劣決定音源分離のための 分離音声のケプストラムスムージング*

☆安齊祐美 (筑波大，NTT 研究所), 荒木章子 (NTT 研究所), 牧野昭二 (筑波大), 中谷智弘 (NTT 研究所), 山田武志 (筑波大), 中村篤 (NTT 研究所), 北脇信彦 (筑波大)

1 はじめに

ブラインド音源分離（BSS）とは，それぞれのセンサが観測した混合信号の情報のみを用いて，各音源の信号を推定する手法である．今回扱う音声信号における BSS の技術は，ハンズフリーテレビ会議システムなど，多くの応用が期待されている．

本研究では，音源信号のスパース性に基づく BSS, 特に [1–4] のような, 時間周波数バイナリマスク（BM）を用いる手法について議論する．時間周波数マスクは，音源の数 N がセンサ数 M よりも多い劣決定問題における BSS を実現する場合に広く用いられているためである．しかし，バイナリマスクを用いた場合，分離信号に時間的な不連続が生じ，ミュージカルノイズと呼ばれるノイズが発生することが知られている．

そこで近年，この問題を解決するため，ミュージカルノイズの発生を抑えた音源分離手法が提案された [5]. この手法では，時間周波数領域のバイナリマスクに時間方向のケプストラムスムージングを行う（Cepstral smoothing of spectral masks, CSM）．CSM は分離音声のミュージカルノイズの低減に効果的であることが示されている．しかし，[5] では結果の一例しか示されておらず，他のミュージカルノイズ低減手法との性能比較も行われていない．

また CSM では，バイナリマスクをケプストラム領域に変換することで音声特性とそれ以外の成分とを区別しているが，これらの特性はバイナリマスクよりも音声信号の方が正確に反映されると考えられる．そこで本研究では，バイナリマスクを用いる手法によって得られた分離音声信号にケプストラムスムージングをかける方法（Cepstral smoothing of separated signals, CSS）を提案し，CSM や他のミュージカルノイズ低減手法との性能比較を行う [6]．CSS によるもう 1 つの利点として，分離信号に直接スムージングをかけるため，シングルチャネル雑音抑圧手法のようなあらゆる音源分離や雑音抑圧の手法に適用可能であることが挙げられる．

2 問題設定

2.1 混合系

実環境において，N 人の音声信号 $s_i\,(i = 1, \ldots, N)$ が M 個のセンサで観測されたとすると，観測信号は

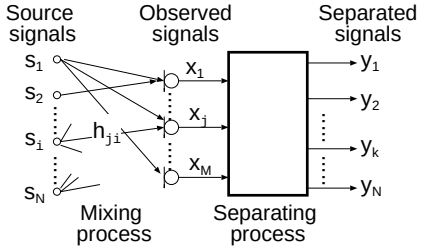


Fig. 1　劣決定 BSS のブロック図. ($N > M$)

畳み込みによって次のようにモデル化できる．

$$x_j(n) = \sum_{i=1}^{N} \sum_{l=1}^{L} h_{ji}(l) s_i(n - l + 1) \;\; (j = 1, \ldots, M) \tag{1}$$

x_j はセンサ j による観測信号，h_{ji} は音源 i からセンサ j へのインパルス応答である（Fig. 1）．BSS の目的は，観測信号 x_j の情報のみを用いて分離信号 y_i を得ることである．本研究では劣決定問題（$N > M$）について議論し，N と M は既知であると仮定する．

今回は時間周波数領域を用いる手法を採用する．これは，時間周波数領域での音声信号は時間領域よりスパースであることや [7]，時間領域での畳み込みは各周波数で積に変形できることを利用するためである．時間周波数領域の観測信号は次のようにモデル化される．

$$X_j(f, m) = \sum_{i=1}^{N} H_{ji}(f) S_i(f, m) \;\; (j = 1, \ldots, M) \tag{2}$$

$H_{ji}(f)$ は音源 i からセンサ j への伝達関数，$S_i(f, m)$ と $X_j(f, m)$ はそれぞれ短時間フーリエ変換（STFT）された原信号と観測信号を表す．f は周波数，m は時間フレーム番号である．

2.2 分離処理

分離には時間周波数バイナリマスク（BM）を用いる手法 [4] を採用する．この手法では，信号はスパースである，つまり各時間周波数スロットにおいて原信号のうち 1 つだけが支配的であると仮定する．この仮定を適用すると，観測信号の位相差によって N 個のクラスタが形成される．個々のクラスタは各音源に対応するため，個々のクラスタに属する時間周波数の観測信号を再構成することで，各信号を分離できる．まず k-means 法で観測信号ベクトル $\mathbf{X}(f, m) =$

---

*Cepstral Smoothing of Separated Signals for Underdetermined Speech Separation by Yumi Ansai (University of Tsukuba and NTT Communication Science Laboratories, NTT Corporation), Shoko Araki (NTT Communication Science Laboratories, NTT Corporation), Shoji Makino (University of Tsukuba), Tomohiro Nakatani (NTT Communication Science Laboratories, NTT Corporation), Takeshi Yamada (University of Tsukuba), Atsushi Nakamura (NTT Communication Science Laboratories, NTT Corporation) and Nobuhiko Kitawaki (University of Tsukuba)

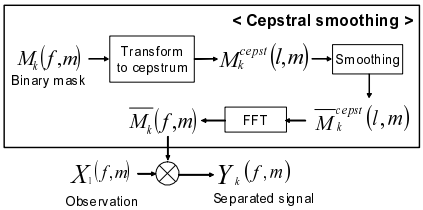


Fig. 2　BM のケプストラムスムージング (CSM) のブロック図

$[X_1(f,m),\ldots,X_m(f,m)]^T$ をクラスタリングし，次式で表されるマスクを生成する [4].

$$M_k(f,m) = \begin{cases} 1 & \mathbf{X}(f,m) \in C_k, \\ M_{min} & \text{otherwise} \end{cases} \quad (3)$$

ここで k は音源番号，C_k は音源 k のクラスタを表す．M_{min} には 0 のような非常に小さい値 (≥ 0) が入る．この BM を観測信号の 1 つに適用して分離音声を生成する．

$$Y_k(f,m) = M_k(f,m)X_j(f,m) \quad (4)$$

最後にこの分離信号 $Y_k(f,m)$ に逆 STFT とオーバーラップアドを適用し，時間領域の分離信号 y_k を得る．

しかし，BM を使用すると分離信号に時間的な不連続が生じ，ミュージカルノイズが発生する．

3　従来手法

本章では BM のケプストラムスムージング（CSM）[5] について概説する．Fig. 2 は CSM のブロック図を示している．

ケプストラム領域のマスクは次式により得られる．

$$M_k^{cepst}(l,m) = DFT^{-1}\{\ln(M_k(f,m))|_{f=0,\ldots,F-1}\} \quad (5)$$

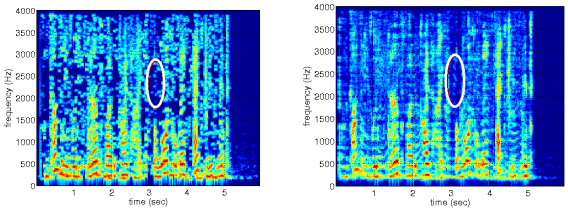
l はケプストラム係数，$DFT\{\cdot\}$ は離散フーリエ変換を意味し，F は変換フレーム長である．$M_k(f,m)$ は (3) と同様で，$M_{min} = 0.01$ とする．この M_k^{cepst} に対して時間方向の再帰的なスムージングを行う．

$$\overline{M}_k^{cepst}(l,m) = \beta_l \overline{M}_k^{cepst}(l,m-1) + (1-\beta_l)M_k^{cepst}(l,m) \quad (6)$$

式 (6) のケプストラムスムージングにおいて，ミュージカルノイズとして知覚される不要なランダムピークと音声特性とを区別するため，ケプストラム係数 l によってスムージング係数 β_l を次のように決定する．

$$\beta_l = \begin{cases} \beta_{env} & \text{if } l \in \{0,\ldots,l_{env}\} \\ \beta_{pitch} & \text{if } l = l_{pitch} \\ \beta_{peak} & \text{if } l \in \{(l_{env}+1),\ldots,F/2\}\backslash\{l_{pitch}\} \end{cases} \quad (7)$$

低次の係数 $l \in \{0,\ldots,l_{env}\}$ では $M_k^{cepst}(l,m)$ が $M_k(f,m)$ のスペクトル包絡を表すため，β_{env} を非常に小さな値にして包絡を保持する．同様に，$M_k(f,m)$ のピッチ周波数を表す $l = l_{pitch}$ における係数 β_{pitch} も比較的小さな値とする．それ以外の係数は $M_k(f,m)$ の微細構造を表し，不要なランダムピークを含んでいる可能性が非常に高いため，大きな値を持つ β_{peak} ($> \beta_{pitch}$) により強いスムージングをかける．



(a) BM　　(b) CSM

Fig. 3　分離信号のスペクトログラム

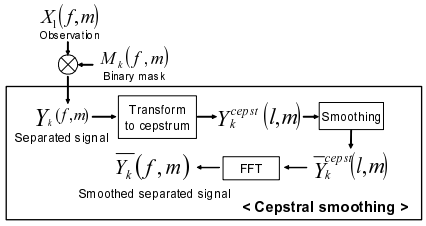


Fig. 4　分離信号のケプストラムスムージング (CSS) のブロック図

時間フレーム m に対し，l_{pitch} は以下の式を満たすケプストラム係数として選ばれる．

$$l_{pitch} = \underset{l}{\text{argmax}}\{M_k^{cepst}(l,m)|l_{low} \leq l \leq l_{high}\} \quad (8)$$

範囲 $\{l_{low}, l_{high}\}$ は音声のピッチ周波数の存在し得る 70〜500Hz に対応するように決定する．

$l > F/2$ の範囲では，DFT における対称性の仮定から $\overline{M}_k^{cepst}(l,m)$ を決定する．そして DFT と指数関数により時間周波数領域に変換し，スムージングされたマスク $\overline{M}_k(f,m)$ を生成する．

$$\overline{M}_k(f,m) = \exp(\text{DFT}\{\overline{M}_k^{cepst}(l,m)|_{l=0,\ldots,F-1}\}) \quad (9)$$

このスムージングされたマスクを使用し，(4) から分離信号を得る．

Fig. 3 は BM と CSM で得られた分離信号のスペクトログラムを示している．BM の場合 (Fig. 3 (a))，孤立のランダムピークが目立ち，多くのミュージカルノイズが発生する．一方 CSM では (Fig. 3 (b))，孤立のランダムピークはスムージングされており，CSM がミュージカルノイズ低減に効果的であることが確認できる．

4　提案手法

CSM では BM をケプストラム領域でスムージングしている．しかし，マスクのケプストラム表現が，(7) が仮定している音声のスペクトル包絡やピッチ周波数の情報を反映できているのかは定かではない．BM

Table 1　文献 [5] に基づくパラメータ

$f_s = 8$kHz	$l_{env} = 16$	$\beta_{env} = 0$
$F = 512$	$l_{low} = 32$	$\beta_{pitch} = 0.4$
$M_{min} = 0.01$	$l_{high} = 228$	$\beta_{peak} = 0.8$

Table 2　β_l の値の組み合わせ

		original	case 1	case 2	case 3
CSM	β_{pitch}	0.4	0.4	0.2	0.2
	β_{peak}	0.8	0.6	0.8	0.6
CSS	β_{pitch}	–	0.4	0.4	0.4
	β_{peak}	–	0.8	0.5	0.4

は 1 と 0 といった 2 値であり，目的音声の存在を推定しているに過ぎない．そこで CSM を参考に，分離信号のケプストラムスムージング（CSS）を提案する．この手法は，分離信号のケプストラム表現が持つ音声特性はマスクから得る特性よりも信頼性があるという前提に基づいている．

Fig. 4 は提案する CSS のブロック図を示している．CSM との相違点は，CSS は BM で得られる分離信号にケプストラムスムージングを適用していることである．

(4) で得られる分離信号 $Y_k(f,m)$ を，次式によりケプストラム領域に変換する．

$$Y_k^{cepst}(l,m) = DFT^{-1}\{\ln(Y_k(f,m))|_{f=0,\ldots,F-1}\} \quad (10)$$

Y_k^{cepst} に時間方向の再帰的なスムージングをかける．

$$\overline{Y}_k^{cepst}(l,m) = \beta_l \overline{Y}_k^{cepst}(l,m-1) + (1-\beta_l) Y_k^{cepst}(l,m) \quad (11)$$

β_l の条件は (7) と同様であるが，l_{pitch} を求める式は次のように変更する．

$$l_{pitch} = \underset{l}{\mathrm{argmax}}\{Y_k^{cepst}(l,m) | l_{low} \le l \le l_{high}\} \quad (12)$$

$l > F/2$ の範囲では DFT の対称性の仮定から $\overline{Y}_k^{cepst}(l,m)$ を決定する．$\overline{Y}_k^{cepst}(l,m)$ に DFT と指数関数を適用して時間周波数領域に変換し，最後に逆 STFT とオーバーラップアドによりスムージングされた時間領域の分離信号 y_k を得る．

5　実験と評価

5.1　実験 1：BM，CSM との比較

提案する CSS の性能を評価し，CSM との比較を行う．CSM で用いるパラメータ β_l は，[5] で示されている値 (Table 1 を参照) に加え，その他の値でも実験を行った．CSS の β_l については，複数の値で実験を行い，今回はその中から 3 組を採用する．これらの β_l の組み合わせを Table 2 に示す．β_{env} はスペクトル包絡を保持するためにどの組み合わせでも 0 とし，スムージングの強さに特に関係している β_{pitch} と β_{peak} を変更した．

5.2　実験 2：その他のミュージカルノイズ低減手法との比較

[5] において，CSM の性能と他のミュージカルノイズ低減手法との性能比較は行われていなかった．そこで本研究では，BM や CSM，そして CSS の性能を以下の 3 つの手法と比較する．

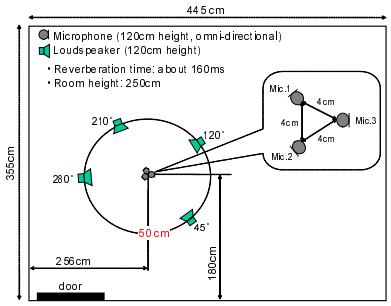


Fig. 5　実験条件

Table 3　5 段階絶対品質尺度

5	musical noise はほとんど気にならない
4	musical noise はあまり気にならない
3	musical noise がやや気になる
2	musical noise がかなり気になる
1	musical noise が非常に気になる

- 原音付加（AO）：BM による分離信号に小音量の観測信号 $X_1(f,m)$ を付加
- 画像処理によるマスクの再構成（MRI）：画像処理的な発想に基づき，分離信号のスペクトル $Y_k(f,m)$ とその周囲の時間周波数スロットを用いてスムージング

$$\begin{aligned}\overline{Y}_k(f,m) &= \tfrac{1}{2}Y_k(f,m) \\ &+ \tfrac{1}{4}\{Y_k(f-1,m)+Y_k(f+1,m) \\ &+ Y_k(f,m-1)+Y_k(f,m+1)\}\end{aligned} \quad (13)$$

- 人の知覚に基づくスペクトル減算法 [8]：臨界帯域幅を用いた知覚特性に基づき BM を変更

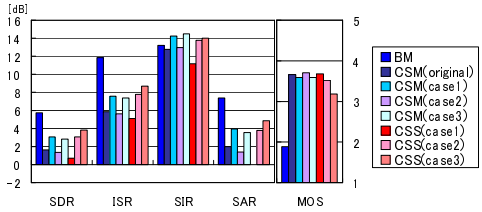
5.3　実験条件

観測データは，音声信号と実験室 (Fig. 5) で録音したインパルス応答との畳み込みにより生成する．実験室の残響時間は約 160msec である．センサ数 $M = 3$，音源数 $N = 4$ である．使用した音声データは男女それぞれ 4 発話の計 8 発話で，音声の組み合わせは 8 通りある．

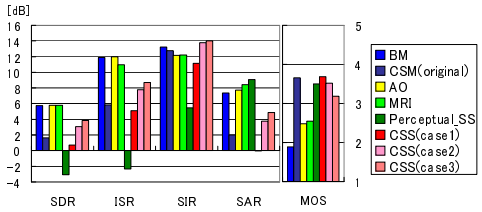
評価には客観評価と主観評価を用いる．客観評価値には [9] で提案された 4 つの歪み尺度を使用する．

- Signal to Distortion Ratio (SDR): 以下 3 つの総合的な歪み
- Source Image to Spatial distortion Ratio (ISR): 線形歪み
- Source to Interference Ratio (SIR): 目的音声以外の音声による歪み
- Sources to Artifacts Ratio (SAR): 非線形歪み

主観評価値には，11 人の聴取者による Mean Opinion Score (MOS) を使用する．MOS はミュージカルノイズの多さに着目したリスニングテストにより 5 段階絶対品質尺度 (Table 3) で評価する．



(a) 実験 1 の結果



(b) 実験 2 の結果

Fig. 6　比較結果

5.4　結果

Fig. 6(a) は CSM と CSS の性能比較を示している．比較のため，ケプストラムスムージングを行わない BM の性能も評価した．CSM，CSS 共に SIR の値は BM と同程度であるが，ISR や SAR の値は BM よりも低い．これらの歪みについては次節で考察する．CSM と CSS を比較すると，適切な β_l を選ぶことで CSS の ISR や SAR が CSM より良い値となることがわかった．これは CSM より CSS の方が正確に音声特性を保持していることを意味する．また，CSS の MOS 値は CSM の値と同等であることもわかる．

Fig. 6(b) は，ケプストラムスムージング手法（CSM や CSS）と 5.2 節で述べたミュージカルノイズ低減手法との比較結果を示している．CSM や CSS の ISR と SAR は他手法より低いが，一方でミュージカルノイズの多さに着目した MOS 値は CSM や CSS の方が高く，他手法よりミュージカルノイズ低減に効果的であることを意味する．

5.5　考察

前節で述べたように，ケプストラムスムージング手法（CSM や CSS）はミュージカルノイズ低減に効果的である．しかし，ISR や SAR の結果から，ケプストラムスムージング手法ではミュージカルノイズとは異なる歪みが生じることがわかった．ケプストラムスムージングではミュージカルノイズだけでなく目的音声の成分も取り除いてしまうことがあり，それによる歪みが生じていた．また，ケプストラムスムージング後は残響のような歪みが加わっているようにも感じられた．これらの歪みが ISR や SAR として評価されている．

実験により，CSS は CSM よりも ISR や SAR の値が良く，MOS 値も同程度の結果であることが分かった．CSS による分離信号を聞いても，CSM より目的音声の抑圧や残響感は少ない．これにより，分離信号のスペクトルへのケプストラムスムージングは CSM よりも効果的に音声特性を保持できることが示された．

それに加えて，Fig. 2 と 4 を見てわかるように，CSS と CSM の計算量は同じである．

6　おわりに

劣決定 BSS において，分離信号をケプストラム領域でスムージングする手法を提案し，その性能を評価した．CSM と比較して，提案法である CSS は音声の歪みが小さく，ミュージカルノイズの低減も可能であることがわかった．これにより，ミュージカルノイズ低減における CSS の有効性が示された．

謝辞　本研究を進めるにあたり，詳細な議論にご協力くださいました，NTT コミュニケーション科学基礎研究所の木下慶介氏に感謝致します．

参考文献

[1] O. Yilmaz and S. Richard, “Blind separation of speech mixtures via time-frequency masking,” *IEEE Trans. Signal Processing*, vol. 52, no. 7, pp. 1830-1847, July, 2004.

[2] N. Roman and D. Wang, “Binaural sound segregation for multisource reverberant environments,” *Proc. ICASSP 2004*, vol. II, pp. 373-376, May 2004.

[3] S. Rickard and O. Yilmaz, “On the W-Disjoint orthogonality of speech,” *Proc. ICASSP 2002*, vol. 1, pp. 529-532, May 2002.

[4] S. Araki, H. Sawada, R. Mukai, and S. Makino, “Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors,” *Signal Processing*, vol. 77, no. 8, pp. 1833-1847, Aug. 2007.

[5] N. Madhu, C. Breithaupt, and R. Martin, “Temporal smoothing of spectral masks in the cepstral domain for speech separation,” in *Proc. ICASSP 2008*, pp. 45-48, Mar. 2008.

[6] Y. Ansai, S. Araki, S. Makino, T. Nakatani, T. Yamada, A. Nakamura and N. Kitawaki, “Cepstral smoothing of separated signals for underdetermined speech separation,” in *ISCAS 2010*（to appear）.

[7] P. Bofill and M. Zibulevsky, “Blind separataion of more sources than mixtures using sparsity of their short-time-Fourier transform,” *Proc. ICA 2000*, pp. 87-92, June 2000.

[8] N. Virag, “Single channel speech enhancement based on masking properties of the human auditory system,” *IEEE Trans. Speech Audio Processing*, vol. 7, no. 2, pp. 126-137, Mar. 1999.

[9] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. P. Rosca, “First stereo audio source separation evaluation campaign: data, algorithms and results,” in *Proc. ICA 2007*, pp. 552-559, Sept. 2007.