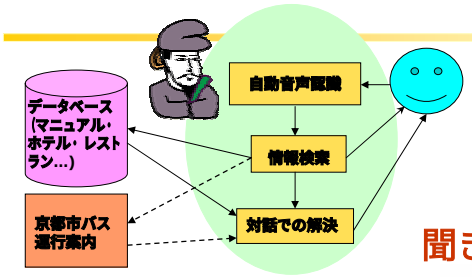
音声情報処理特論 2003年11月18日

聖徳太子コンピュータ
ーー混合音の分離を目指してーー

奥 乃 博

京都大学 大学院情報学研究科
知能情報学専攻
知能メディア講座 音声メディア分野
http://winnie.kuis.kyoto-u.ac.jp/~okuno/
okuno@i.kyoto-u.ac.jp, okuno@nue.org

実環境での音声メディア研究



音声により要求を聞き分ける

聞き分ける聖徳太子ロボット

音楽を聞き分け・探索

混合音を聞き分ける・擬音語認識

目次

1. 混合音の分離
2. 調波構造による音響ストリーム分離
3. 音源方向による音響ストリーム分離
4. 画像情報統合による音響ストリーム分離

本講義での目標・立場

1. 混合音など一般的な音の理解
   - CASA (音環境理解、Computational Auditory Scene Analysis)
   - 音声認識システムは、単一話者の声を想定
2. 人工知能研究の立場から
   - 信号処理中心ではなく、記号処理(音の表現) を中心に
   - 統計的なアプローチだけでなく、情報統合で
3. ソーシャルインターラクションを志向
   - カクテルパーティ効果：アクティブ知覚との統合
   - 社会性を持ったインターラクション
   - 聖徳太子コンピュータ：同時に10人の訴えを聞く

聴覚の重要性

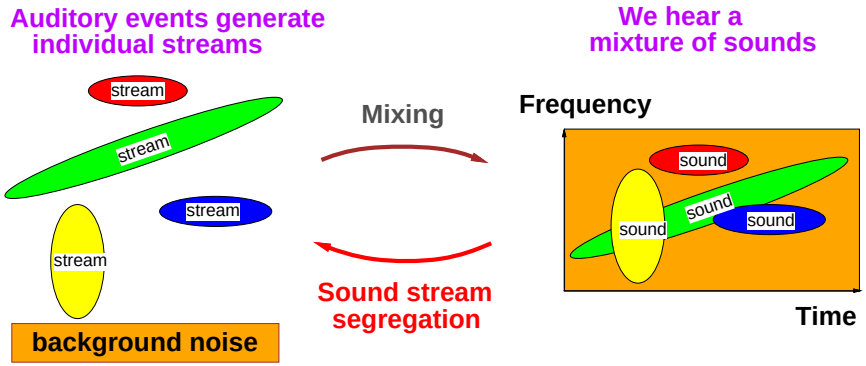
聴覚は人間にとって最も重要な感覚である。言語によるコミニケーションが聴覚によって成立することは容易に理解されるが、「ヒトは聴覚によってのみ言語を獲得し、そこに文化が生まれ、継承される。**書かれた言語は目によって伝承されるが、話す言葉は耳からしか得られない。**話し言葉があって書く言葉が生まれる」ことを、多くの人が理解していないのは残念なことである。

鈴木淳一、小林武夫共著『耳科学 --- 難聴に挑む』(中公新書1598, 2001)

刺激と反応

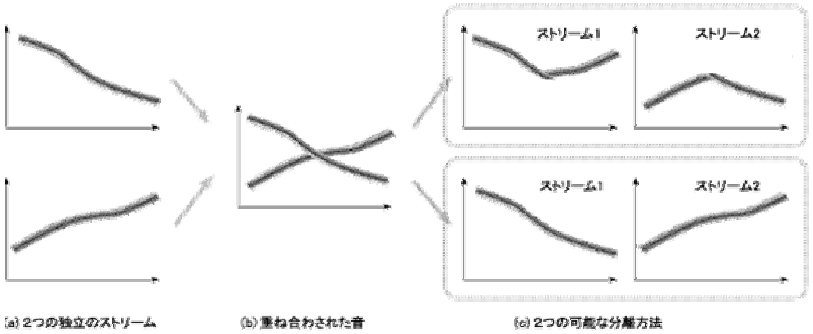
1. **傾聴** (listening) **や注視** (looking)は、**聞こえる** (hearing) **や見える** (seeing) と違い、より能動的、主体的な行動である。
2. 傾聴や注視では、注意が移るきっかけはそれぞれの感覚情報だけでなく、環境から得られる複数の感覚情報に基づいたマルチモーダル情報の影響が大きい。
3. 刺激と反応は、身体を有したシステムで実証する必要がある。

どのようにして私たちは音を聞くのか



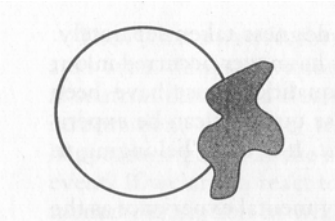
単一の音ではなく、混合音を聞いている。

音響ストリーム分離における曖昧性



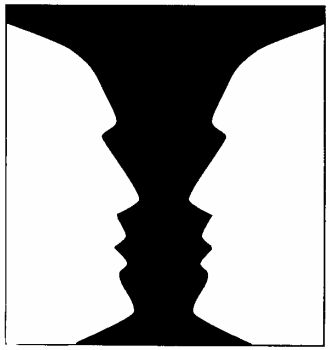
曖昧性を解消するためにさまざまな特徴や他の情報を利用する必要がある。

ストリーム化 ― 部分の帰属性



円とシミの境界線はどちらに属する？

ストリーム化 ― 排他的割り当て



The Principle of Exclusive Allocation
(地と図との問題)

ストリーム化 ― 変化を検出

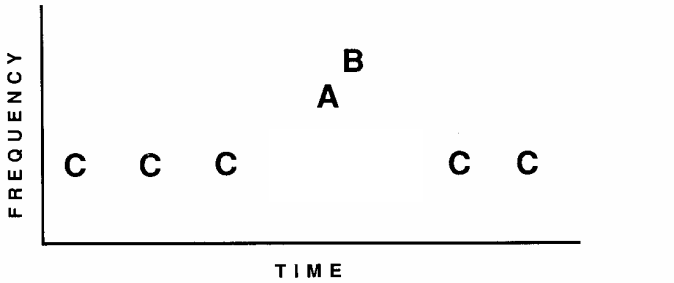


Figure 1.7
A tone sequence of the type used by Bregman and Rudnicky (1975).

A B の順序は分かるか？
(Old-Plus-New Heuristics)

ストリーム化 ― 周波数の差

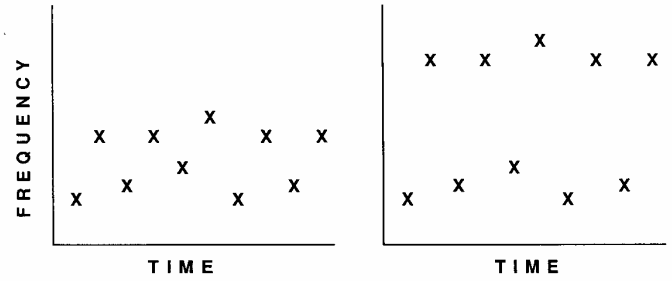
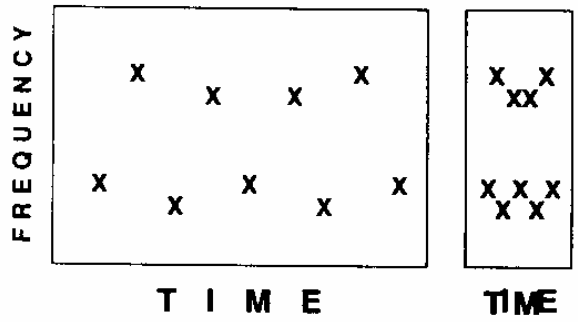


Figure 1.9
Stream segregation is stronger when the frequency separation between high and low tones is greater, as shown on the right.

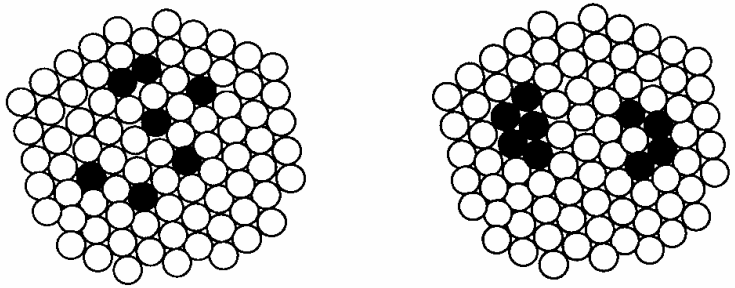
ストリーム分離はどちらが強いか？

ストリーム化 ー 時間間隔



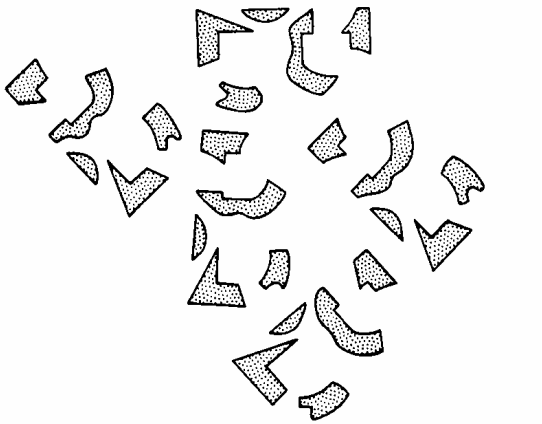
ストリーム分離はどちらが強いか？

Gestalt Principles



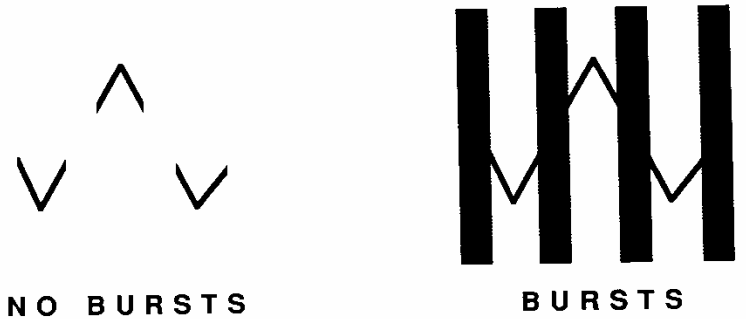
Forces of attraction により perceptual organization が生ずる。
(全体主義 vs. 還元主義)

Gestalt Principle of Closure



この文字は？ Occlusion（見え隠れ）？

Gestalt Principle of Closure

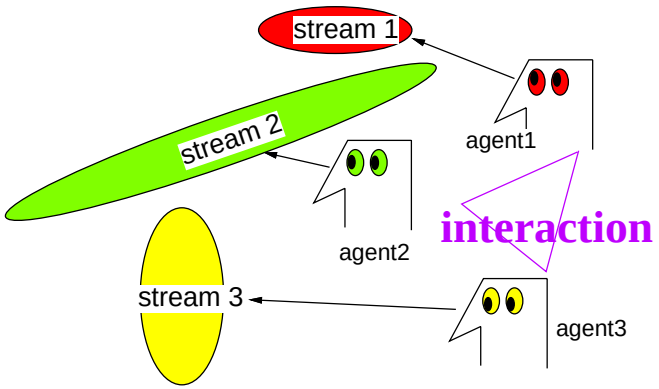


Occlusion（見え隠れ）の鍵があると。
Auditory Induction （音素修復）

音を抽出するための特徴

- 低レベルの音の特徴
  - 音の立ち上がり (オンセット) 、立下り (オフセット) 、パワー、調波構造 (基本周波数の音とその整数倍音) 、変調 (AM，FM) 、音源方向、音源の距離
- 音源の特徴
  - 音源のモデル (音声、楽音、動物の鳴き声) 、音源の種類 (ヘヤードライヤ、電話のベル)
  - 音源の個数
  - リズム、和音の遷移

マルチエージェントによる分離



各エージェントは自分が追跡する音響ストリームに集中し、相互作用を通じて調整。

人は皆同じように感じ取るのか

同じ情景を描いた2つの絵の違いは？

ゴヤ「1805年5月3日」

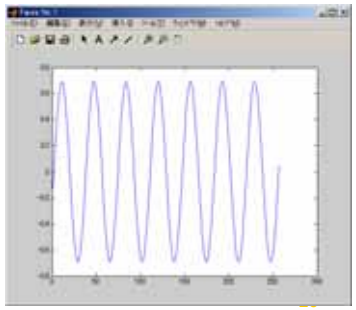
マネ「マクシミリアン皇帝の処刑」

調波構造：基本的な音の表現

- 基本周波数 $A(t)$

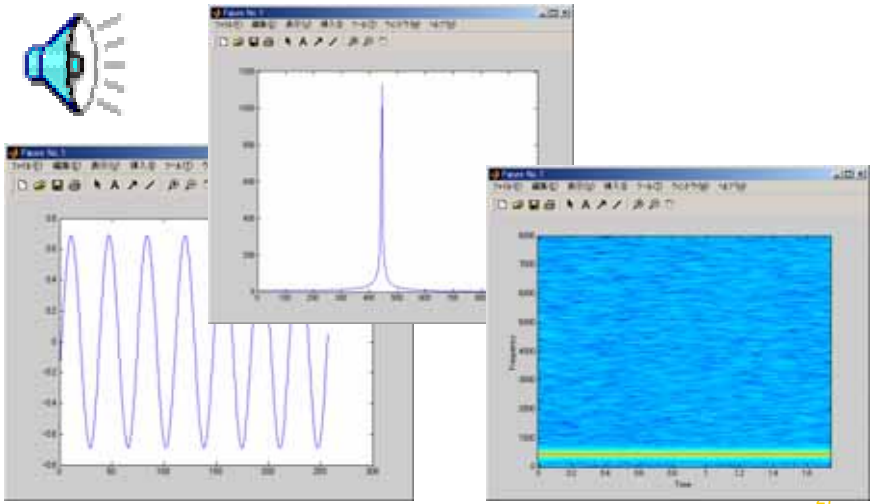
$$A(t) = \sum_i a_i \sin(\omega_0 i t + \theta_i)$$

- $i=1$ は基本周波数、他は倍音と呼ぶ
- ω_0 は基本周波数
- a_i は振幅
- θ_i は位相
- 440Hzの純音



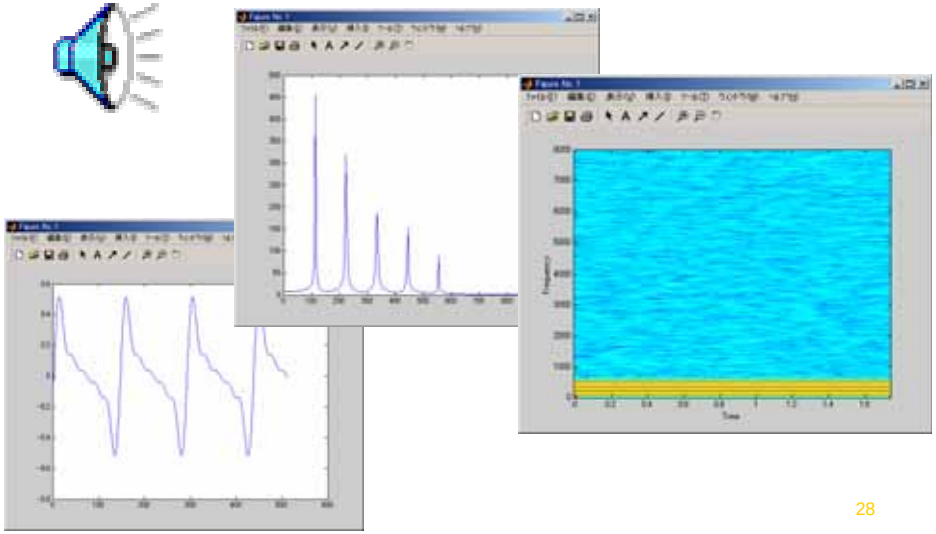
440Hzの純音を信号処理をすると

- 波形、フーリエ変換、スペクトル



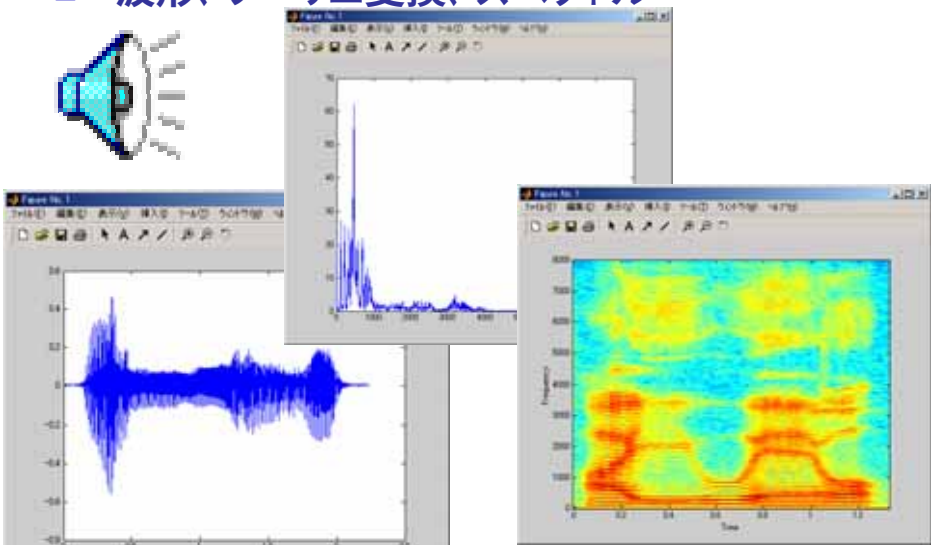
110Hzの調波構造を持つ音

- 波形、フーリエ変換、スペクトル

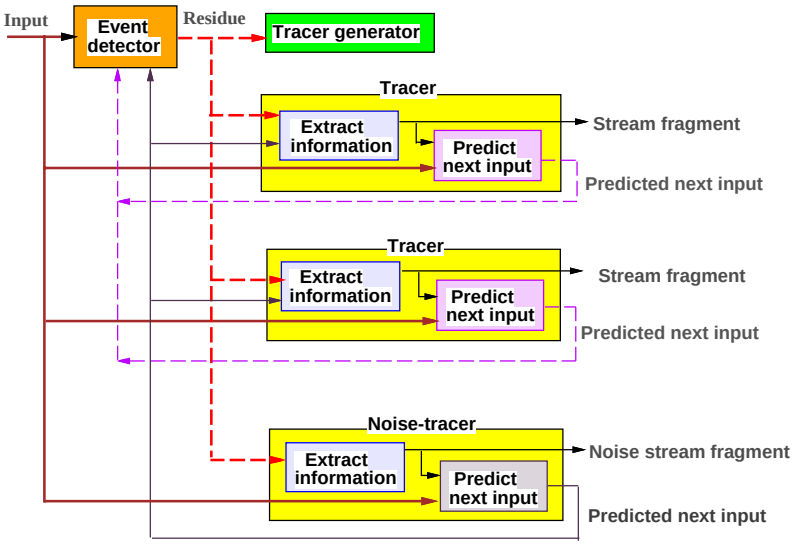


発話「あいうえお」の信号解析

- 波形、フーリエ変換、スペクトル



残差駆動型アーキテクチャ



残差駆動型調波構造ストリーム分離

HBSS (Harmonics-Based Stream Separation)

1. Event detector が予測信号を入力信号から減算し、得られた残差をtracer generator に与える。
2. 残差が閾値より大きければ、tracer generator は、新たな音を発見したものとして tracer を生成。
3. Tracer generator が残差の中に調波構造を発見すると、tracer を生成。調波構造がない場合には、定常雑音を除去する noise tracer を生成。
4. Tracer は調波構造断片を抽出するとともに、次の入力信号 (混合音) 中の調波構造を予測し、生成した予測信号を event detector に渡す。

調波構造断片のグルーピング

1. Tracer は次式を最大にする周波数を予想

$$E(\omega) = \sum_{k=1}^{\sharp of\ harmonic\ series} \left\| \sum_{t} x(t) \exp(-jk\omega t) \right\|^2$$

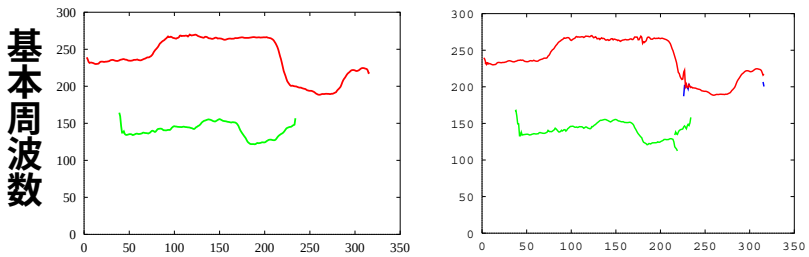
2. 調波構造断片に含まれる基本周波数 ω_0 が近いものを次々につないでいく

$$A(t) = \sum_{i} a_i \sin(\omega_0 i + \theta_i)$$

3. 漸進的に分離を実行。
4. 音源数は予め与えておく必要がないし、音源数が動的に変化してもよい。

HBSSによる音響ストリーム分離

3つの音を演奏： 入力混合音、
分離音 (女性の発話、男性の発話)

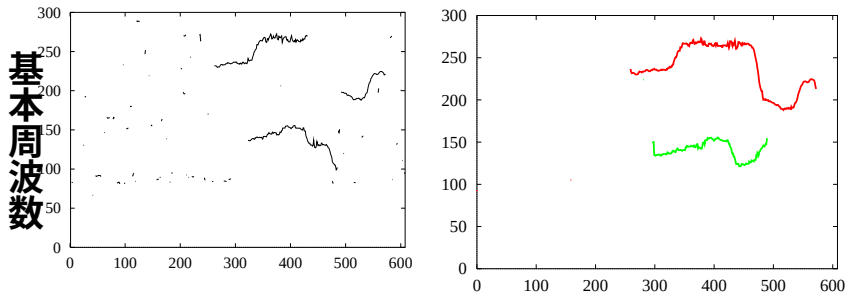


入力混合音

HBSSによる分離音

Noise Tracer の効果

3つの音を演奏： 入力混合音、
分離音 (女性の発話、男性の発話)



Noise tracerなし
(デモなし)

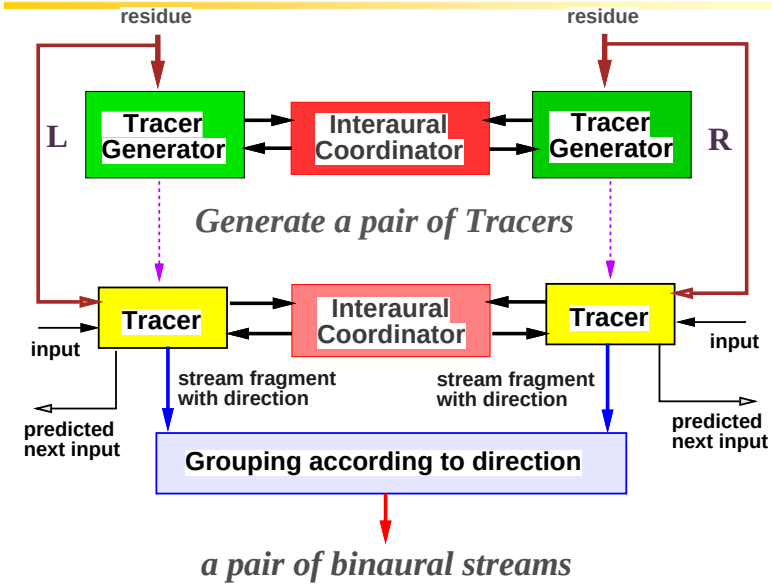
HBSSによる分離音

HBSSの能力

1. 音源数を仮定しない。
2. 音源数が変化してもよい。
3. 漸進的に音を分離。
4. モノラル音では、前述した分離の曖昧性ができないことがある。

方向情報を使用し、分離精度向上を狙う
バイノーラル音（頭に組み込まれたマイクロフォン） Binaural HBSS (BiHBSS)

BiHBSS: 方向情報を取り込む

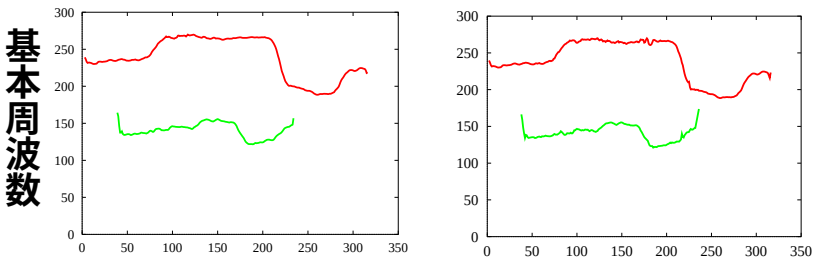


方向情報を求めるには取り込む

1. 左右のHBSSで同じ基本周波数を持つ調波構造ストリームを検出
2. 見つけた1対のストリームに対して、
   - IPD（両耳間位相差）
   - IID（両耳間強度差）
3. 頭の形から、
   - IPD（両耳間位相差） は1500Hz位まで
   - IID（両耳間強度差） は1500Hz以上で

BiHBSSによる音響ストリーム分離

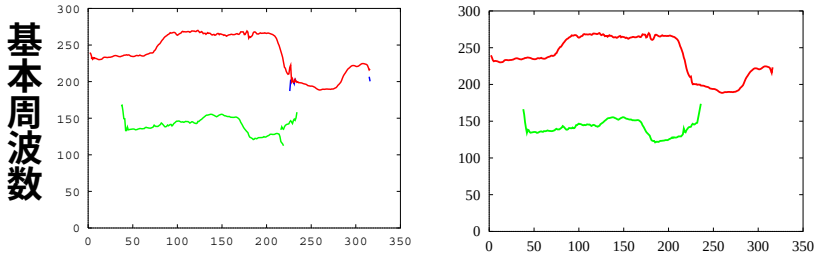
3つの音を演奏： 入力混合音、
分離音（女性の発話、男性の発話）



入力混合音　　BiHBSSによる分離音

HBSSとBiHBSSの性能比較

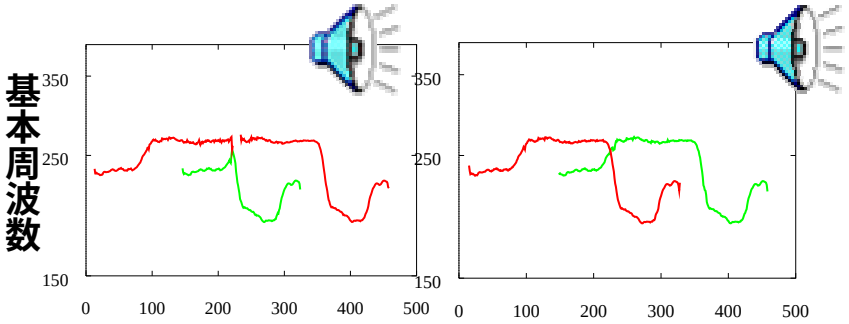
3つの音を演奏： 入力混合音、
分離音 (女性の発話、男性の発話)



HBSSによる分離音　BiHBSSによる分離音

BiHBSSによる音響ストリーム分離

2種の3つの音を演奏： 入力混合音、
分離音 (女性の発話、男性の発話)



HBSSによる分離音　BiHBSSによる分離音

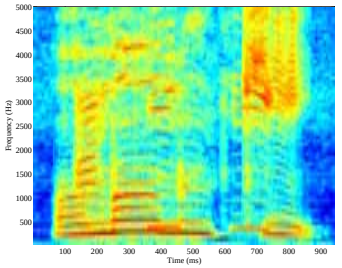
音声ストリームの分離

1. BiHBSSによる調波構造ストリーム分離
   - 母音、有声子音は調波構造を持つ
   - 無声子音は×
2. 無声子音の抽出は難しい。

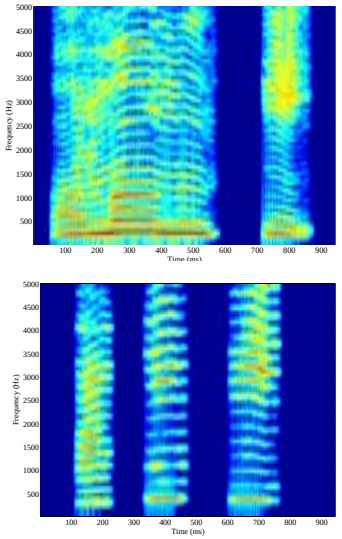
1. 残差はほとんど調波構造がふくまれないはず。
2. 残差を無声子音の代用とする

音声ストリームの分離:第1段階

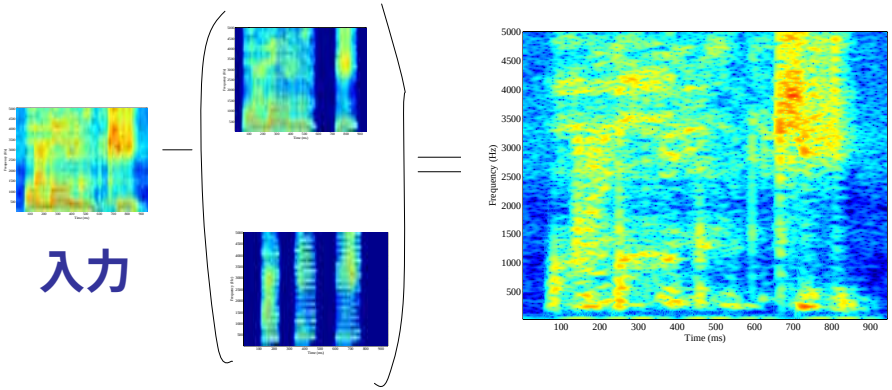
BiHBSSで分離



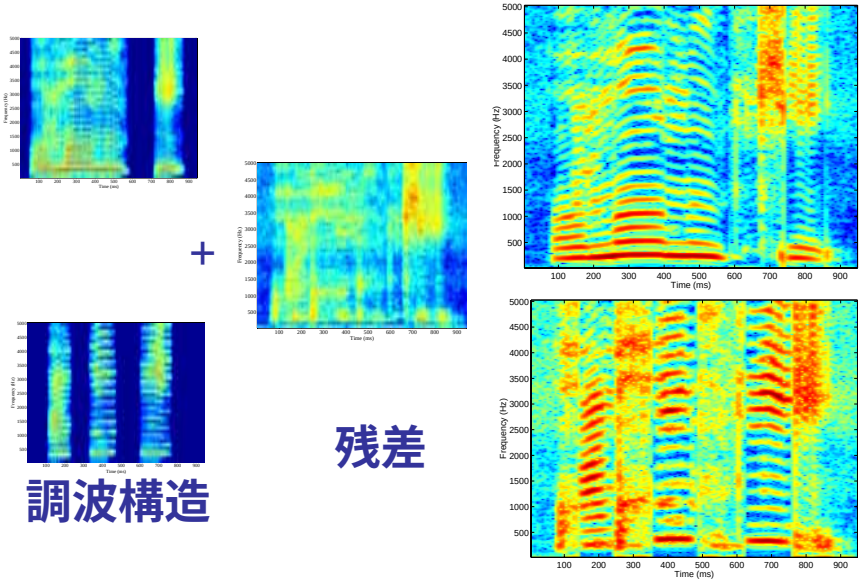
入力混合音



音声ストリームの分離:第2段階

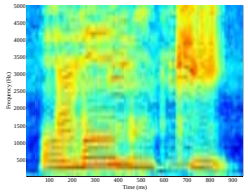


音声ストリームの分離:第3段階

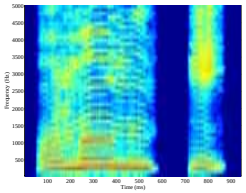


音声ストリームの分離:デモ

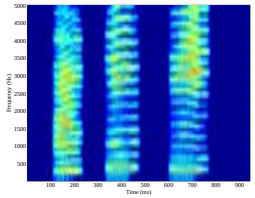
1 入力混合音



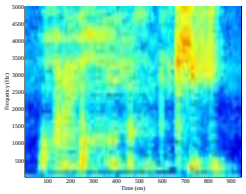
2 分離されたomomuki



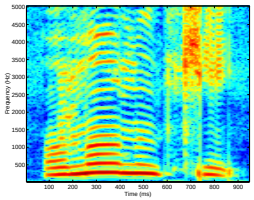
3 分離されたakiti



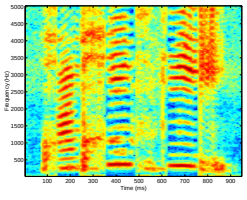
4 残差



5 分離された「おもむき」



6 分離された「あきち」



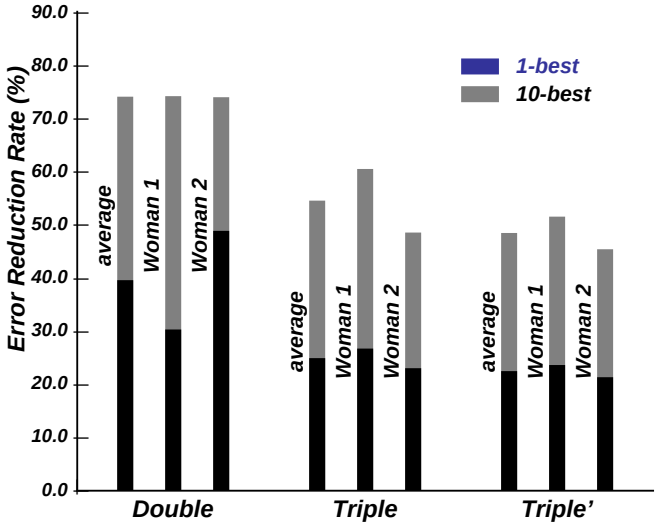
音声ストリーム分離の評価

1. 音声認識システムで単語発話を評価

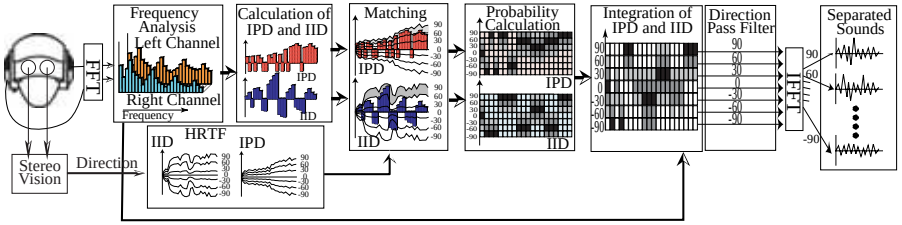
	第1話者	第2話者	第3音
Double	女性1	女性2	
Triple	女性1	女性2	弱いノイズ
Triple’	女性1	女性2	強いノイズ

2. 隠れマルコフモデル (HMM)による自動音声認識システムで評価
3. 日本語の単語５００組で評価。学習データと評価データは独立。

音声ストリーム分離の評価

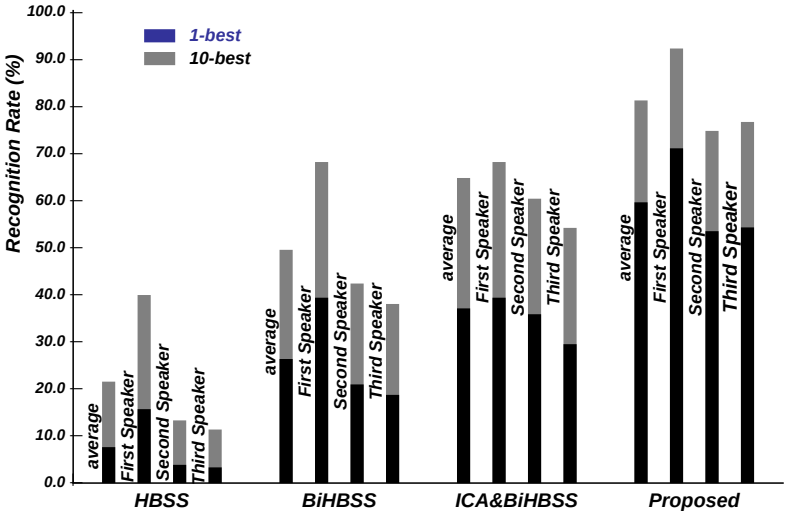


方向情報による画像との統合



1. サブバンド（FFTのポイント）毎に処理
2. IPD(両耳間位相差）とIID(両耳間強度差）の組合せ
3. 画像処理からIPD、IIDの予想値と実際の値との間で仮説推論

モーダリティ増加による音声ストリーム分離への効果



音楽分離システムとの統合

1. 異なる音響ストリーム分離システムを統合。
2. 音楽と音声