

# 音や映像から情報を引き出す

日本電信電話株式会社  
コミュニケーション科学基礎研究所  
主幹研究員・特別研究員

柏野 邦夫

---

2009年7月31日

# 当 研究チームの目指すもの

---

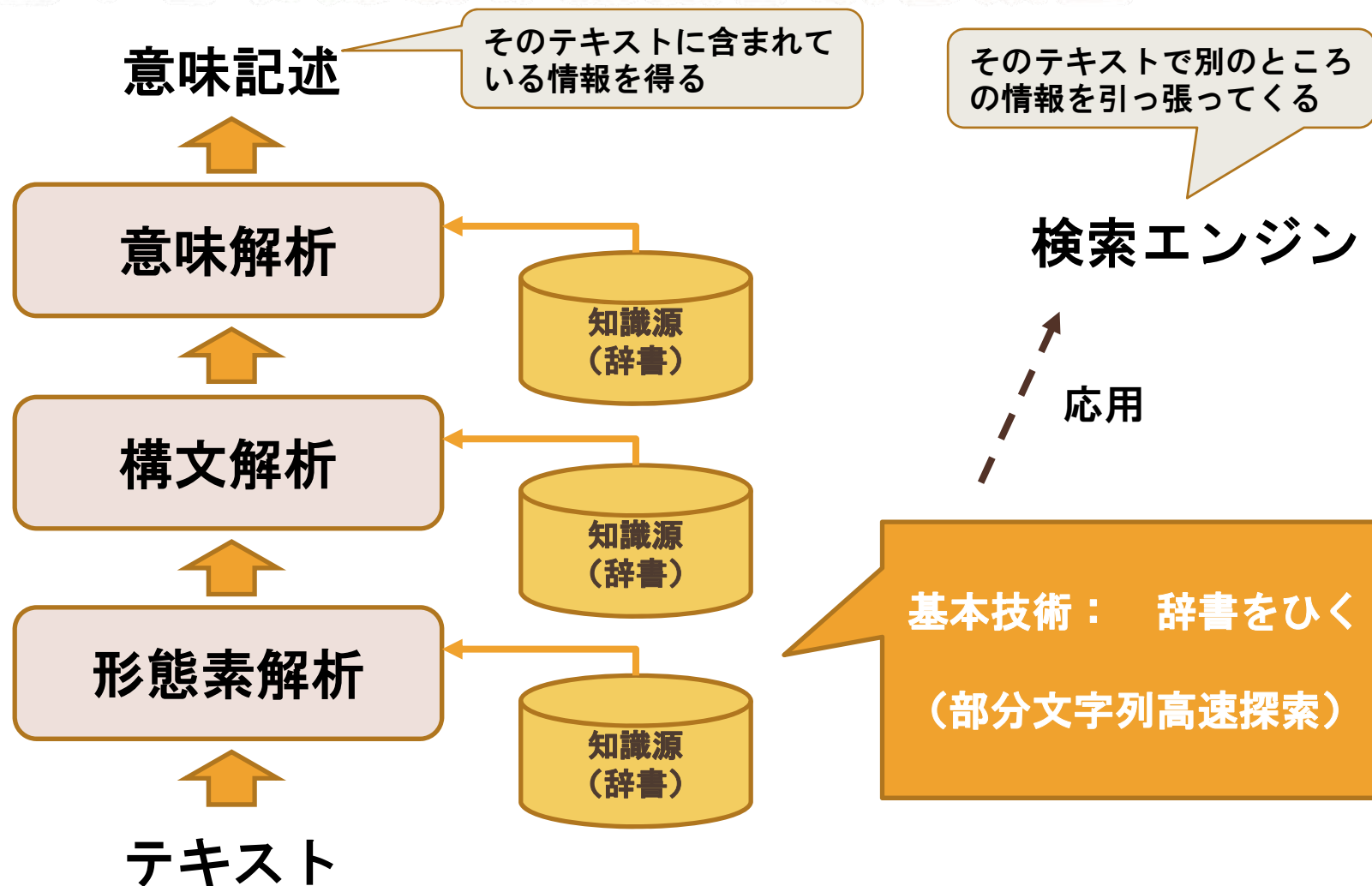
膨大な音・映像メディアを自在に検索し、操り、  
楽しむ環境の実現

(1) メディア処理の体系的建設

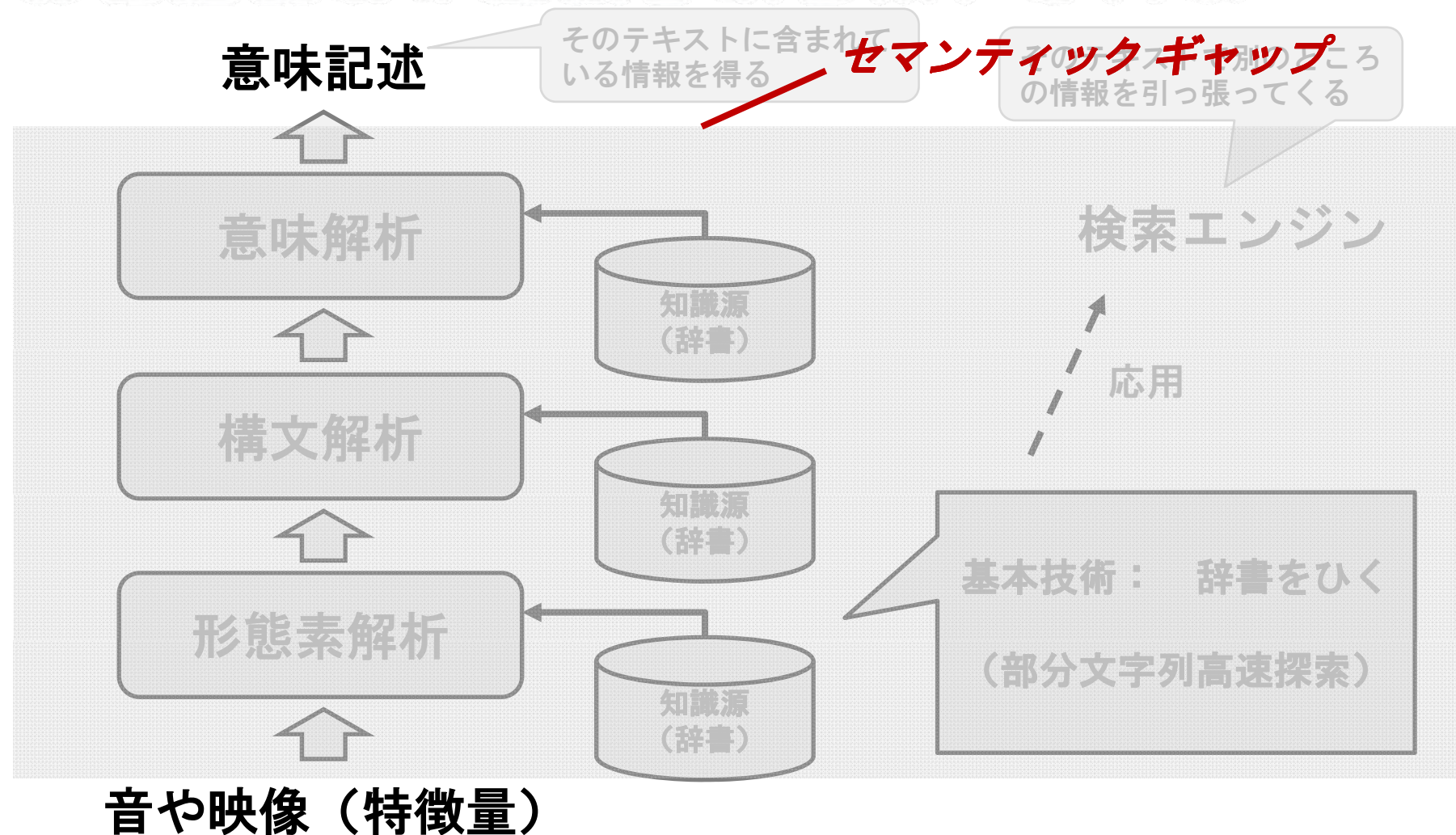
(2) 実世界における技術の実証

(3) 上記2つの相互刺激と相互貢献

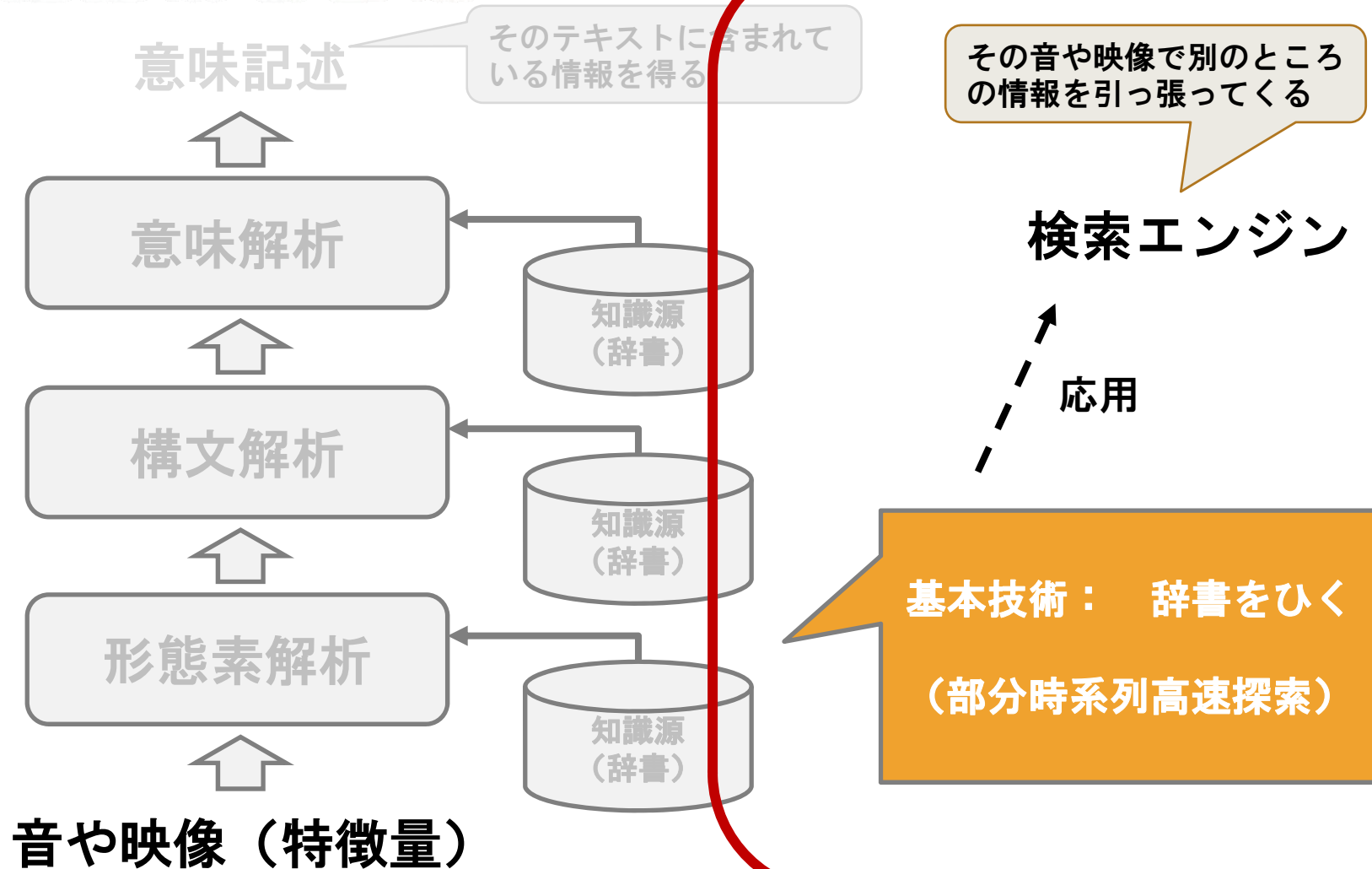
# 参考：ことばから情報を引き出す



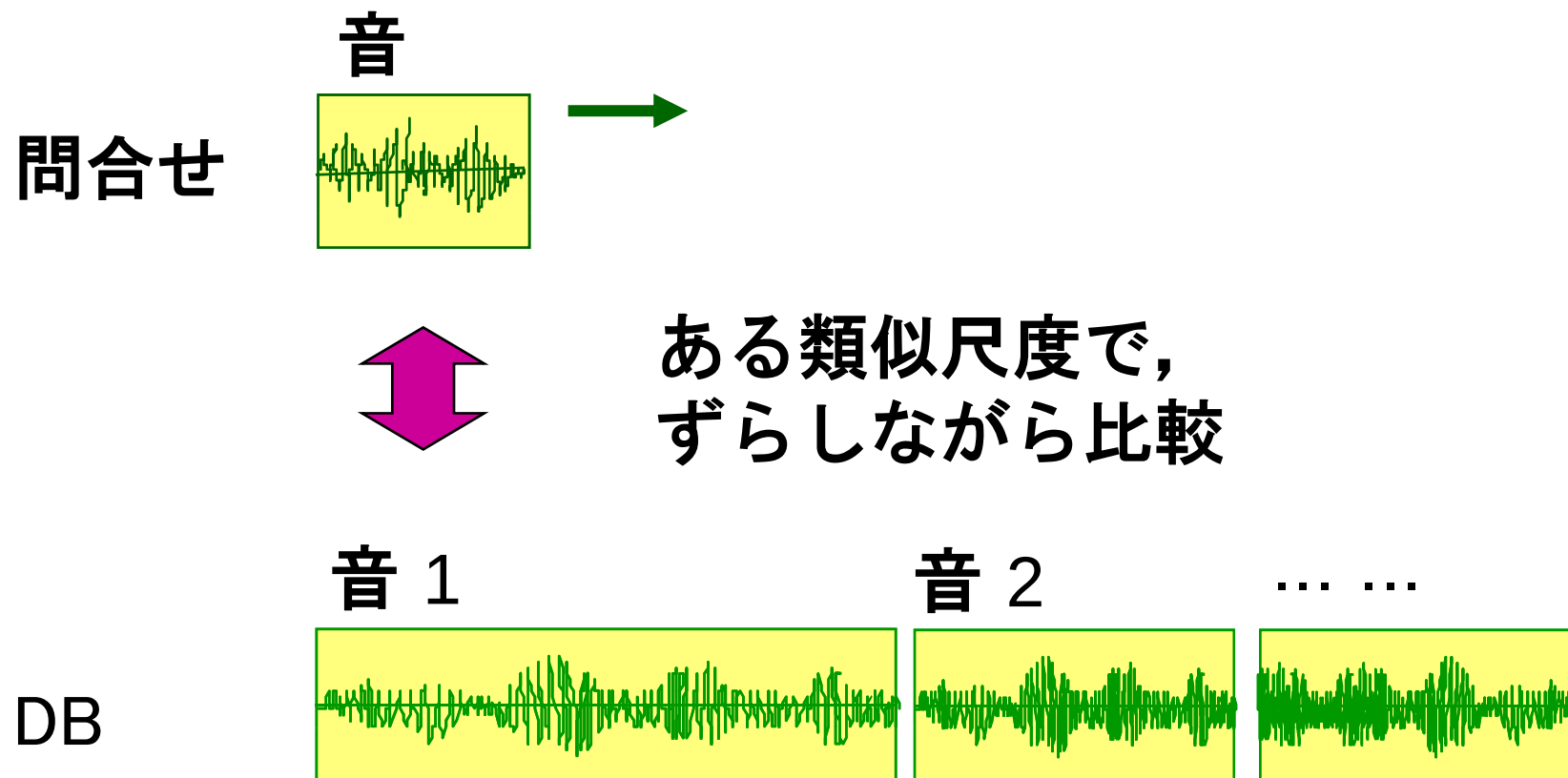
# 音や映像から情報を引き出したいが・・・



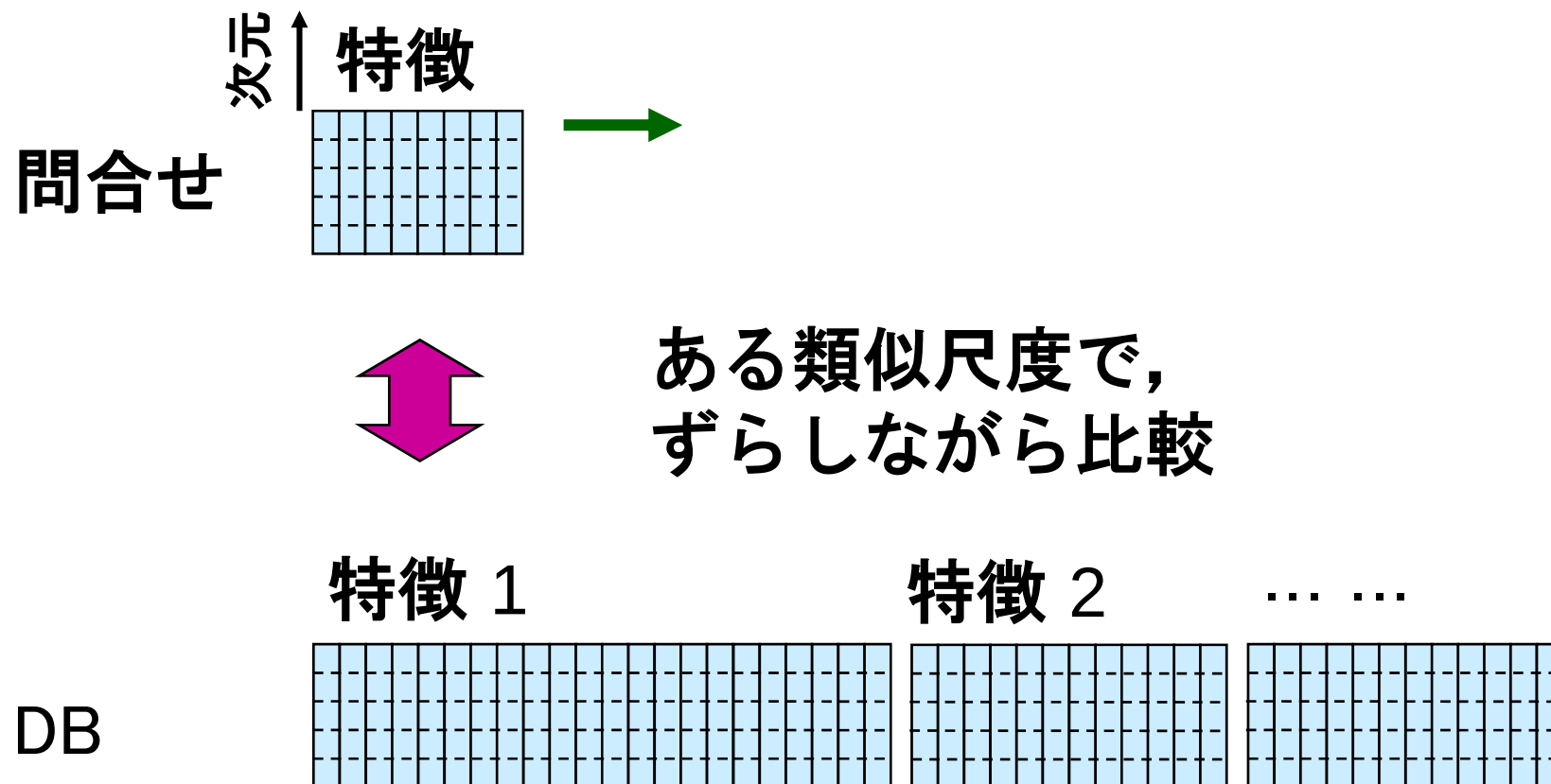
# 低抽象度の処理



# 部分時系列探索

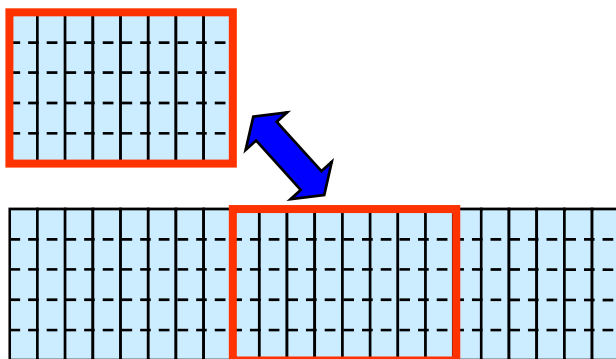


# 部分時系列探索



# 部分時系列探索の従来方法

## (1) 領域の照合



### - 2値スペクトルの照合

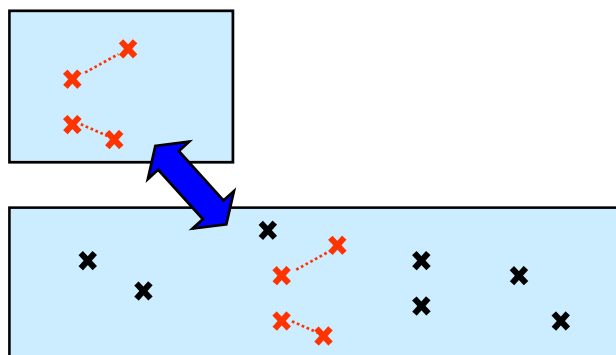
e.g. [Haitsma 2002]

### - ヒストグラム照合

e.g. [Kashino 1999]

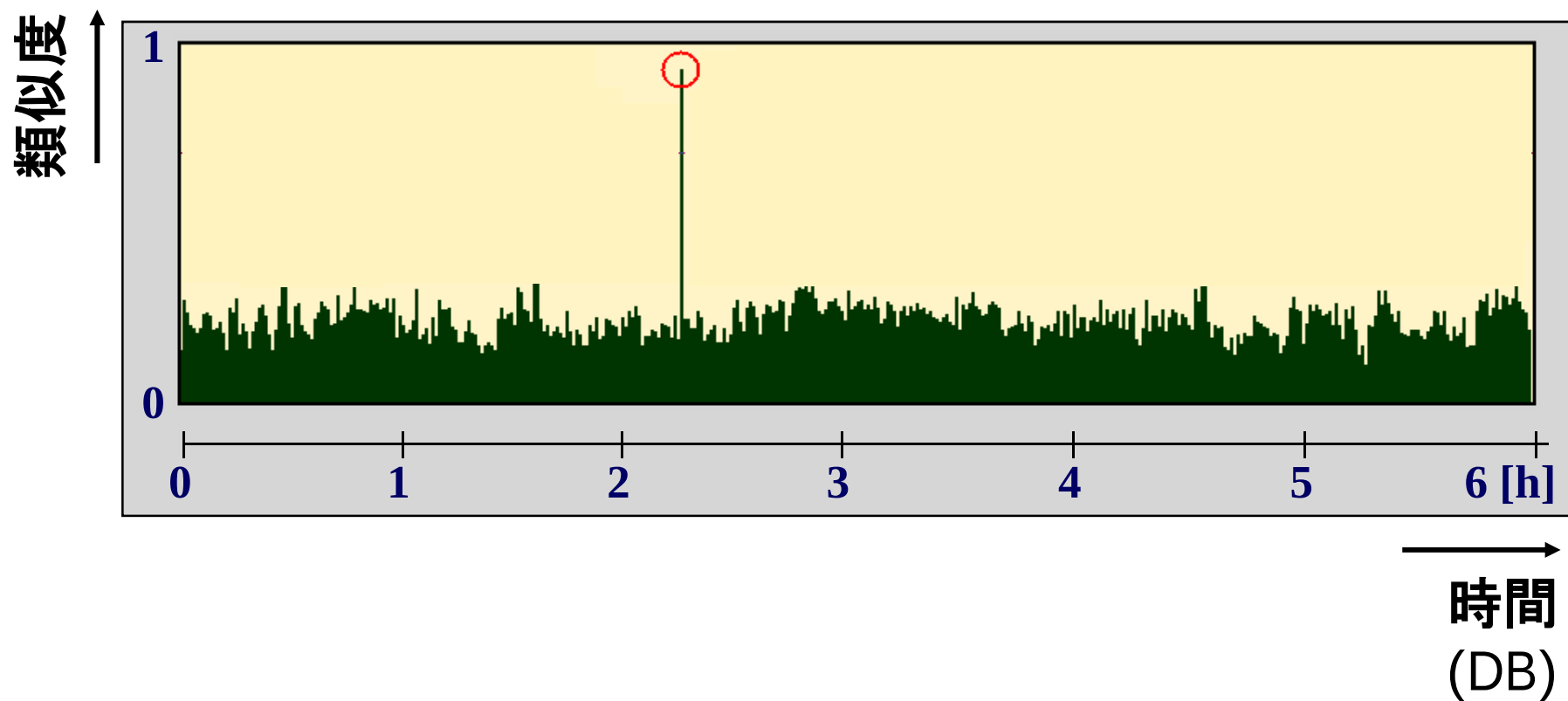
## (2) 特徴点の照合

e.g. [Wang 2003]



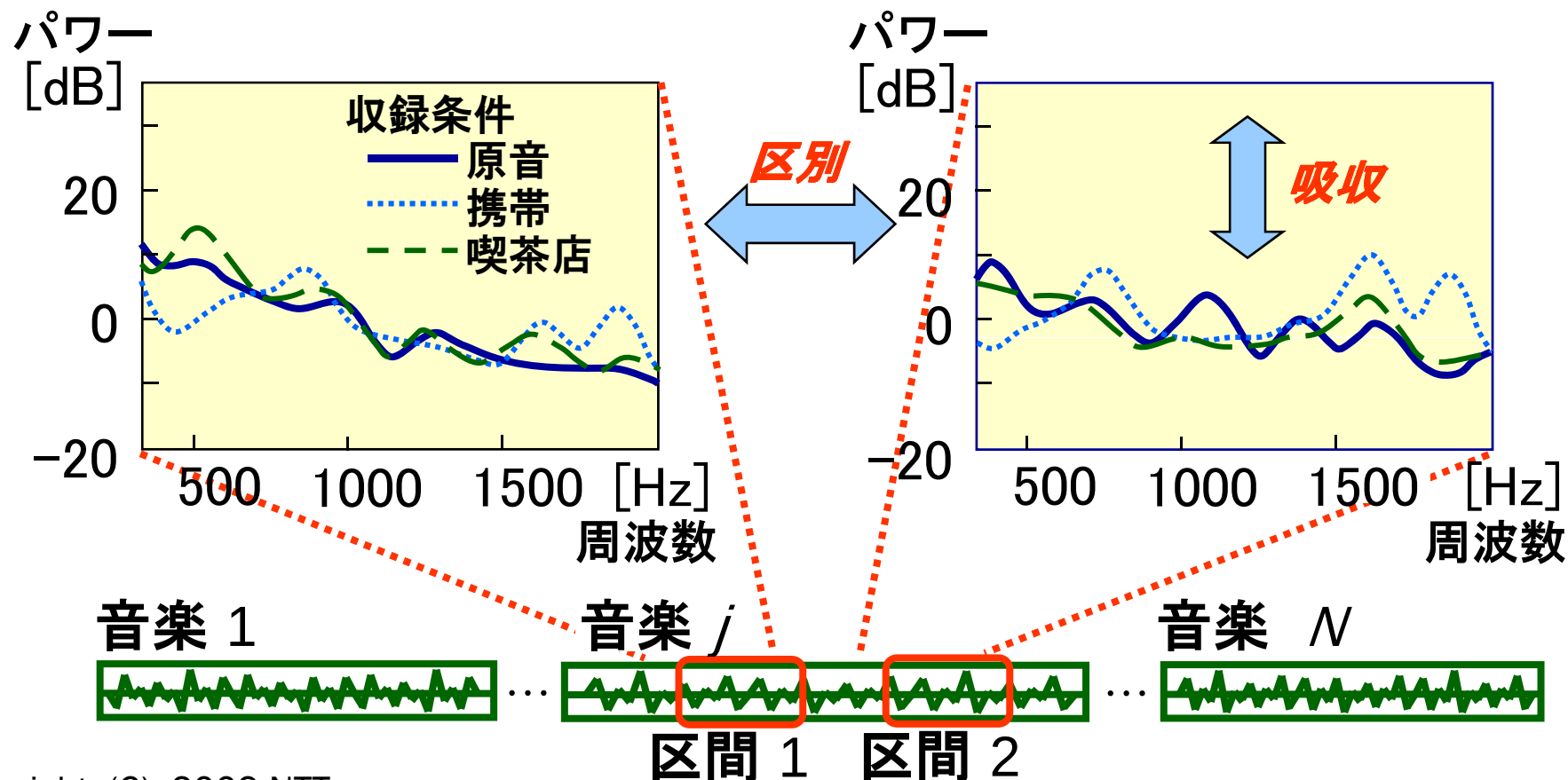


# 部分時系列探索の例



# 部分時系列探索の課題

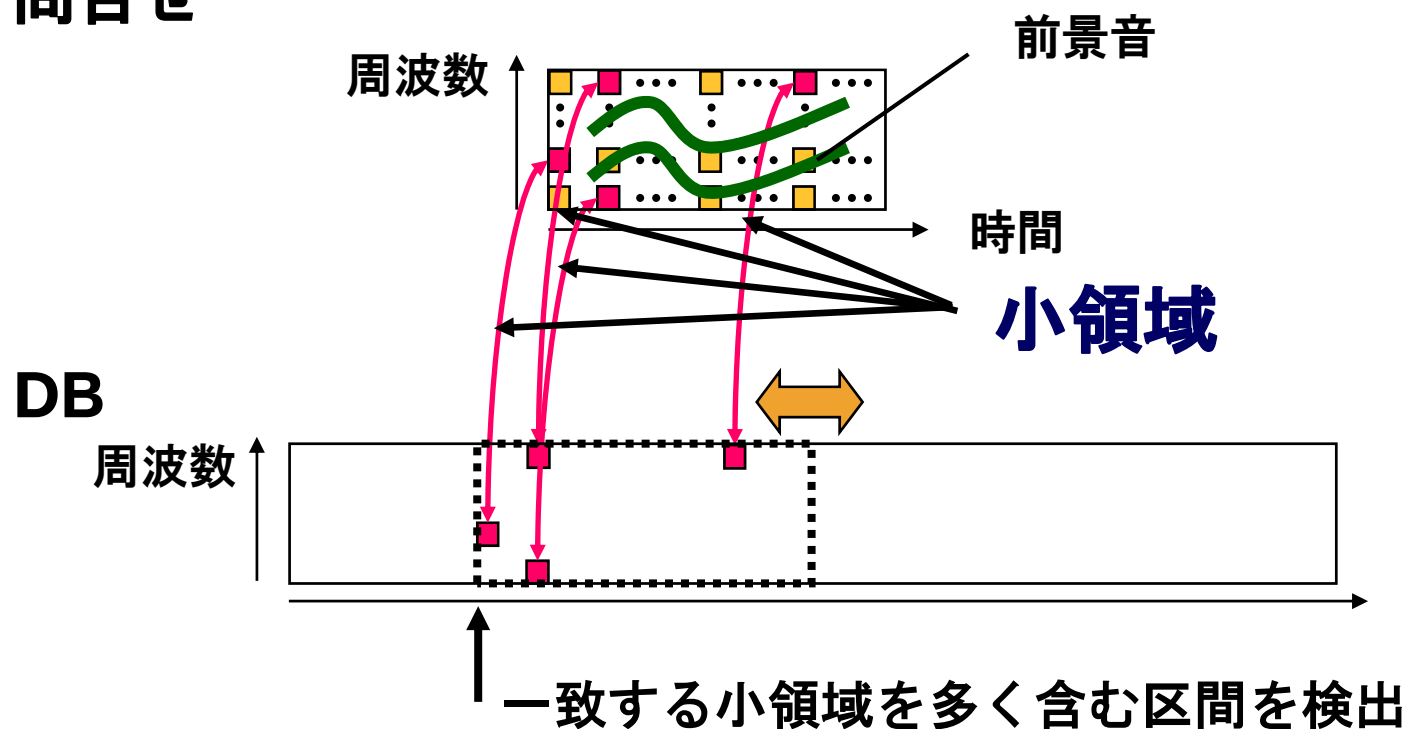
- (1) 高速な処理
- (2) ロバストな処理 (加法性・乗法性の雑音に対して)



# DAL (DIVIDE AND LOCATE) 法の原理

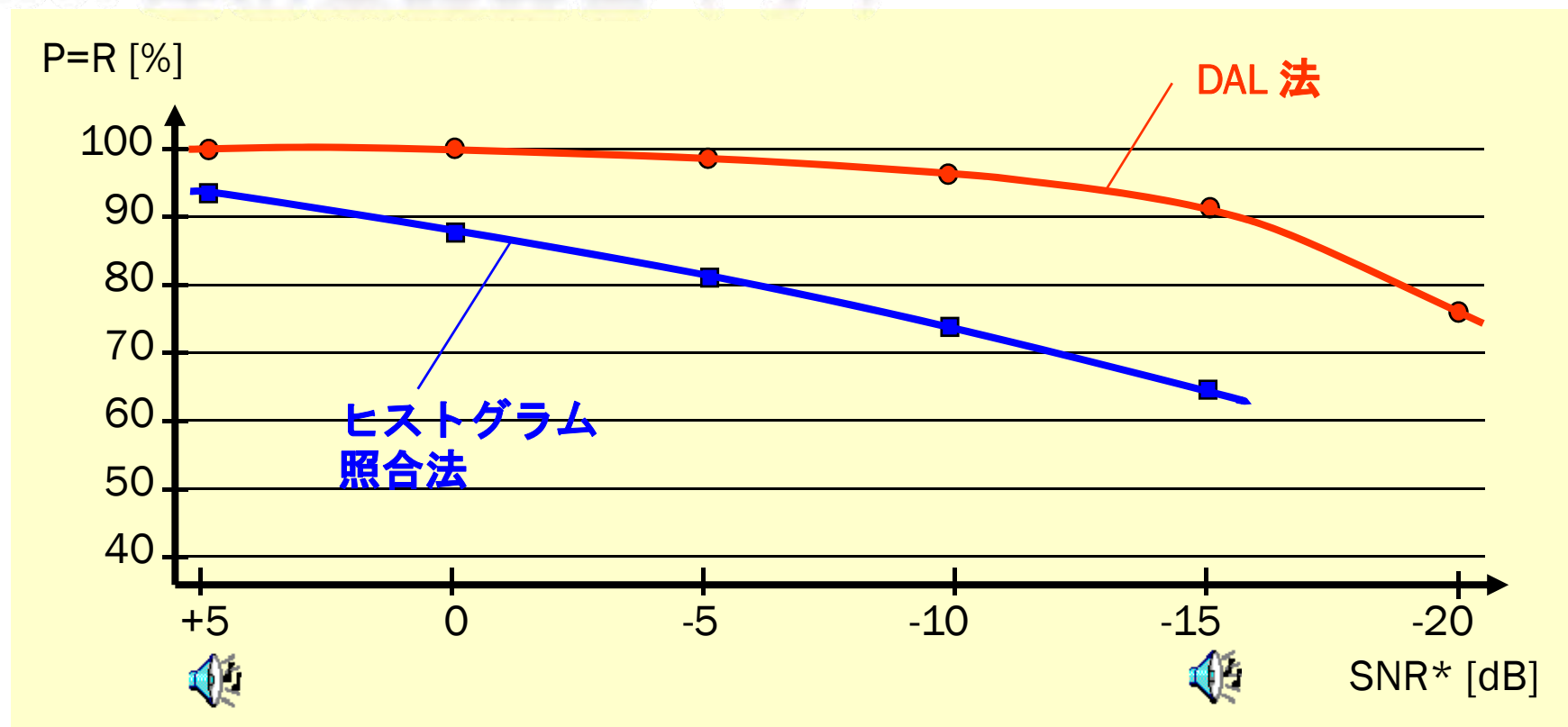
[永野 2004]

## 問合せ



小領域については解像度を下げても、  
多くの小領域を用いることで高精度が得られる

# DAL法の実験結果（１）

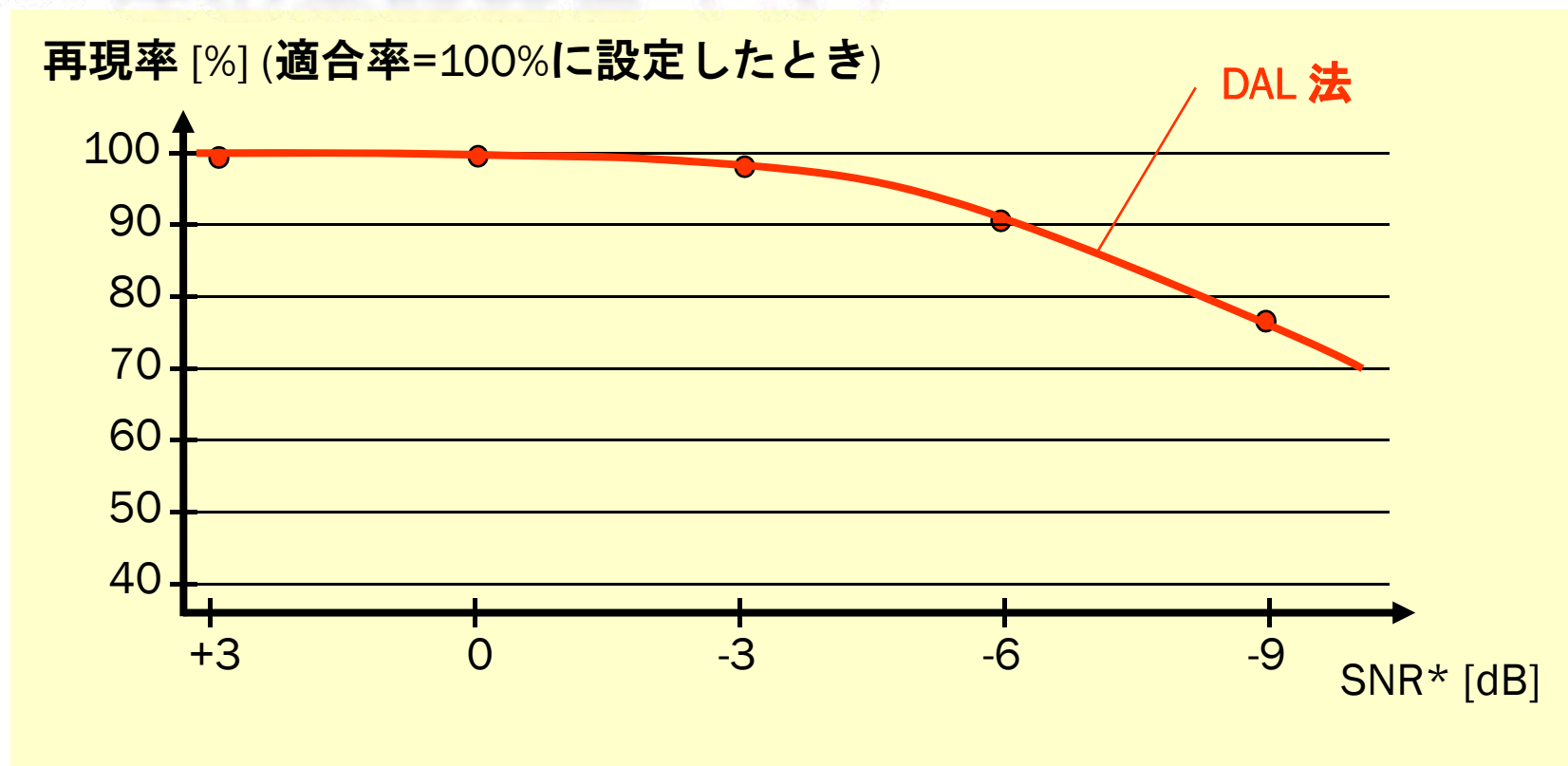


問合せ: 15秒のクリーンな音楽断片

DB: 30分間（音楽＋ナレーション）

\*  $SNR = 10 \log (\text{音楽平均パワー} / \text{ナレーション平均パワー})$

## DAL法の実験結果（２）



問合せ: 5秒の音（音楽＋ナレーション）

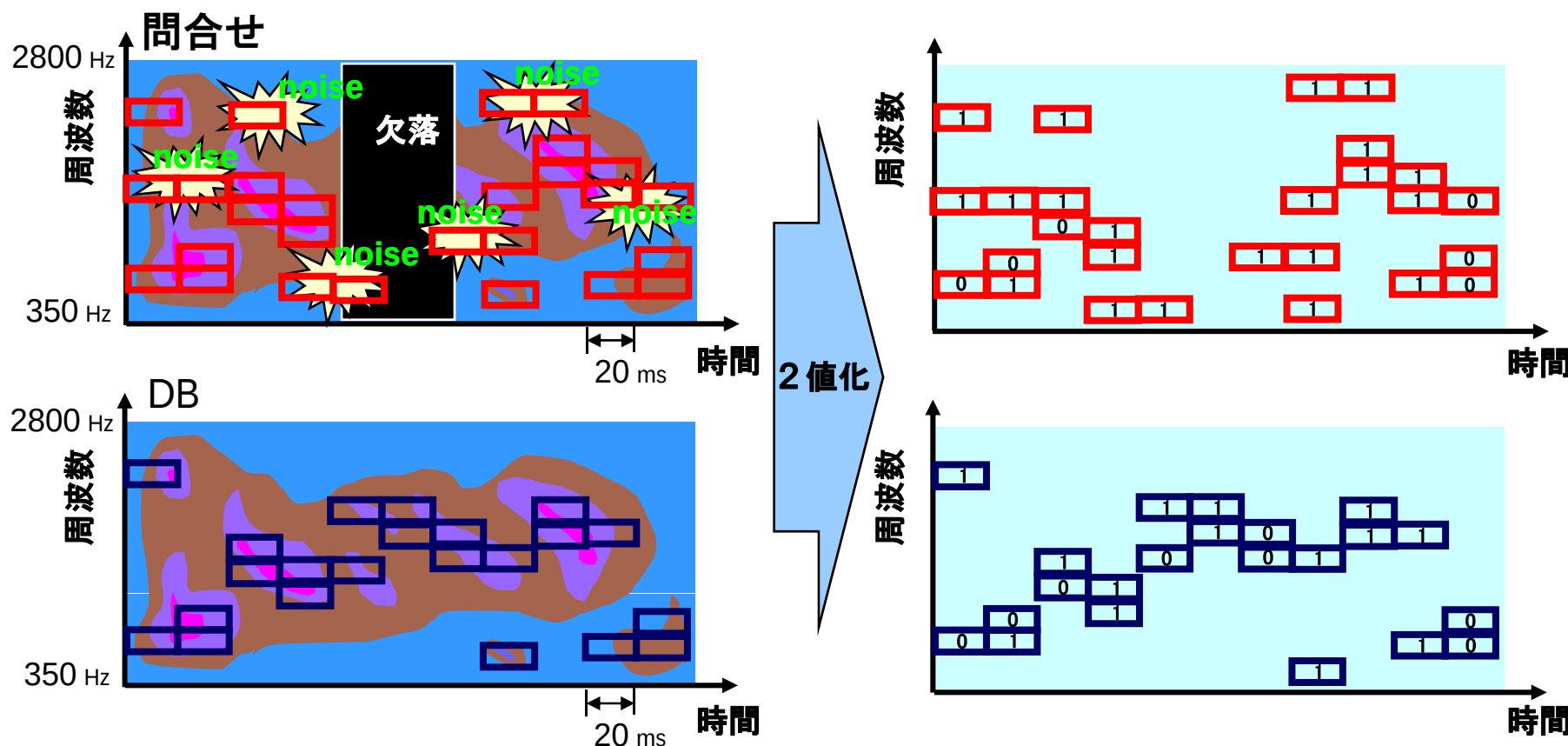
DB: 約58万曲

\*  $SNR = 10 \log (\text{avg. power of music} / \text{avg. power of narration})$

# BAM (BINARY AREA MATCHING) 法の原理

重要領域だけに着目した2値照合

[黒住 2007]



重要な領域だけを用いることでロバスト性が向上

# BAM法の実験風景



# BAM法の実験結果（１）

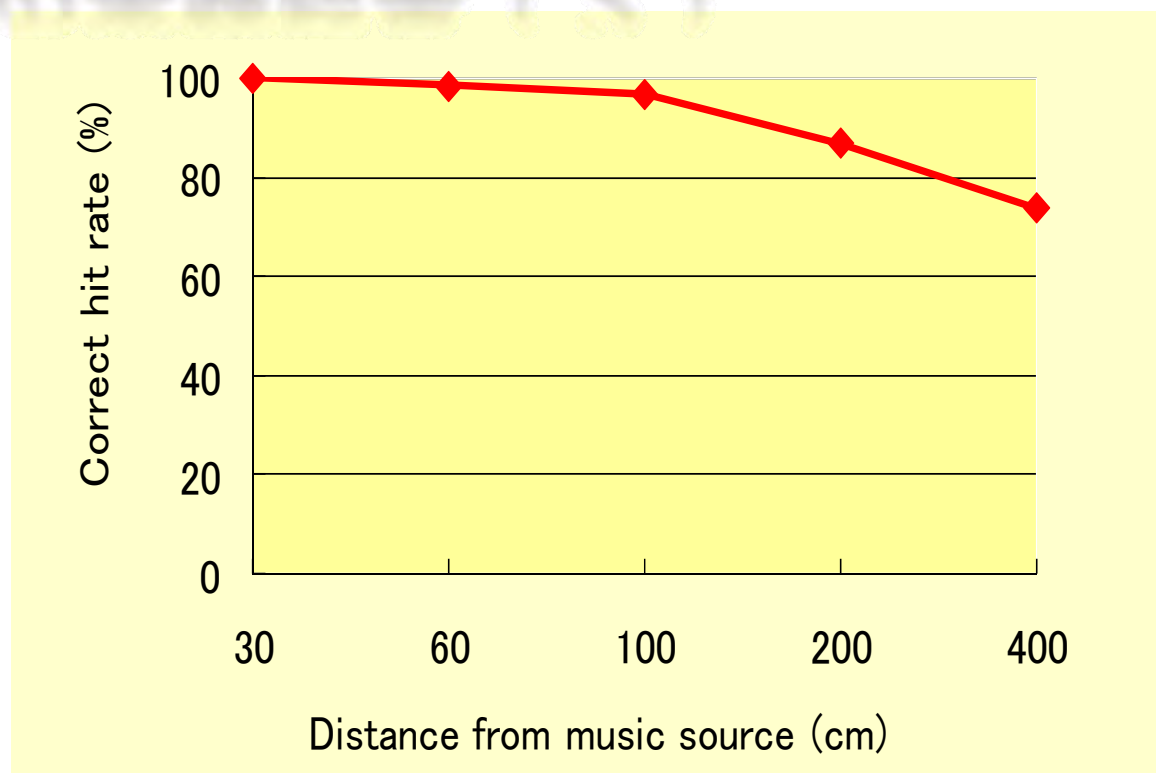
---

|           | P=R [%] |
|-----------|---------|
| ヒストグラム照合法 | 15.9    |
| BAM法      | 93.4    |

- 問合せ: 15 秒, DB: 20 分間.
- PDC携帯, 騒がしい喫茶店内で収録
- スピーカと携帯電話との距離=0.75 m



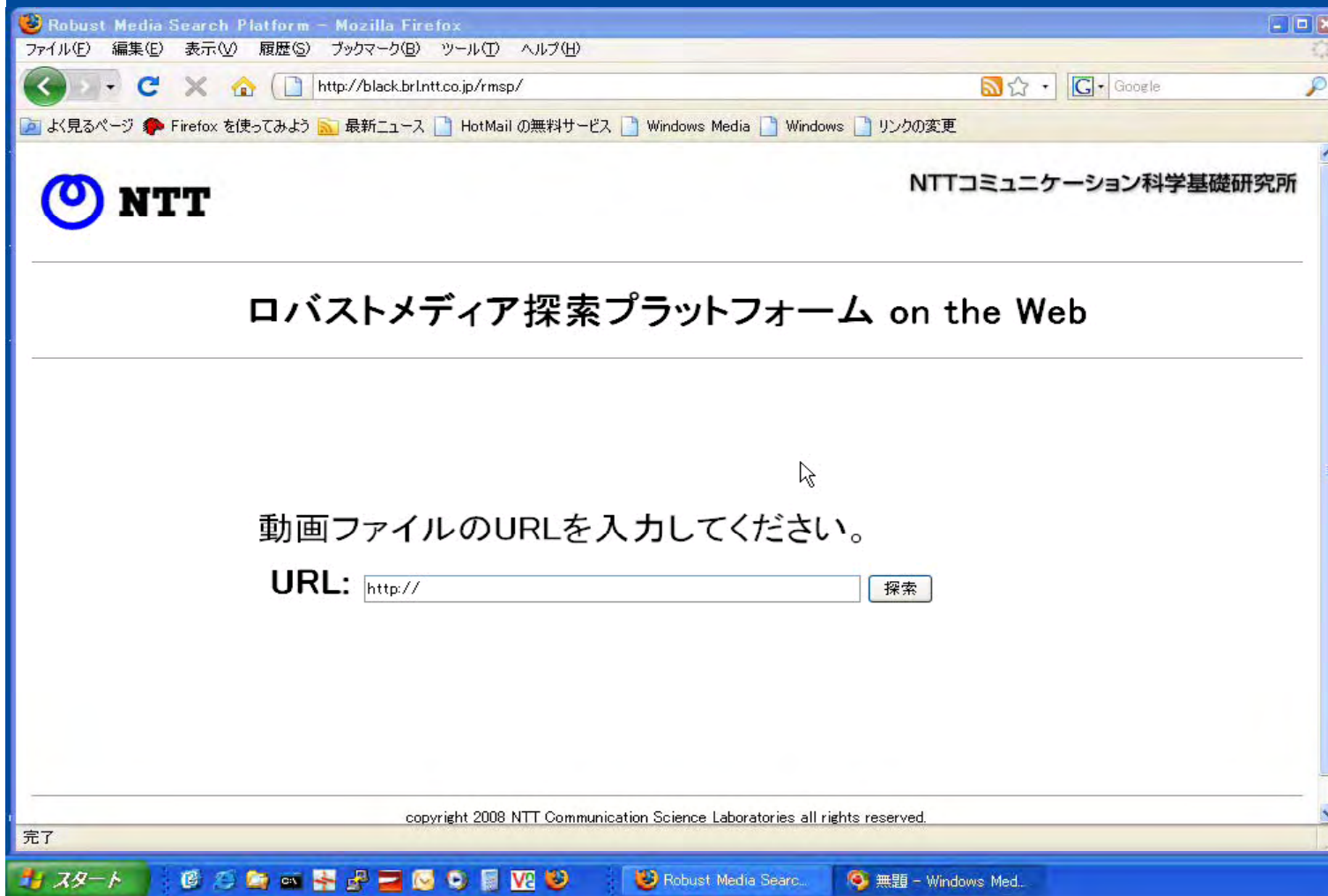
## BAM法の実験結果（２）



- 問合せ: 15 秒, DB: 約23万曲
- 音楽: 72dB SPL, 雑音: 66dB SPL (音源から 1 mの距離において)
- 携帯電話 61機種の平均値  
(correct hit rate = 正答を得た回数 / 試行数)

[Kashino 2007]

# ネット上コンテンツ特定への応用



# ネット上コンテンツ特定における要請

こんな音・映像を特定するのが重要：

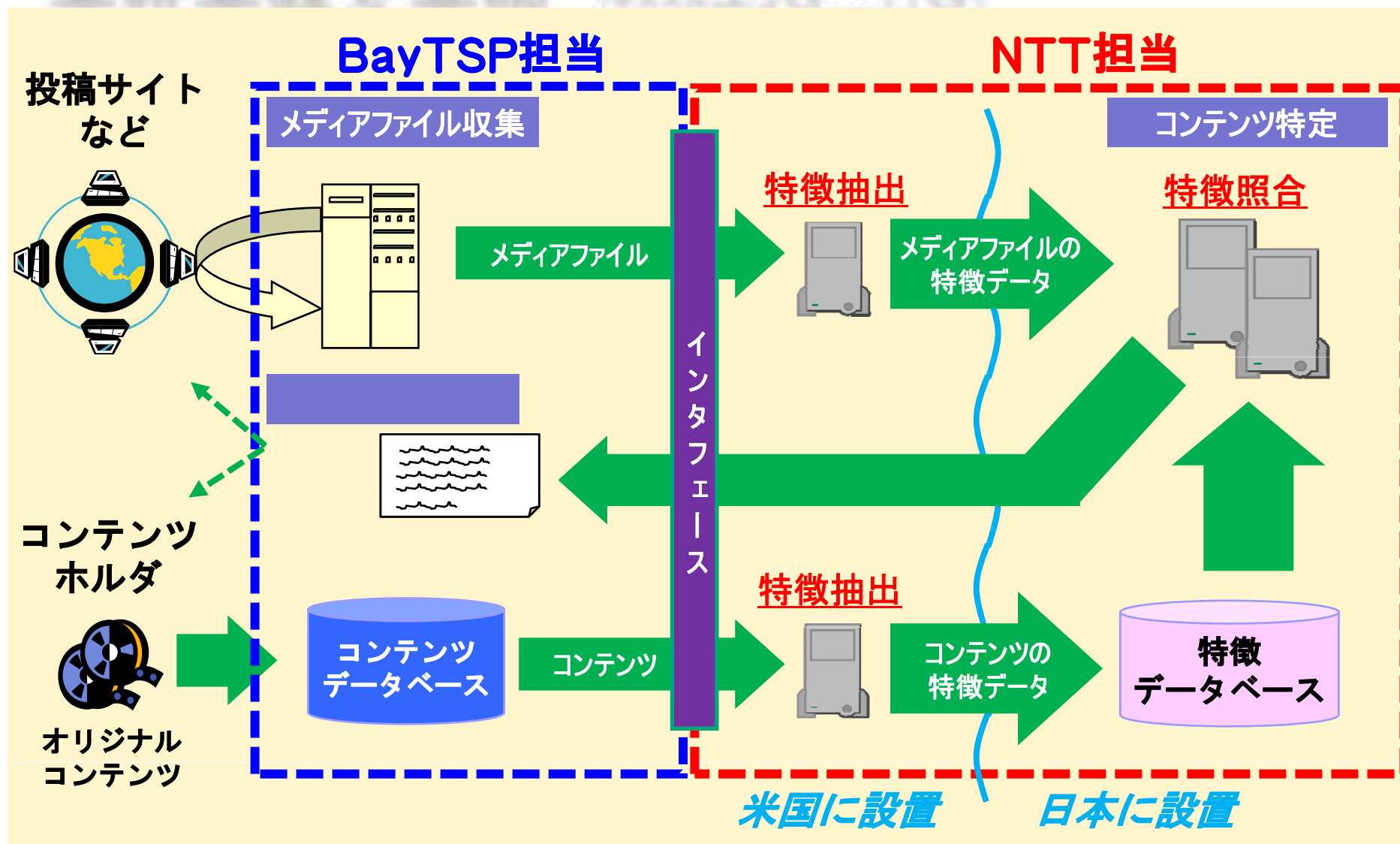
- (例1) **圧縮**などによって品質劣化した映像
- (例2) **字幕**などによって一部が隠された映像
- (例3) 吹き替えなどの**別音声**が重なった音
- (例4) **こま切れ**の使用



検出

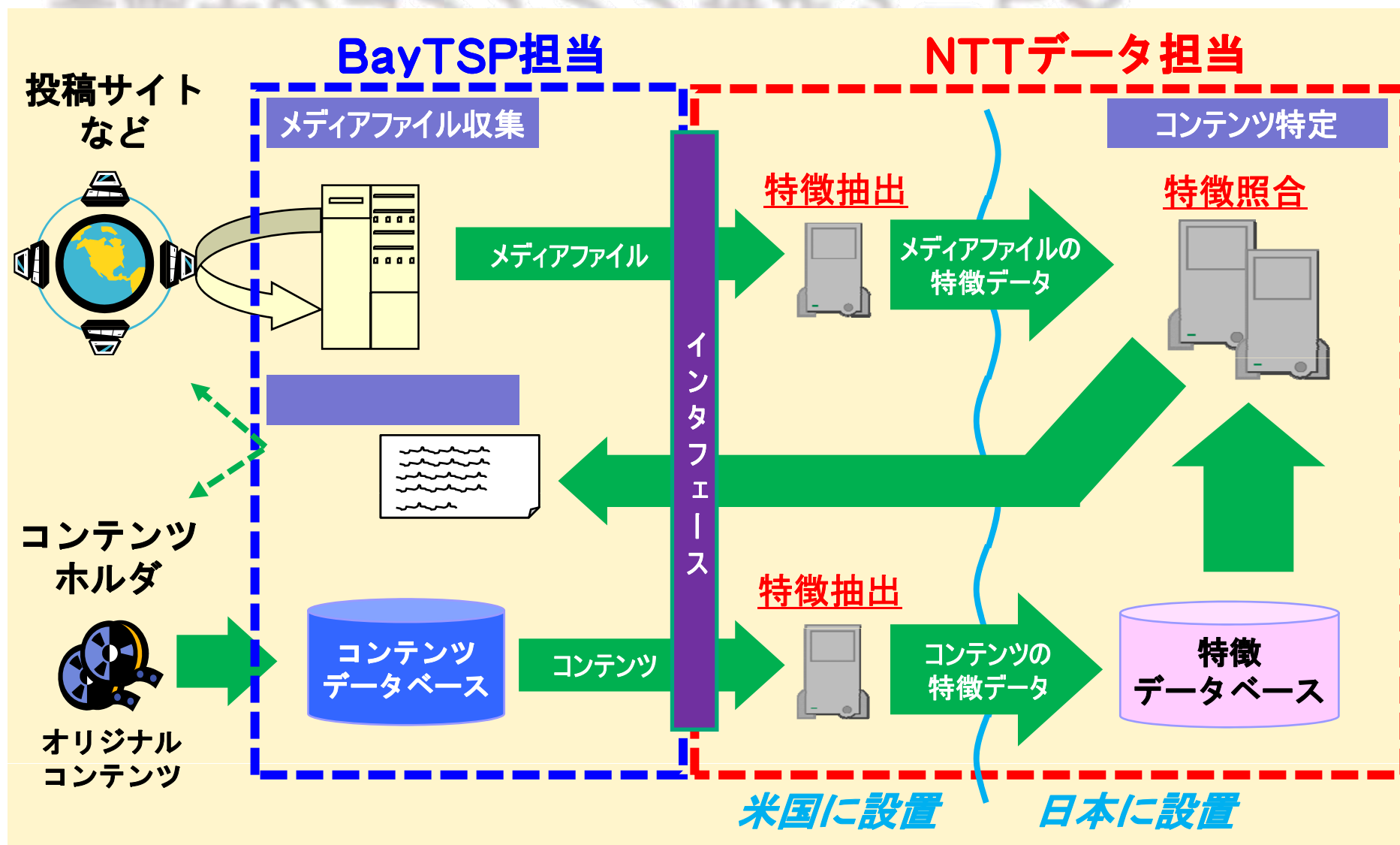


# 実証実験を実施 (2008年4月～11月)

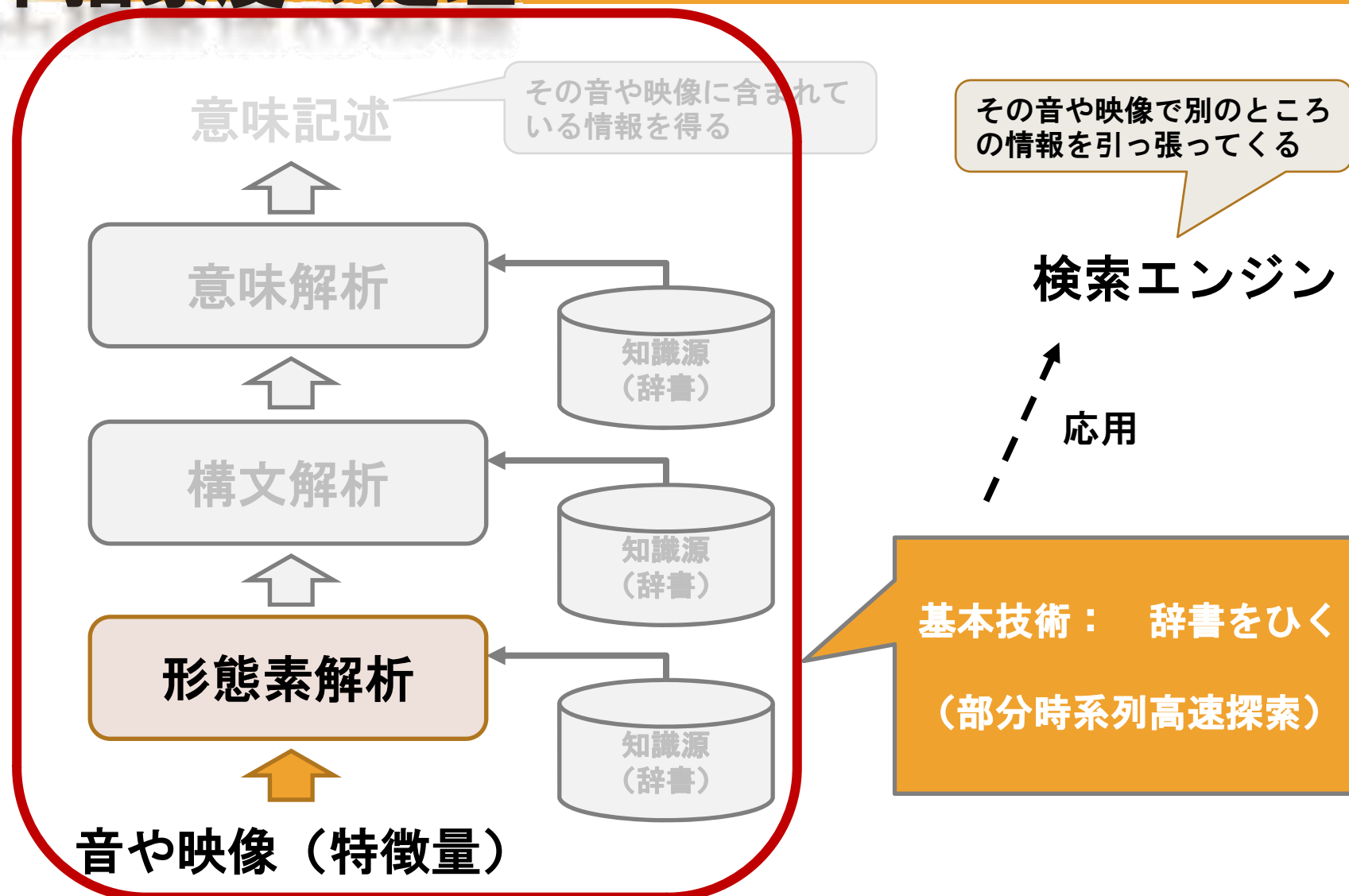


# 実施中のコンテンツ特定サービス

[NTTD 2008]



# 中抽象度の処理



# 音や映像の形態素？

---

## (1) 辞書で定義

例：適応型混合テンプレートで楽器音同定

## (2) データから導き出す

### 重要3性質

顕著 *salient*

反復 *repetitive*

共起 *collocative*

今回 紹介

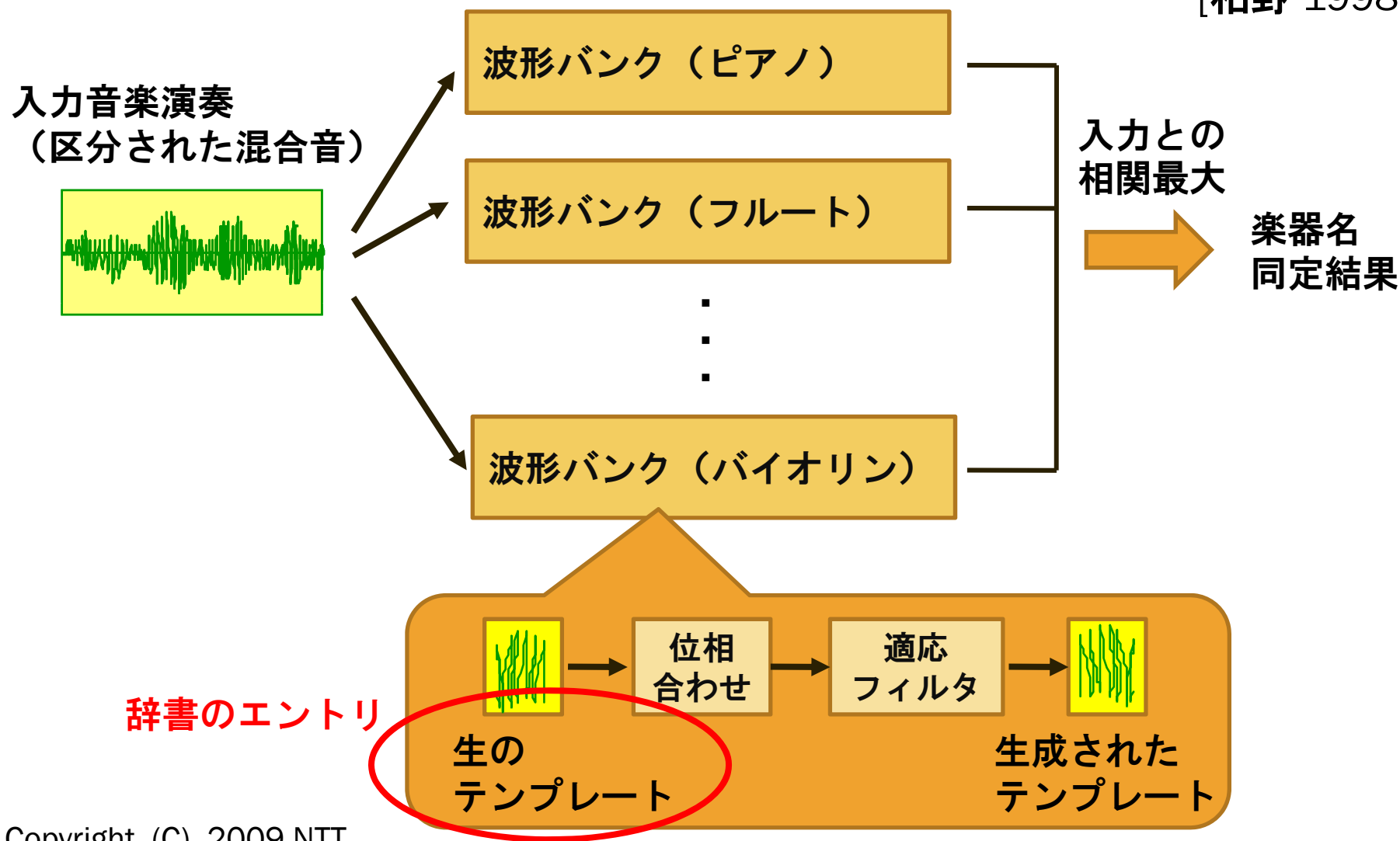


①

K- 24

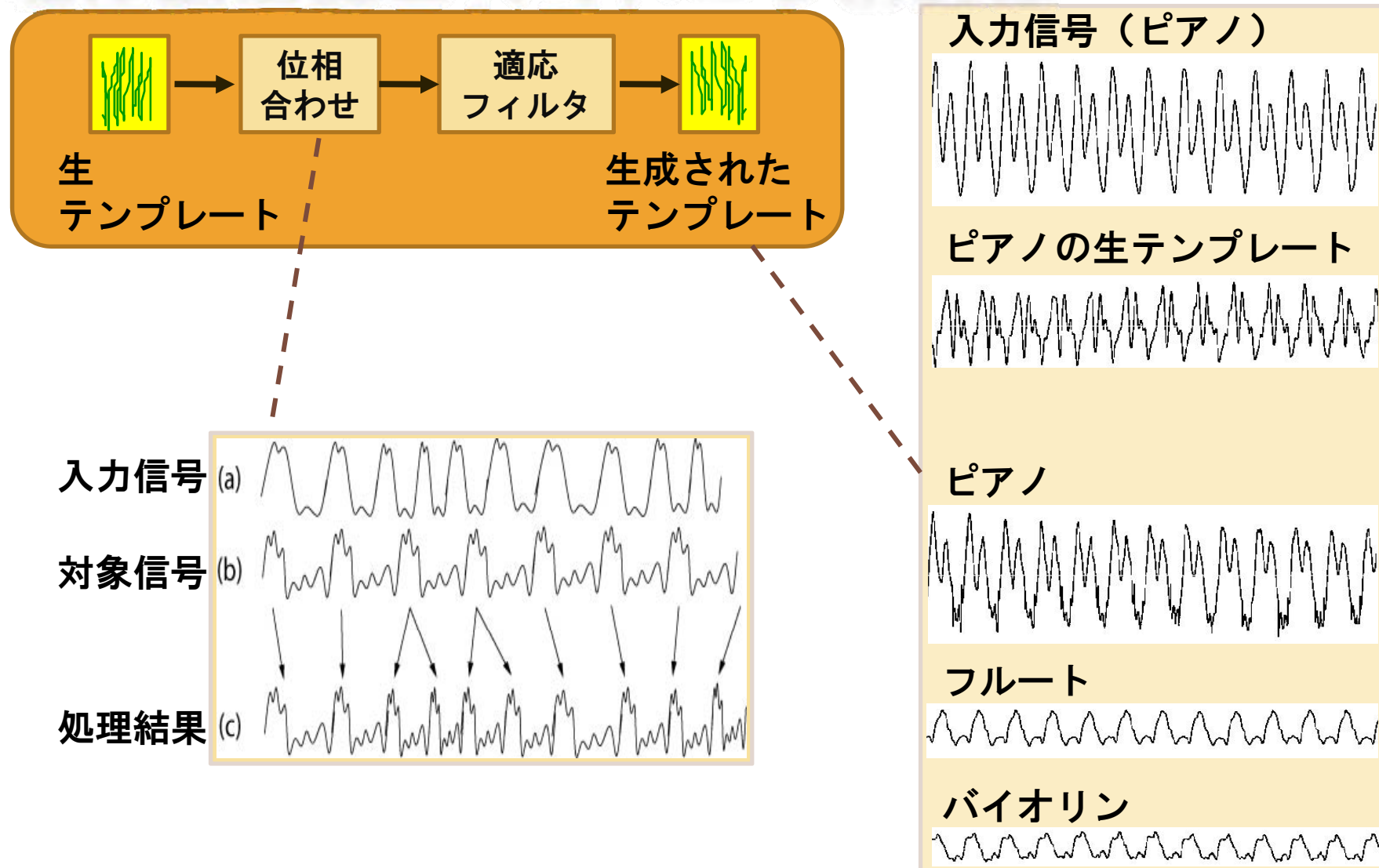
# 適応型混合テンプレートで楽器音同定

[柏野 1998]





# 適応型混合テンプレートの動作



## 顕著度による物体自動検出

---

### (1) 注目位置推定 [Kimura 2008]

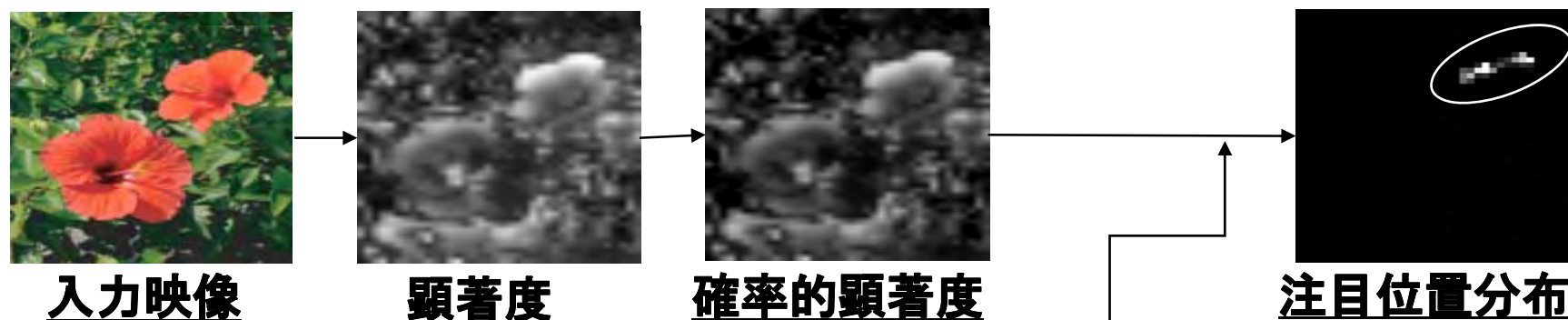
視覚特性に基づく処理で算出した顕著性から  
注目すべき箇所を特定

### (2) 注目領域抽出 [福地 2009]

注目箇所にある, 「もの」と思われる領域を  
自動的に抽出

# 注目位置推定 [Kimura 2008]

視覚機構の非決定的特性※を反映した処理により、  
人間が注目しやすい箇所を特定



初期視覚特性に基づく処理で  
「顯著度」を算出

確率的に変動する顯著度を逐次的に推定

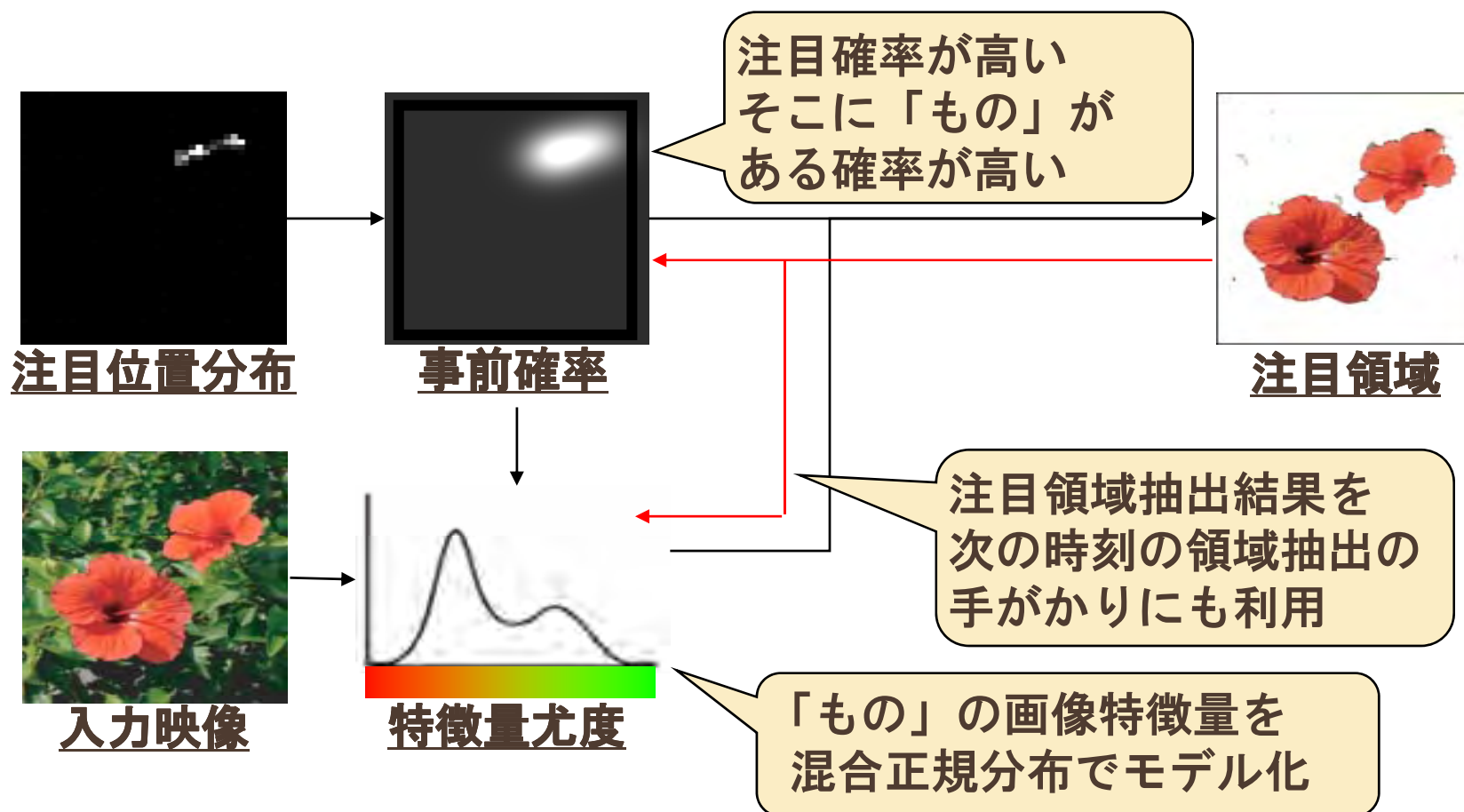
移動させる？  
留まる？  
視線移動戦略

動的ベイジアンネットワークで推定

(※ 同一映像を同一人物が見ても、  
時と場合で注目位置が異なる)

# 注目領域抽出 [福地 2009]

注目しやすい箇所の画像特徴量に基づき、  
「もの」と思われる領域を自動抽出



# 複素NMFによる音の分解

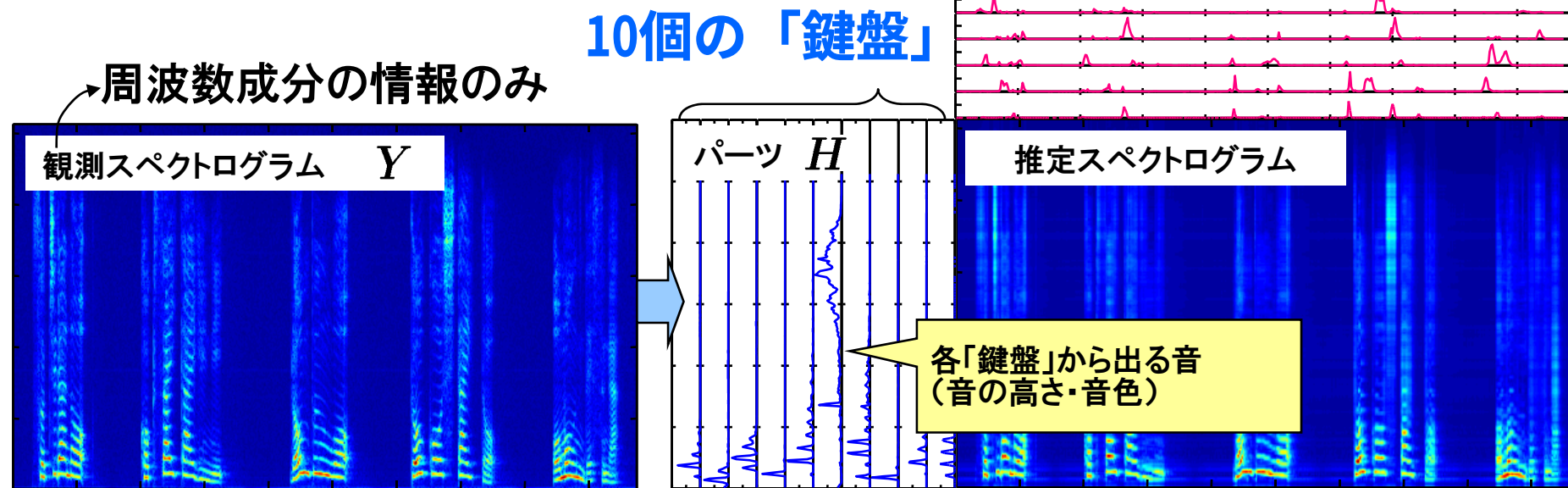
[Kameoka 2009]

実世界音響信号を一種の「鍵盤楽器」から生成されたものと捉え、  
そのもとで、各「鍵盤」が、押されたとき  
どんな音を出すか、いつからいつまで  
押されたか、を観測信号から推定



設計図 U

各「鍵盤」が弾かれる  
タイムテーブル  
(発話・演奏のリズム／抑揚)



# NMF（既存手法）

**非負値行列因子分解**  $Y \simeq HU$  ( $H \succeq 0, U \succeq 0$ )  
(Nonnegative Matrix Factorization; NMF)

**画像処理分野で誕生**

[Lee 1999]

**スペクトログラムに直接適用**

**→ 音声分離・強調・復元への応用**

[Smaragdis 2003]

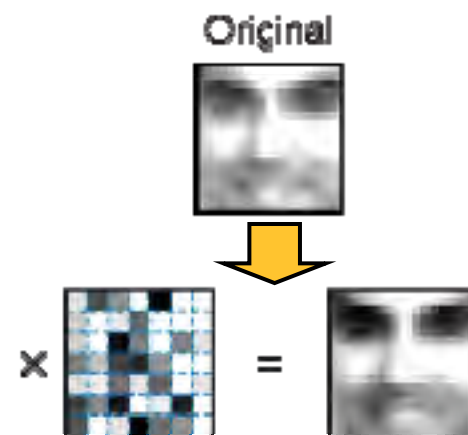
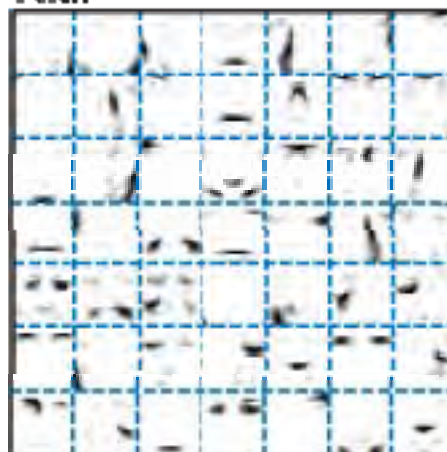
**効率的な反復アルゴリズムが存在**

[Lee 2000]

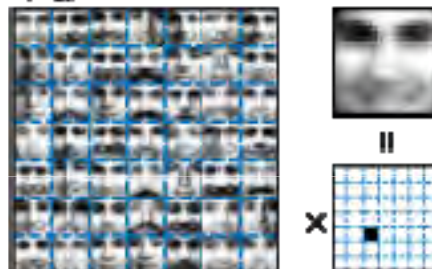
$$H_{k,i} \leftarrow H_{k,i} \frac{[YU^T]_{k,i}}{[HUU^T]_{k,i}}$$

$$U_{i,j} \leftarrow U_{i,j} \frac{[H^T Y]_{i,j}}{[H^T HU]_{i,j}}$$

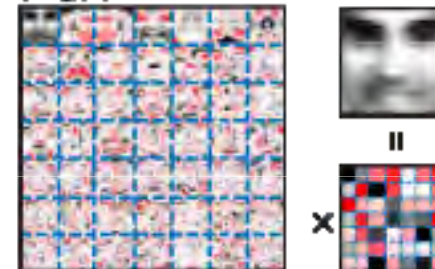
**NMF**



**VQ**



**PCA**





# 複素NMF: 音響信号に適合した新たな枠組

[Kameoka 2009]

## NMFの観測モデル

観測スペクトログラム

$$Y \simeq HU$$

周波数  
時刻

$Y_{\omega,t} \simeq \sum_i H_{\omega,i} U_{i,t}$   
( $H_{\omega,i} \geq 0, U_{i,t} \geq 0$ )

$i$  番目の  
振幅スペクトルパターン

$i$  番目の  
振幅スペクトルパターンの  
アクティベーションの大きさ

## 提案モデル

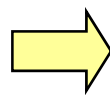
$i$  番目の  
振幅スペクトルパターン

時不変 時変

$$Y_{\omega,t} \simeq \sum_i \underbrace{H_{i,\omega}}_{\text{時不変}} \underbrace{U_{i,t}}_{\text{時変}} e^{j\phi_{i,\omega,t}}$$

( $H_{\omega,i} \geq 0, U_{i,t} \geq 0$ )

時不変な振幅スペクトル「部品」  
時変のアクティビティ



時不変な振幅スペクトル「部品」  
時変のアクティビティ/位相スペクトル

# 展望

---

## 要素技術の成熟

「形態素」解析に基づく新たな応用が出現

「中抽象度」表現の開拓  
一致検索の幅が拡大

## 物量が増大

ニーズも増大

メディア資源の活用法が重要に



# 文献（１）

---

- [Haitsma 2002] J. Haitsma and T. Kalker: "A Highly Robust Audio Fingerprinting System," Proc. of Int. Conf. Music Information Retrieval (ISMIR) , Oct. 2002.
- [Kashino 1999] K. Kashino, G. Smith, and H. Murase: "Time-Series Active Search for Quick Retrieval of Audio and Video," Proc. of Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol.6, pp.2993-2996, Mar. 1999.
- [Wang 2003] A. Wang: "An Industrial Strength Audio Search Algorithm," Proc. of Int. Conf. Music Information Retrieval (ISMIR) , Oct. 2003.
- [永野 2004] 永野, 柏野, 村瀬: "多数の小領域スペクトログラムの探索に基づく背景音楽の高速探索法", 電子情報通信学会論文誌, vol.J87-D-II, no.5, pp.1179-1188, May 2004.
- [黒住 2007] 黒住, 永野, 柏野: "実環境で収録された映像断片をキーとする一致映像探索," 電子情報通信学会 論文誌 D-II, VOL.J90-D, no.8, pp.2223-2231, Aug. 2007.
- [Kashino 2007] K. Kashino, A. Kimura, H. Nagano, and T. Kurozumi: "Robust Search Methods for Music Signals Based on Simple Representation," Proc. of Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol.IV, pp.1421-1424, Apr. 2007.
- [NTTD 2008] NTTデータ 報道発表: "インターネット上の音楽や映像を高速・高精度に特定可能なコンテンツモニタリングサービスの提供を開始,"  
<http://www.nttdata.co.jp/release/2008/120100.html>, Dec. 2008.

## 文献（２）

---

- [柏野 1998] 柏野, 村瀬: "適応型混合テンプレートを用いた音源同定 - 音楽演奏への適用 -," 電子情報通信学会論文誌, vol.J-81-D-II, no.7, pp.1510-1517, Jul. 1998.
- [Kimura 2008] A. Kimura, D. Pang, T. Takeuchi, J. Yamato, and K. Kashino: "Dynamic Markov Random Fields for Stochastic Modeling of Visual Attention," Proc. of Int. Conf. Pattern Recognition (ICPR), Dec. 2008.
- [福地 2009] 福地, 宮里, 木村, 高木, 大和: "グラフコストの逐次更新を用いた映像顕著領域の自動抽出," 電子情報通信学会技術報告（パターン認識・メディア理解）, May 2009.
- [Kameoka 2009] H. Kameoka, N. Ono, K. Kashino, S. Sagayama: "Complex NMF: A New Sparse Representation for Acoustic Signals," Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Apr. 2009.
- [Lee 1999] D. Lee and H. Seung: "Learning the Parts of Objects by Non-Negative Matrix Factorization," Nature, vol. 401, pp.788-791, 1999.
- [Smaragdis 2003] P. Smaragdis and J. Brown: "Non-Negative Matrix Factorization for Polyphonic Music Transcription," Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pp.177-180, Oct. 2003.
- [Lee 2000] D. Lee and H. Seung: "Algorithms for Non-negative Matrix Factorization," Proc. Neural Information Processing Systems Conference (NIPS), pp. 556-562, Nov. 2000.