

聖徳太子コンピュータ —混合音処理・ロボット聴覚—

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



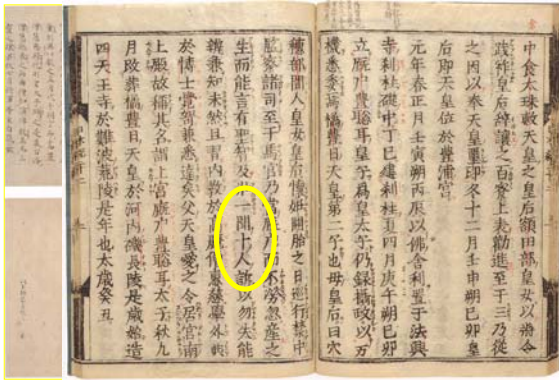
第1回目の目次 (Introduction)

1. 混合音処理
 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
- 第2回: 音源定位・音源分離
 - 第3回: ロボット聴覚
 - 第4回: 音楽情報処理・擬音語認識

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

聖徳太子が聞いたのは7人か10人?



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。言語によるコミュ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。

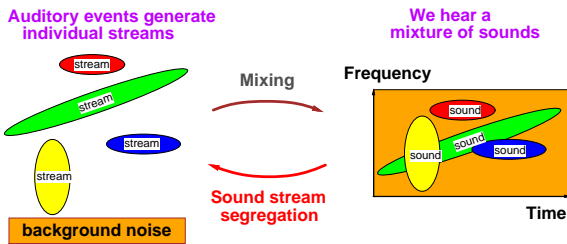
鈴木淳一、小林武夫共著『耳科学 —— 難聴に挑む』
(中公新書1598, 2001)

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。

7

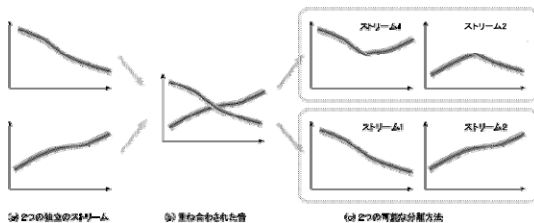
どのようにして私たちは音を聞くのか



単一の音ではなく、混合音を聞いている。

8

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。

9

音響ストリーム分離における曖昧性

17 Failure of crossing trajectories to cross perceptually.

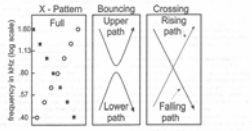
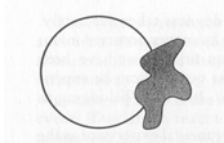


FIGURE 17
 The stimulus in this example is shown in Panel 1. It shows the lower half. For example, when listeners follow the right path, they track the ascending sequence in the crossing point and then shift over to the ascending sequence and follow it upward (Panel 2). This is called a "bouncing" pattern, because the subject's voice seems to bounce away from the crossing point. It is contrasted with a "crossing" pattern in which one sequence that rises or falls, and is followed right through the crossing point. The two possible link sequence patterns, rising and falling, are shown in Panel 3.

聴者は Bouncing のように聞こえ、 Crossing として聞くことは難しい。

10

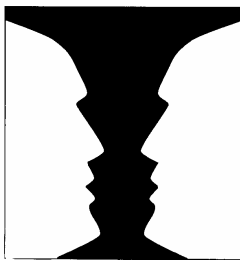
ストリーム化 — 部分の帰属性



円とシミの境界線はどちらに属する？

11

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)

12

ストリーム化 — 変化を検出

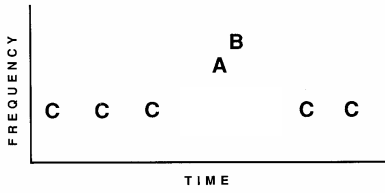


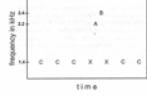
Figure 1.7
A tone sequence of the type used by Bregman and Rudnický (1975).

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

ストリーム化 — 変化を検出

16 The release of a two-tone target by the capturing of interfering tones.

An experiment by Bregman and Rudnický (1975) asked whether stream can be caused by attention or by a grouping mechanism. The strategy of the experiment was to cause a grouping mechanism to capture attention away from stream, thereby showing that the grouping mechanism was not part of attention. This demonstration is very subtle. The figure shows a number of tones, each followed by a tone. These tones A and B are to be judged for their order—ascending or descending in pitch. The order shown in the figure is ascending (A-B), but in the Bregman-Rudnický experiment, it could be either ascending (A-B) or descending (B-A).



When the A-B or B-A pair is played in isolation, it is easy to judge the order of A and B. This is demonstrated in the first part of the demonstration which gives the pitch of the order (B-A, as a standard), then in the order A-B (reversing the comparison). In the second part of the demonstration, we make this order demonstration very difficult by inserting the A-B or B-A pair by two instances of tone C (1,482 Hz), so that the A and B tones and four tone responses (A-B-A-B). Despite the fact that we are being asked to discriminate the tones A-B pair in the tone and four tone responses, the addition of the two interfering tones makes the task very difficult. This probably because the four tones form a higher order unit in which the A and B tones lose their precedence, in favor of the X tones, which fall at the perceptually prominent beginning and end positions. In the second part of the demonstration, a two-tone standard (A-B) is followed by a four tone test sequence (A-B-X-A-B), although A-B is in the same order as the tone and four tone responses, it is very hard to judge their order.

The X's are represented as an interference stream. Frequency varies with the C's, separating A and B from the stream. This generally isolates A and B from interference by the X's. Accordingly, in the first part of the demonstration, when a two tone standard (B-A) is followed by the full four tone sequence that contains the C's, it is fairly easy to tell that the A-B order is different in the four sequence.

Like Demonstration 15, this one illustrates the fact that there is a competition in the grouping of tones, so that adding more tones to a sequence can change the perceptual organization. In the present demonstration, we do so. This change in grouping changes the likelihood of some of the tones, and hence their perceptibility.

Technical details. The tones were sinusoidal, 57 notes in duration, including 7 notes that are 7 notes long (tones, with 7 notes between them). To make the tones appear equally valued, the amplitude of the 2,600 Hz tone was multiplied by .6, and that of the 2,000 Hz tone by .5, relative to that of the 2,200 Hz tone.

Reading: The experiment by Bregman & Rudnický (1975) is discussed in ASA-98 pp. 14, 153, 1, 148, 149, 150, 151, 444, 4, 453, 473. The most general effects of competition on perceptual grouping are discussed in ASA-98 pp. 145-171, 424, 524, 526, 505-511, 515, 481.



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

ストリーム化 — 周波数の差

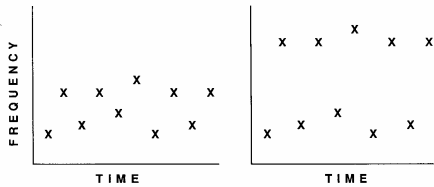
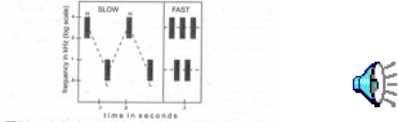


Figure 1.9
Stream segregation is stronger when the frequency separation between high and low tones is greater, as shown on the right.

ストリーム分離はどちらが強いのか？

ストリーム化 — 時間間隔

14 Stream segregation of high and low bands of noise.



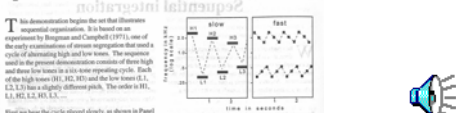
This demonstration shows that noise bands of tones, centered at different frequencies, will segregate into tones that vary in time. The concept is shown in an experiment by Deutsch and Campbell (1975), who were able to hear what noise bands with larger head edges. The tones can be combined with high and low tones in a single repeating cycle. The pitch is constant. The pitch is constant. The pitch is constant.

ストリーム分離はどちらが強いか？

19

ストリーム化 — 時間・周波数の差

1 Stream segregation in a cycle of six tones.

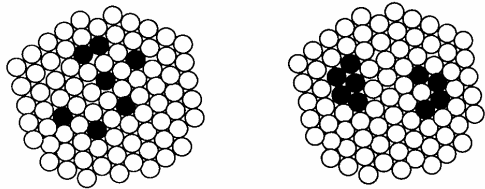


This demonstration begins the set that illustrates sequential organization. It is based on an experiment by Deutsch and Campbell (1975), one of the early examinations of stream segregation that used a cycle of alternating high and low tones. The experiment used in the present demonstration consists of three high and low tones in a six-tone repeating cycle. Each of the high tones (H1, H2, H3) and the low tones (L1, L2, L3) has a slightly different pitch. The order is H1, L1, H2, L2, H3, L3.

ストリーム分離はどちらが強いか？

20

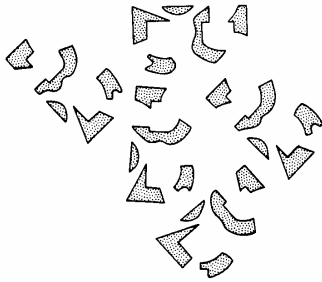
Gestalt Principles



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

21

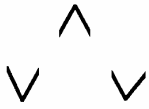
Gestalt Principle of Closure



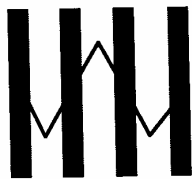
この文字は？ Occlusion (見え隠れ)？

25

Gestalt Principle of Closure



NO BURSTS



BURSTS

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

27

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

28

これまでの研究

1. 調波構造・音源方向を使った音響ストリーム分離
2. マルチエージェントシステムによる構成
3. 音声ストリームの分離
4. 音声と音楽の混合音の分離
5. 音響処理と画像処理との統合
6. ロボット聴覚
7. 実時間複数話者追跡
8. ソーシャルインタラクション

29

人は皆同じように感じ取るのか

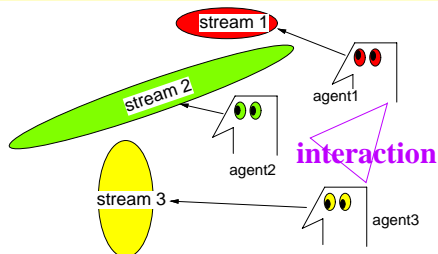
同じ情景を描いた2つの絵の違いは？



ゴヤ「1805年5月3日」 マネ「マクシミリアン皇帝の処刑」

30

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。

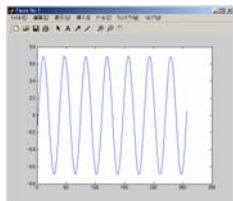
34

調波構造: 基本的な音の表現

- 基本周波数 $A(t)$

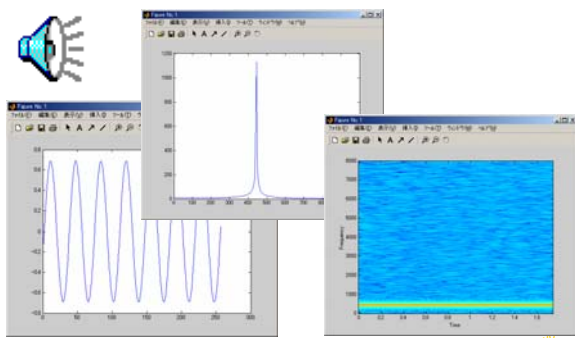
$$A(t) = \sum_i a_i \sin(\omega_0 i t + \theta_i)$$

- $i=1$ 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



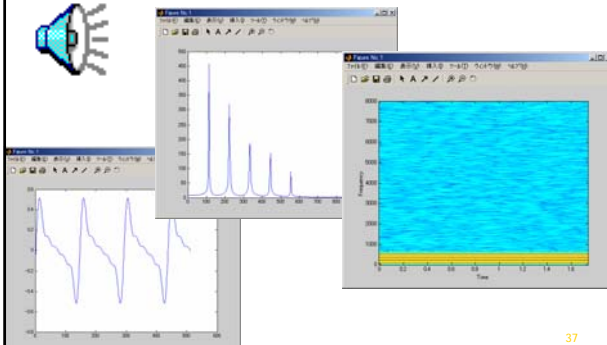
440Hzの純音を信号処理をすると

- 波形、フーリエ変換、スペクトル



110Hzの調波構造を持つ音

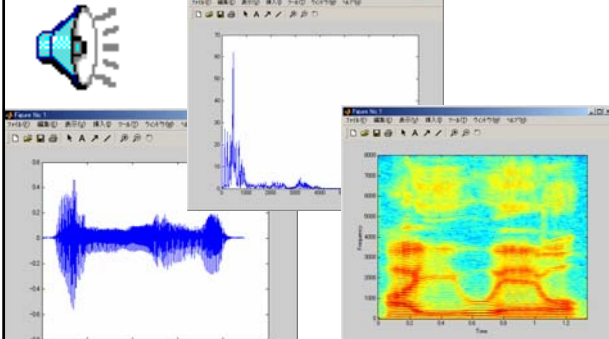
- 波形、フーリエ変換、スペクトル



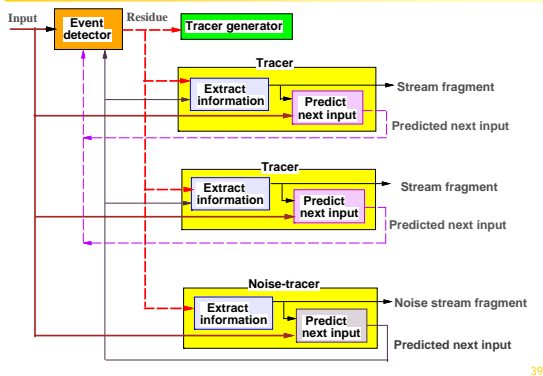
37

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



39

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。

40

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

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


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

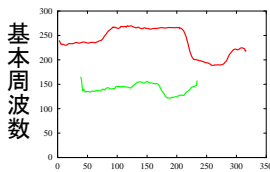
$$A(t) = \sum_i a_i \sin(\omega_0 i + \theta_i)$$

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。

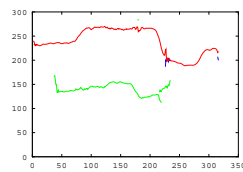
41

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

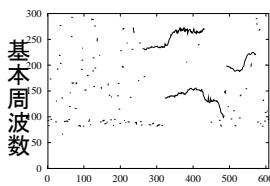


HBSSによる分離音

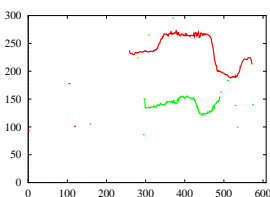
42

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

43

HBSSの能力

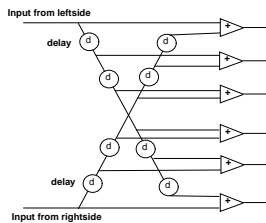
1. 音源数を仮定しない。
2. 音源数が増えたり減ってもよい。
3. 漸進的に音を分離。
4. モノラル音では、前述した分離の曖昧性ができないことがある。

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)

44

音源定位のJeffressのモデル

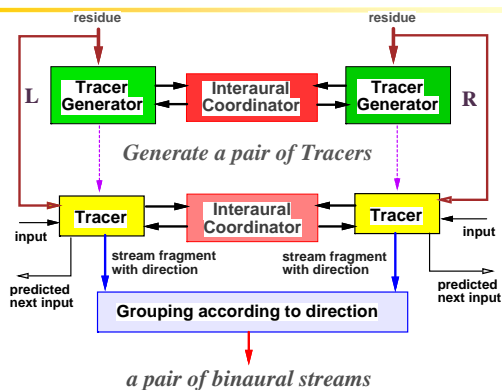
- 左右の耳に到着する時間差 (ITD, Interaural Time Difference, Delay Time of Arrival (DTOA))



- 同じ音という判断は
- 頭の形の影響は

45

BiHBSS: 方向情報を取り込む



46

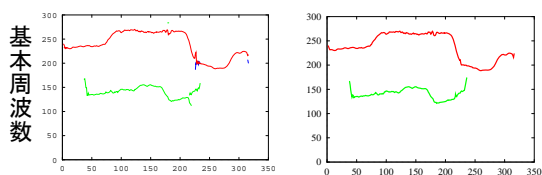
方向情報を求めるには取り込む

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

47

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)

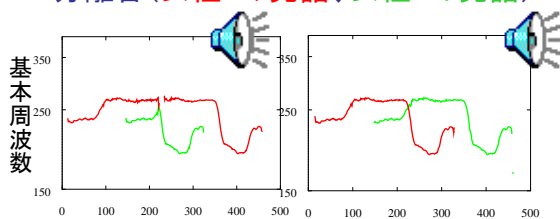


HBSSによる分離音 BiHBSSによる分離音

49

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1の発話、女性2の発話)



HBSSによる分離音 BiHBSSによる分離音

50

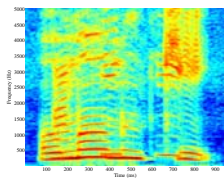
音声ストリームの分離

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

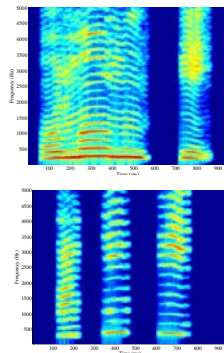
51

音声ストリームの分離: 第1段階

BiHBSSで分離

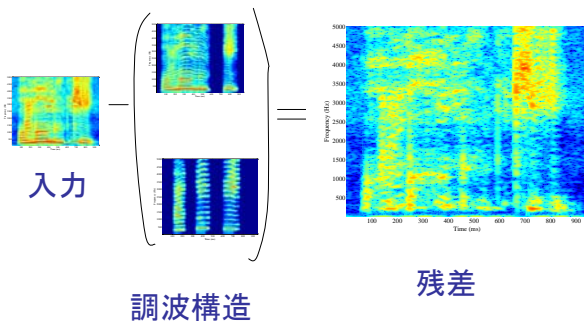


入力混合音



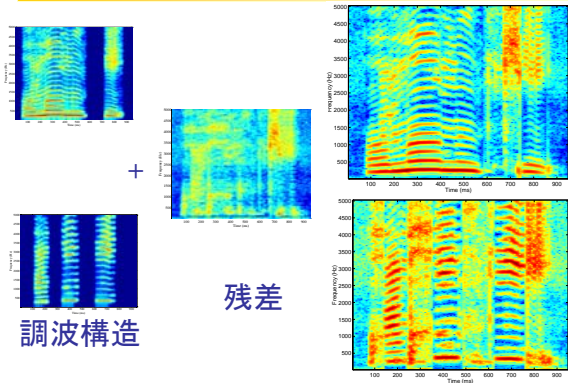
52

音声ストリームの分離: 第2段階

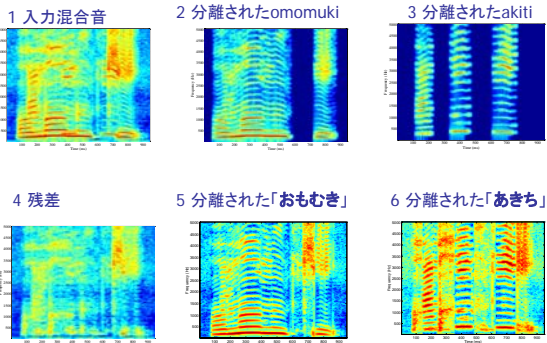


53

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

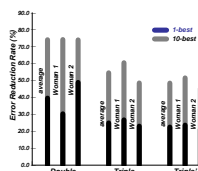


55

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

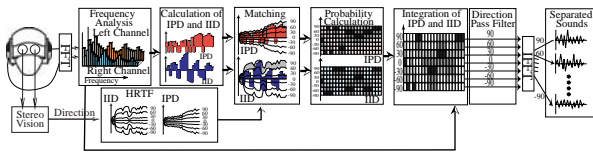
	第1話者	第2話者	第3音
Double	女性1	女性2	
Triple	女性1	女性2	弱いノイズ
Triple'	女性1	女性2	強いノイズ



- 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
- 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。

56

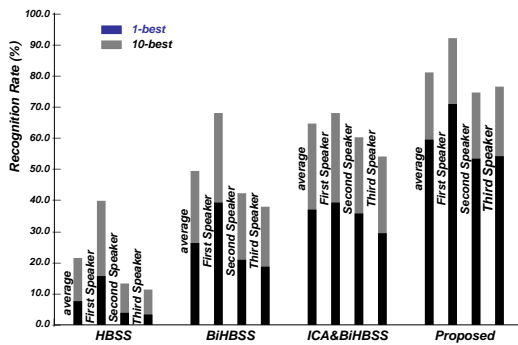
方向情報による画像との統合



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

58

モーダリティ増加による音声ストリーム分離への効果



59

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



60