音声情報処理特論 2003年12月9日

# 画像との情報統合による混合音の分離・認識・ロボットの行動制御

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



## 目次



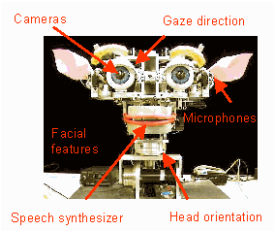


## ソーシャルインターラクション

- 感情を表出
  - 顔、声の調子
- 身振り・手振り
  - 知覚と結びついた挙動
  - 指差しによる共通の注意を喚起
- 常識が通じる
  - 非言語的レベル：声・顔・しぐさ、言語的レベル
- 五感が使える
  - 「見分ける」、「聞き分ける」、「読み分ける」、「言い分ける」、・・・



## Kismet (MIT AIL) の顔の表情





Interest


Calm


Angry


Happy


Sad


Surprise


Disgust

# ロボット聴覚の課題

1. アクティブオーディション
2. 音一般の認識・理解
3. 階層的な情報統合
4. 実時間処理
5. 注意の制御



# 階層的情報統合（画像・音響）

1. 音響信号・映像信号
2. 音素(phoneme)・口形素（viseme）
3. 音源定位・３Ｄ位置
4. 話者認証・顔認証（識別・照合）
5. 感情（顔の表情・音声の表情）
6. 話の内容
7. …



# 口形素（viseme）

- 日本語　１３個
- 米語　２１個
- ドイツ語　１２個
- フランス語　２１個

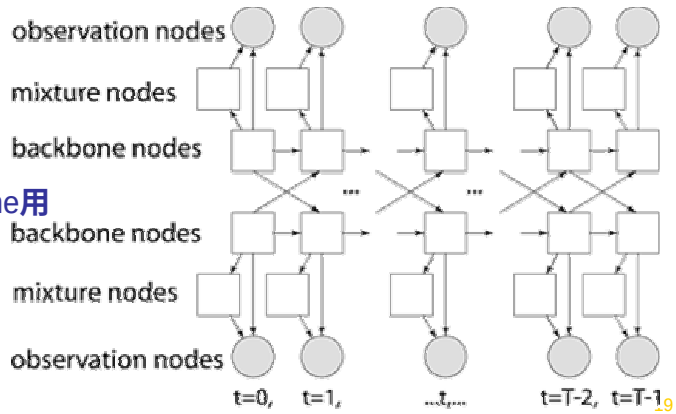
Table 4　Mouth shape symbols for the Japanese vowels and consonants.

| Phonetic symbol | | Mouth shape symbol | |
|---|---|---|---|
| [a] [e] [o] [u] [i] | → | *a* *e* *o* *u* *i* | Shape of the mouth of each of the five vowels |
| [p] [b] [m] | → | *p*: | Shape of the mouth of the lip closure |
| [Φ] [w] | → | *w*: | Shape of the mouth of the lip narrowing |
| [t] [d] [n] | → | *t*: | Quick upward and downward movement of the tongue tip |
| [r] | → | *r*: | Lower side of the tongue |
| [ts] [dz] [s] [z] | → | *s*: | Constriction by attaching the tongue tip to the alveolar |
| [tS] [dʒ] [ɲ] [S] [ʒ] [ç] | → | *sy*: | Constriction by attaching the tongue to the hard-palate |
| [j] | → | *y*: | Transition from the shape of the mouth similar to [i] |
| [k] [g] [ŋ] [h] | → | *Vf*: | Shape of the mouth of the following vowel |

# 読唇術（Lip Reading）

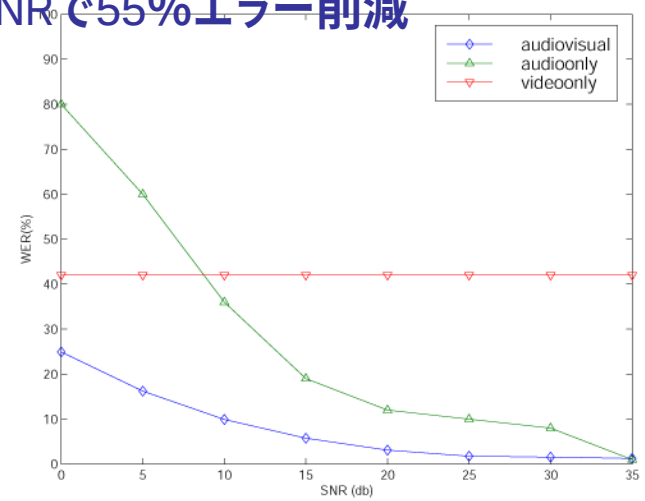
1. Audio-Visual Speech Recognition
2. Coupled HMM（Intel の OpenCV、IBMも同様）

Phoneme用



# Intel の Coupled HMM の性能

- XM2VTS データベース (295話者)
- 0dbのSNRで55%エラー削減




# 表情･感情認識の統合

■顔表情 (Ekman &Frisen)

- Facial Action Coding System： 44個のAction Unit (AU)
- Pleasure, Fear, Sadness, Disgust, Anger, Surprise に関係するのは、AU1 (眉の内側を上げる) , 2 (眉の外側を上げる) , 4 (眉を下げる) , 5, (上瞼を上げる) 6 (頬を上げる) , 7 (瞼をしかめる) , 9 (鼻に皺) , 10 (上唇を上げる) , 12 (唇端を引き上げる) , 15 (唇端を下げる) , 17 (顎を上げる) , 20 (唇を横に引張る) , 25 (顎を下げずに唇を開く) , 26 (顎を下げて唇を開く)

■音声からの感情認識 (Fernald vs Johnston & Scherer)

- Intensity, F0 floor/mean, F0 variability, Sentence contour, High frequency energy, Speech and articulation rate を
- Stress, Anger/rage, Fear/panic, Sadness, Joy/elation, Boredom にマップ
  - ➢Pitch mean/variance, Maximum/minimum pitch, Pitch range, Delta pitch mean, Absolute delta pitch mean, energy mean/variance/range, Maximum/minimum energy を
  - ➢Approval, Attention, Soothing, Neutral, Prohibition

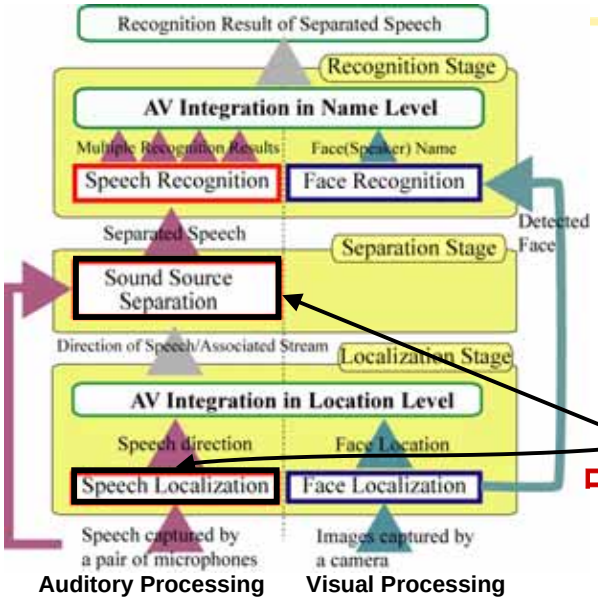


# 実時間情報統合の課題

1. 情報統合の単位
2. 同期の仕組み
3. 分散処理



# AV除法統合によるロボット聴覚システム

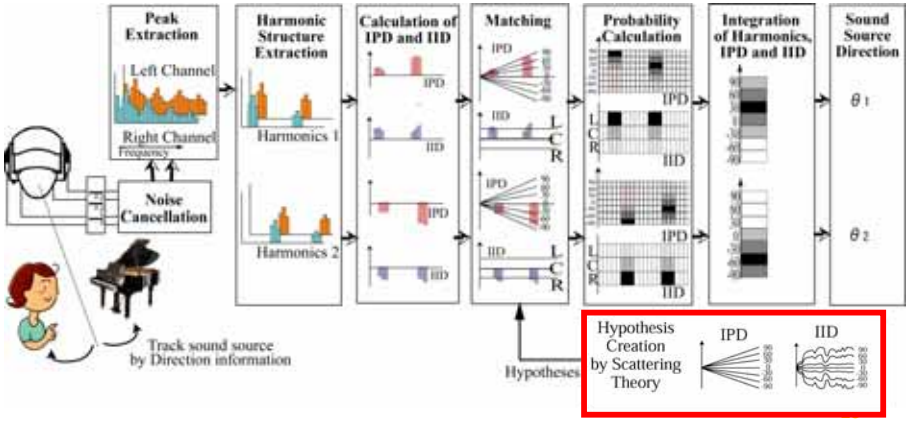



- 3つのステージ
  - 定位
  - 分離
  - 認識
- 各ステージに視聴覚統合
  - 位置レベル
  - 名前レベル

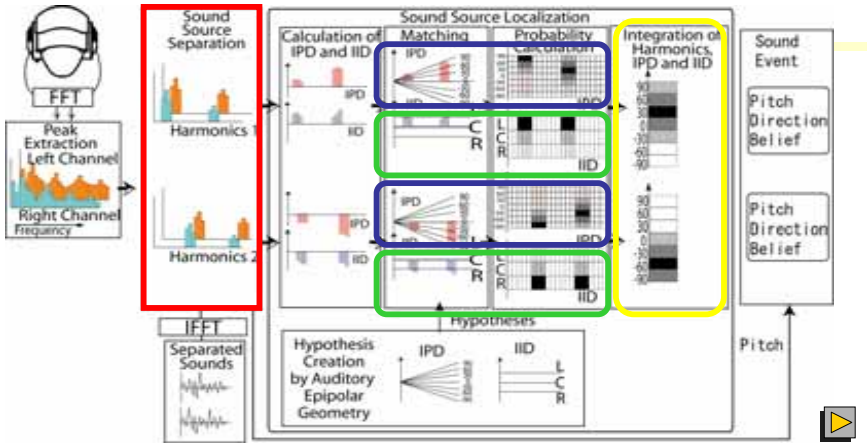
ロボット頭部音響モデルの使用

# 複数の音響情報を統合した音源定位

- 倍音構造に注目した音のグルーピング
- IPD (両耳間位相差) とIID (両耳間強度差) を利用した仮説生成・照合
- IPDとIIDの確信度を Dempster-Shafer結合則で統合し、最大確信度をもつ方向情報を出力



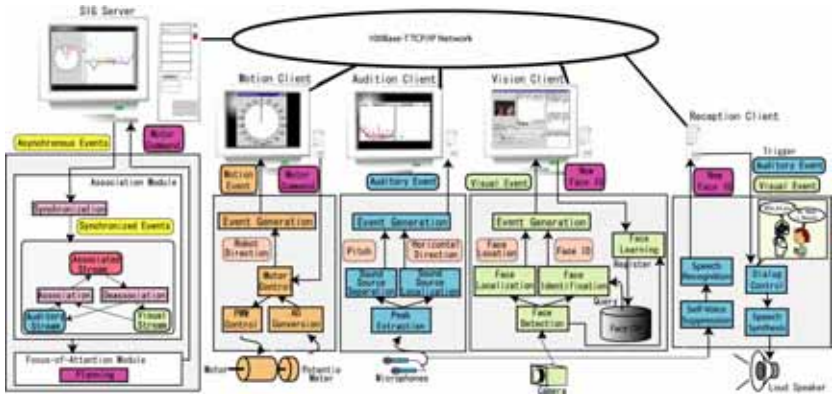

# 仮説推論による複数音源定位




- 倍音構造に注目した音源分離
- IPD (両耳間位相差) と聴覚エピポーラ幾何による仮説生成・照合 (1500Hz以下)
- IID (両耳間強度差) を利用した右左の判断 (1500Hz以上)
- IPD と IID を Dempster-Shafer で統合し確信度をつき方向情報を算出
  →ロバスト性の確保



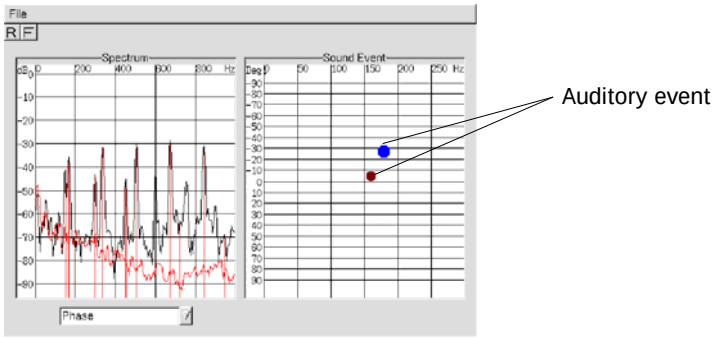
# 実時間複数話者追跡システム




ステレオ画像・顔認識・音源同定・モータ制御の情報統合
注意制御・音声合成による対話

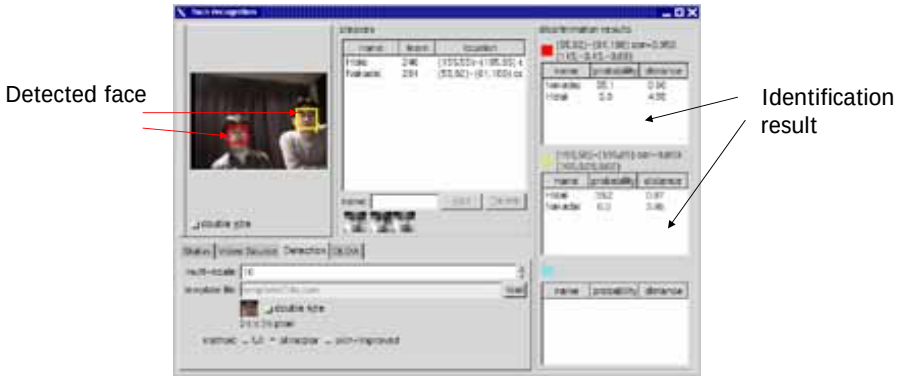


# 音響エベント




- X and Y axes mean pitch and direction.
- Diameter of each event is proportional to belief factor.
- Multiple events are detected by sound source separation.
- A event has 20 best candidates on sound direction.

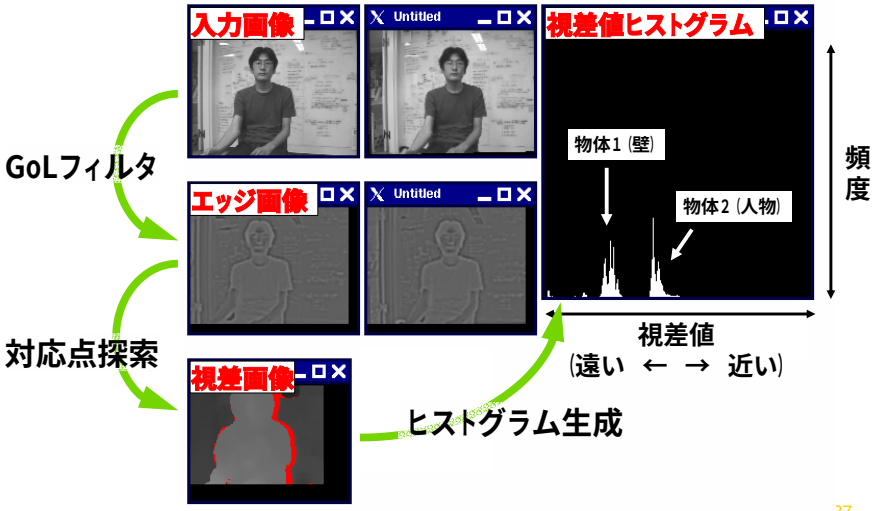
# 顔認識モジュール




- Multiple face detection by skin-color extraction and correlation based pattern matching
- Face identification by Online LDA
- Face localization with coordinate conversion
- Event has 5-best candidates on face ID and face location

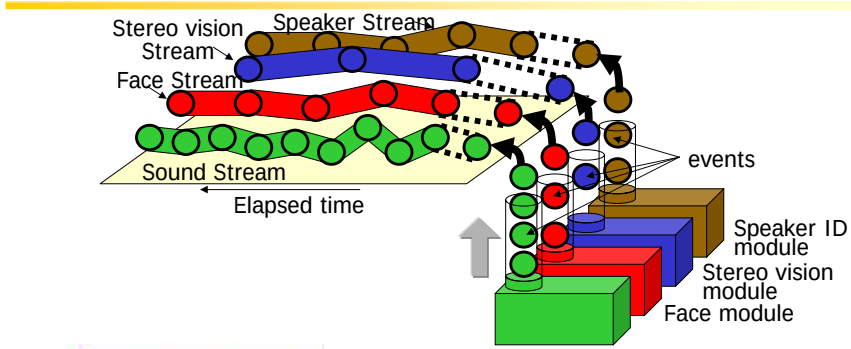


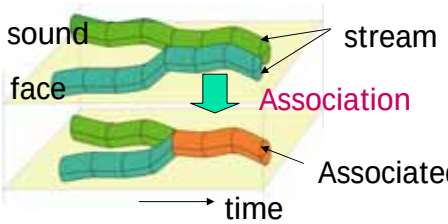
# ステレオ視による物体定位






# ストリーム形成とアソシエーション

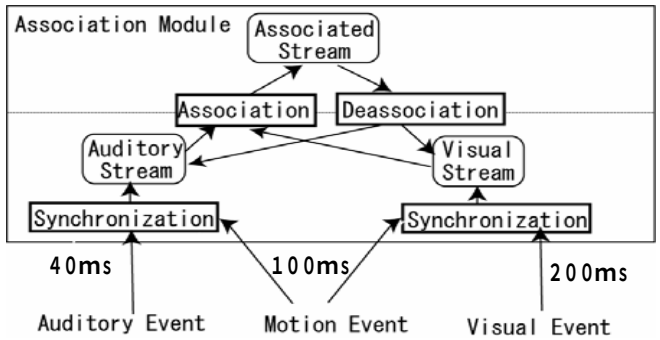



角度が近い（±10°）か距離が近ければ、アソシエーションする



# アソシエーションモジュール




- イベントの同期（2秒の短期記憶使用、100ms周期で同期、500msの遅延）
- ストリームの生成（聴覚イベント±10度、視覚イベント40cm以内のものを時間方向に接続）
- アソシエーションストリームの生成（視聴覚ストリームが1秒間で500ms以上近いと判断されたとき）

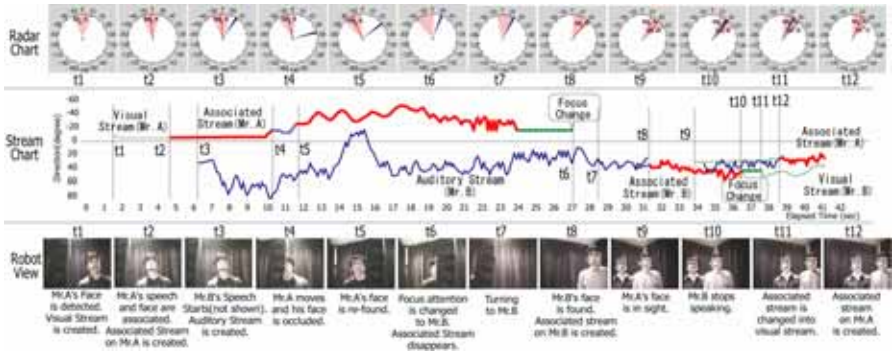
## 物体定位におけるセンサ情報の特徴比較

| | 有効範囲 | 存在判定 | 定位精度 | 定位次元 | 個人識別 |
|---|---|---|---|---|---|
| 音源定位 | 全方位 | △<br>(発声が前提) | △<br>(ばらつき多い) | 1 | × |
| 顔認識 | 45度<br>(視野角依存) | △<br>(正面顔のみ) | ○<br>(顔の大きさを仮定) | 3 | ○ |
| ステレオ視 | 45度<br>(視野角依存) | ◎ | ◎ | 3 | × |



## トラッキングにおける統合の効果

2つのチャートがocclusion や視野外の話し手をうまくトラックできていることを示している。





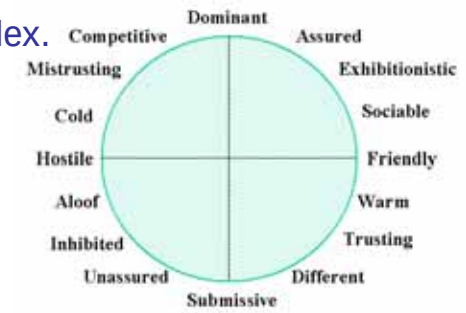

## 注意制御 (Focus-of-Attention)

- タスク指向･ソーシャル性指向
- 受付ロボットは、タスク指向：
  顔を同定し、認識し、認識結果に基づいて音声応答をし、音声認識、さらには顔データベースの更新を行う；
  *associated stream > visual > auditory*
- コンパニオンロボットは、ソーシャル性指向
  新たな音のするほうに顔を向ける；
  *auditory stream > associated > visual*



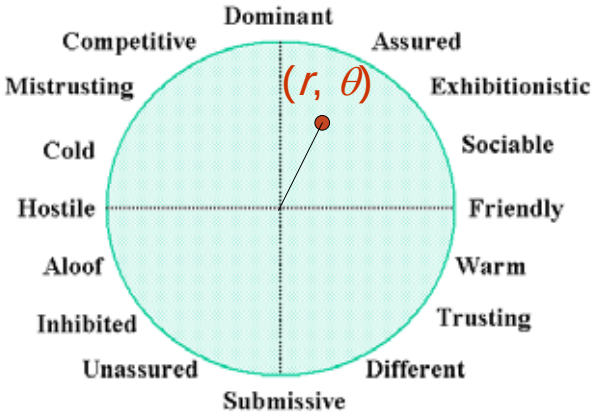
## Interpersonal Theoryによる個性

1. Interpersonal theory based on two dimensions
   - Dominant vs. submissive, friendly vs. hostile
   - Big five theory is not appropriate for the current humanoid robots
   - Dominant/submissive, friendly, conscientious, emotional stable, open
2. Interpersonal circumplex.



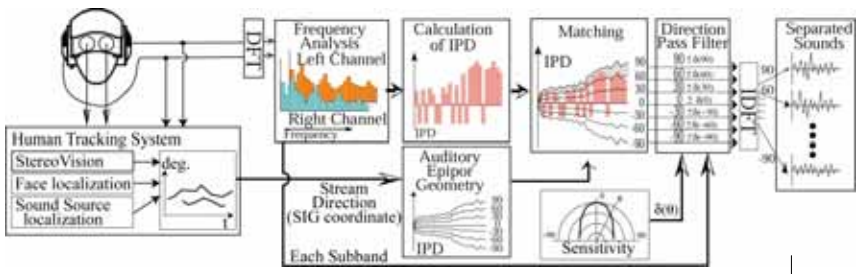
# 注意制御における個性

- Personality is represented as a point $(r, \theta)$ in the Interpersonal Circumplex.

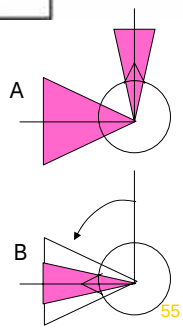





# アクティブ方向通過型フィルタ



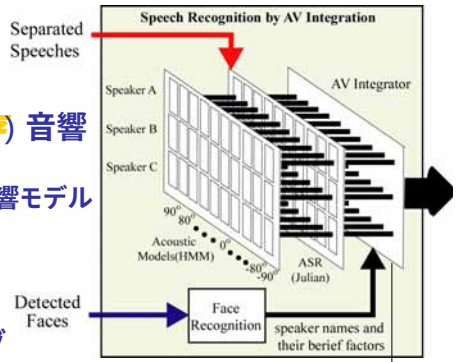

- 聴覚用エピポーラ幾何と両耳間位相差 (IPD) による音源分離
- ストリームの位置情報を利用
  → より精度の高い位置情報が得られる
- 音源方法に依存した適応的な感度制御(A)
- アクティブな動作による感度向上(B)






# 分離音の認識

- 顔認識と音声認識の統合
- 分離音の認識
  - •複数の方向･話者依存(DS依存) 音響モデル
    - 17 方向×3話者 -> 51 DS依存音響モデル
    - 51 音声認識システムの並列実行
    - 語彙数は150 語
- 顔認識
  - •一般的なテンプレートマッチング
  - •オンライン線形判別分析
  - •確信度付の候補を生成



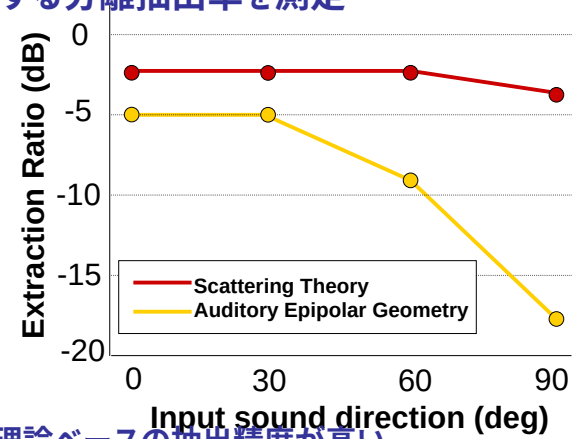

$$V(p_e)=\left(\sum_d r(p_e,d)v(p_e,d)+\sum_p r(p,d_e)v(p,d_e)-r(p_e,d_e)\right)P_v(p_e)$$

$$v(p,d)=\begin{cases}1 & if \quad \mathrm{Res}(p,d)=\mathrm{Res}(p_e,d_e),\\ 0 & if \quad \mathrm{Res}(p,d)\neq\mathrm{Res}(p_e,d_e).\end{cases}$$



# 散乱理論による音源分離の性能向上

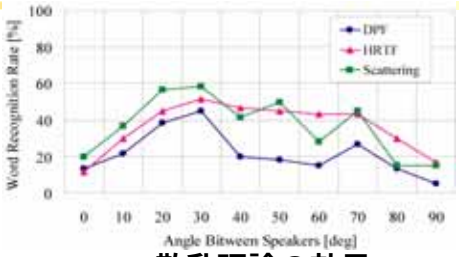
- 100Hz の調波構造音 (100Hz – 3kHz) の分離抽出
- 音源方向に対する分離抽出率を測定



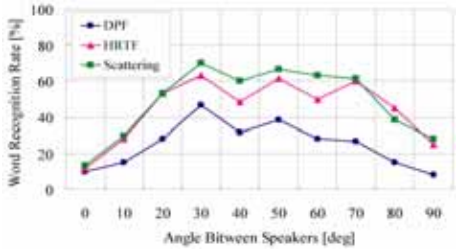

- 全体的に散乱理論ベースの抽出精度が高い
- 横方向の音に対するロバスト性の向上

# 散乱理論による音声認識の性能向上

- 三話者同時発話の音声認識
- スピーカ間の角度を0-90度まで変化させ認識実験
- 20回認識実験での三話者各々の平均認識率
- 20% の認識率向上 ← 音源定位･分離の向上による
- 視聴覚統合により認識率が10〜20%向上




散乱理論の効果

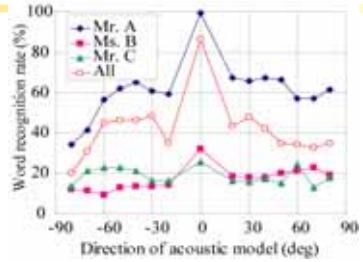



視聴覚統合の効果

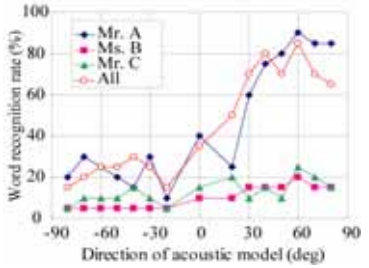


# DS依存音響モデルによる単語認識率

- WRR is the highest (> 80%), when the **speaker name** and **direction of an acoustic model** are coincident with those of input.
- WRR tends to be high, when either **speaker name** or **direction** is coincident.




Mr.A and 0 deg.



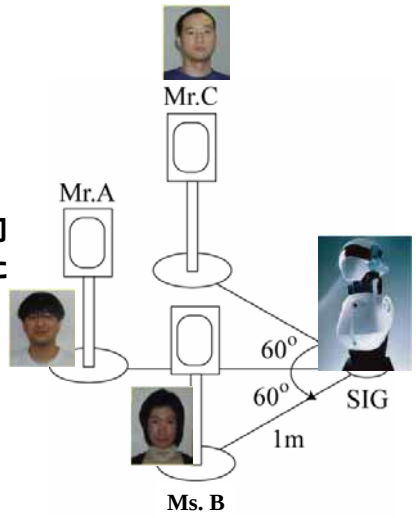

Mr. A and 60 deg.



# 3話者同時発話の認識

**基本的なシナリオ**

1. ロボット (SIG)が質問する.
2. 3人がいっせいに答える.
3. SIG はそれぞれの音声を定位･分離･認識する
4. SIG はそれぞれのスピーカに向いてそれぞれの人が何を喋ったかを当てる.






# まとめ

1. 混合音からの音源分離は聴覚処理の基本。
2. 人間とロボットとのインターラクションでは、ソーシャルインターラクションとアクティブパーセプションが本質。
3. 環境からの情報取得には、音響と画像を統合した実時間複数話者トラッキングシステムが重要な役割を果たす。
4. 注意制御と組合せたアクティブパーセプションはソーシャルインターラクションの本質。
5. 受動的なソーシャルインターラクションの実例を示し、その効果を確認。