NTT

NTT Communication Science Laboratories

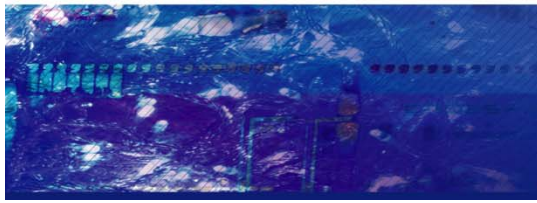
音学シンポジウム 2013年5月12日

音声生成過程と信号観測過程のモデルに基づくマルチチャンネル音声強調

中谷智広

NTTコミュニケーション科学基礎研究所

Contributed by NTT CS研 信号処理研究グループのメンバ
伊藤信貴，吉岡拓也，木下慶介，荒木章子，藤本雅清，エスピ ミケル，
デルクロア マーク，久保陽太郎，小川厚徳，堀貴明，中村篤



目次

- 研究のモチベーション
  - 実世界で目的音声を聞き分け理解する技術：音声強調+音声認識
- 目的音声を聞き分ける技術（=音声強調）
  - 『生成モデル』の考え方に基づく音声強調
  - 代表的な音声強調手法
- 音声強調＋音声認識の応用例
  - 生活雑音環境下における遠隔発話音声認識
  - 複数人会話の音声認識

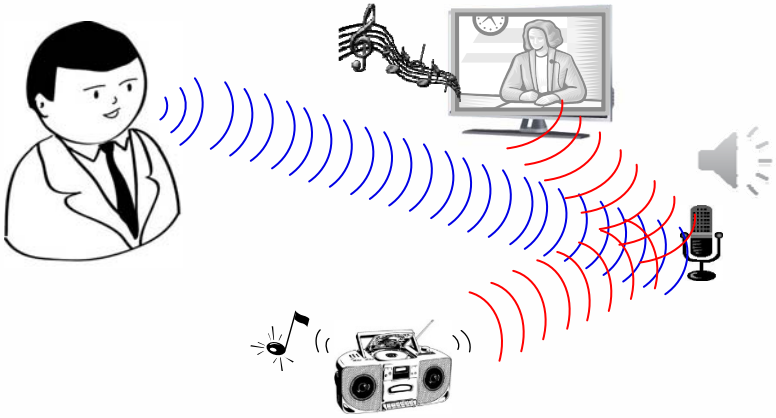
音声インタフェース　―現状と未来―

▪ 現状

例：ボイスサーチ，
しゃべってコンシェル

- マイクの近くで話す
- 雑音が少ない

▪ 未来



どんな場所でも、話している内容を理解できる

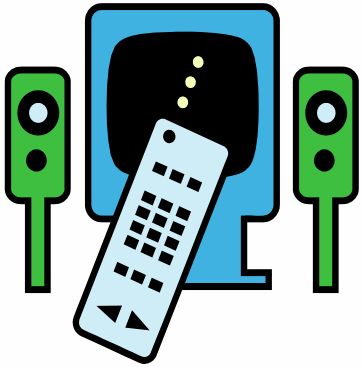
- マイクから離れてもOK
- 目的音声以外に音があってもOK

研究のモチベーション

様々な音が聞こえる実世界で，
人の声を聞き分け・理解する
コンピュータの実現

- 人どうしなら不自由なく会話できる環境での音声の認識（利用環境を選ばない技術）
- 音声を聞き分ける能力（音声強調）
  ＋ロバスト音声認識／話し言葉認識

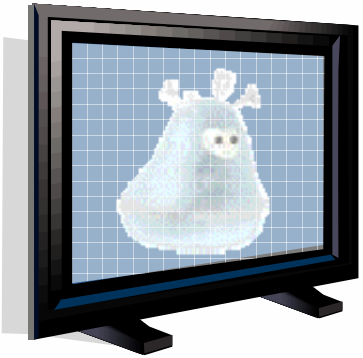
音声を聞き分け理解する技術の応用先



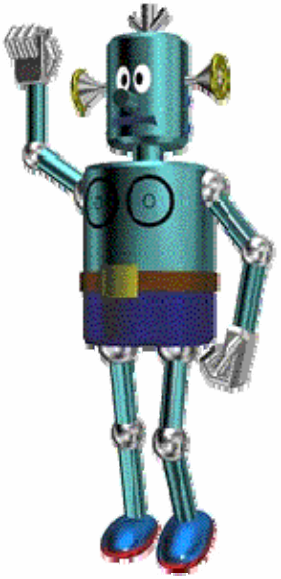
家電・住宅機器の操作

会議録の
自動作成

字幕の
自動付与



インタラクティブ・エージェント



共生型ロボット

音声(言語)は人間にとって最も基本的な情報伝達手段

本日のトーク：二つの研究シナリオ

- **生活雑音環境下での遠隔コマンド音声認識**
  - 人がコンピュータに話す音声の認識
  - 応用例：カーナビ，家電・ゲーム機の操作…

観測信号　　音声強調結果

→ **人間に比肩する音声認識性能**

- **複数人会話の音声認識**
  - 人と人の会話，誰がいつ何を話したのかを認識
  - 応用例：議事録・要約，オンライン会話支援…

複数人会話の音声認識
(複数人・遠隔・自由会話音声認識)

観測信号

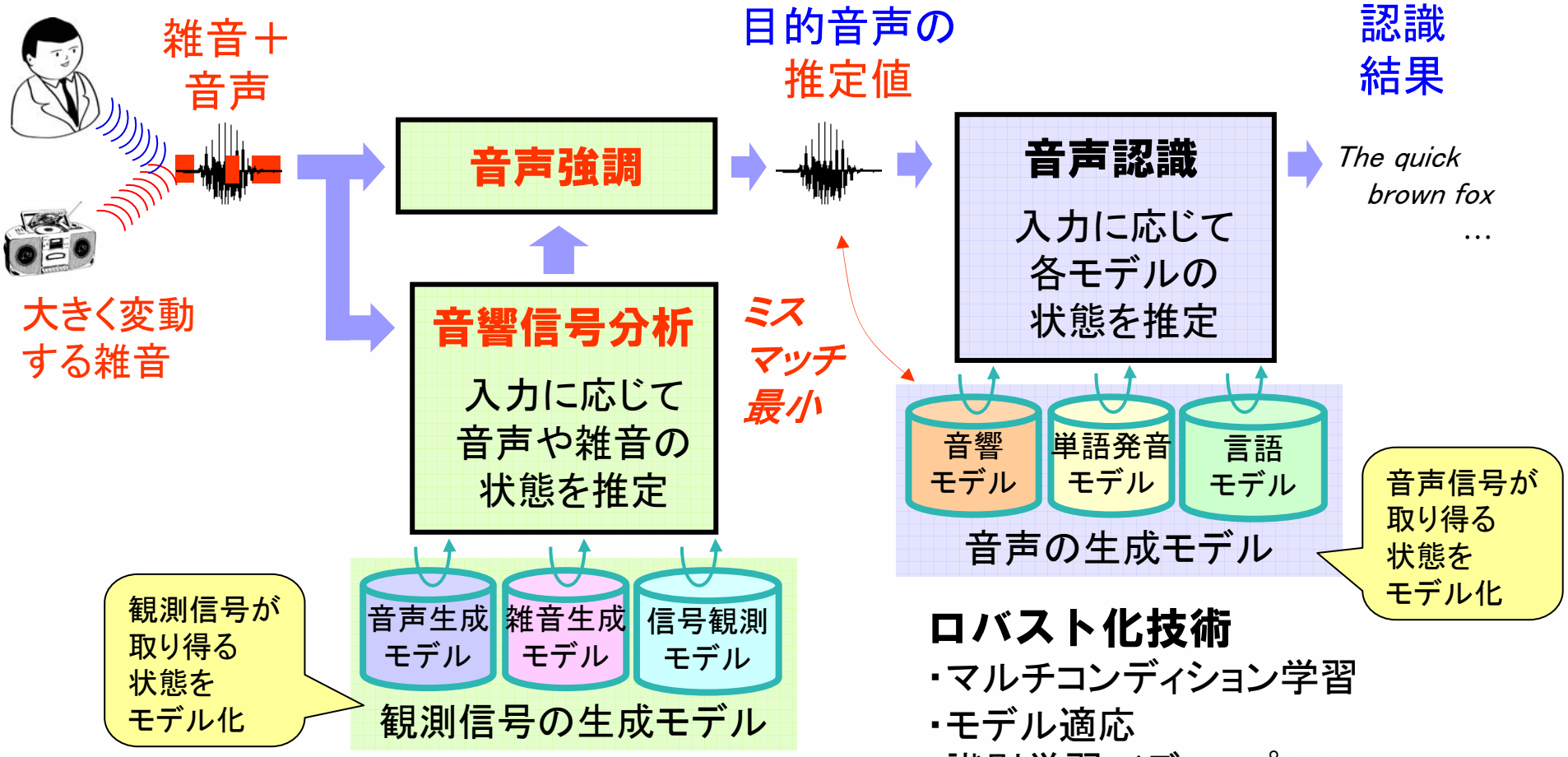
音声強調
結果

テーブルマイクを用いてヘッドセット並の音声認識性能

目次

- 研究のモチベーション
  - 実世界で目的音声を聞き分け理解する技術：音声強調+音声認識
- 目的音声を聞き分ける技術（=音声強調）
  - 『生成モデル』の考え方に基づく音声強調
  - 代表的な音声強調手法
- 音声強調＋音声認識の応用例
  - 生活雑音環境下における遠隔発話音声認識
  - 複数人会話の音声認識

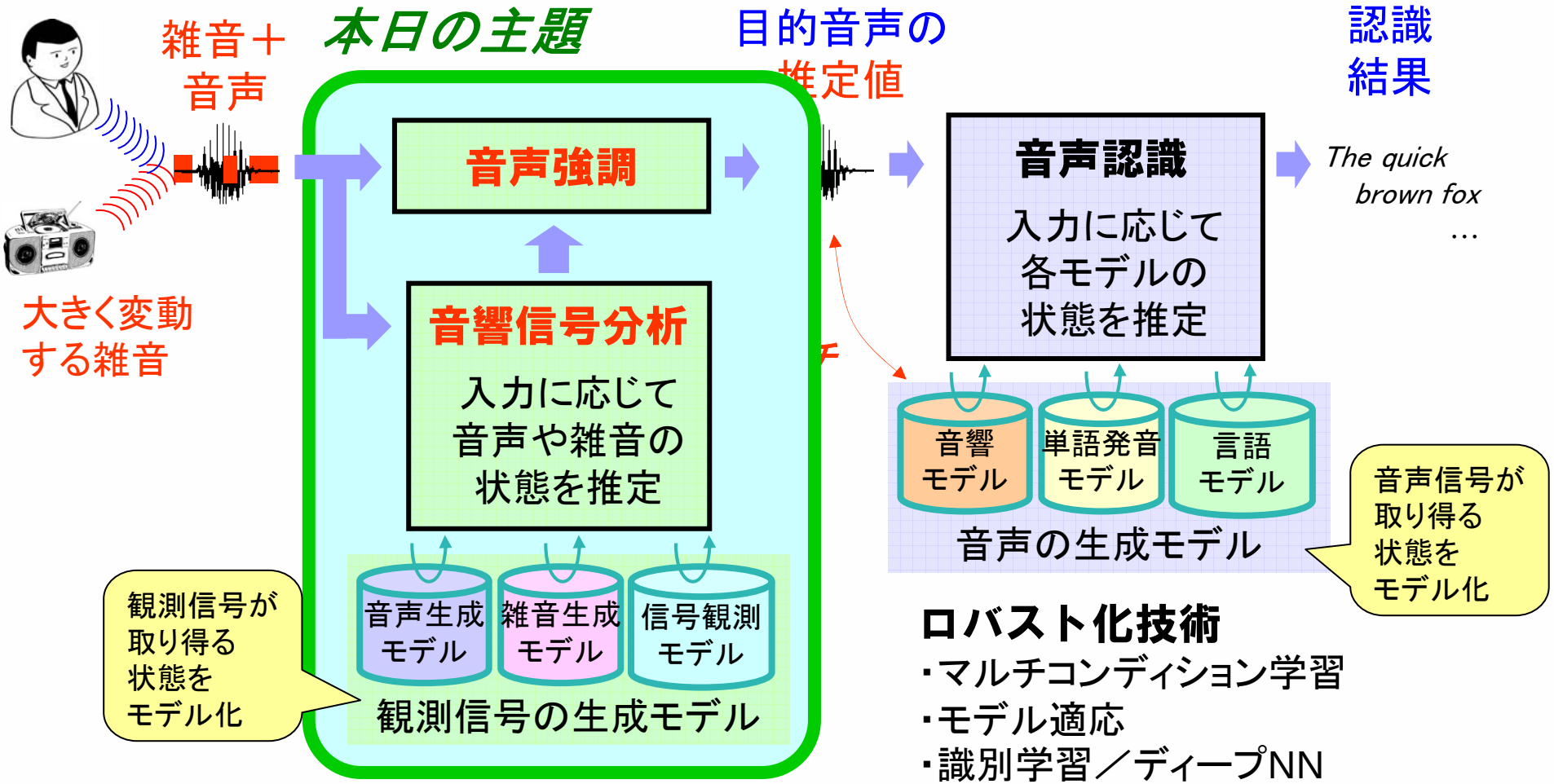
生成モデルの考え方に基づく音声強調＋音声認識



ロバスト化技術

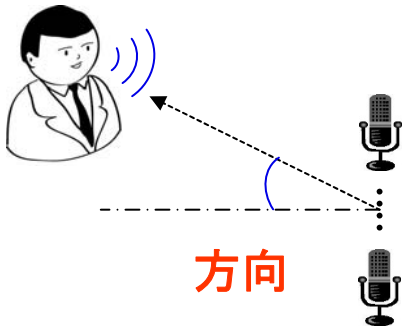
- マルチコンディション学習
- モデル適応
- 識別学習／ディープNN

生成モデルの考え方に基づく音声強調＋音声認識



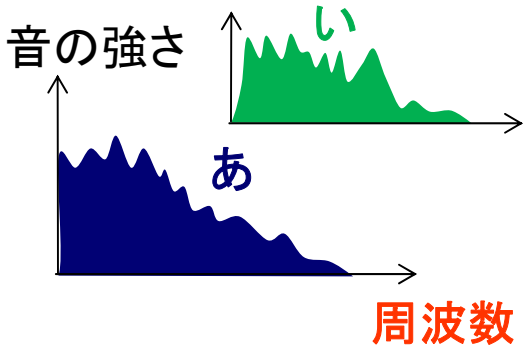
生成モデルで表現する音声・雑音の状態の例

聞こえる方向
（空間情報）



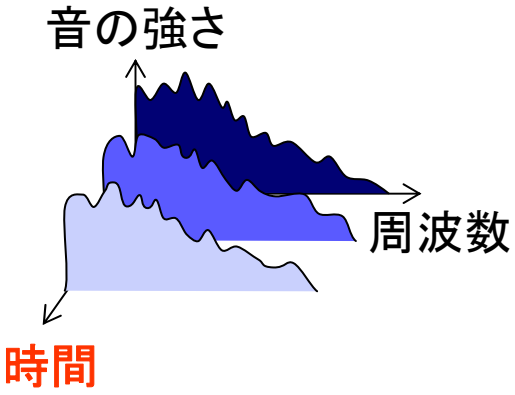
特定の**方向**から聞こえる

スペクトルパタン
（周波数情報）



音素（「あ」「い」など）ごとに、特有の**スペクトルパタン**

時間変化パタン
（時間情報）



特有の**時間変化パタン**を持つ

その他：音声区間 [藤本, 2008]，調波構造 [亀岡, 2007], F0パタン [亀岡, 2010], 発話内容 [Rennie, 2010] など

生成モデルに基づく音響信号分析の考え方 (1/2)

分析すべき音シーン　：単一音源（雑音なし）の場合

音声生成モデル
$p(\mathbf{s}|\theta)$

生成（順問題）

逆問題

観測信号の特徴量時系列
$\mathbf{x} = \{x_t\}$

$p(\mathbf{x}|\theta) = p(\mathbf{s} = \mathbf{x}|\theta)$

θ ：音声の状態（音源方向，スペクトルパタンなど）を表すパラメータ

$p(\mathbf{s}|\theta)$：音声特徴量に関する確率密度関数

最尤推定*1

$$\hat{\theta} = \arg\max_{\theta} p(\mathbf{x}|\theta)$$

*1: 他にも，ベイズ推定，MAP 推定など

生成モデルに基づく音響信号分析の考え方 (1/2)

分析すべき音シーン　：単一音源（雑音なし）の場合

音声生成モデル
$p(\mathbf{s}|\theta)$

生成（順問題）

逆問題

観測信号の特徴量時系列
$\mathbf{x} = \{x_t\}$

$p(\mathbf{x}|\theta) = p(\mathbf{s} = \mathbf{x}|\theta)$

最尤推定*1

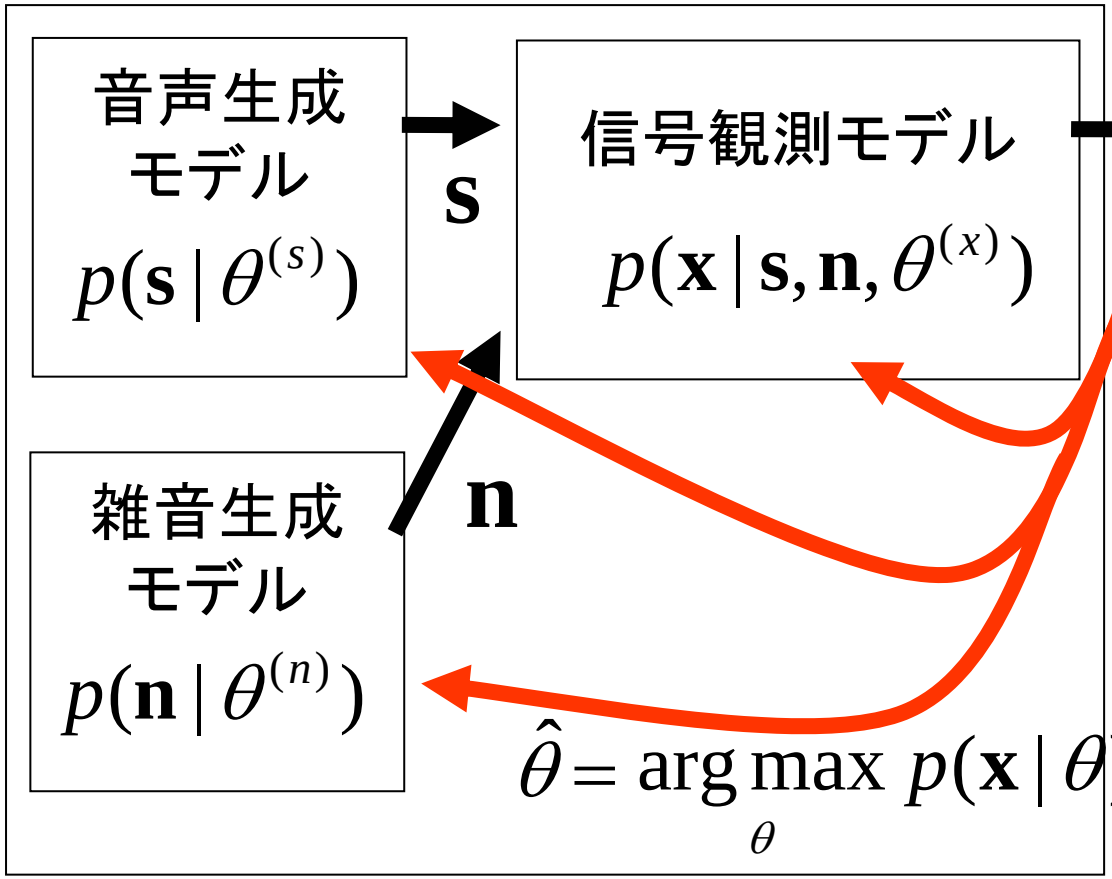
$\theta)$

代表的な応用事例：
線形予測に基づく最尤スペクトル推定 [板倉 1968]
隠れマルコフモデルによる音声モデル [Juang 1985他]
他多数

定，

生成モデルに基づく音響信号分析の考え方 (2/2)

分析すべき音シーン　：複数音源の場合



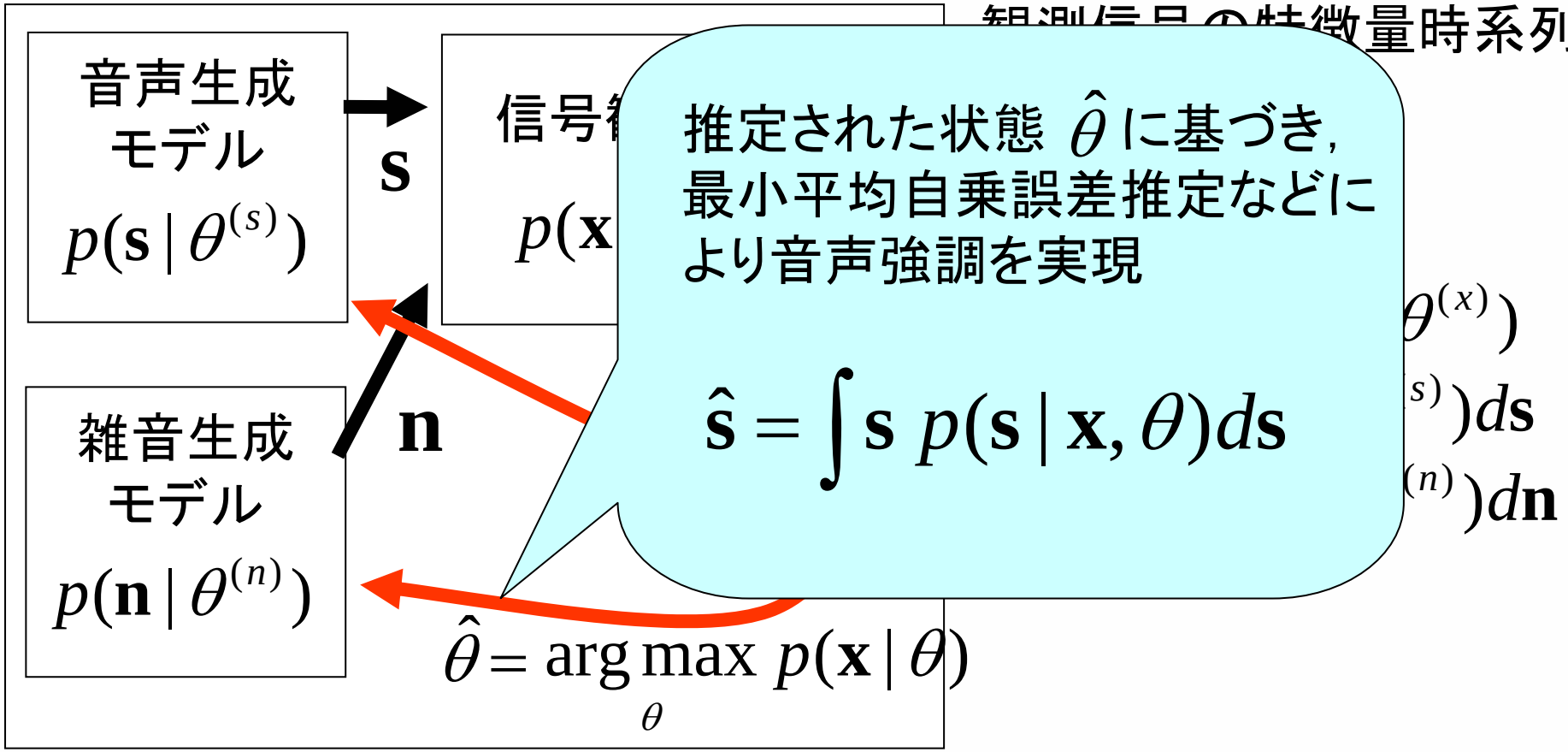
観測信号の特徴量時系列

$$\mathbf{x} = \{x_t\}$$

$$p(\mathbf{x}|\theta) = \iint p(\mathbf{x}|\mathbf{s},\mathbf{n},\theta^{(x)}) \times p(\mathbf{s}|\theta^{(s)})d\mathbf{s} \times p(\mathbf{n}|\theta^{(n)})d\mathbf{n}$$

生成モデルに基づく音響信号分析の考え方 (2/2)

分析すべき音シーン　：複数音源の場合



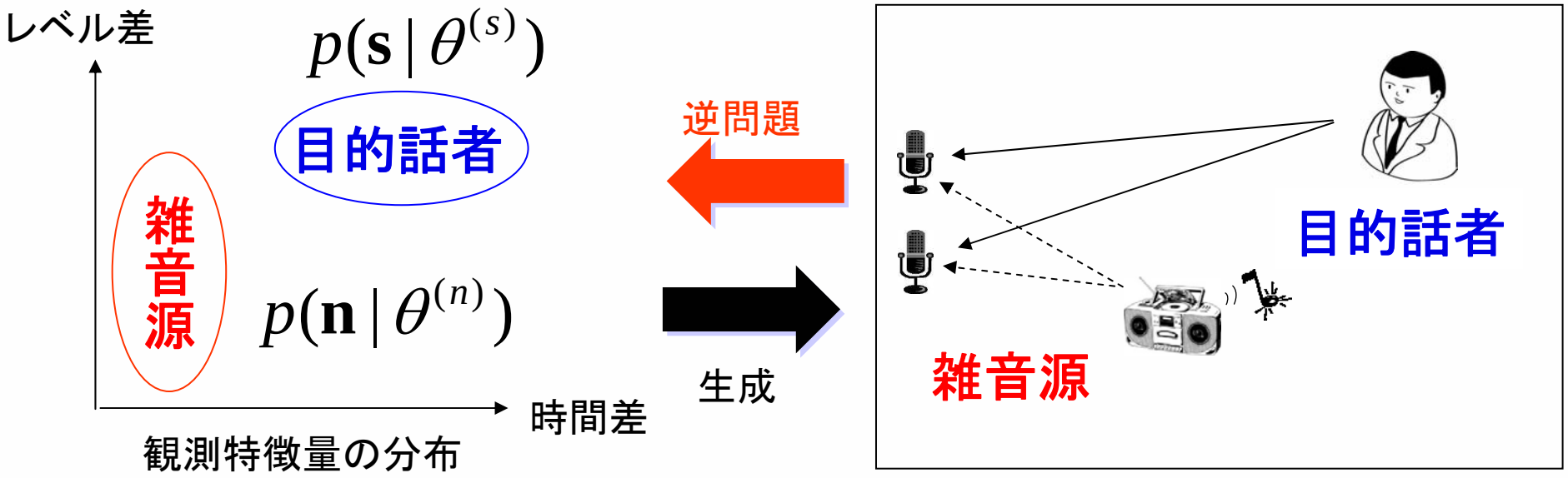
生成モデルによる音響信号分析：特徴・課題

- 音源の状態から観測信号を予測する観測信号生成モデルの構築は，比較的容易（**順問題**）
  - **様々な手がかり**を導入できる
- 音響信号分析は，観測信号生成モデルのパラメータ推定問題として定式化できる（**逆問題**）
  - ただし，多くの場合，非線形最適化問題となり，**効率的に解けるとは限らない**

生成モデルに基づく代表的な音声強調の手法

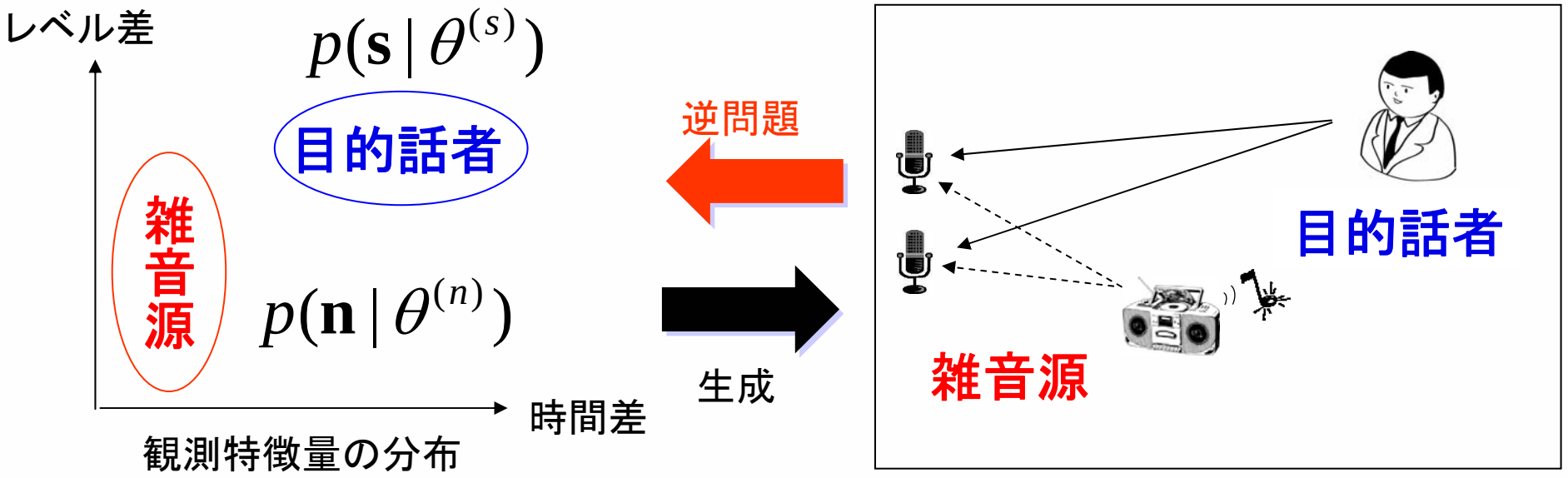
手がかり	音源の方向	スペクトルパタン	時間変化パタン
代表的な手法	・独立成分分析（ICA） ・音源方向クラスタリング	・非負値行列分解 (NMF) ・Factorial モデル	・隠れマルコフモデル (HMM) ベース ・事例ベース
特長	生成モデルの事前学習が不用 →ブラインド処理が可能	生成モデルの事前学習が必要 →音声・雑音のパタンの分布をあらかじめ制限できる	

音源方向クラスタリングに基づく音声強調 [Yilmaz, 2004]



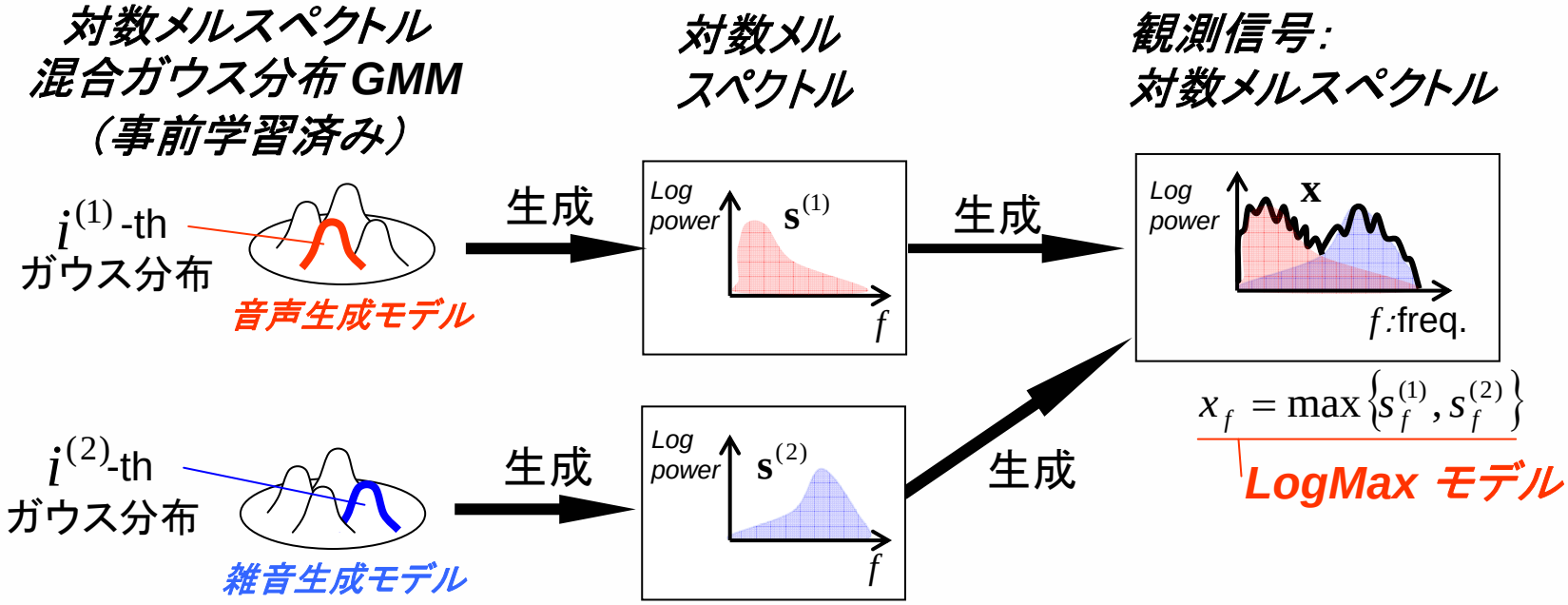
事前に音源方向等が不明でも，観測信号のみから
生成モデルを推定できる（ブラインド推定可能）

音源方向クラスタリングに基づく音声強調 [Yilmaz, 2004]



音源分離技術評価キャンペーンSiSEC 等において，
世界トップレベルの性能（信号対歪み比 SDR 等）
を達成 ([澤田 2012] 他)

Factorial モデルに基づく音声強調 [Roweis 2003]

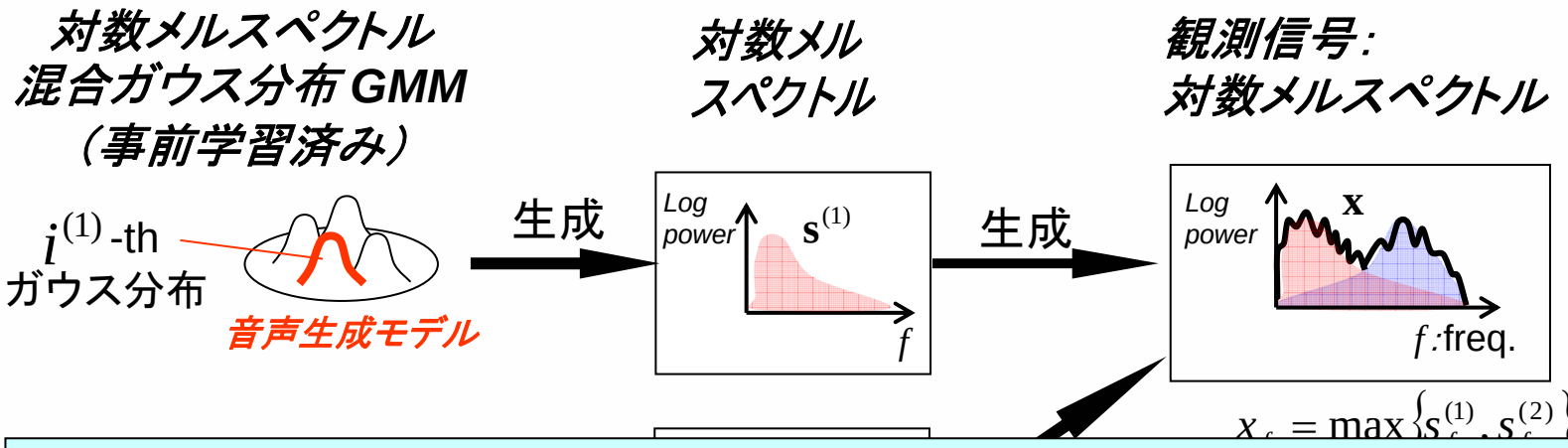


音響信号分析（逆問題） $\mathbf{i} = \arg\max_{\mathbf{i}} p(\mathbf{x}, \mathbf{i})$

観測信号を生成する確率値の高いガウス分布の組 $\mathbf{i} = \{i^{(1)}, i^{(2)}\}$ を推定

⇒ EMアルゴリズムにより効率的な推定が可能

Factorial モデルに基づく音声強調 [Roweis 2003]



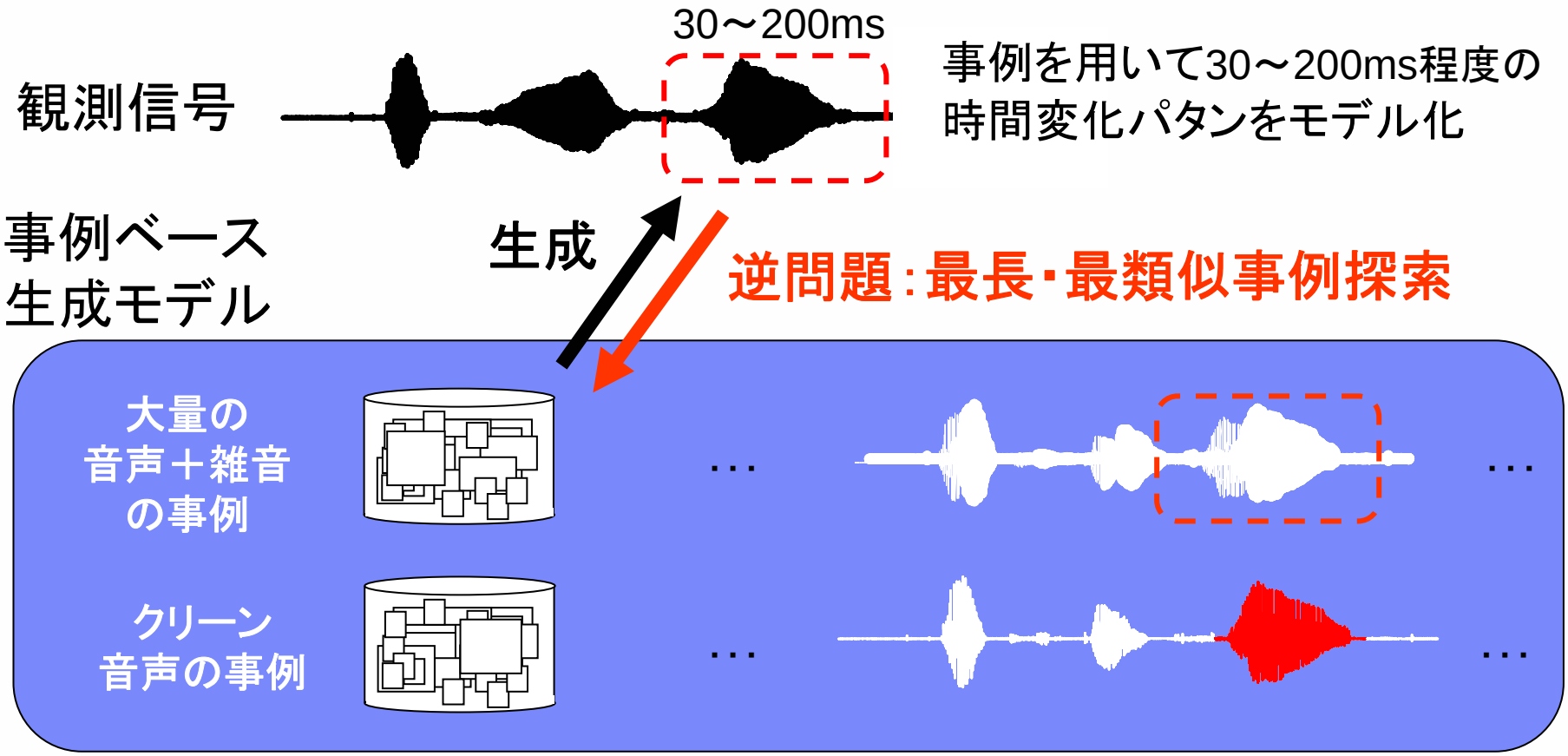
モノラル混合音声の分離・**音声認識**評価キャンペーン
[Cooke 2010]においてトップスコアを達成 [Rennie 2010]

$$\mathbf{i} = \arg\max_{\mathbf{i}} p(\mathbf{x}, \mathbf{i})$$

観測信号を生成する確率値の高いガウス分布の組 $\mathbf{i} = \{i^{(1)}, i^{(2)}\}$ を推定

⇒ EMアルゴリズムにより効率的な推定が可能

事例ベース音声強調 [Ming, 2011, Kinoshita, 2011]

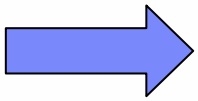


クリーン事例に基づきウィナフィルタを設計し音声強調を実現

生成モデルに基づく音声強調法のまとめ

- 音源の方向，スペクトルパタン，時間変化パタンなどに基づき，様々な効果的な方法が提案されている

	音源方向クラスタリング	Factorial モデル／事例ベース
長所	・ブラインド処理が可能 →ボトムアップ的処理に有利	・スペクトルパタン，時間変化パタンの分布をあらかじめ規定できる →音声認識との相性○
短所	・同じ方向の音を区別できない ・音声認識との相性？	・雑音と音声のパタンが近いと区別が困難 ・事前学習→収録環境への依存性大

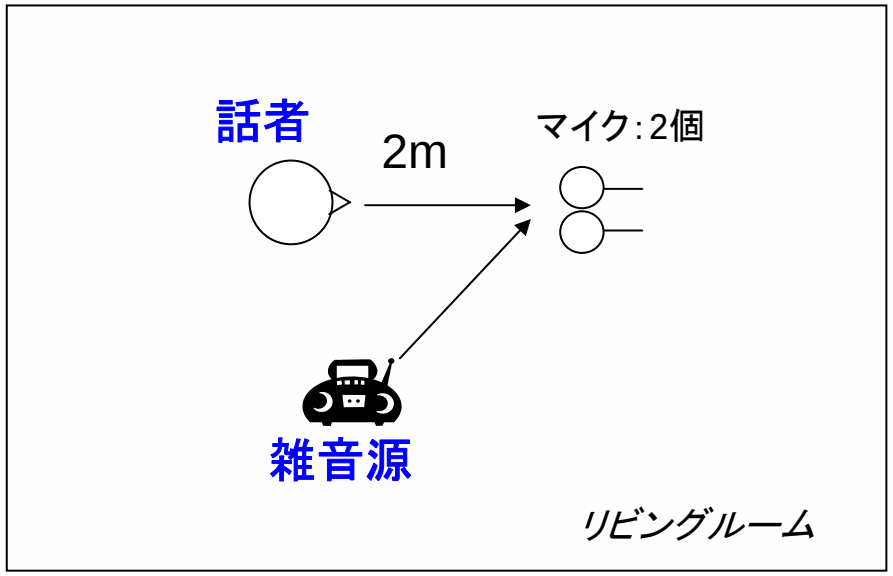


各生成モデルの長所を活かすことで，
より信頼性の高い音声強調が実現できる期待

目次

- 研究のモチベーション
  - 実世界で目的音声を聞き分け理解する技術：音声強調+音声認識
- 目的音声を聞き分ける技術（=音声強調）
  - 『生成モデル』の考え方に基づく音声強調
  - 代表的な音声強調手法
- 音声強調＋音声認識の応用例
  - 生活雑音環境下における遠隔発話音声認識
  - 複数人会話の音声認識

生活雑音環境下での遠隔発話音声認識 CHiME タスク



- コマンド文※の音声認識
  - 生活環境で録音された雑音を含む
  - 特定話者音声認識
  - オフライン処理
- マルチマイク（2マイク）録音
  - 話者位置は固定
  - 事前学習データ有（音声・雑音）

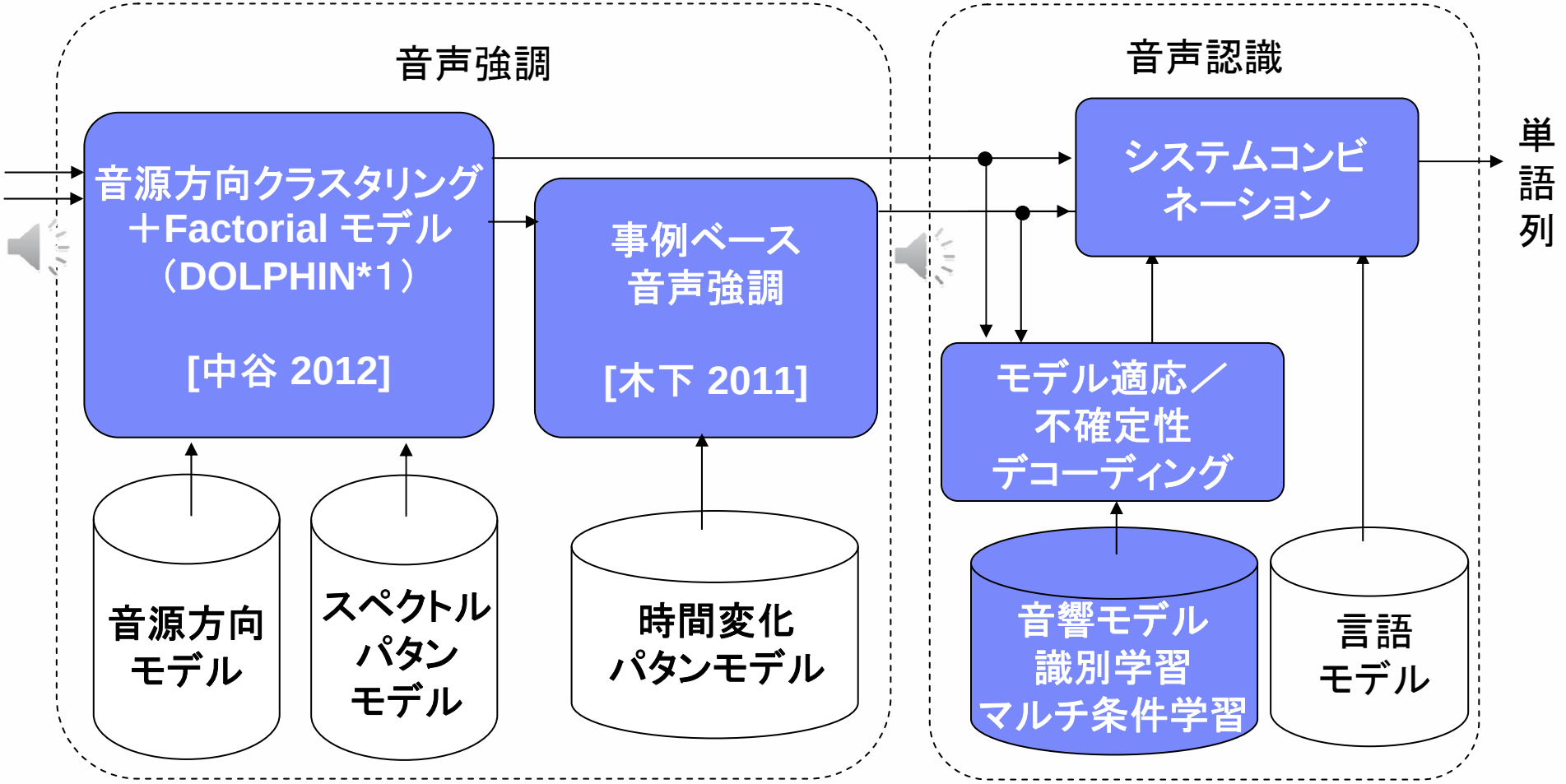
※ コマンド文の例（赤字で示した部分のみの認識率を評価）

<コマンド 4通り><色 4通り><前置詞 4通り>**<アルファベット 25通り><数字 10通り>**<副詞 4通り>

例：“Bin red by F 8 please”
“Lay green with G 4 again”
⋮

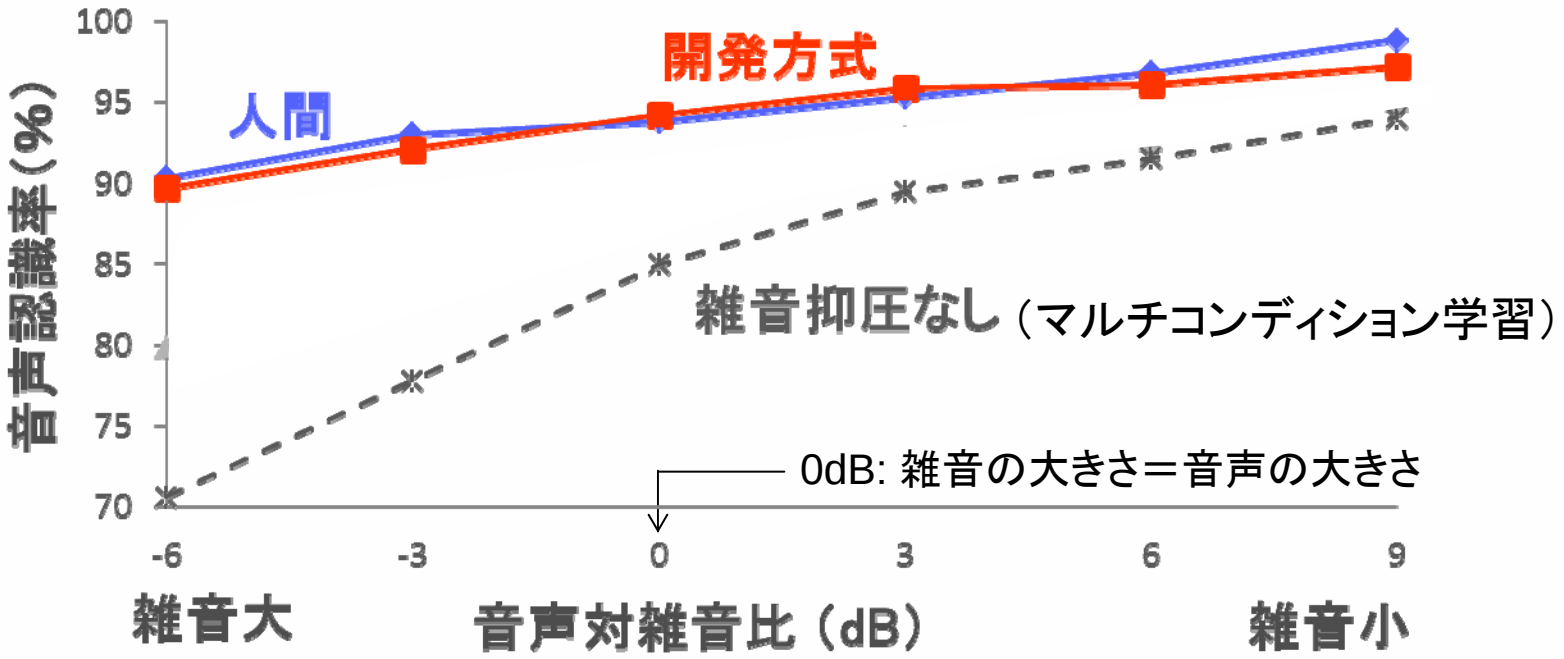
クリーン　観測信号（含雑音）

NTT開発方式の構成



*1) **DO**minance based **L**ocational and **P**ower-spectral c**H**aracteristics **IN**tegration

キーワード認識正解率



人間に比肩する性能を達成（現時点で世界トップ性能）

複数人会話の音声認識
(複数人・遠隔・自由会話音声認識)

- 会話音声認識
  - 比較的フォーマルな自由会話
    例:「仕事でのストレスについて」
  - オフィスルーム／防音室
  - 不特定話者音声認識
  - オフライン処理
- マルチマイク(8マイク)録音
  - 話者数は4(固定)
  - テーブル周囲に着席
    (方向は不明)

NTT開発方式の構成

8ch
観測
信号

音声強調

①音源方向
クラスタリングに
基づくMVDR
[Souden 2012]

②①＋DOLPHIN

③残響除去 [吉岡,
2012]＋ICA
+DOLPHIN

音声認識

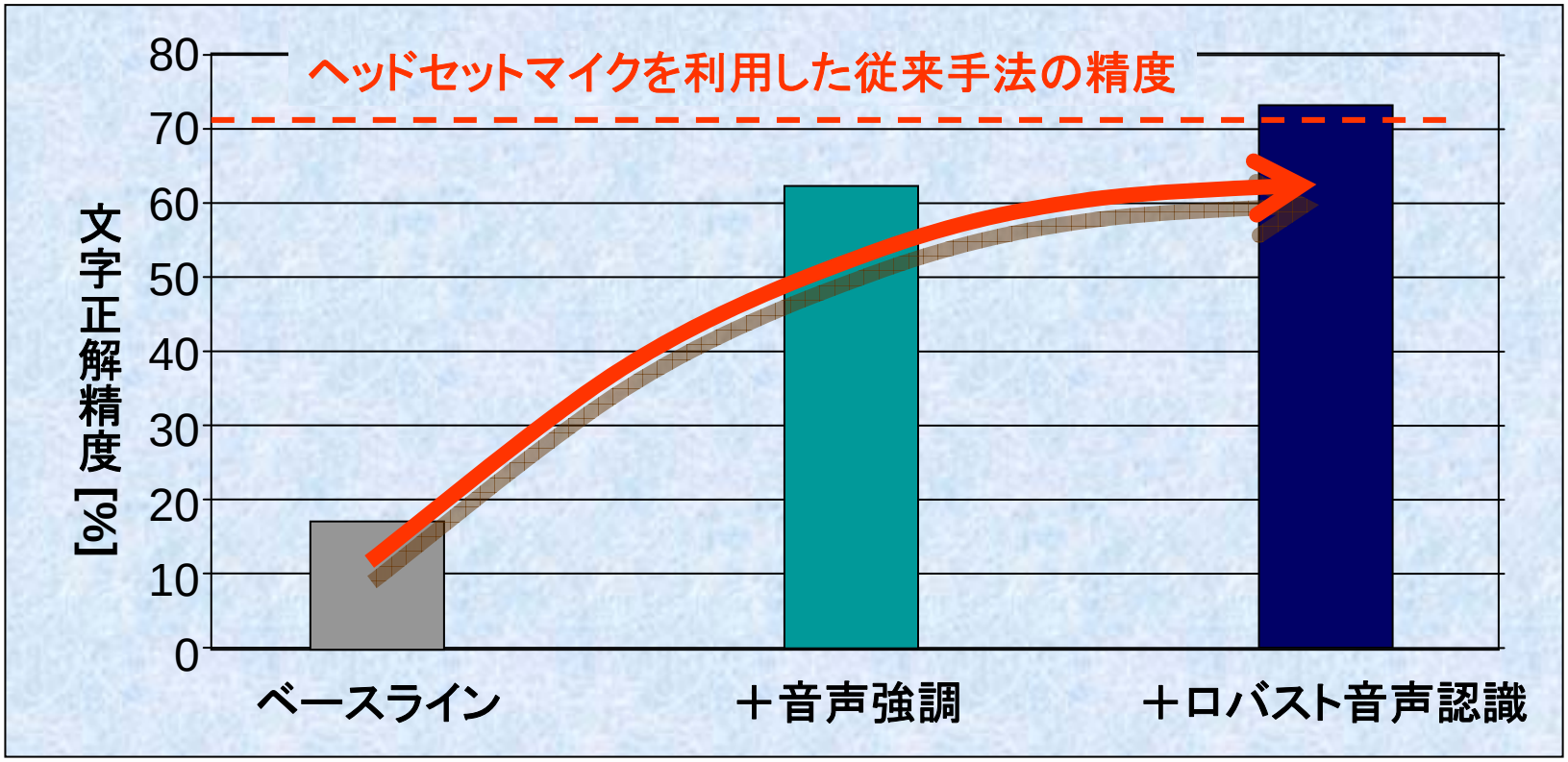
システム
コンビネーション

モデル適応

音響・言語一体モデル
（WFST+識別学習)

Aさん: いいのかなー
Bさん: うーん
Cさん:
…

音声認識性能



ヘッドセットマイクを利用した従来手法を凌ぐ認識精度

まとめ

- 生成モデルに基づき，観測信号中の各構成音の状態を推定し，目的音声を強調するアプローチを紹介
- 音源方向・スペクトルパタン・時間変化パタンに基づく音声強調＋最先端音声認識技術により，実世界音声認識に応用
- 今後の課題
  - 大量の収録データからの教師なし学習
  - モデル構造（音源数・音源の種類など）の自動獲得