ブラインド音源分離に基づく環境適応型の音メディア処理とその実用化

奈良先端科学技術大学院大学
情報科学研究科・音情報処理学講座
猿渡　洋

(2008年7月)

本日の講演内容

- ブラインド音源分離 (BSS) モジュールの紹介
- 従来BSSの問題点について概説
- 新しいICA(SIMO-ICA) に基づく二段BSS手法の解説
  - 汎用DSPによる実装例を紹介
  - 音声認識結果、対話システムへの導入例を紹介
- まとめ

こんな機械に知性を感じる？

……

君の名前は何？
何が出来るの？

えっ、人間型ロボットなのにスイッチで入力するの？

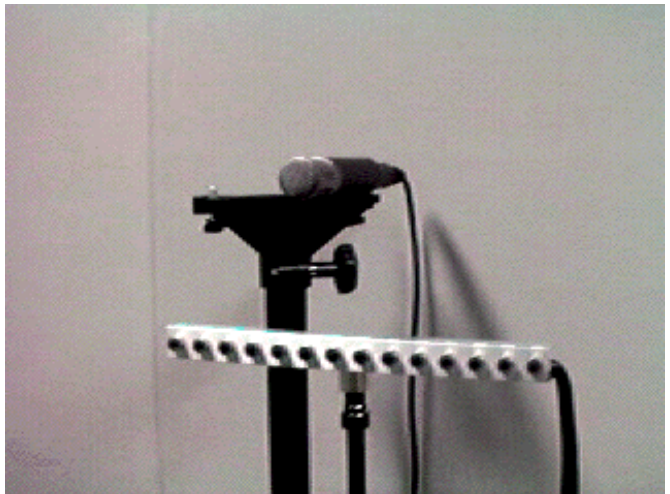
音声利用に秘められた可能性

- **音声は人間が持つ最も原始的なコミュニケーション手段**
  - 特別な道具は必要ない
  - 誰でも使える意思伝達手段
  - コンピュータやロボットとのコミュニケーション手段の一つとして昔から注目されてきた
- **機械と人間のインタラクションにおける音声メディア**
  - 知性を感じさせる機械・ロボットを実現するには音声対話メディアが必須である
  - 必要とされる機能
    - 必要な音を聞き分ける（**雑音抑圧・音源分離**）
    - ユーザの声を言語として認識理解する（**音声認識**）
    - ユーザに情報を発信する（音声合成）

音を聞き分ける耳：マイクロホンアレー

実際、人間も2つの耳で聞くことによって、音の方向や複数音の聞き分けを行っている

- **音声処理での一例：マイクロホンアレー**
  - **複数のマイクによって得られた複数の受音信号のなかから、必要な情報（目的音声）のみを取り出す装置**



<u>期待される応用</u>

- **高性能な hands-free 通信**
- **雑音にロバストな音声認識**

⇒ではどういうアルゴリズム（ソフト）が必要なのか？

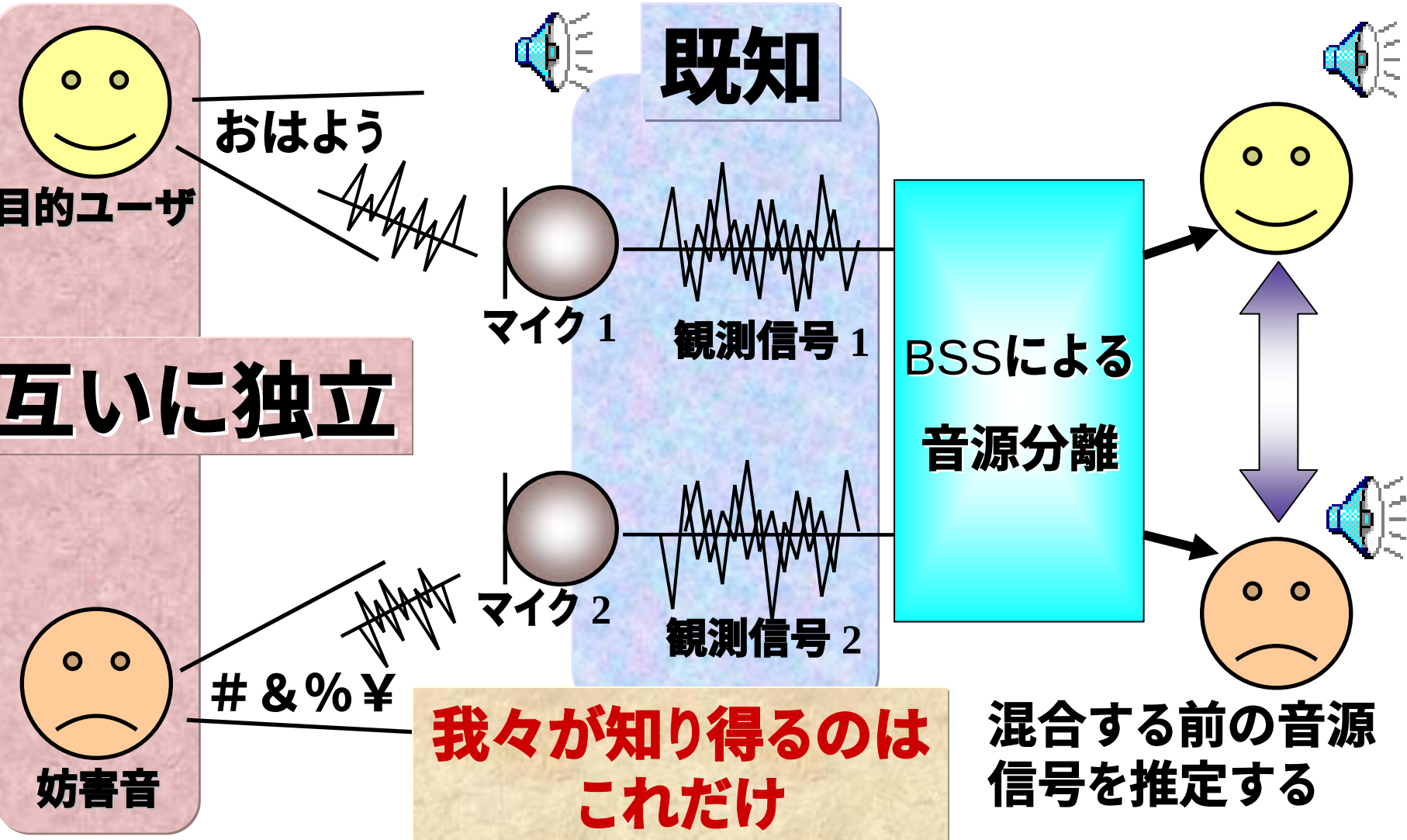
アルゴリズム：ブラインド音源分離

- **Blind Source Separation (BSS)**
  - 複数の音源信号が混合されて観測された場合、観測信号のみから音源信号を推定する処理
  - 一切の事前情報を必要としない、つまり音の種類、音環境の変化、マイク特性などに依存せず、音を分離できる技術

2つの代表的なアルゴリズム

- **バイナリマスク**
  - 各音源の時間空間分布の違いを利用
- **独立成分分析** (ICA)
  - 「**目的音が互いに独立**」という仮定のみを利用
  - 「**互いに独立**」とは、「**お互いに関係が無い**」ということ
  - 統計的な情報処理に基づく信号処理の一種で、**人間の脳に近い**情報分類処理

BSS とは？


既知
おはよう
目的ユーザ
互いに独立
マイク 1
観測信号 1
BSSによる
音源分離
マイク 2
観測信号 2
＃＆％￥
妨害音
我々が知り得るのは
これだけ
混合する前の音源
信号を推定する

ポケットサイズ音源分離処理ユニット

- ＢＳＳが**実時間で**稼動
- 演算量の多いICAを間欠的に動作させ、消し残り誤差は後段のポスト処理で高速かつ効率的に除去（**２段構成**）
- 2005年秋に神戸製鋼所と共同で開発
- 4chマイクアンプとDSP（TI社製C６７）から構成されるBSS専用モジュール
- バッテリー（単3×2）込みでも150g以下
- 任意のマイクロホンを接続することが可能（マイク補正不要）

プレスリリース

- 2005年10月、神戸製鋼と共同でBSSモジュールをプレスリリース
- 各社新聞紙、雑誌（日経ビジネス等）、テレビにて報道された。

テレビ東京　ワールドビジネスサテライト2005年10月24日

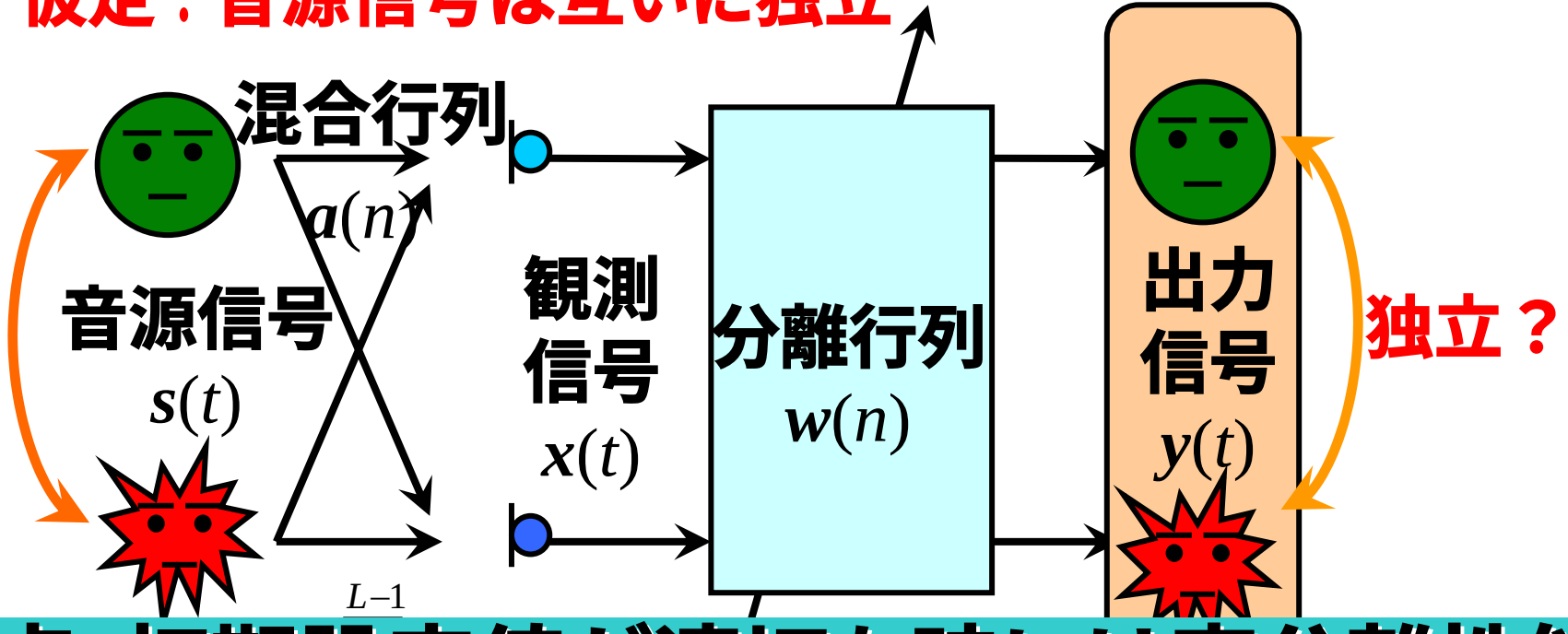
本技術開発の背景・目的

- ＢＳＳの実用化における重要なファクタ
  - 演算量（少ないほど高速であり実用的）
  - 分離性能（高いほど音声認識率向上に貢献）
- 今までの音源分離技術
  - 処理は重いが音質は良い（例：独立成分分析）
  - 処理は軽いが音質はそれなり（例：バイナリマスク）

リアルタイム処理可能な軽さと
高音質を兼ね備えた分離技術が目標

従来技術1：独立成分分析 (ICA)

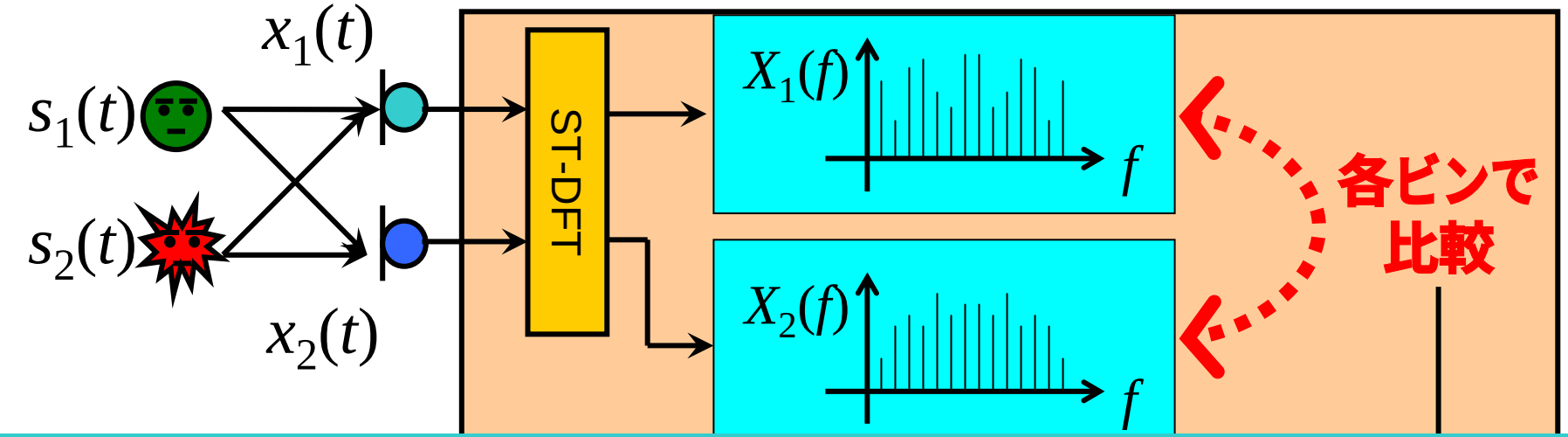
仮定：音源信号は互いに独立



利点：初期設定値が適切な時には高分離性能

欠点：初期値に敏感、演算量多い

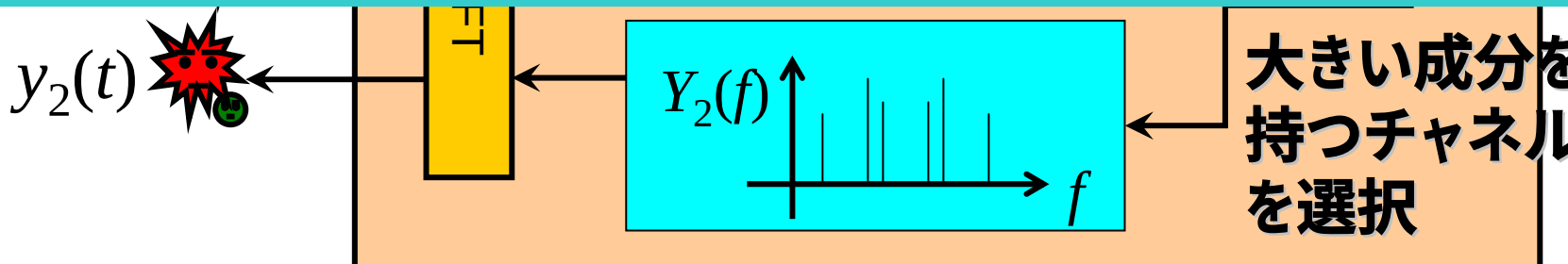
従来技術2：Binary Mask Processing



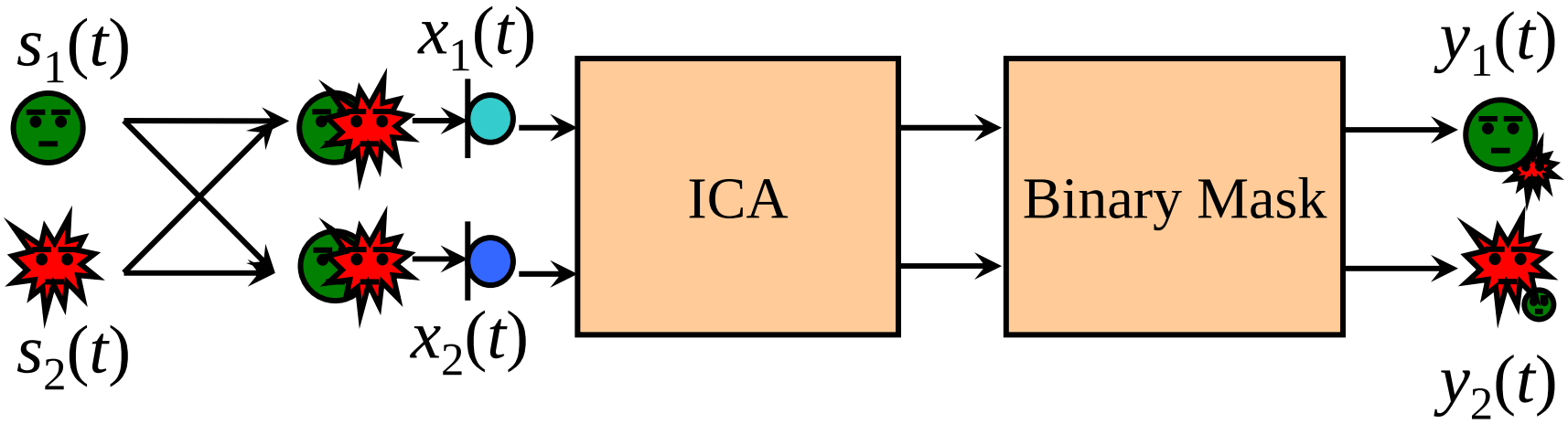
利点：演算量が少なく実時間処理向き

欠点：音源の配置に制約（左右に広がっている必要あり）

音源自体に制約（同一周波数帯域にて重複なし）



従来法3：単純接続



問題点：

ＩＣＡはＳ１、Ｓ２それぞれを個別に出力

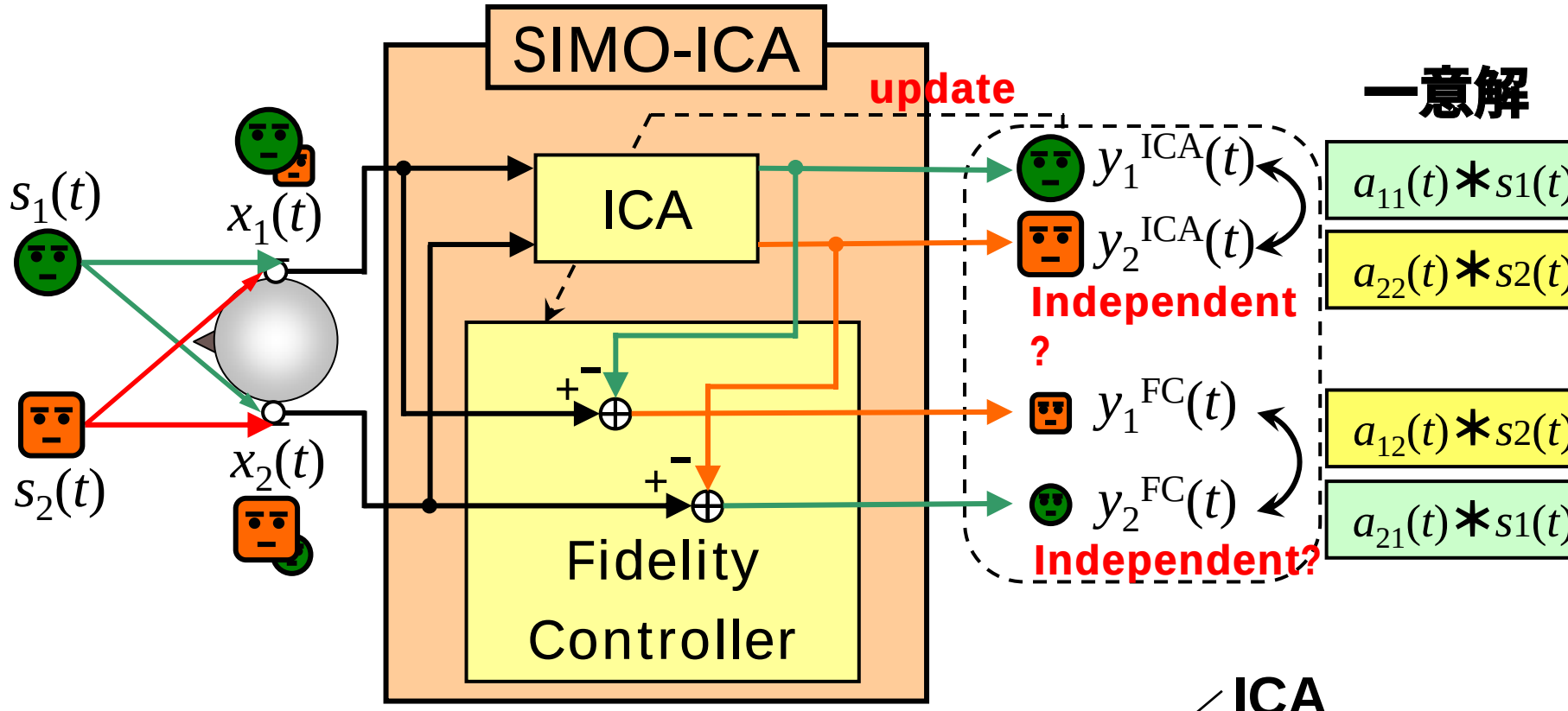
同一周波数帯域内に2つの信号が零でない成分を持つ場合
バイナリ・マスクは正しく判定・抽出できない

従来の音源分離技術のまとめ

	欠点	利点
従来法 1 Binary Mask	◆音源信号間にスパース性が必要 ◆歪が生じる	◆高速動作
従来法 2 ICA	◆反復学習に時間がかかる	◆スパース性不要 ◆歪が少ない
従来法 3 ICA + Binary Mask	◆スパース性が必要 ◆高速動作に不向き ◆歪が生じる	

スパース性を必要とせず、なおかつ歪の少ない高速ブラインド音源分離手法が必要

新しいICA: SIMO-ICA (高谷, 猿渡 他 2003)



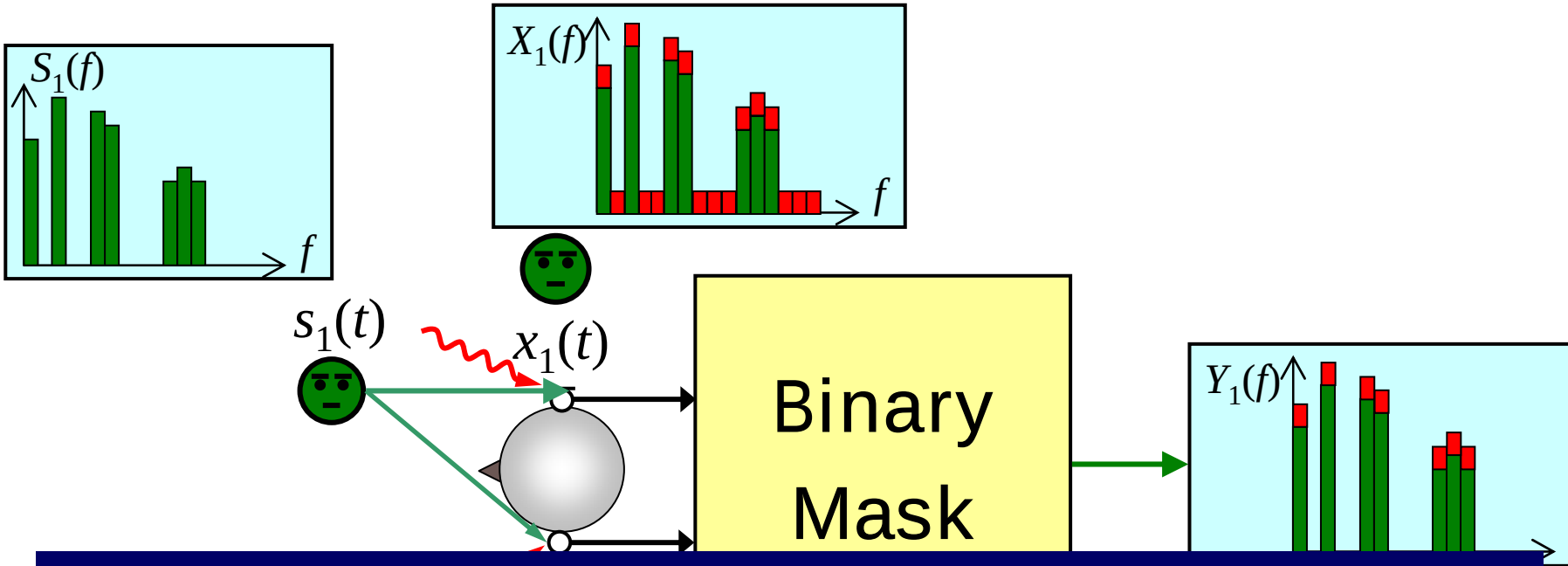
ICA

$$\mathbf{W}_{(\mathrm{ICA}l)}^{[i+1]}(f) = \mathbf{W}_{(\mathrm{ICA}l)}^{[i]}(f) - \eta\left[\text{off-diag}\left\langle \phi\left(\mathbf{Y}_{(ICAl)}^{[i]}(f,t)\right)\mathbf{Y}_{(ICAl)}^{[i]}(f,t)^{\mathrm{H}}\right\rangle_m\right]\mathbf{W}_{(\mathrm{ICA}l)}^{[i]}(f)$$

FC

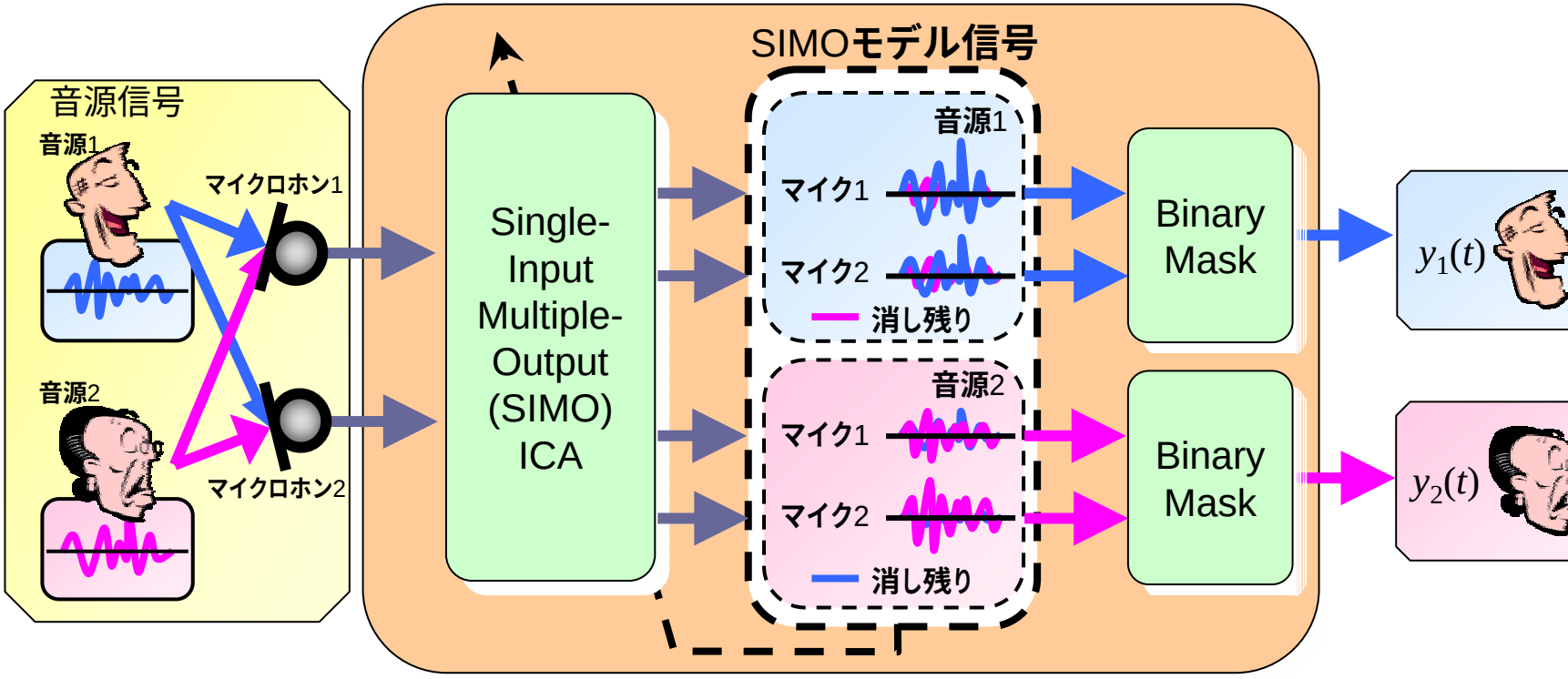
$$-\text{off-diag}\left\{\left\langle \phi\left(\mathbf{X}(f,t)-\sum_{k=1}^{L-1}\mathbf{Y}_{(ICAk)}^{[i]}(f,t)\right)\left(\mathbf{X}(f,t)-\sum_{k=1}^{L-1}\mathbf{Y}_{(ICAk)}^{[i]}(f,t)\right)^{\mathrm{H}}\right\rangle_m\right\}\left(\mathbf{I}-\sum_{k=1}^{L-1}\mathbf{W}_{(ICAk)}^{[i]}(f,t)\right)$$

着眼点：SIMO信号に対するバイナリマスク？



このようなバイナリマスク処理は、目的音成分が不在の時間-周波数帯域において、加法的（ただし少量）な残留雑音成分を除去することが出来る。

提案法：二段BSSアルゴリズム

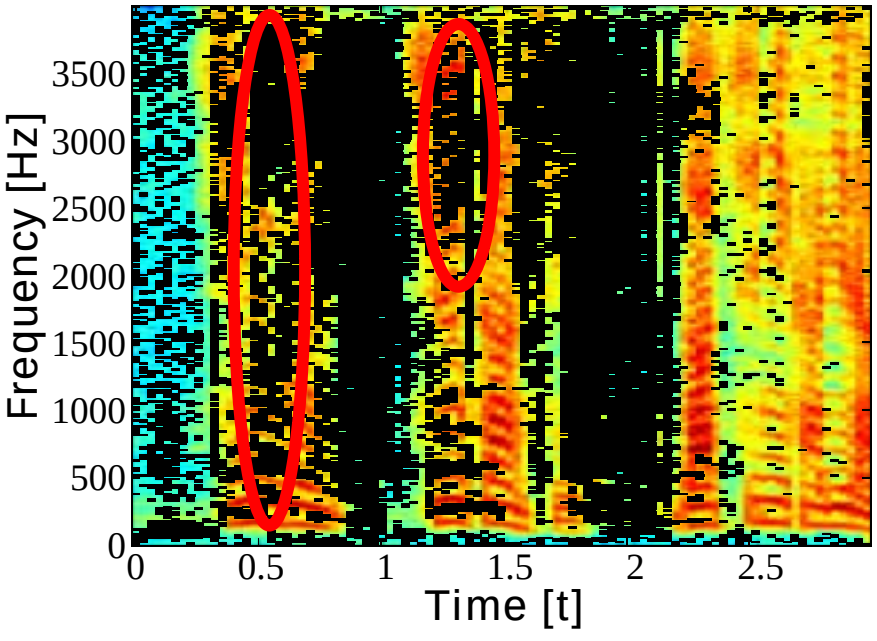


- 前段に Binary Mask にとって都合の良い信号を出力するSIMO-ICA
- 後段に処理の高速な Binary Mask
  - Binary mask が必要とする前提条件を ICA により解決
  - ICA の学習不足により生じる消し残りを Binary Mask により除去
  - 前段ＩＣＡは間欠的に稼動し、後段バイナリマスクは実時間で稼動
    → トータルではリアルタイムに動く

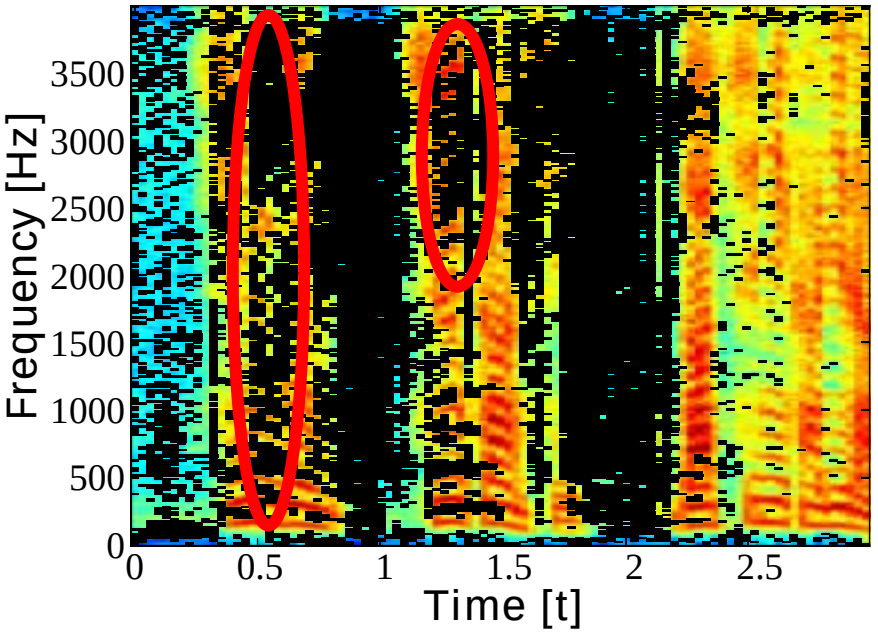
時間-周波数上でのマスク動作 (単純接続)

■マスク(信号を除去)

$Y_1^{ICA1}(f,t)$

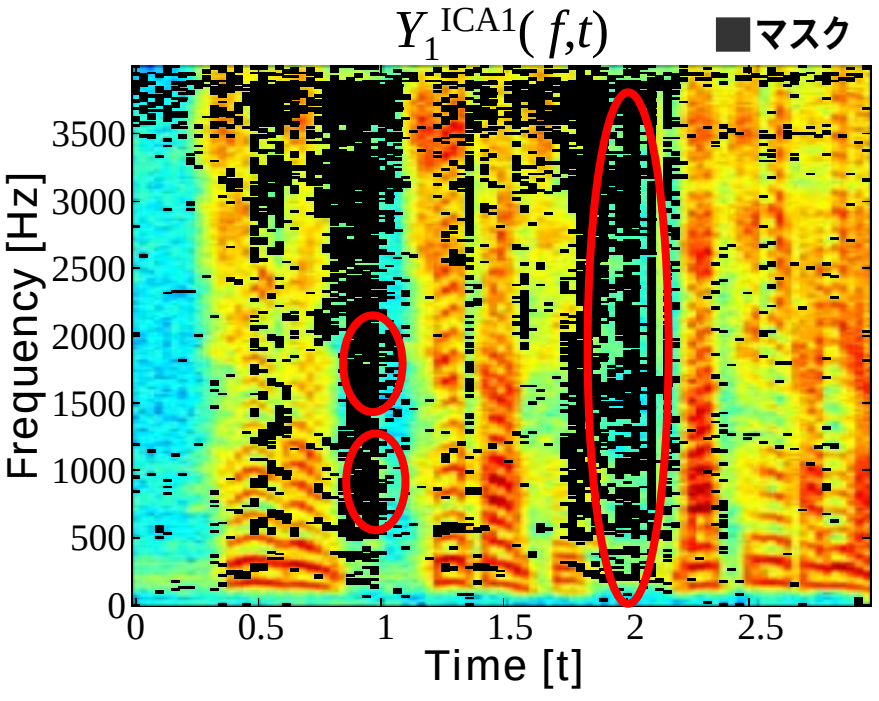


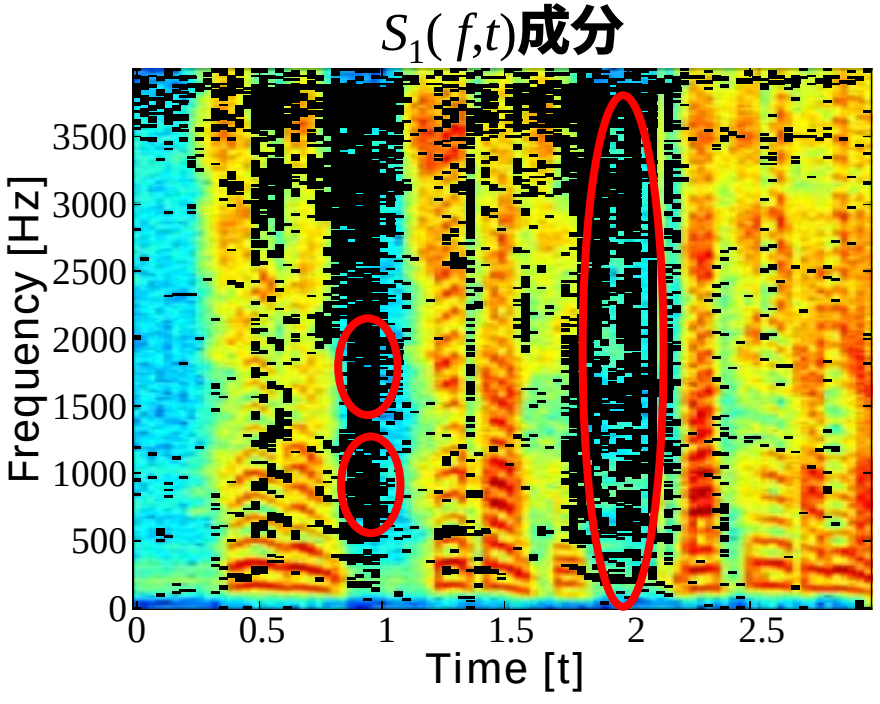
$S_1(f,t)$成分



S_1 成分が大きくマスクされている
⇒分離音に歪みが発生

時間-周波数上でのマスク動作（提案法）



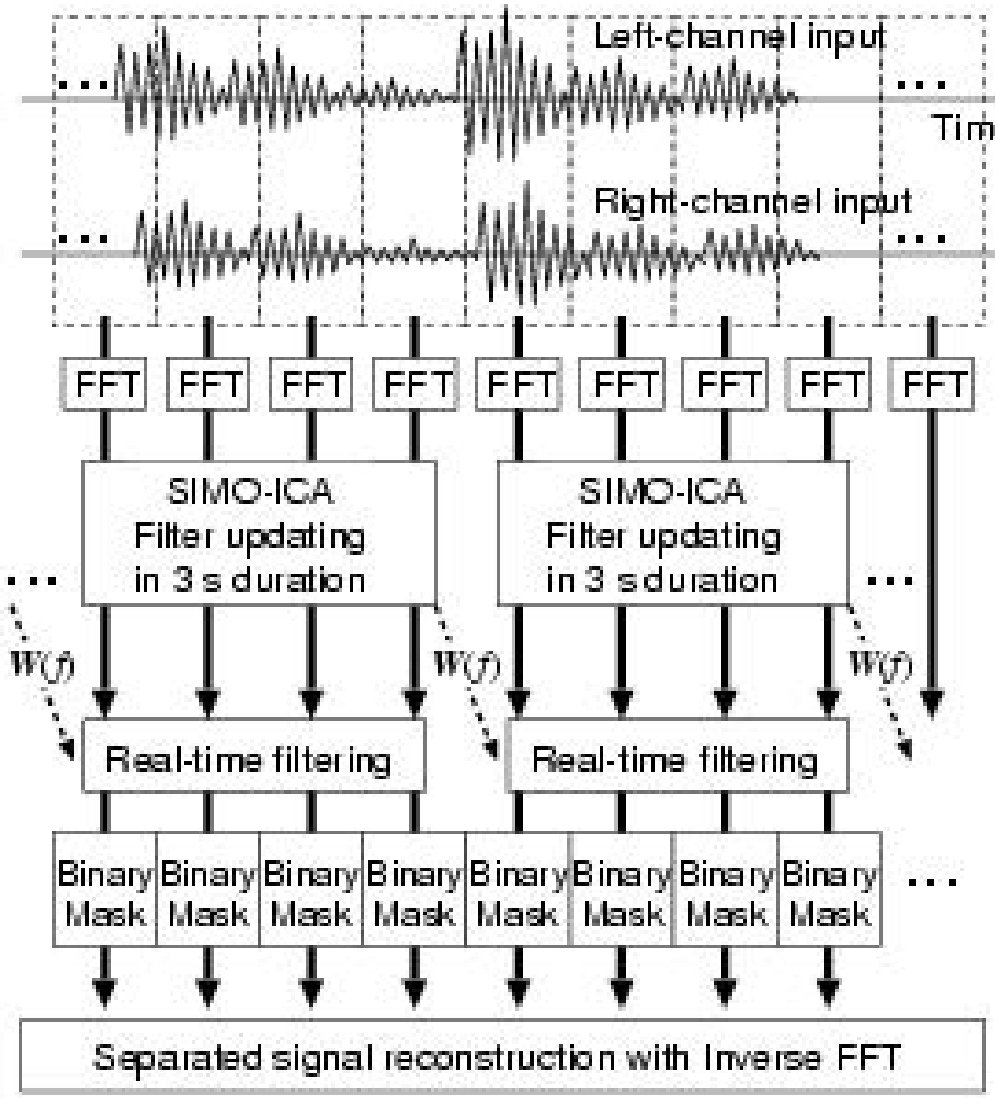


※消し残りを効率的にマスクしている

※目的音に対する歪は少ない

リアルタイム実装

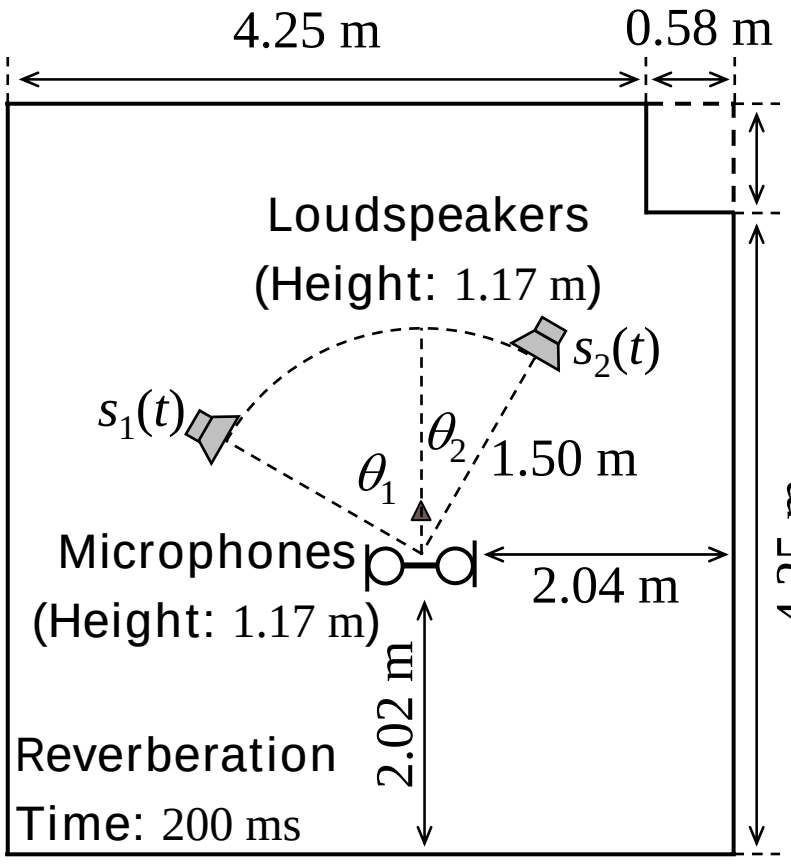
- **実際の処理フロー**
  - **ICA部は処理量が多いので、数秒に一回のみフィルタを更新する。過去において最適な係数を現在のフレームに適用**
  - **バイナリマスク部は常時稼動し、ICA部で生じた誤差成分を除去する。**



トータルとして、精度を落とすことなくリアルタイム性を確保可能

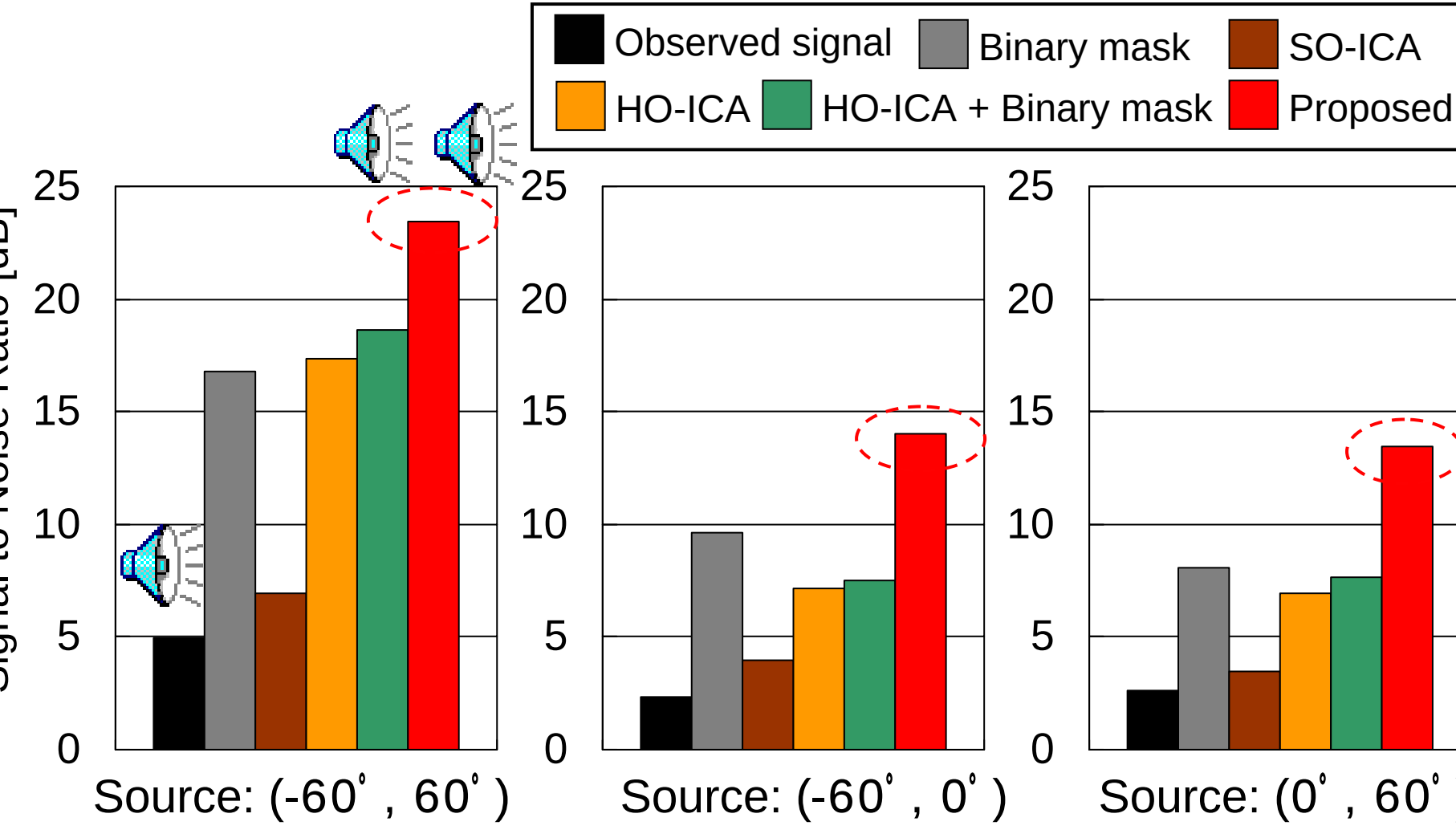
実験条件

Reverberation	200 ms
Sources	Speech/Stationary Noise (3 seconds)
Sampling frequency	8 kHz
Filter length	Binary mask: 512 taps ICA: 1024 taps
Source DOA (θ_1, θ_2)	(-60°, 60°), (-60°, 0°), (0°, 60°)
Initial filter	NBF steered to (-15 °, 15 °)
Evaluation score	Signal to Noise Ratio [dB]



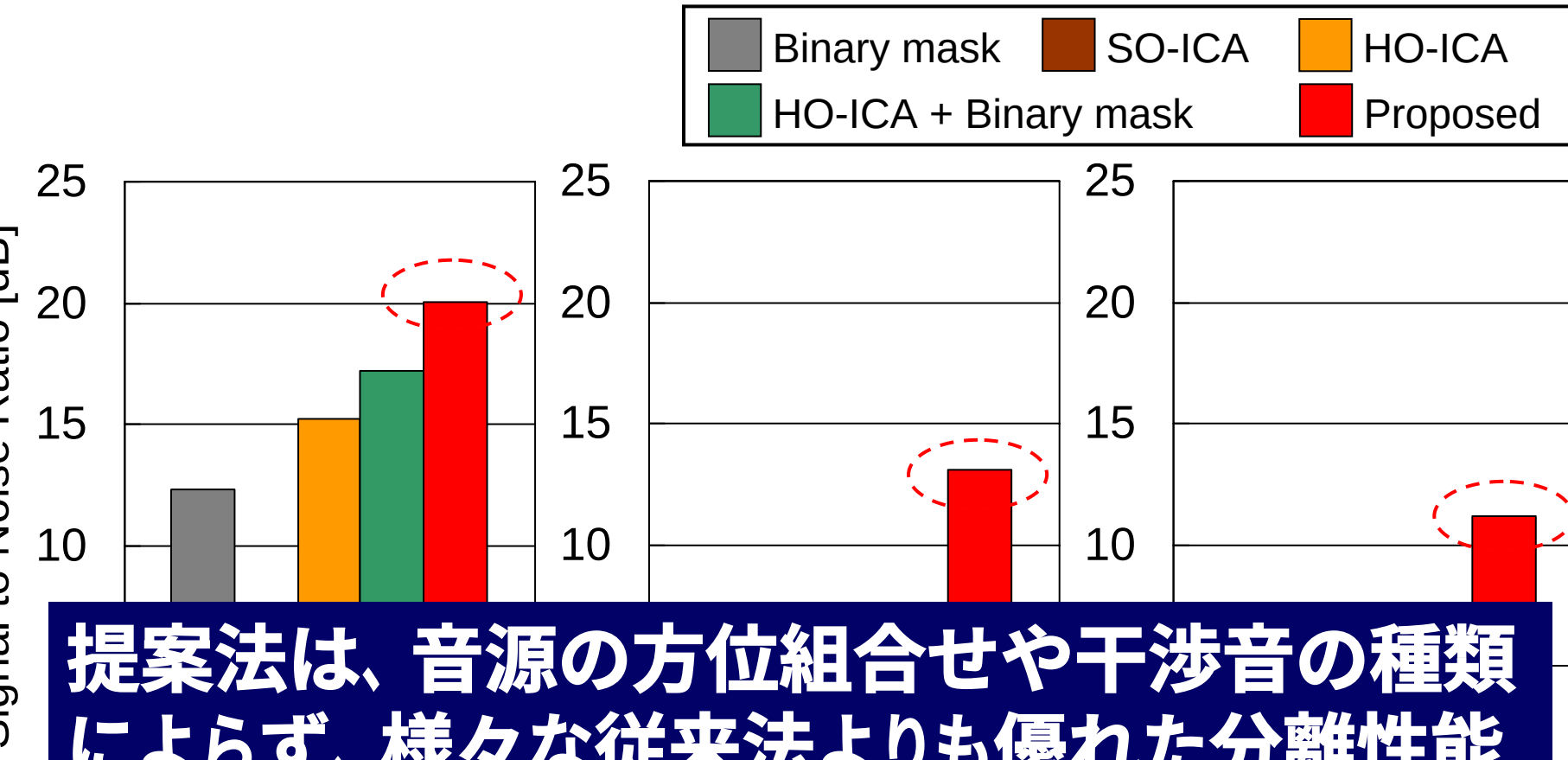
音声＆音声の分離結果

- 全て12通り話者組合せの平均値



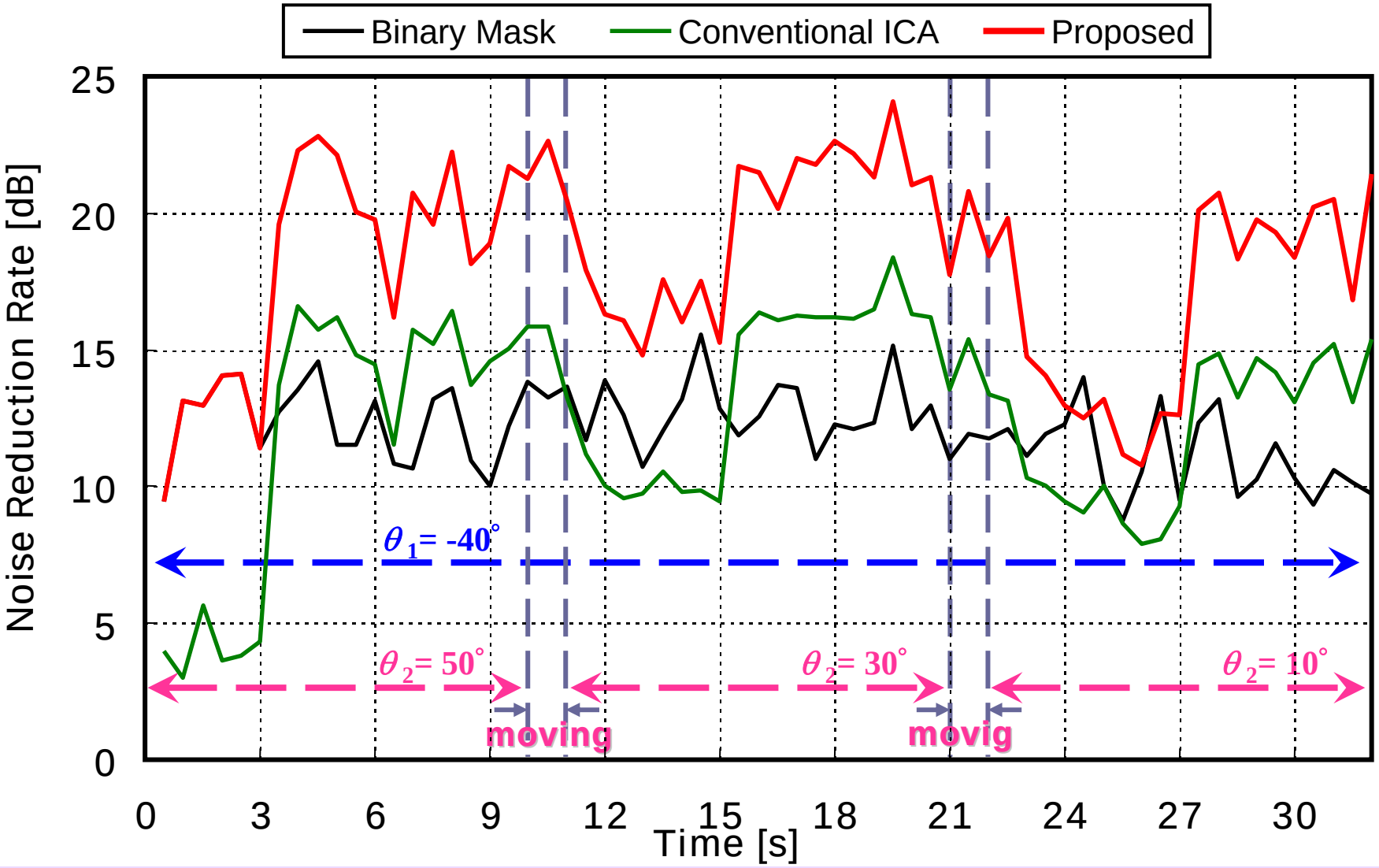
音声＆雑音の分離結果

・全て12通り話者組合せの平均値



提案法は、音源の方位組合せや干渉音の種類によらず、様々な従来法よりも優れた分離性能を示している。

移動音源の分離結果



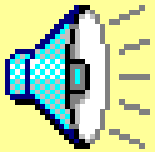
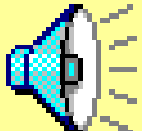
移動音源に関しても提案法は優れた分離性能を示す

IEEE MLSP2007 音源分離コンペ

- 未知の環境で計測された音楽と音声信号の混合音がWEBで公開され、それを分離するタスク
- 世界中（北米、イギリス、フランス、ドイツ、北欧、日本）から参加応募があり、「誰のBSSアルゴリズムが世界で最も性能が良いか？」を競った。

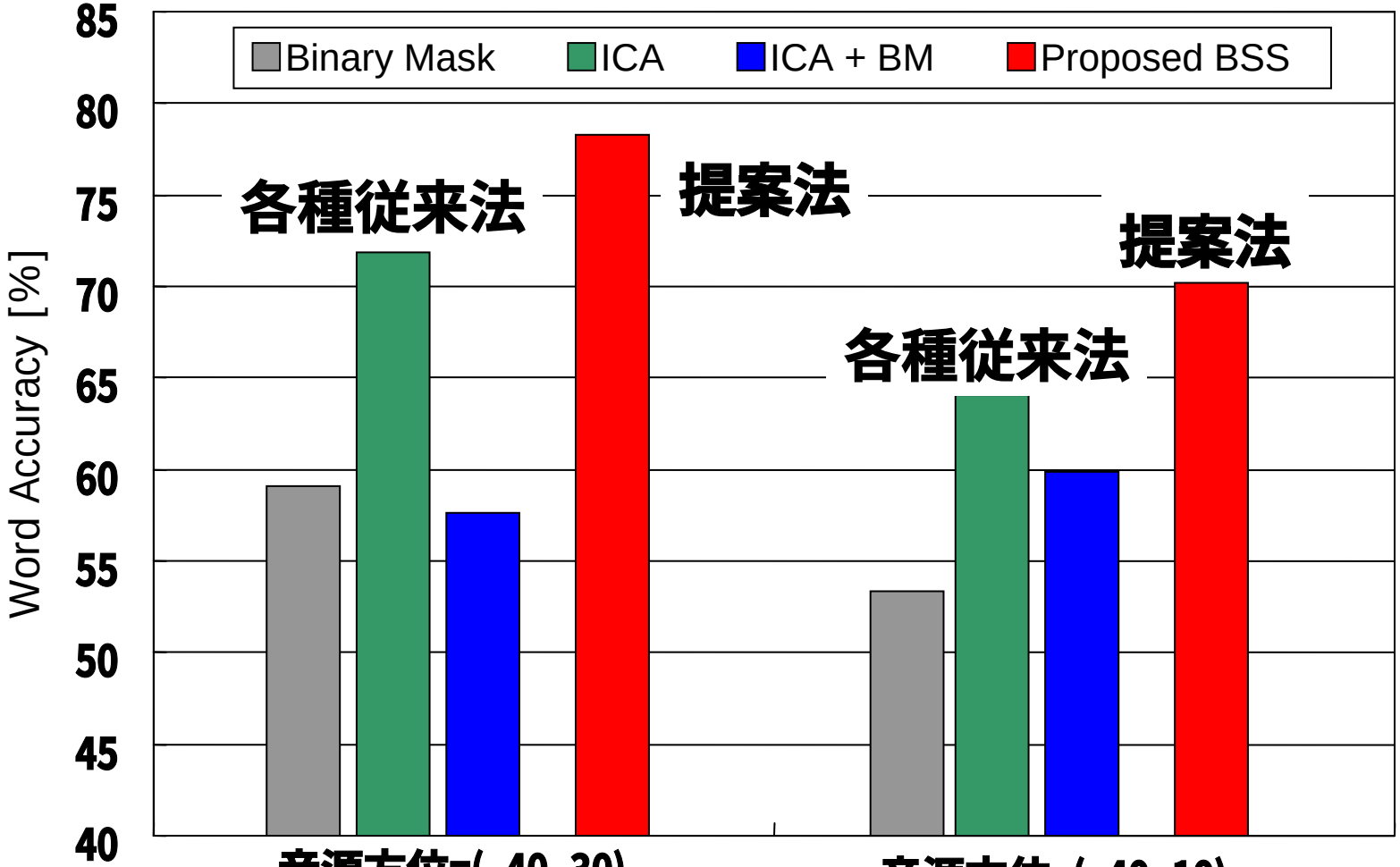
- 我々の「二段BSS法」が**一等賞**を受賞した。

  提出した分離結果:

  - 分離音声 
  - 分離音楽 

音声認識による評価 (音声＆音声)

- 大語彙音声認識タスクによる評価 (JNASデータベース、JULIUS (PTM)による認識、音響モデルはクリーンモデル)
- 本実験においては16 kHzサンプリングデータを取り扱った



音源分離機能付きロボットヘッドの試作

- ダミーヘッドとBSSを組み合わせることにより構成
  (2005年　NAISTと神戸製鋼所の共同開発)
- 離れた場所のユーザ音声を認識する対話システム
  (音声案内システム) の前処理としてデモ可能

音楽が鳴り響いている中でも対話可能

複数の話者がいても対話可能

まとめ

- 従来BSSの問題点について概説
- 新しいICA(SIMO-ICA) に基づく二段BSS手法を紹介
  - バイナリマスク処理と組み合わせれば、実時間動作を可能としながらも高い分離性能を得ることができる
  - 汎用DSPによる実装例を紹介
  - 音声認識結果、対話システムへの導入例を紹介
- 現在、様々な実環境 (例：暗騒音のある通常室内、駅など) における評価および改良を実施中

音源分離マイク（実用化バージョン）

- 神戸製鋼所より2007年から販売
- 標準で１６ kHzサンプリングをサポート
- すでに「音監視システム」などへの導入実績あり