日本音響学会誌 68 巻 2 号（2012），pp. 74–85

論　　文

43.60.−c

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[*1,*2]　荒木章子[*1]　牧野昭二[*2,*3]　中谷智広[*1]
山田武志[*2]　中村　篤[*1]　北脇信彦[*4]

［要旨］ 本論文では，音源信号のスパース性に基づき，時間周波数バイナリマスク（BM）を用いる音源分離手法におけるミュージカルノイズの低減を目的とした，分離音声のケプストラムスムージング（CSS）を提案する。CSS は，近年提案されたスペクトルマスクのケプストラムスムージング（CSM）で用いられるケプストラム領域でスムージングする考え方と，ケプストラム表現による音声特性の保持の制御という観点では，マスクではなく BM によって得られた分離音声を直接スムージングする方が好ましいという仮説とに基づいている。また，従来法（CSM）や提案法（CSS）と他のミュージカルノイズ低減手法の性能を実験により比較する。CSS では，CSM と同程度のミュージカルノイズ低減性能を有し，更に目的音声の歪の小さい分離信号が得られた。

キーワード　ブラインド音源分離，スパース性，バイナリマスク，ミュージカルノイズ，ケプストラムスムージング，分離音声信号
Blind source separation, Sparseness, Binary mask, Musical noise, Cepstral smoothing, Separated speech signals

1. はじめに

ブラインド音源分離（BSS）[1] 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。音声信号を対象とした 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。BSS 技術としては，独立成分分析（ICA）を利用する方法 [2] や，音源信号のスパース性に基づく方法（e.g, [3]）が知られている。前者は一般に，線形フィルタを用いて音源分離を行うことから非線形歪のない分離が可能であるが，音源数 N がセンサ数 M を越える劣決定問題に適用することはできない。一方後者は，劣決定問題における BSS を実現することから広く検討が行われており，特に各時間周波数ごとに占有的な音源成分を抽出する時間周波数バイナリマスク（BM）を用いる手法は，音声の音源分離に関して数多くの報告がある [3–15]。しかし BM 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非線形歪が発生することが知られている。そこで本論文では，音源信号のスパース性に基づく BSS，特に時間周波数バイナリマスク（BM）を用いる手法における非線形歪の問題について議論する。

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るための音源分離手法が提案された [16]。この手法では，BM に対して時間方向のケプストラムスムージングを行う（CSM: Cepstral Smoothing of spectral Masks）。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。CSM では，ケプストラム領域においてはスペクトル包絡や基本周波数といった音声特性とそれ以外の成分が区別し易くなることを利用し，BM をケプストラム領域に変換している。ミュージカルノイズの原因の一つとされるスペクトログラム上の孤立点は高次ケプストラム成分として現れるため，CSM はマスクの高次ケプストラムをスムージングすることで，ミュージカルノイズを低減している。しかし，スペクトル包絡や基本周波数などの音声特性がケプストラム領域で保存されることは，音声信号に対しては知られているものの，バイナリマスクのケプストラム表現が分離対象音声の音

---

* Cepstral smoothing of separated signals for underdetermined speech separation,
by Yumi Ansai, Shoko Araki, Shoji Makino, Tomohiro Nakatani, Takeshi Yamada, Atsushi Nakamura and Nobuhiko Kitawaki.

*1 NTT コミュニケーション科学基礎研究所
*2 筑波大学大学院システム情報工学研究科
*3 筑波大学生命領域学際研究センター
*4 筑波大学
（問合先：牧野昭二　〒305–8577 つくば市天王台 1–1–1 筑波大学生命領域学際研究センター）

（2011 年 2 月 3 日受付，2011 年 8 月 3 日採録決定）

声特性を保持しているかどうかは必ずしも自明ではない。また, [16] 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。

そこで本論文では，BM に比べケプストラム領域で音声特性をより保持すると期待される，BM による分離音声信号に対してケプストラムスムージングをかける手法（CSS: Cepstral Smoothing of separated Signals）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を比較する実験を行う [17]。この実験から，CSM と比較して，CSS は目的音声の歪が小さく，ミュージカルノイズ低減も可能であることを明らかにする。また，他手法よりもミュージカルノイズ低減に効果的であることも明らかにする。これにより，ミュージカルノイズ低減と音声特性の保持における CSS の有効性を示す。

本論文の構成を以下に示す。第 2 章では，BM を用いた BSS の問題設定について述べる。第 3 章では，従来法である BM のケプストラムスムージング（CSM）について概説する。第 4 章では，分離音声のケプストラムスムージング（CSS）の提案を行う。第 5 章では，提案法（CSS）と従来法（CSM）や他のミュージカルノイズ低減手法との性能を比較する実験について説明し，実験結果の提示と考察を行う。第 6 章では，結論として本論文のまとめを行い，今後の課題を挙げる。

2. 問 題 設 定

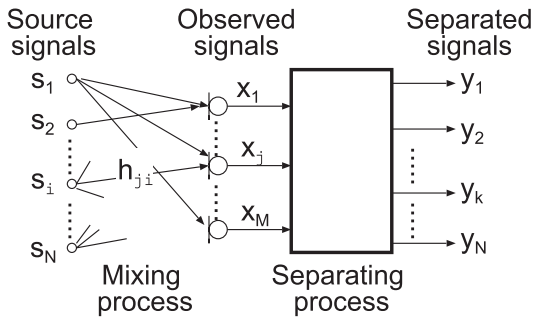
2.1 混 合 系

複数の音声信号 $s_i\,(i = 1, \ldots, N)$ が M 個のセンサで観測されたとすると，センサ j による観測信号 $x_j (j = 1, \ldots, M)$ は畳み込みによって次のようにモデル化できる。

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

h_{ji} は音源 i からセンサ j へのインパルス応答, l はインパルス応答のタップ数，n は時刻である（図–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。本論文では劣決定問題（$N > M$）について議論し，音源数 N とセンサ数 M は既知であると仮定する。

ここでは，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, 18–21]，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，時間周波数領域手法を採用する。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。



図–1 劣決定 BSS のブロック図（$N > M$）

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

$H_{ji}(f)$ 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 [3, 6] を採用する。この手法では，原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一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，2 個のセンサの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。

具体的には，まず k-means 法で観測信号ベクトル $\boldsymbol{X}(f, m) = [X_1(f, m), \ldots, X_M(f, m)]^{\mathrm{T}}$ をクラスタリングし，次式で表されるマスクを生成する [6]。

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

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 に近い非常に小さい値 (> 0) が入る。この BM を観測信号の一つに適用して分離音声を生成する。

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

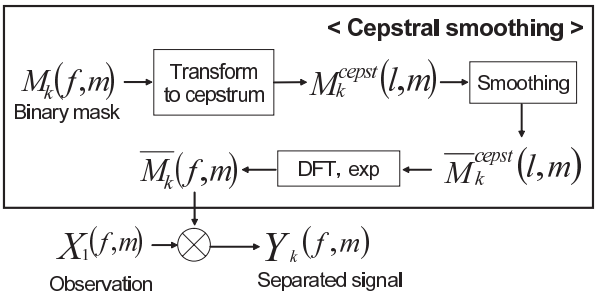
最後にこの分離信号 $Y_k(f, m)$ に逆 STFT とオーバラップアドを適用し，時間領域の分離信号 y_k を得る。

しかし, BM は時間的に非定常なマスクであり, 非線形な処理であるため, 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という問題がある [15]。

3. 従 来 法

3.1 CSM の概要

本節では BM のケプストラムスムージング（CSM）



図–2 BM のケプストラムスムージング（CSM）のブロック図

[16] について概説する。図–2 に CSM のブロック図を示す。

ケプストラム領域のマスクは次式により得られる。

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

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 は周波数ビンの数である。$M_k(f,m)$ 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。

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

ミュージカルノイズの原因の一つとされるスペクトログラム上の孤立点はスペクトル上ではランダムピークとして現れ，ケプストラム領域では高次ケプストラム成分となる。そこで，式 (6) のケプストラムスムージングにおいて，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 l によってスムージング係数 β_l を次のように設定する。

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

ここで F はケプストラム係数の数であり，式 (5) の周波数ビンの数 F と等しいものとする。

文献 [16] では，低次の $l \in \{0,\ldots,l_{\mathrm{env}}\}$ では $M_k^{\mathrm{cepst}}(l,m)$ が $M_k(f,m)$ のスペクトル包絡を表すと想定し，β_{env} を非常に小さな値にしてスムージングをほとんど行わない。同様に，$M_k(f,m)$ の基本周波数を表す $l = l_{\mathrm{pitch}}$ 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 $\beta_{\mathrm{peak}}\,(>\beta_{\mathrm{pitch}})$ により強いスムージングをかける。

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の次数として選ばれる。

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

範囲 $\{l_{\mathrm{low}}, l_{\mathrm{high}}\}$ は音声の基本周波数の存在し得る 70〜500 Hz に対応するように決定する。ここで，[16] では，有声区間と無声区間を区別することなく式 (8) により l_{pitch} を求めている。そのため無声区間における l_{pitch} にも式 (7) のとおりスムージングが行われることになるが，該当する係数は一つでありその影響は限定的である。

$l > F/2$ の範囲では，DFT における対称性の仮定から $\overline{M}_k^{\mathrm{cepst}}(l,m)$ 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。

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

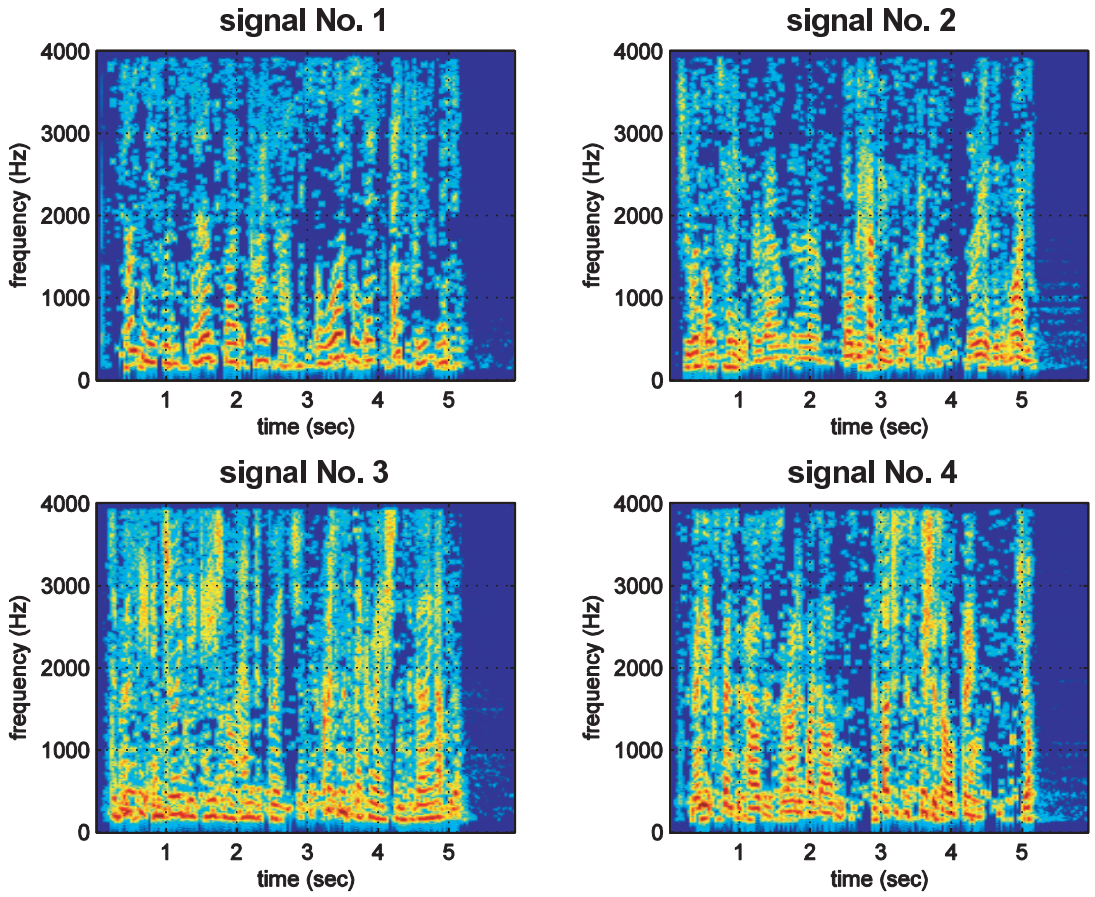
このスムージングされたマスクを使用し，次式により分離信号を得る。

$$\overline{Y}_k(f,m) = \overline{M}_k(f,m) X_j(f,m) \quad (10)$$

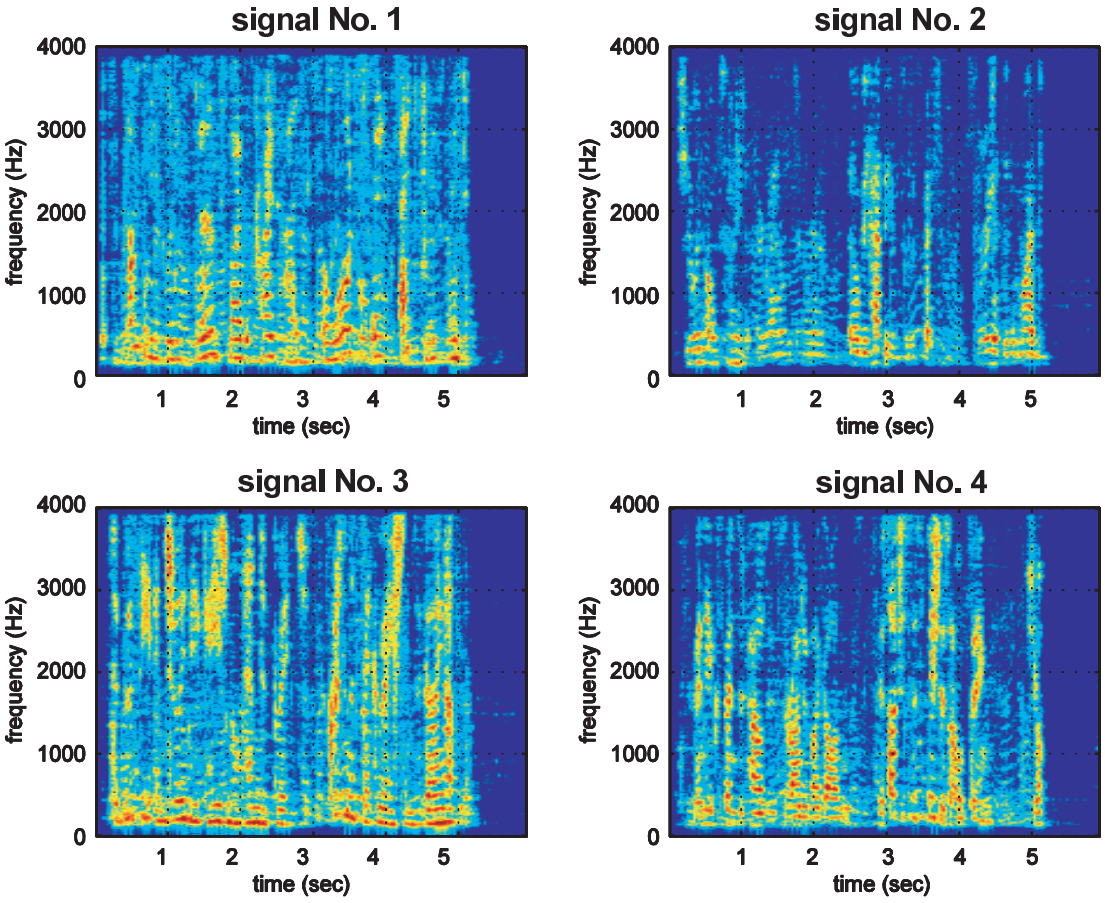
3.2 CSM の有効性と問題点

図–3 に BM と CSM で得られた分離信号のスペクトログラムを示す。これらの音声は，センサ数 $M = 3$，音源数 $N = 4$ の観測信号に対して，BM と CSM でそれぞれ音源分離を行って得られた 4 種類の分離信号である。BM では（図–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（図–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。

しかし，CSM では BM をケプストラム領域でスムージングしているが，BM は 1 と $M_{\min}$ といった 2 値であり，目的音声の存在の有無を各時間周波数で指定しているに過ぎない。図–4 に，スムージング前後のマスクのケプストラムの時間変化を図示したもの（以降，ケプストログラムと呼ぶ）を示す。ここでは原信号 (A) と BM による分離信号 (B)，及びスムージング前後のマスク (C)(D) のケプストログラムを示している。マスク M_k^{cepst} のケプストログラム（図–4(C)）は，原音声（図–4(A)）や BM による分離音声（図–4(B)）のケプストログラムとは異なり，成分が低次に集中しない。すなわち，音声のケプストラム表現とマスクのケプストラム表現は大きく異なっていることから，(7) のようにマスクに対して低ケフレンシー領域の値を保持しても，目的音声のスペクトル包絡を保持することは難しいと考えられる。

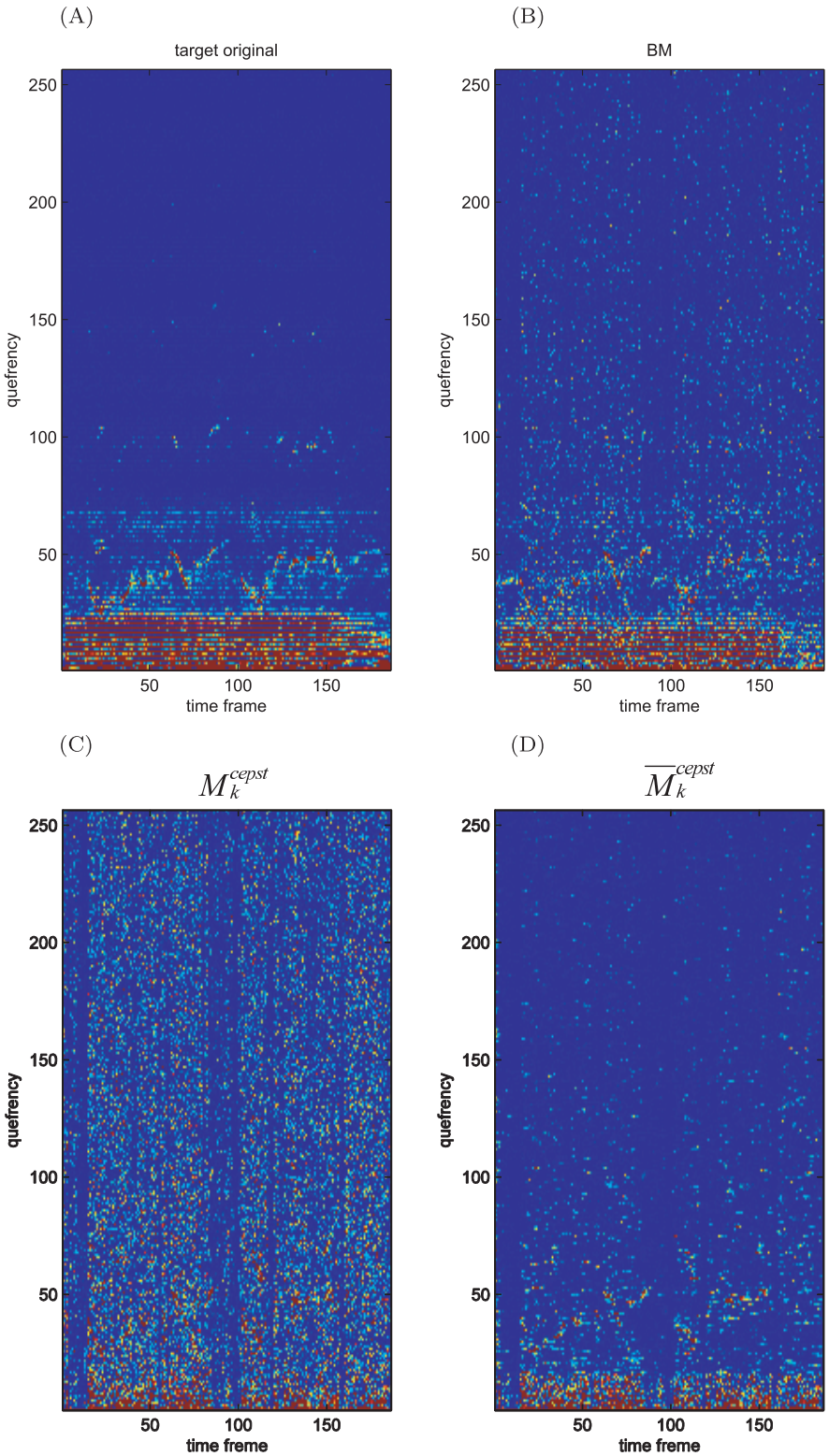


(A) BM



(B) CSM

図–3　分離信号のスペクトログラム

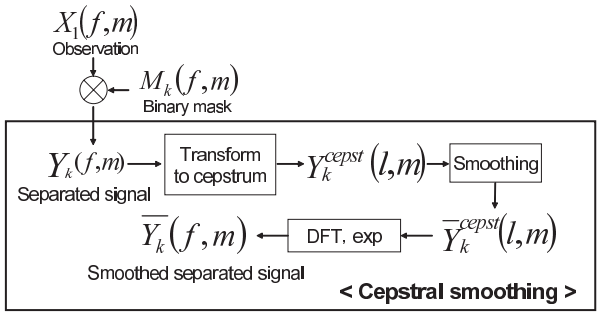


図–4　音声信号と CSM におけるマスクのケプストログラムの比較
(A) 目的音声の原信号，(B) BM による分離音声，(C) CSM におけるマスクのスムージング前，(D) CSM におけるマスクのスムージング後 のケプストラムの時間変化。横軸は時間フレーム，縦軸はケプストラム次数を表す。

4. 提　案　法

4.1 CSS の利点

3.2 節において，CSM の用いるマスクのケプストラム表現では，マスクのスペクトル包絡を保持しても，それが目的音声のスペクトル包絡を保持することにはならない，という問題点を指摘した。そこで本節では，マスクではなく，音声信号そのもののケプストラム表現を用いたケプスロラムスムージングを提案する。すなわち提案する CSS では，マスクではなくマスクに



図–5 分離信号のケプストラムスムージング（CSS）のブロック図

よって得られた分離音声に対してケプストラムスムージングを行う。これにより，低次ケプストラムの保持により分離音声のスペクトル包絡を保持しながら，高次ケプストラムのスムージングによってミュージカルノイズの低減ができると期待される。なお，高次ケプストラムのスムージングにより，音声のスペクトルの微細構造が崩れることが懸念される。この問題，すなわちミュージカルノイズ低減の度合いとスペクトル微細構造の保持の程度については，4.2 節に記載のとおり，パラメータ β_{peak} にて調整可能である。

図–5 に提案する CSS のブロック図を示す。図–2 と比較すると，分離音声にケプストラムスムージングをかけている点が異なる。これにより，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期待される。また，CSS の計算量は CSM と同じである。

4.2 CSS の処理の概要

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

$$Y_k^{\mathrm{cepst}}(l,m) = DFT^{-1}\{\ln(Y_k(f,m))|_{f=0,\ldots,F-1}\} \tag{11}$$

Y_k^{cepst} に時間方向の再帰的なスムージングをかける。

$$\overline{Y}_k^{\mathrm{cepst}}(l,m) = \beta_l \overline{Y}_k^{\mathrm{cepst}}(l,m-1) + (1-\beta_l) Y_k^{\mathrm{cepst}}(l,m) \tag{12}$$

ここで β_l の条件は (7) と同様である。$l \in \{(l_{\mathrm{env}}+1),\ldots,F/2\}\backslash\{l_{\mathrm{pitch}}\}$ の範囲における (12) によるスムージングはミュージカルノイズ低減を目的としているが，同時に目的音声のスペクトル微細構造の短時間変動を失う可能性もある。すなわちミュージカルノイズの低減の程度とスペクトル微細構造の保持の程度にはトレードオフの関係があり，これを β_{peak} により制御していると考えることができる。また，l_{pitch} を求める式は次のように変更する。

$$l_{\mathrm{pitch}} = \underset{l}{\mathrm{argmax}}\{Y_k^{\mathrm{cepst}}(l,m) | l_{\mathrm{low}} \le l \le l_{\mathrm{high}}\} \tag{13}$$

表–1 文献 [16] に基づくパラメータ

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

表–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

本稿では [16] と同様に，l_{pitch} の推定の際の有声区間と無声区間の区別は行わない。

$l > F/2$ の範囲では DFT の対称性の仮定から $\overline{Y}_k^{\mathrm{cepst}}(l,m)$ を決定する。$\overline{Y}_k^{\mathrm{cepst}}(l,m)$ に，(9) と同様に 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 $\overline{y}_k$ を得る。

5. 実　　験

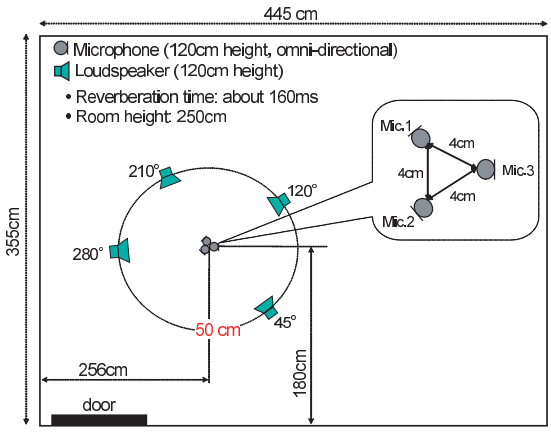
5.1 実験 1：CSS と CSM との比較

実験 1 では，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，[16] で示されている値（**表–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，その中から結果が優れている 3 組を採用した。これらの β_l の組み合わせを**表–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：CSS・CSM とその他のミュージカルノイズ低減手法との比較

文献 [16] 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。そこで実験 2 では，BM を用いる手法や CSM，そして CSS の性能を，以下の四つのミュージカルノイズ低減手法と比較する。

- 原音付加（AO）：BM による分離音声に小音量の観測信号 $X_1(f,m)$ を付加。BM やスペクトル減算法等による出力音の聴感上の非線形歪の低減を目的に，よく取られる簡便な方法である。
- 収縮化・膨張化処理 [22]（MRI）：BM に対して画像処理における収縮化と膨張化処理を適用し，スペクトルの孤立点を除去。具体的には，まず $M_k(f,m)$ の 4 近傍（$M_k(f-1,m)$, $M_k(f+1,m)$, $M_k(f,m-1)$, $M_k(f,m+1)$）に一つでも 0 が



図–6 実験室の大きさとセンサや音源の配置場所

あれば $M_k(f,m)$ を 0 としたあと（収縮化），更に $M_k(f,m)$ の 4 近傍に一つでも 1 があれば $M_k(f,m)$ を 1 とする（膨張化）。

- 画像処理的アプローチによるマスクの再構成（MRI_2）：画像処理的な発想に基づき，分離音声のスペクトル $Y_k(f,m)$ とその周囲の時間周波数スロットから次式のようにスムージングを行う。

$$\overline{Y}_k(f,m) = \frac{1}{2}Y_k(f,m) + \frac{1}{4}\{Y_k(f-1,m) + Y_k(f+1,m) + Y_k(f,m-1) + Y_k(f,m+1)\} \quad (14)$$

- 人の知覚に基づくスペクトル減算法 [23]（Perceptual_SS）：BM にて推定した雑音スペクトル（ここでは雑音と目的話者音声以外の話者音声を含む）を観測スペクトルから減算する方法。その際，スペクトル減算におけるオーバサブトラクション係数やフロアリング係数等をヒトの知覚特性に基づき決定し [23]，聴感上の非線形歪を低減させる。

5.3 実験条件

観測データは，音声信号と実験室（図–6）で測定したインパルス応答との畳み込みにより生成した。実験室の残響時間は約 160 ms，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各 4 名，各 1 発話の英語音声計 8 発話であり，この中から 4 個の音声を 8 通りの組み合わせで選択する。サンプリング周波数 f_s 及び DFT のフレーム長 F は表–1 のとおりであり，シフト長は $F/2$ である。

評価には客観評価と主観評価を用いた。客観評価値には [24] で提案された四つの歪尺度を使用した。

- Signal to Distortion Ratio (SDR): 音声対全歪比
- Source Image to Spatial distortion Ratio (ISR): 音声対線形歪比
- Source to Interference Ratio (SIR): 音声対目的音声以外の音声による歪比

表–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

- Sources to Artifacts Ratio (SAR): 音声対非線形歪比

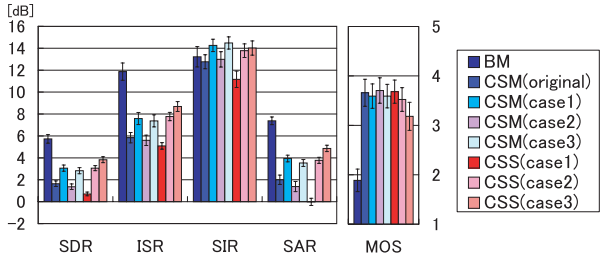
ここで SDR は ISR・SIR・SAR をすべて含んだ総合的な評価値となっており [24]，SIR は分離性能の，ISR と SAR はそれぞれ線形・非線形歪の評価値である。また，数値が大きいほど性能が良いことを示している。なお，単位は dB である。

主観評価値には，11 名の日本人聴取者による Mean Opinion Score (MOS) を使用した。11 名の内訳は，男性 7 名（20 代 5 名，30 代及び 50 代各 1 名），及び女性 4 名（20 代 3 名，30 代 1 名）である。各聴取者は，ミュージカルノイズの量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（表–3）で評価した。事前の聴取者への教示事項として，ミュージカルノイズ及び表–3 を説明し，幾つかの原音・BM 及び CSS 処理音声をランダムに例示した。ここでミュージカルノイズの程度と表–3 の尺度の対応づけについての教示は特に行わず，聴取者の判断で点数を付けるよう依頼した。評価に用いた聴取音声は，前述のすべての音声組み合わせに対する BM, CSM, CSS 及び 5.2 節の 4 手法すべての処理結果と幾つかの原音声であり，これらをランダムに提示した。リスニングテストは，防音室におけるヘッドホン受聴により行った。

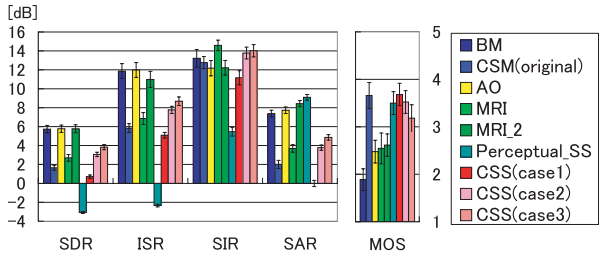
5.4 結　　果

図–7(a) に従来法（CSM）と提案法（CSS）の各歪値と MOS を示す。比較のため，ケプストラムスムージングを行わない BM を用いる手法の性能も評価した。CSM, CSS ともに SIR の値は BM を用いる手法と同程度であるが，ISR や SAR の値は BM を用いる手法よりも低い。これについては次節で議論する。また，CSM（original [16]）と CSS を比較すると，適切な β_l（例えば CSS（case2））を選ぶことで CSS の ISR や SAR が CSM（original）より良い値となることが分かった。これは CSM より CSS の方が正確に目的音声のスペクトル包絡や基本周波数を保持できることを意味する。また，CSS の MOS 値は CSM の値と同等であることも分かる。図–8 に，case2 の場合の β_l で CSS を適用した分離信号のスペクトログラムを示す。図–3(A) と比較して，孤立のランダムピークがスムージングされており，ミュージカルノイズが減っていることが分かる。

図–7(b) は，ケプストラムスムージング手法（5.1 節と同一の 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量に着目した 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が分かる。



(a) 実験 1 の結果。誤差範囲は 95%信頼区間。



(b) 実験 2 の結果。誤差範囲は 95%信頼区間。(BM, CSM, CSS の結果は実験 1 と同一)

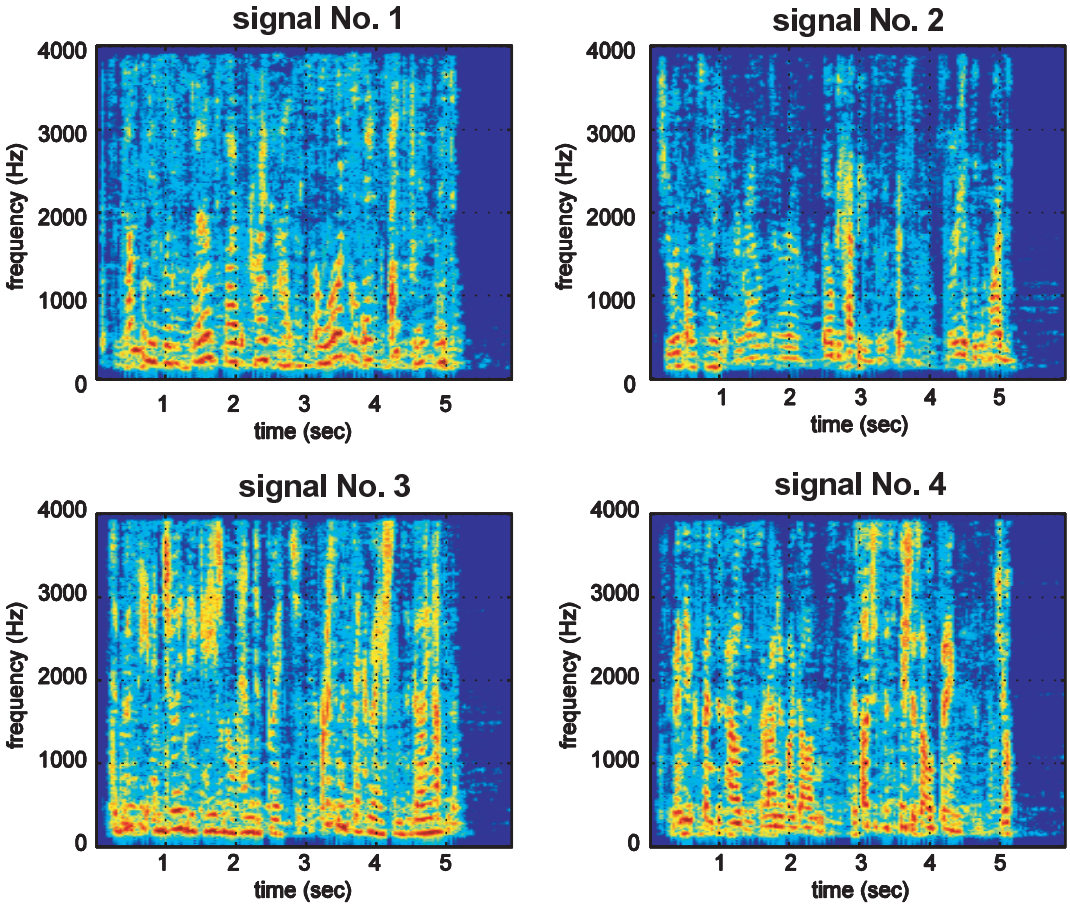
図–7　各歪値と MOS の比較結果

このように，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。また，図–7(a) より，CSS の性能とパラメータ β_{peak} との関係を読み取ることができる。すなわち，β_{peak} が大きな値の場合（case1）には MOS 値が高く ISR や SAR が低い，すなわちミュージカルノイズが低減され信号歪は大きい結果が得られる。一方 β_{peak} が小さな値の場合（case2, case3）には信号歪は小さいものの，ミュージカルノイズが顕著となる。これは 4.2 節に述べたミュージカルノイズの軽減の程度とスペクトル微細構造の保持の程度のトレードオフを示している。

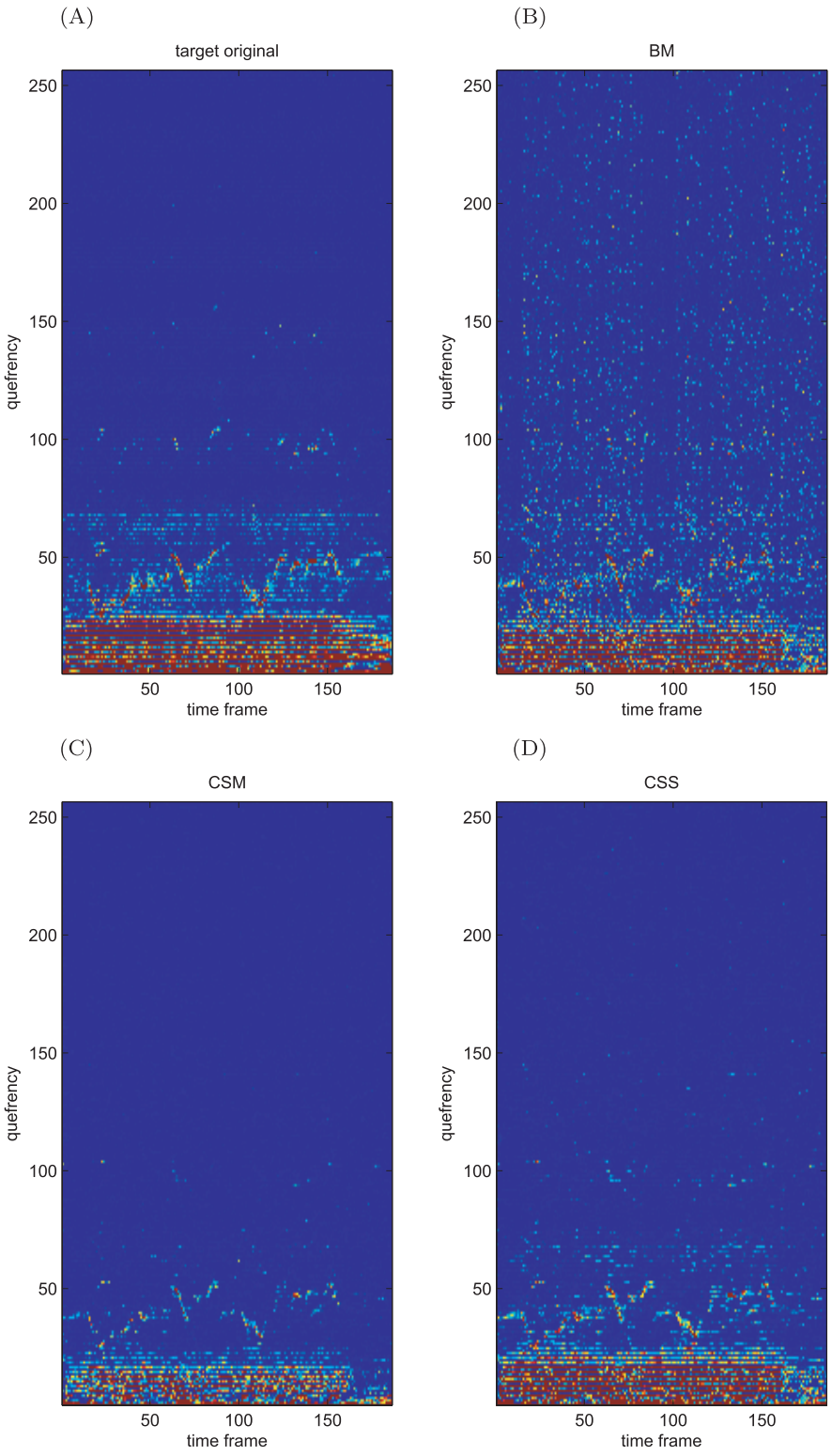
5.5 考　　察

前節で述べたとおり，提案法である CSS は CSM より高い ISR や SAR を持つことが分かった。また，CSS 及び CSM 法はミュージカルノイズ低減に効果的であることも示された。

しかし上述したとおり，CSS 及び CSM 法は BM 法（図–7(a)）や 5.2 節に述べたミュージカルノイズ低減手法（図–7(b)）と比較して ISR や SAR が低くなることから，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が生じることが分かった。実際に著者らが聴取したところでは，ケプストラムスムージ



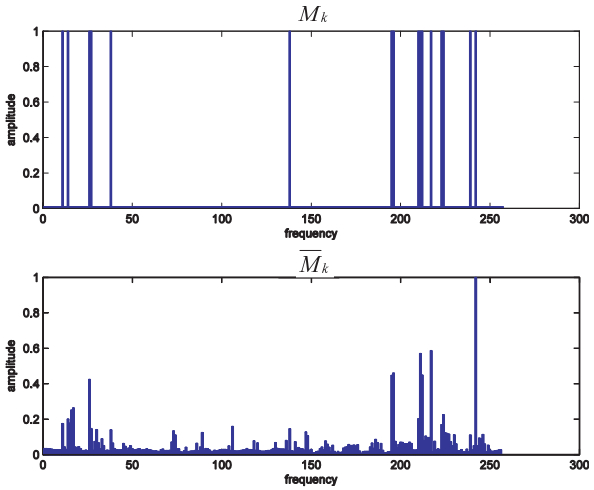
図–8　CSS による分離信号のスペクトログラム



図–9　各音声信号のケプストログラム
(A) 目的音声の原信号，(B) BM による分離音声，(C) CSM による分離信号，(D) CSS による分離信号のケプストラムの時間変化。横軸は時間フレーム，縦軸はケプストラム次数を表す。

ング後は残響のような歪が加わっているようにも感じられ，これらの歪が ISR や SAR が低くなる原因であると考えられる。また，前節において，CSS（case2）は CSM（original）よりも ISR や SAR の値が良く，MOS 値も同程度の結果であった。この CSS（case2）による分離信号の聴感上の特徴としては，CSM より目的音声の抑圧感や残響感は少ないということが挙げられる。

そこで本節では，ミュージカルノイズ低減と目的音声の歪について，ケプストラム表現の側面から考察する。図–9 に，目的音声の原信号と，BM，CSM，CSS



図–10 CSM におけるマスクの周波数特性（横軸は周波数 bin 番号）

によるそれぞれの分離信号のケプストログラムを示す。

まず目的音声の原信号と BM による分離信号について，その高次成分に着目する。図–9 上段のケプストログラムより，BM による分離信号では高次のケフレンシーにも成分が現れている。低次ケプストラムについては大きな変化は見られないことから，ミュージカルノイズは高次ケプストラムに現れていると考えられる。

次に，BM と CSM による分離信号について比較する。図–9(B) と (C) を見ると，CSM による分離信号では低次ケプストラム，特に包絡成分に当たる 0 ～ l_{env} の成分が減少している。ここで，CSM におけるマスクのある時間フレームでの周波数特性のスムージング前後での変化を図–10 に示す。BM は，1 と $M_{\min}$ の 2 値のみで中間値を持たない（図–10 上）。音声のスパース性が完全に成り立っているならば，分離処理 (4) によって BM と観測信号を掛け合わせても，分離信号のスペクトル包絡は変形しないことになる。しかし CSM のケプストラムスムージングによって，マスクのスペクトル包絡が変形しているのが分かる（図–10 下）。CSM の分離信号を得る演算 (10) では，スムージングされてスペクトル包絡が変形したマスクと観測信号を掛け合わせるため，分離音声のスペクトル包絡にも変形が起こると考えられる。つまり，3.2 節で指摘したように，マスクの低次ケプストラム成分を保持したとしても，目的音声のスペクトル包絡を保持していることにはならないと言える。CSM における目的音声の歪は，スペクトル包絡に関するこれらの問題が原因であると考えられる。一方，CSM ではマスクの高次ケプストラムにスムージング処理を行うだけで，目的音声自体のスムージングは行わない。そのため，目的音声の微細構造には直接的な影響を与えておらず，微細構造の変化による歪は CSS と比べて小さいと考えられる。

最後に，BM と CSS による分離信号を比較する。図–9 の (B) と (D) から，CSS では低次ケプストラム成分が BM による分離信号のものとほぼ変わらずに保持されていることが分かる。更に原信号 (A) と比較しても，低次ケプストラム成分は十分保持できていると言える。ここから，CSS の方が CSM よりも正確に目的音声のスペクトル包絡成分を保持できていると言える。

しかしその一方で，CSS では目的音声の高次ケプストラムに直接スムージングの処理をしているため，目的音声の微細構造が変形してしまうことが推測される。CSS では，この微細構造の変化によって目的音声に歪が生じていると考えられる。以上より，CSM と CSS における目的音声の聴感上の違いは，このような歪の原因の違いによるものと考えられる。

6. おわりに

本論文では，劣決定 BSS において分離音声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。従来法である CSM と比較して，提案法である CSS は音声の歪が小さく，ミュージカルノイズの低減も可能であることが分かった。ケプストラムスムージングを用いない，他のミュージカルノイズ低減手法と比較しても，ミュージカルノイズ低減に効果的であることが分かった。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。また，CSS の利点として，CSS と CSM の計算量は同じであること，分離音声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。今後の課題としては，CSS のケプストラムスムージングにおいて，目的音声の微細構造に与える影響を小さくする工夫を考案することが挙げられる。

謝　辞

本研究を進めるにあたり，詳細な議論をいただきました，NTT コミュニケーション科学基礎研究所の木下慶介博士，筑波大学生命領域学際研究センターの寺澤洋子博士，東京大学の宮部滋樹博士に感謝いたします。

文　献

[1] S. Haykin, Ed., *Unsupervised Adaptive Filtering, Volume I: Blind Source Separation* (Wiley, New York, 2000).

[2] A. Hyvärinen, J. Karhunen and E. Oja, *Independent Component Analysis* (John Wiley & Sons, New York, 2001).

[3] Ö. Yilmaz and S. Richard, “Blind separation of speech mixtures via time-frequency masking,” *IEEE Trans. Signal Process.*, 52, 1830–1847 (2004).

[4] N. Roman and D. Wang, “Binaural sound segregation for multisource reverberant environments,” *Proc. ICASSP 2004*, Vol. II, pp. 373–376 (2004).

[5] S. Rickard and Ö. Yilmaz, “On the W-disjoint orthogonality of speech,” *Proc. ICASSP 2002*, Vol. 1, pp. 529–532 (2002).
[6] S. Araki, H. Sawada, R. Mukai and S. Makino, “Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors,” *Signal Process.*, **77**, 1833–1847 (2007).
[7] P. Bofill and M. Zibulevsky, “Blind separation of more sources than mixtures using sparsity of their short-time Fourier transform,” *Proc. ICA 2000*, pp. 87–92 (2000).
[8] A. Jourjine, S. Rickard and Ö. Yilmaz, “Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures,” *Proc. ICASSP 2000*, Vol. 5, pp. 2985–2988 (2000).
[9] M. Aoki, M. Okamoto, S. Aoki, H. Matsui, T. Sakurai and Y. Kaneda, “Sound source segregation based on estimating incident angle of each frequency component of input signals acquired by multiple microphones,” *Acoust. Sci. & Tech.*, **22**, 149–157 (2001).
[10] S. Rickard, R. Balan and J. Rosca, “Real-time time-frequency based blind source separation,” *Proc. ICA 2001*, pp. 651–656 (2001).
[11] N. Roman, D. Wang and G.J. Brown, “Speech segregation based on sound localization,” *J. Acoust. Soc. Am.*, **114**, 2236–2252 (2003).
[12] S. Araki, S. Makino, A. Blin, R. Mukai and H. Sawada, “Blind separation of more speech than sensors with less distortion by combining sparseness and ICA,” *Proc. IWAENC 2003*, pp. 271–274 (2003).
[13] J.M. Peterson and S. Kadambe, “A probabilistic approach for blind source separation of underdetermined convolutive mixtures,” *Proc. ICASSP 2003*, Vol. VI, pp. 581–584 (2003).
[14] S. Araki, S. Makino, A. Blin, R. Mukai and H. Sawada, “Underdetermined blind separation for speech in real environments with sparseness and ICA,” *Proc. ICASSP 2004*, Vol. III, pp. 881–884 (2004).
[15] S. Araki, H. Sawada, R. Mukai and S. Makino, “Blind sparse source separation with spatially smoothed time-frequency masking,” *Proc. IWAENC 2006* (2006).
[16] N. Madhu, C. Breithaupt and R. Martin, “Temporal smoothing of spectral masks in the cepstral domain for speech separation,” *Proc. ICASSP 2008*, pp. 45–48 (2008).
[17] Y. Ansai, S. Araki, S. Makino, T. Nakatani, T. Yamada, A. Nakamura and N. Kitawaki, “Cepstral smoothing of separated signals for underdetermined speech separation,” *Proc. ISCAS 2010*, pp. 2506–2509 (2010).
[18] 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 (2000).
[19] A. Blin, S. Araki and S. Makino, “Blind source separation when speech signals outnumber sensors using an sparseness mixing matrix combination,” *Proc. IWAENC 2003*, pp. 211–214 (2003).
[20] Y. Izumi, N. Ono and S. Sagayama, “Sparseness-based 2ch BSS using EM algorithm in reverberant environment,” *Proc. WASPAA*, pp. 147–150 (2007).
[21] 和泉洋介, 小野順貴, 嵯峨山茂樹, “スパースな混合モデルに基づく雑音・残響環境下の劣決定 BSS,” 信学総大, pp. 58–59 (2008).
[22] 山口　亮, 金田　豊, “雑音抑圧信号処理におけるミュージカルノイズ改善の検討,” 音講論集, pp. 619–620 (2004.3).
[23] N. Virag, “Single channel speech enhancement based on masking properties of the human auditory system,” *IEEE Trans. Speech Audio Process.*, **7**, 126–137 (1999).
[24] E. Vincent, H. Sawada, P. Bofill, S. Makino and J.P. Rosca, “First stereo audio source separation evaluation campaign: Data, algorithms and results,” *Proc. ICA 2007*, pp. 552–559 (2007).

安齊　祐美

平 21 筑波大学第三学群情報学類卒。平 23 同大大学院システム情報工学研究科コンピュータサイエンス専攻博士前期課程修了。在学中，音声認識，ブラインド音源分離に関する研究に従事。

荒木　章子

平 12 東大大学院・工・計数修士課程了，同年日本電信電話㈱入社。実環境における音響信号処理の研究に従事。現在，同社コミュニケーション科学基礎研究所主任研究員。博士（情報科学）。日本音響学会，IEEE，電子情報通信学会会員。

牧野　昭二

昭 56 東北大大学院修士課程了。同年日本電信電話公社入社。以来，NTT 研究所において，電気音響変換器，音響エコーキャンセラ，ブラインド音源分離などの音響信号処理の研究に従事。工博。現在，筑波大学生命領域学際研究センター教授。IEEE Distinguished Lecturer。IEEE Fellow。電子情報通信学会 Fellow。

中谷　智広

平 03 京大大学院・工・応用システム科学修士課程了，同年日本電信電話㈱入社。実環境における音響信号処理の研究に従事。現在，同社コミュニケーション科学基礎研究所主幹研究員。博士（情報学）。日本音響学会，IEEE，電子情報通信学会会員。

山田　武志

平 11 奈良先端大博士後期課程了。同年，筑波大学講師。現在，同准教授。音声認識，音環境理解，多チャネル信号処理，メディア品質評価の研究に従事。博士（工学）。IEEE，電子情報通信学会，情報処理学会，日本音響学会各会員。

中村　篤

昭 62 九大大学院・工・情報修士課程了。同年日本電信電話㈱入社。平 6 国際電気通信基礎技術研究所（ATR）出向。平 12 日本電信電話㈱復帰。音声認識，音声メディア情報処理，及び学習理論応用の研究に従事。現在，同社コミュニケーション科学基礎研究所主幹研究員/信号処理研究グループリーダ。博士（工学）。日本音響学会，電子情報通信学会会員。IEEE シニアメンバ。

北脇　信彦

昭 44 東北大・工・電子卒。昭 46 同大・院・修士課程了。同年日本電信電話公社入社。平 9 筑波大学大学院教授。工博。現在，筑波大学教授（国際戦略室長）・名誉教授。通信品質，音声符号化，音声認識，音源分離などの研究に従事。IEEE Fellow，信学会フェロー。