,\omega_c) = \frac{1}{C(\omega)} \exp\left(-\frac{(\omega-\omega_c)^2}{2B(\omega_c)^2}\right) \qquad (9)$$

のようなフィルタバンクにより実験を行なった。ただし $B(\omega_c)$ は周波数と帯域幅の関係を決定する関数，$C(\omega)$ は式 (4) を満足するための規格化関数である。

$B(\omega_c)$ としてはいろいろな関数系が考えられるが，ここでは，信号処理でよく用いられる定帯域幅フィルタバンク (以下定 BW) と定 Q フィルタバンク (以下定 Q) がカバーされるような関数で最も簡単な

$$B(\omega_c) = 2\pi b_0 + b_1\omega_c \qquad (10)$$

を用いた。フィルタ形状には規格化関数 $C(\omega)$ の影響が含まれるため，$B(\omega_c)$ は厳密には帯域幅そのものにはならないが，おおよそ $b_1 = 0$ の場合には帯域幅 b_0[Hz] の定 BW，$b_0 = 0$ の場合には Q 値 $1/b_1$ の定 Q，それ以外の場合には $\omega_c/2\pi \simeq b_0/b_1$[Hz] を境にそれらが滑らかに移り替わるようなフィルタバンクを表わす。

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 2 人女性 2 人計 4 名の話者が発話した音素バランス文 50 文，計 200 文を選んで 1 セットとし，実験の再現性を確かめるために話者が異なるデータ 2 セットを用いた。これらに対し，フィルタバンクパラメータを，b_0 を 2[Hz] 刻み，b_1 を 0.005 刻みで変化させ，スパース性の指標 P の変化を調べた。P を得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，

$$E[a^k] = \frac{1}{N}\sum_i\sum_n\sum_m |f_i(nT, m\Delta\omega_c)|^k \qquad (11)$$

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，$\Delta\omega_c/2\pi = 8000/2048$[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。

3.3 実験結果

2 つのデータセットに対して実験を行なった結果を Fig. 1 に示す。各セルの色は，そのセルの左下の格子点でのスパース性の指標 P の値を表わしている。P は信号の無音区間の長さ等にも左右されるため，2 つのデータセットでそのレンジは変化しているが，フィルタバンクパラメータによる変化の傾向は同様であり，安定した結果が得られていることが確認できる。Table 1 に，スパース性最大 (P 最小) を与えたパラメータの組を示す。定 BW 条件とは $b_1 = 0$，定 Q 条件とは $b_0 = 2$[Hz] (厳密には 0[Hz] とすべきであるが，数値的安定性を得るために除いた) を表わす。混合条件は探索した全パラメータが対象である。定 BW 条件の場合，2 つのデータセットで最適値がややずれているが，これは b_0 の離散化の影響も含まれており，ほぼ $b_1 = 12$[Hz] 辺りが最適値を与えている。この結果は先行研究における短時間 FFT の場合の最適窓幅の値ともおおよそ符合する。また，スパース性の値としてはそれほど大きくは変わらないが，$b_1 = 0.02$(Q 値 50 程度) の定 Q や，$(b_0, b_1) = (6, 0.01)$ の混合フィルタバンクの方が，定 BW よりもスパース性が大きいことが確認できる。今後は，残響環境下でのスパース性の変化等についても検証する予定である。

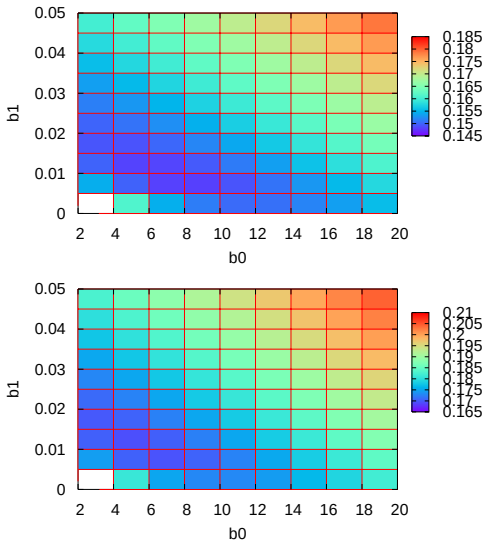

Fig. 1 フィルタバンクパラメータ (b_0, b_1) とスパース性の指標 P の関係

Table 1 各条件でスパース性最大を与える (b_0, b_1, P)

条件	セット 1	セット 2
定 BW	(14, 0.00, 0.1517)	(12, 0.00, 0.1744)
定 Q	(2, 0.02, 0.1495)	(2, 0.02, 0.1701)
混合	(6, 0.01, 0.1476)	(6, 0.01, 0.1690)

謝辞 本研究の一部は科学研究費補助金・萌芽研究（課題番号 17650045）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] M. Baeck et al., Proc. DAFx-03, Sep., 2003.
[2] Ö. Yilmaz et al., IEEE Trans. on SP, vol. 52, no. 7, pp. 1830–1847, 2004.
[3] A. Blin, et al., IEICE Trans. Fundamentals, vol. E88-A, no. 7, pp. 1693-1700, July 2005.
[4] 荒木他, 音講論 (秋), pp. 591-592, 9 月, 2005.
[5] 井上他, 音講論 (秋), pp. 619-620, 9 月, 2004.
[6] 戸上他, 音講論 (秋), pp. 625-626, 9 月, 2005.