音声情報処理特論 2005年11月29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

1

目次

2

ロボットのカンブリア紀大爆発

- 生命:カンブリア紀(BC5億7千万年～5億年)
- 多様な種が出現
- Brooks: "**Cambrian Intelligence: The Early History of the New AI**"
- ヒューマノイドロボットの多様化

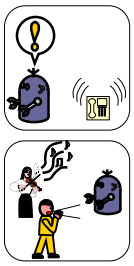
人間との共生にはソーシャルインタラクションが不可欠

ロボットの知覚処理、特に聴覚機能は、ソーシャルインタラクションが可能なほどロバストではない

7

人間と共生するロボットでは…

- **日常的環境で使用されるロボット**
  - 人間と音声によるインタラクション
  - 音源が見えなくても抽出すべき
  - 雑音下でも特定の音に注目できるべき（カクテルパーティ効果）



ロボットにとって聴覚は本質的な機能

8

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」、・・・

11

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

12

アクティブオーディションの課題

1. 混合音など一般的な音の理解
   - CASA (音環境理解、Computational Auditory Scene Analysis)
2. センサ情報統合
   - 信号レベル
   - シンボリックレベル
3. ノイズキャンセル
   - 動作中のモータノイズは避けられない。
   - 一般にマイクはモータの近くにあるため、ノイズは比較的大きな音として収音されてしまう。

13

アクティブオーディションシステムの設計

- 調波構造による音源分離
- 実IPD(両耳間位相差)と音響エピポーラによるIPDとの照合による仮説生成 (< 1500Hz)
- IID(両耳間強度差)による左右の判別 (> 1500Hz)
- IPDとIIDをDempster-Shaferにより統合しロバストな処理の実現

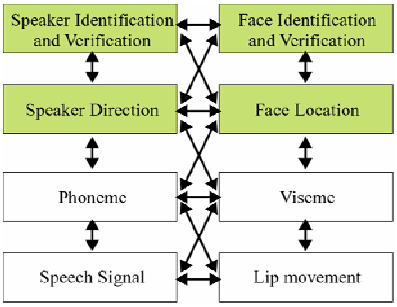
14

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. …

15

階層的情報統合（画像・音響）



16

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.

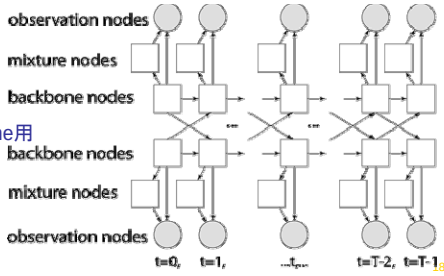
Phonetic symbol	Mouth shape symbol	
[a] [e] [o] [u] [i]	→ [a e o u i]	Shape of the mouth of each of the five vowels
[p] [b] [m]	→ *p*:	Shape of the mouth of the lip closure
[Φ] [w]	→ *w*:	Shape of the mouth of the lip narrowing
[t] [d] [n]	→ *t*:	Quick upward and downward movement of the tongue tip
[r]	→ *r*:	Lower side of the tongue
[ts] [dz] [s] [z]	→ *s*:	Constriction by attaching the tongue tip to the alveolar
[tS] [dZ] [n] [S] [Z] [j]	→ *sy*:	Constriction by attaching the tongue to the hard-palate
[j]	→ *y*:	Transition from the shape of the mouth similar to [i]
[k] [g] [0] [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）

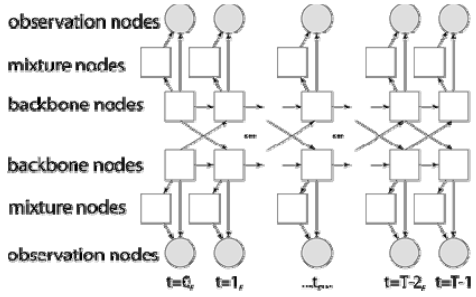
Phoneme用

Viseme用



18

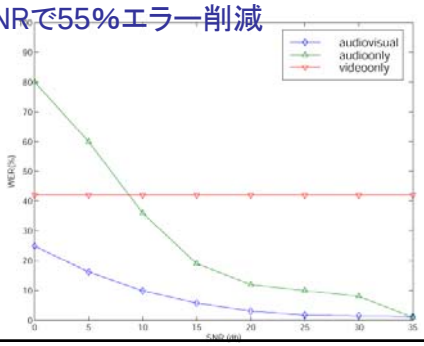
Coupled HMM for AV 統合



19

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

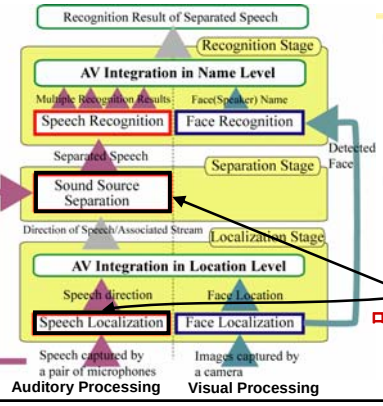
22

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

23

AV情報統合によるロボット聴覚システム



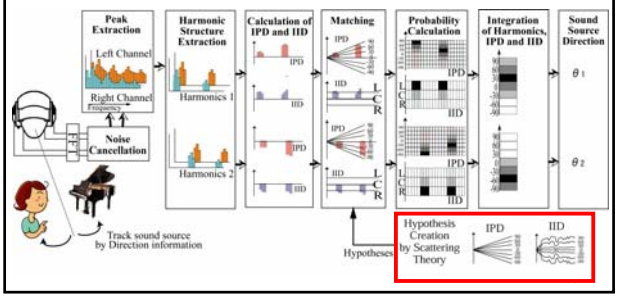
- 3つのステージ
  - 定位
  - 分離
  - 認識
- 各ステージに視聴覚統合
  - 位置レベル
  - 名前レベル

ロボット頭部音響モデルの使用

25

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（IPDは1500Hz以下、IIDは1500Hz以上）

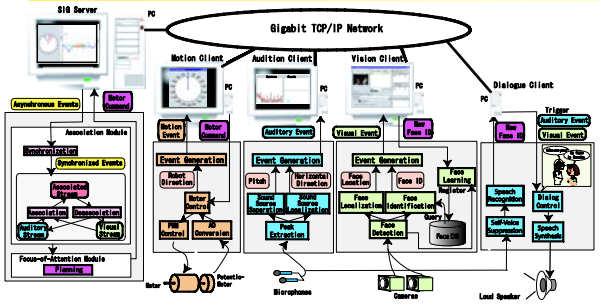


実時間複数話者追跡システム

1. 挙動による有効性の検証には実時間処理が不可欠。
2. アクティブパーセプションでは、動作が悪影響を及ぼすので、個々の感覚の曖昧性を情報統合により解消する必要がある。
3. 階層的な情報統合による曖昧性の解消。

27

実時間複数話者追跡システム

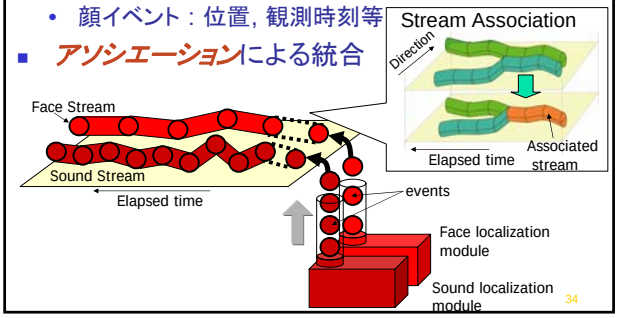


ステレオ画像・顔認識・音源同定・モータ制御の情報統合
注意制御・音声合成による対話

28

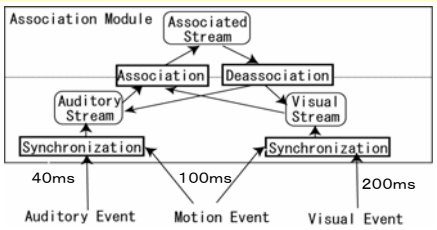
視聴覚の方向情報統合

- *ストリーム：イベント*を時間方向に接続したもの
  - 音イベント：方向，観測時刻，ピッチ等
  - 顔イベント：位置，観測時刻等
- *アソシエーション*による統合



34

アソシエーションモジュール

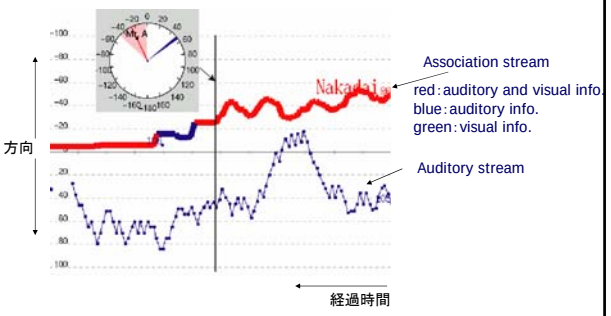


- イベントの同期（2秒の短期記憶使用、100ms周期で同期、500msの遅延）
- ストリームの生成（聴覚イベント±10度、視覚イベント40cm以内のものを時間方向に接続）
- アソシエーションストリームの生成（視聴覚ストリームが1秒間で500ms以上近いと判断されたとき）

36

ストリーム形成

Radar and stream chart



37

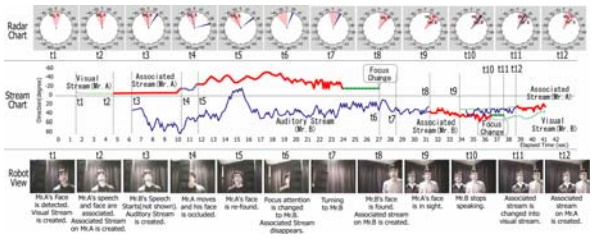
物体定位におけるセンサ情報の特徴比較

	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	△（発声が前提）	△（ばらつき多い）	1	×
顔認識	45度（視野角依存）	△（正面顔のみ）	○（顔の大きさを仮定）	3	○
ステレオ視	45度（視野角依存）	◎	◎	3	×

40

トラッキングにおける統合の効果

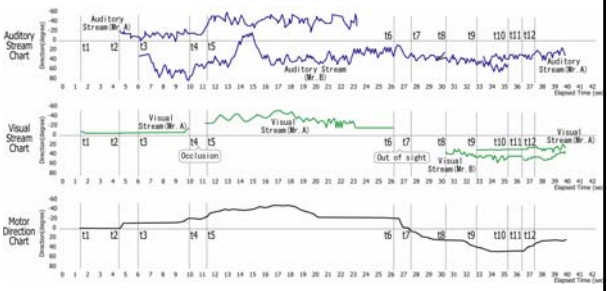
2つのチャートがocclusion や視野外の話し手をうまくトラックできていることを示している。



41

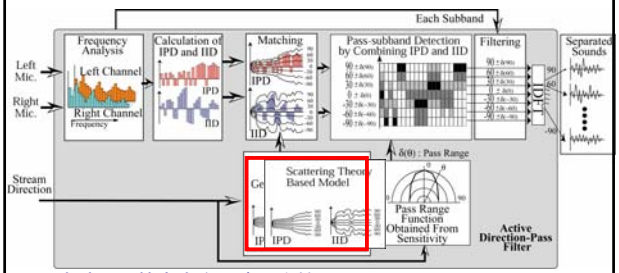
トラッキングの正確さ

アソシエーションによりトラッキングの曖昧性が解消



42

音源分離：アクティブ方向通過型フィルタ



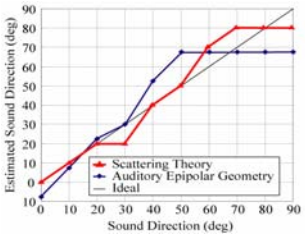
- 実時間で特定方向の音源を抽出
- 両耳間位相差（IPD)、両耳間強度差 (IID) – 2つのマイク
- 入力方向情報からIPD, IID仮説の生成
- 聴覚中心窩ベースの通過帯域制御
- 選択条件：$|\ IPD_{Input} - IPD_{Hypo}(\theta)\ | < \delta_{IPD}(\theta)$
  $|\ IID_{Input} - IID_{Hypo}(\theta)\ | < \delta_{IID}(\theta)$

44

実験1: 音源定位

- 100Hz の調波構造音 (100Hz – 3kHz) の定位

音源定位結果

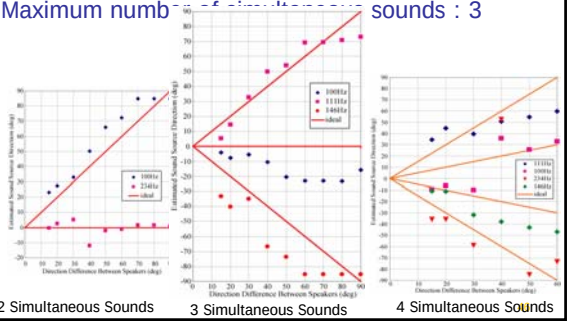


- 50度までは、同程度の精度
- 50度以上になると散乱理論の精度が高い。

45

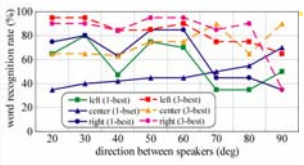
Performance of Sound Localization

- Sound localization : sound mixture of harmonics 100Hz, 111Hz, 146Hz and 234Hz
- Maximum numb[illegible] sounds : 3

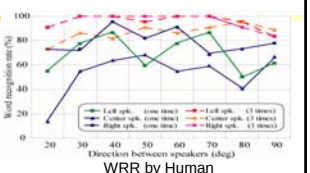


2 Simultaneous Sounds　3 Simultaneous Sounds　4 Simultaneous Sounds

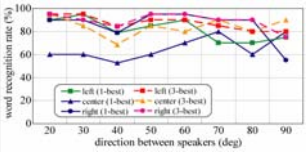
視聴覚統合の効果



WRR (baseline)



WRR by Human

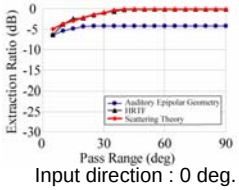


WRR improved by AV integration

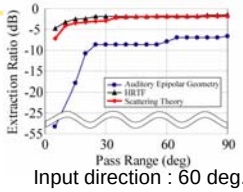
- 3-best の結果が、1-best と同じぐらいにまで向上
- 同じ環境内では、人間のパフォーマンスを超える結果.

47

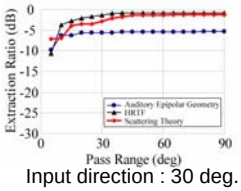
Evaluation 2: Sound Extraction



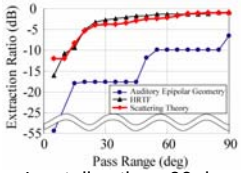
Input direction : 0 deg.



Input direction : 60 deg.



Input direction : 30 deg.



Input direction : 90 deg.

48

IPD、IIDによる確信度の算出

- IPDによる確信度
  - 聴覚エピポーラ幾何によるIPD仮説との距離 *d(θ)* を算出し、正規分布を仮定して、確信度に変換する

$$d(\theta) = \frac{1}{n_{th}} \sum_{i=0}^{n_{th}-1} \frac{(\Delta\varphi_h(\theta, H(i)) - \Delta\varphi_s(i))^2}{H(i)}$$

$$B_{\mathrm{IPD}}(\theta) = \int_{-\infty}^{\frac{d(\theta)-m}{\sqrt{\frac{\pi}{n_{th}}}}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx$$

nth: 1500Hz 以下の倍音数
H(i)：*i* 番目の倍音の周波数

- IIDによる確信度
  - 各倍音のIIDの総和から左右、正面方向を判断し、下表に定義する確信度を割り当てる

$$S_I = \sum_{i=n_{th}}^{n-1} \Delta I_s(H(i))$$

	θ	$90° \sim 35°$	$30° \sim -30°$	$-35° \sim -90°$
S_I	+	0.35	0.5	0.65
	-	0.65	0.5	0.35

Dempster-Shafer理論

- ベイズ理論ではうまく表せない人間の主観にかかわる確信度を表すことが可能 [Dempster 67]
- Dempster の結合則は独立な証拠から推論された基本確率を結合できる

$$m(A_k) = \frac{\sum_{A_{1i} \cap A_{2j} = A_k} m_1(A_{1i})\, m_2(A_{2j})}{1 - \sum_{A_{1i} \cap A_{2j} = \phi} m_1(A_{1i})\, m_2(A_{2j})}$$
$$(A_k \neq \phi)$$

A1i, A2j：焦点要素
m1, m2：基本確率

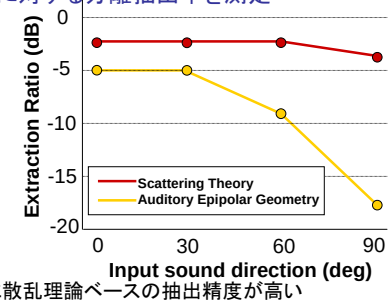
$$B_{\mathrm{IPD+IID}}(\theta) = m(A_1) = 1 - (1 - B_{\mathrm{IID}}(\theta))(1 - B_{\mathrm{IPD}}(\theta))$$

- IPD, IID の統合に利用

50

実験2: 音源分離

- 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

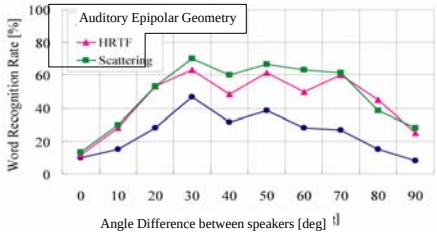


- 全体的に散乱理論ベースの抽出精度が高い
- 横方向の音に対するロバスト性の向上

51

実験3:三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回認識実験による、三話者各々の認識率の平均



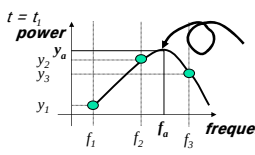
Speech recognition rate

- 20% の認識率の向上 ← 音源定位・分離の向上による

52

ピッチの抽出

- Spectral subtraction（残差駆動）方式
- 高速で高精度
  - BiHBSS より 200 倍高速
  - FFTの離散性による周波数分解能の曖昧性解消

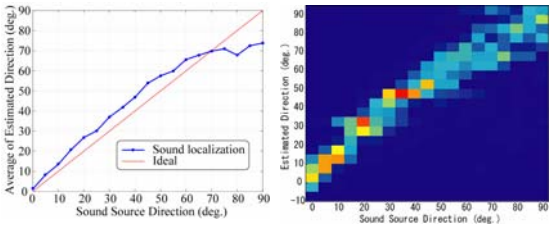


$$y = A_0 \sin(2\pi f_a t_1 + \Phi_0)$$

$$f_a = f_2 - \frac{-|y_2| + 2|y_3|}{T(|y_2| + |y_3|)}$$

53

Direction-dependent Sensitivity of DPF



Distribution of sound source localization (average)

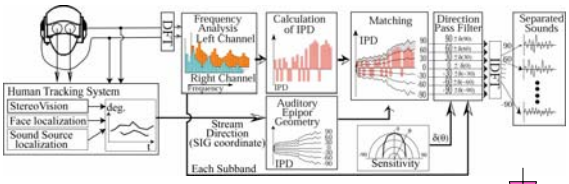
Distribution of sound source localization (variance)

- The error of localization is bigger for the periphery.
- Phenomenon similar to ***visual fovea***

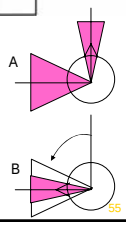
→ *auditory fovea* in Direction-Pass Filter

54

アクティブ方向通過型フィルタ



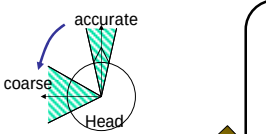
- 聴覚用エピポーラ幾何・散乱理論と両耳間位相差（IPD）による音源分離
- ストリームの位置情報を利用
  → より精度の高い位置情報が得られる
- 音源方法に依存した適応的な感度制御(A)
- アクティブな動作による感度向上(B)

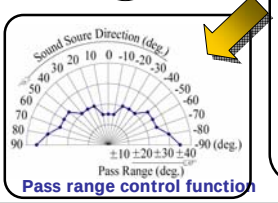


55

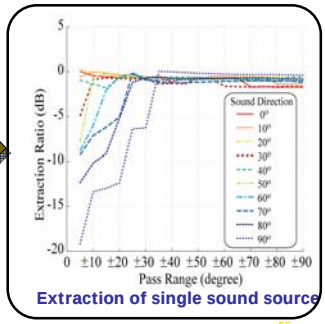
Auditory Fovea based Pass Range Control

- Optimal pass range depends on sound source direction.





Pass range control function

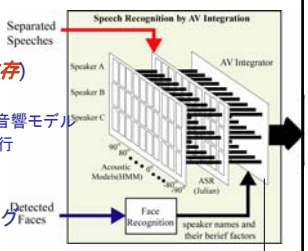


Extraction of single sound source

56

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
- 統合方法
  - 一種のBagging法

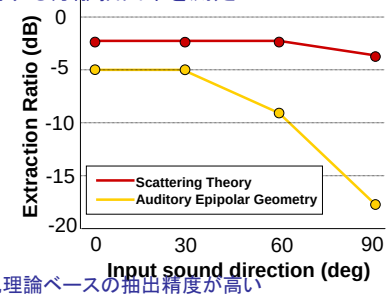


$$V(p_e)=\left(\sum_d r(p_e,d)v(p_e,d)+\sum_p r(p,d_e)v(p,d_e)-r(p_e,d_e)\right)P_t(p_e)$$

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

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

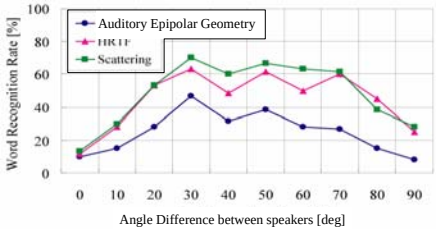


- 全体的に散乱理論ベースの抽出精度が高い
- 横方向の音に対するロバスト性の向上

60

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

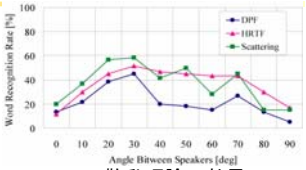


- 20% の認識率向上 ← 音源定位・分離の向上による

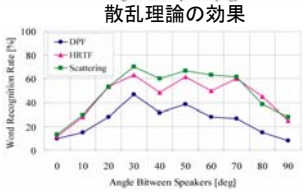
61

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

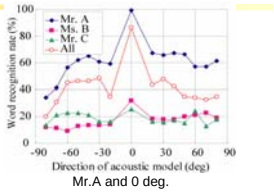


視聴覚統合の効果

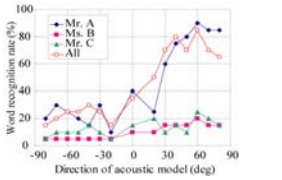
62

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.



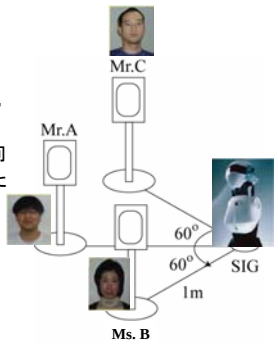
Mr. A and 60 deg.

63

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.



64

三話者同時発話認識（英語版）

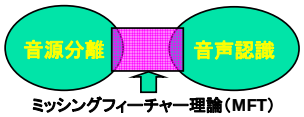
Case 1: 音声認識でエラーがない場合

69

音源分離と音声認識の統合

音源分離と音声認識双方の処理にとって親和性の高いインタフェース

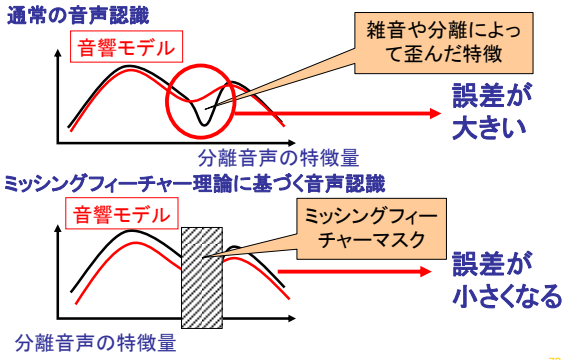
- 実時間動作が可能な音源分離
  - 音源分離エラーによる歪や雑音への対応
- 音声認識
  - 環境に特化しない音響モデル



クリーンな音声で学習した**単一の音響モデル**を利用し，歪，雑音には**MFT**で対応

72

ミッシングフィーチャー理論 (MFT) とは



73

実験に用いたヒューマノイド

3体のヒューマノイド ASIMO, SIG2, Replie を利用

ヒューマノイド	耳介	外装	頭部の形状
ASIMO	なし	硬質	角ばった球形
SIG2, Replie	人間から型をとったシリコン製	シリコン	人間に類似した形状（卵型）

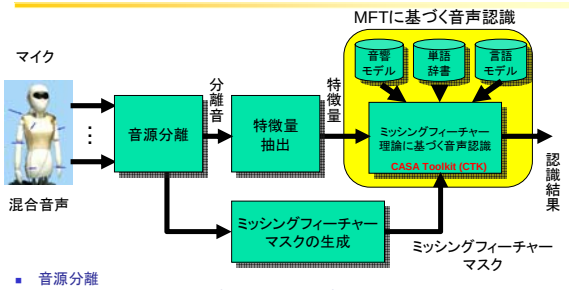

ASIMO


SIG2


Replie

75

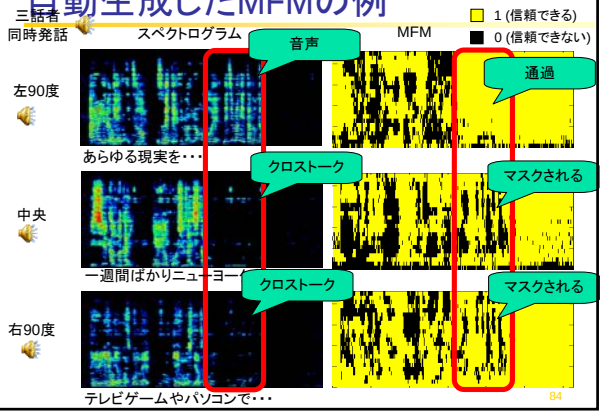
音源分離と音声認識の統合



- 音源分離
  - マイクから入力された混合音を分離して，分離音を出力
- ミッシングフィーチャーマスク生成
  - 音源分離から特徴量，雑音，歪等の情報を得て，ミッシングフィーチャーマスクを生成
- MFTに基づく音声認識
  - スペクトル特徴量，ミッシングフィーチャーマスクを入力とし音声認識を行う

76

自動生成したMFMの例



84

三話者同時発話認識デモ

Robovie-R2
8本のマイク

85

注意制御 (Focus-of-Attention)

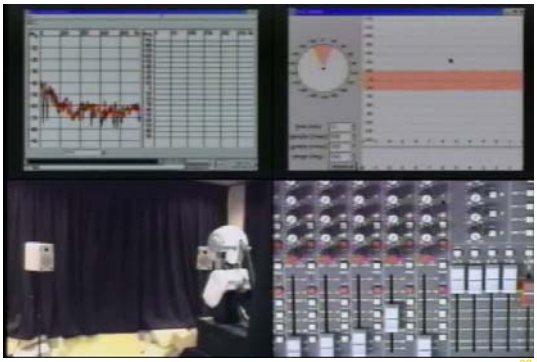
- タスク指向・ソーシャル性指向
- 受付ロボットは、タスク指向：
  顔を同定し、認識し、認識結果に基づいて音声応答をし、音声認識、さらには顔データベースの更新を行う；
  associated stream > visual > auditory
- コンパニオンロボットは、ソーシャル性指向
  新たな音のするほうに顔を向ける；
  auditory stream > associated > visual

87

4話者へ注意を向ける

91

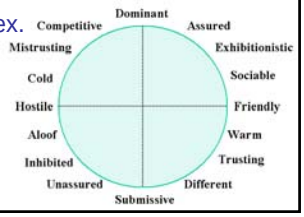
ステレオバランス変化への追従



92

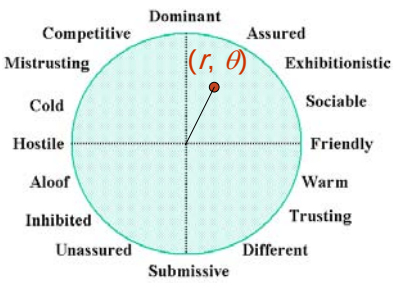
Interpersonal Theoryによる個性

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, θ) in the Interpersonal Circumplex.



94

コミュニケーションにおける距離の概念

- 近接学 (Proxemics)
  人間は、相手との関係性に応じて互いの距離を変化させるという社会心理学の理論
  - 親密距離（intimate distance） ～45cm
    - 身体接触を伴う、非常に親しい者同士の距離
  - 個人距離（personal distance） 45cm～120cm
    - 比較的親しい者同士の距離
  - 社会距離（social distance） 120cm～360cm
    - 親しくない・面識のない者同士の距離
  - 公共距離（public distance） 360cm～
    - 個人的関係のない者同士（講演者と聴衆etc）の距離

97

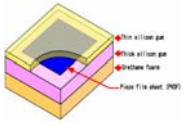
複雑な挙動の設計

- **距離に着目：近接学（Proxemics）**
  - **親密距離 (~45cm)**
  - **個人距離 (~1.2m)**
  - **社会距離 (~3.6m)**
  - **公共距離 (3.6m~)**
- **距離に関する入出力装置**
  - **無指向性スピーカ**
  - **超指向性スピーカ（20度）**
  - **肌センサー**
- **個性・親密度に基づいた挙動**
- **自己のセンサー能力を知った挙動**

98

距離に応じたモダリティ

- **ステレオビジョン**
  - 頭部2カメラの視差画像による距離計測 **ピエゾセンサー**
- **皮膚センサ（親密距離）**
  - 圧力速度を認識するセンサ
  - 上半身（頭部含む）19箇所
  - 多層構造による柔らかさの実現
    →「*触る・叩く・撫でる*」が識別可



肌センサー

- **超指向性スピーカー（社会・公共距離）**
  - スピーカーの正面方向20度にのみ音場を作成

  →距離に関係なく、特定の人に話しかける

パラメトリックスピーカ

99

距離による挙動選択の例（動画）

1. Aさん挨拶
   （社会距離）
   →Aさんに会釈
2. Bさん挨拶
   （親密距離）
   →Bさんに挨拶
3. Aさん呼びかけ
   （社会距離）
   →接近を促す
4. Aさん接近
   →声かける

ステップ2と4は無指向性のスピーカーを利用
ステップ3は超指向性スピーカーを利用

100

親密度による挙動選択の例（動画）

1. Aさん挨拶（親密距離）
   →Aさんに挨拶
   →Aさん親密度増加
2. Bさん挨拶（社会距離）
   →Bさんに会釈
   →Aさんに向く
3. Aさん撫でる（親密距離）
   →Aさん親密度増加
4. Bさん声かけ（社会距離）
   →無視
5. Aさん叩く（親密距離）
   →Aさん親密度減少
   →拒否行動・回避

ステップ3と5で皮膚センサを利用

102

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。

104

今後の課題

1. ソーシャルインターラクションのためにより多くの知覚チャンネルの提供
   - More robust sound source separation and speech recognition
   - Speaker identification and verification by interacting with face identification and verification
   - Motion-based focus-of-attention control
2. より能動的なソーシャルインターラクションの実現
   - Take initiative in dialogue
   - Stop speech output by interception of new sound
   - Allow decay of belief for unseen objects
3. *Without measurement, there is no science.*

105

レポート課題

- レポートは3題のうち、2題選択回答

1. 柏野さんの講演（情報学展望）のまとめと感想（10ptで A4 5ページ以上）
2. 音源定位・音源分離・分離音認識について2つ以上の技法を詳細に報告（10ptで A4 5ページ以上）
3. 音声に限らず一般の音に対して，他のメディアとの情報統合によりどのような機能が実現できるのか，よりロバストな処理が可能になるか，について論ぜよ．（10ptで A4 5ページ以上）。

- レポートの締切は12月21日（水）
- 提出先は10号館レポートボックス

106