

音環境理解とロボット聴覚

情報学展望1A

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

工学部情報学科計算機科学コース(兼任)

<http://winnie.kuis.kyoto-u.ac.jp/~okuno/>

okuno@i.kyoto-u.ac.jp



1

究極の目標：聖徳太子ロボット



- 人と共生するには多様な混合音の聞き分けが重要
 - 音源定位・分離・認識
 - 楽音・環境音
- 発達論的なコミュニケーション機能の獲得
 - signal-to-symbol変換
 - パラ言語・情報統合
- 複数話者の同時発話認識

4

講師自己紹介

- 東京大学教養学部基礎科学科卒業
- 電電公社入社。26年間半、NTT基礎研究所を主に勤務
- スタンフォード大学客員研究員、東京大学工学部電子工学科客員助教授
- 1998/10 科学技術振興事業団 北野共生システムプロジェクトに退職参加
- 1999/4 東京理科大学理工学部情報科学科教授に採用
- 2001/4 京都大学大学院情報学研究科教授に採用
- 工学部情報科学計算機科学コース兼任
- 知的システムの設計・構築のための *enabling 技術*
- Lispマシンの基礎研究から商用化
- 音環境理解、ロボット聴覚、音楽情報処理などの研究
- NTT研究所のネットワーク構築(1983年頃)
- JUNETのボランティア(1984年頃)
- NTT製LispマシンをARPANet(Stanford Univ.)に接続作業(1987年)
- 日本最初のHTML文書(日本国憲法)を作成(1992年)
- 日本の挑戦「インターネットの夜明け」Yahoo! Japan 提供第6話・第9話に登場 (京大も)

2

音メディア処理の課題

- 音を利用したコミュニケーション・音環境理解のための基礎技術の研究開発
- 混合音の処理(聞き分ける)
 - 音源分離・音源定位・分離音の認識
 - 物理量 vs 心理量
- 信号・記号変換(人工知能の根源的な課題)
 - 音声認識、音楽認識、
 - 環境音・動物の鳴声⇒擬音語認識
- シンボルグラウンディング(記号の意味付け)
 - パラ言語、言語の獲得
- ロボット聴覚
- ソーシャルインタラクション

6

本日の話題

1. なぜ音がロボットやシステムでテキストや画像ほどには使われないのか。
2. 人と同じようにロボットやシステムが聞くことができるのか。
3. 音を聞くときにどのような機能が必要か。
4. ロボットやシステムは現在どれだけのことが聞き分けられるのか。
5. 音楽は現在どの程度聞き分けられるのか。
6. 今後の展開

3

Computational Auditory Scene Analysis

CASA (音環境理解) understands auditory environments by sounds.

- Computer modeling of hearing and listening
- Recognize general sounds including speech, music, environmental sounds and their mixture
- Basic functions for CASA
 - Sound source localization (音源定位)
 - Sound source separation (音源分離)
 - Separated sound recognition (分離音認識)

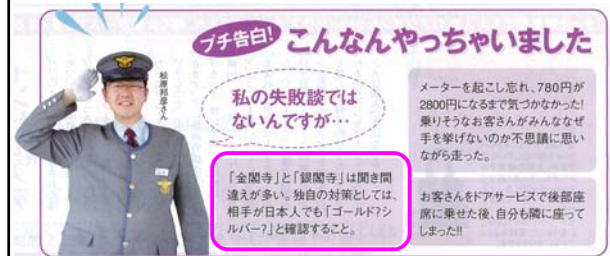
ASA(聴覚による情景分析)



Albert Bregman

音声対話システムの課題

- システムが聞き間違え
- 人が聞き間違え



本日の話題

1. なぜ音がロボットやシステムでテキストや画像ほどには使われないのか.
2. 人と同じようにロボットやシステムが聞くことができるのか.
3. 音を聞かるときにどのような機能が必要か.
4. ロボットやシステムは現在どれだけのことが聞き分けられるのか.
5. 音楽は現在どの程度聞き分けられるのか.
6. 今後の展開

12

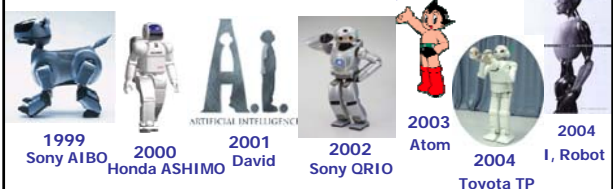
音声認識組み込みシステム

こちら葛飾区亀有公園前派出所@秋本治、秋田書店



ロボット聴覚研究の目的・背景

- 近年のロボット研究の目覚ましい発展



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ボバトではない

13

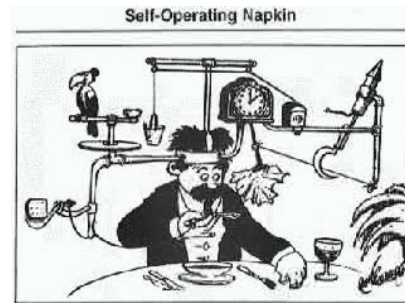
聴覚の重要性

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)

14

身体性を使った音研究： 定性的な処理だけでは不十分



動機：音情報の氾濫

情報爆発にも
なっていない

■ 現状認識：

- 音は氾濫、しかし、技術は単一発話音声およびごく少数の音を認識する。量的爆発の促進
- 音楽、環境音、雑音などなし
- 混合音から特定音の抽出が困難
- 1956年夏Dartmouth会議でAI誕生。しかし、音響信号記号変換の研究の進展はほとんどなし。

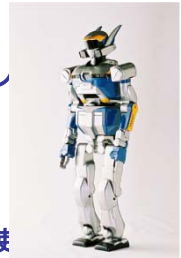
■ 望ましいのは：

- いろいろな音の認識（センサネット）
- LifeLogによる音の記録と分析（explosion sound）
- 音と画像で行動の認識し、挙動へ応用
- 自分の出す音を抑制したい（音の自己モデル獲得）

15

ロボット聴覚の現状

1. 耳がない。なかった。
2. 主に人につけたマイクロフォンで会話。
3. 喋ることはできる。
4. 愛知万博での受付ロボットでは外付けのマイクロフォンを使用。



20

研究のテーマを分かりやすく言うと

1. 聞き分ける

- 事前知識最小の「聞き分け」機能
- ラピッドプロトタイピングによる「聞き分け」
- 市販CD音楽、実環境下での「聞き分け」
- 実時間処理を目指す高速「聞き分け」

2. 身体性

- 身体の音響的特徴を活用した「聞き分け」
- 自己生成音のモデル化・抑制

3. 音を使った発達論的コミュニケーション

- 対話能力の自己獲得の工学的モデル化
- シンボル・グラウディングによる言語獲得

現実のロボットの音声によるインタラクション



21

既存のロボットのマイクロフォンは

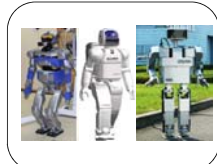
QLIO SDR-4XII

- 7本のマイクロフォン
- 内1本は内部雑音除去用
- 音源定位は行う
- 音源分離は行わず



ASIMO

- 2本、音源定位のみ
- HRP-2(産総研・川田)
- 耳はなかった。今は8本



23

混合音を取り扱うには

1. 入力装置, マイクロフォンは
2. 音源定位(音源方向)
3. 音源分離(特定の音を分離)
4. 分離音の認識
5. 実時間処理・情報統合
6. アクティブオーディション
 1. 自分の出す音を抑制する機能が必要
 2. 自己の音に関するモデル
 3. 肌センサーも同様
7. 定位・分離・認識に基づいたロボットの挙動設計

28

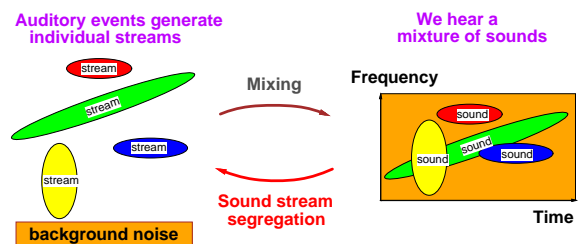
WIRED Magazine

情報学展覧1A 2007年5月18日

July issue of 2004



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。

31

本日の話題

1. なぜ音がロボットやシステムでテキストや画像ほどには使われないのか。
2. 人と同じようにロボットやシステムが聞くことができるのか。
3. 音を聞くとときにどのような機能が必要か。
4. ロボットやシステムは現在どれだけのことが聞き分けられるのか。
5. 音楽は現在どの程度聞き分けられるのか。
6. 今後の展開

25

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

32

混合音からの情報抽出時の仮定

1. Sparseness

- in 周波数領域
- in 時間領域
- in 空間領域

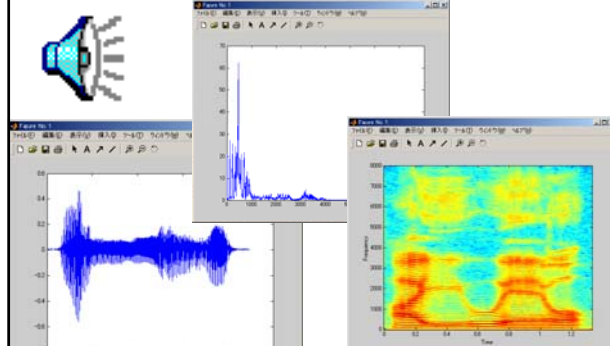
2. 情報統合

- 時間統合
- 空間統合
- マルチモーダル・マルチメディア統合
- システム間統合

33

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

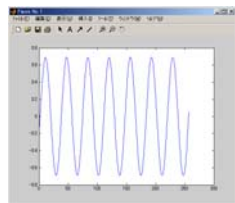


調波構造: 基本的な音の表現

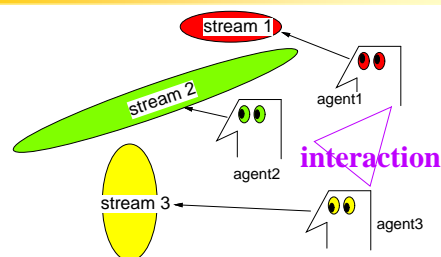
■ 基本周波数 $A(t)$

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

- $i=1$ 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



マルチエージェントによる分離

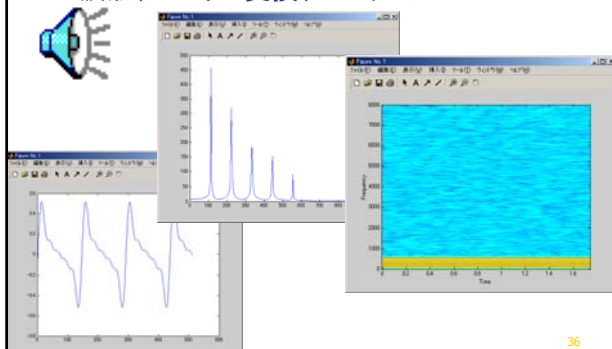


各エージェントは自分が追跡する音響ストリームに集中し、相互作用を通じて調整。

39

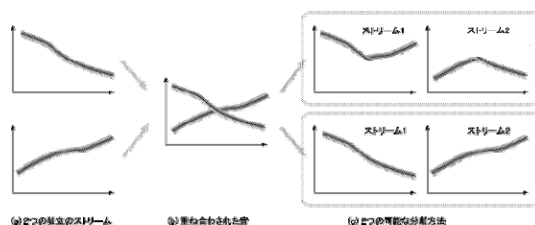
110Hzの調波構造を持つ音

■ 波形、フーリエ変換、スペクトル



36

音を聞き分けるときに生ずる曖昧さ

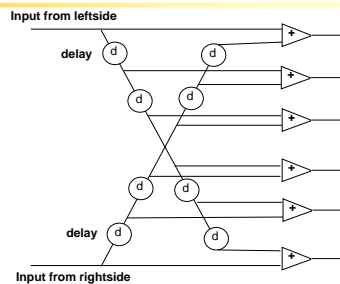


曖昧性を解消するためにさまざまな特徴や他の情報を利用する必要がある。

47

Jeffressモデル: 人間の音源定位

時間差による
モデル化

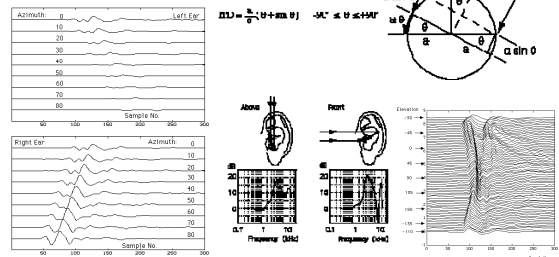


- 両耳間時間差 (Interaural Time Difference)

48

頭部音響伝達関数 (HRTF)

- Rayleigh卿の理論
- Head-shadow効果



51

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

49

外装の音響測定

- 無響室で測定 (日東紡音響エンジニアリング)

- 四方の壁、天井、床 → 吸音材 (グラスウール)
- 突起状の形 → 吸音しやすい形状。



Anechoic room

125Hz以上の周波数域では、
反響が無い部屋

52

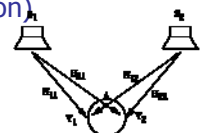
音源定位に関する特徴量

- 両耳間時間差 (Interaural Time Difference)
- 両耳間位相差 (Interaural Phase Difference)
- ◆ 両耳間レベル差 (Interaural Level Differ.)
- ◆ 両耳間振幅差 (Interaural Amplitude Differ.)
- ◆ 両耳間強度差 (Interaural Intensity Differ.)
- これらの特徴と方向情報との対応は？
ITD, IPD & ILD, IAD, IID ⇔
Azimuth & elevation

50

HRTFの近似

1. 水平方向の近似
 - 頭部の形状
 - 上半身の回折 (diffraction)
 - 肩の反射 (reflection)
2. 垂直方向の近似
 - 耳介 (pinnae) の反射
3. クロストークキャンセルステレオ
 - Sweet spot



$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \quad \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \end{bmatrix} = \begin{bmatrix} \hat{H}_{11} & \hat{H}_{12} \\ \hat{H}_{21} & \hat{H}_{22} \end{bmatrix}^{-1} \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}$$

54

頭部の音響モデル

■ 頭部伝達関数 (HRTF)

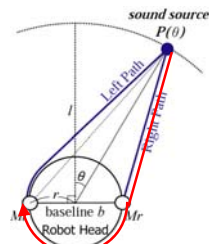
- 両耳間位相差 (IPD)、両耳間強度差 (IID) を取得可能
- 計測に時間がかかる・離散関数

■ 聴覚エピソード幾何

- 水平方向の定位
- IPD を計算的に推定可能
- 高周波、音の回り込みが未考慮

■ 散乱理論

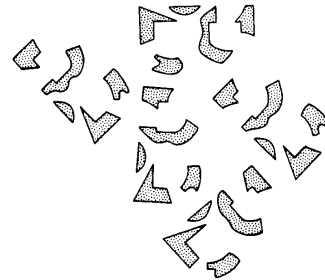
- 水平方向の定位
- IPD と IID の計算的な推定



$$IPD : \Delta\varphi = \frac{2\pi f}{v} \times r(\theta + \sin \theta)$$

55

Gestalt Principle of Closure

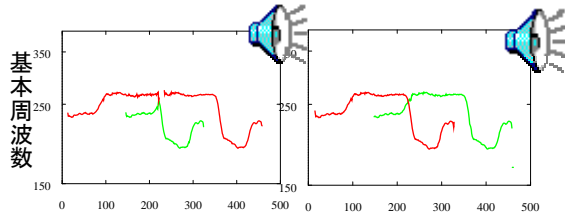


この文字は？ Occlusion (見え隠れ)？

69

方向情報による曖昧性解消

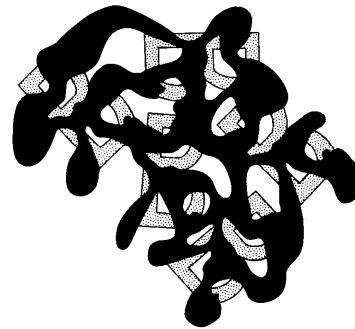
2種の3つの音を演奏： 入力混合音、
分離音(女性の発話、女性の発話)



調波構造だけ 調波構造 & 方向情報使用

58

Gestalt Principle of Closure



70

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。

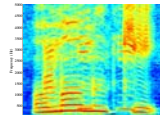
1. 残差はほとんど調波構造がふくまれないはず。
2. 残差を無声子音の代用とする

65

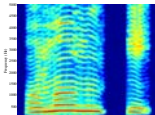
音声ストリームの分離:デモ



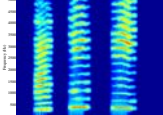
1 入力混合音



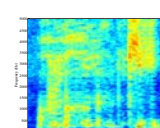
2 分離されたomomuki



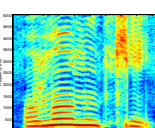
3 分離されたakiti



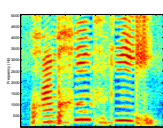
4 残差



5 分離された「おもむき」



6 分離された「あきち」



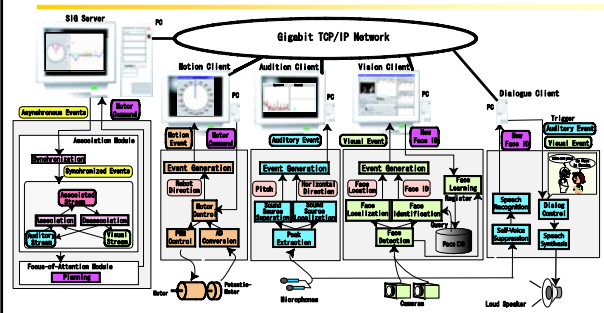
71

本日の話題

1. なぜ音がロボットやシステムでテキストや画像ほどには使われないのか。
2. 人と同じようにロボットやシステムが聞くことができるのか。
3. 音を聞くとときにどのような機能が必要か。
4. ロボットやシステムは現在どれだけのことが聞き分けられるのか。
5. 音楽は現在どの程度聞き分けられるのか。
6. 今後の展開

75

実時間複数話者追跡システム



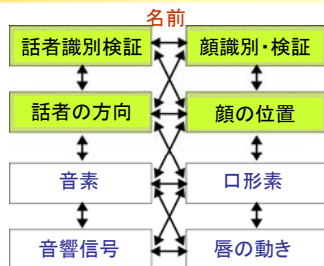
ステレオ画像・顔認識・音源同定・モータ制御の情報統合
注意制御・音声合成による対話

78

階層的な視聴覚情報統合の目標

■ 様々なレベルの視聴覚統合

- 信号レベル
- 音素-口形素レベル
- 位置レベル
- 名前レベル



■ 実装: 位置と名前レベルの視聴覚統合 読唇術、感情認識

76

実時間話者追跡によるインタラクション



受付嬢



コンパニオン役



声のする方へ向く



バランス制御で音を追跡

79

表情・感情認識の統合

■ 顔表情 (Ekman & Friesen)

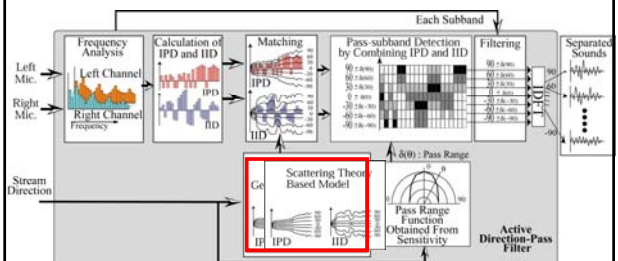
- Facial Action Coding System: 44個のAction Unit(AU)
- Pleasure, Fear, Sadness, Disgust, Anger, Surprise に関係するのは, AU1 (眉の内側を上げる), 2(眉の外側を上げる), 4(眉を下げる), 5,(上瞼を上げる) 6(頬を上げる), 7(瞼をしかめる), 9(鼻に皺), 10(上唇を上げる), 12(唇端を引き上げる), 15(唇端を下げる), 17(顎を上げる), 20(唇を横に引張る), 25(顎を下げて唇を開く), 26(顎を下げて唇を開く)

■ 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

77

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: $|IPD_{Input} - IPD_{Hypo}(\theta)| < \delta_{IPD}(\theta)$
 $|IID_{Input} - IID_{Hypo}(\theta)| < \delta_{IID}(\theta)$

88

混合音分離の例 (元の信号)



89

三話者同時発話認識技術の進展



92

混合音からの分離音



90

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

$$r(p, d) = \left(\sum_i r(p_i, d_i) v(p_i, d_i) + \sum_i r(p_i, d_i) v(p_i, d_i) - r(p_i, d_i) \right) / P_i(p_i)$$

$$v(p, d) = \begin{cases} 1 & \text{if } \text{Res}(p, d) = \text{Res}(p_i, d_i) \\ 0 & \text{if } \text{Res}(p, d) \neq \text{Res}(p_i, d_i) \end{cases}$$

103

本日の話題

- なぜ音がロボットやシステムでテキストや画像ほどには使われないのか。
- 人と同じようにロボットやシステムが聞くことができるのか。
- 音を聞くとときにどのような機能が必要か。
- ロボットやシステムは現在どれだけのことが聞き分けられるのか。
- 音楽は現在どの程度聞き分けられるのか。
- 今後の展開

91

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

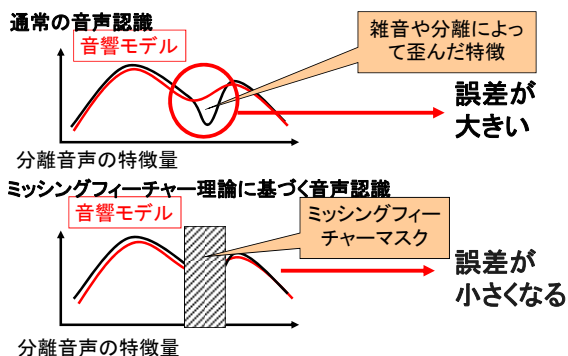


ミッシングフィーチャー理論(MFT)

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

106

ミッシングフィーチャー理論 (MFT) とは

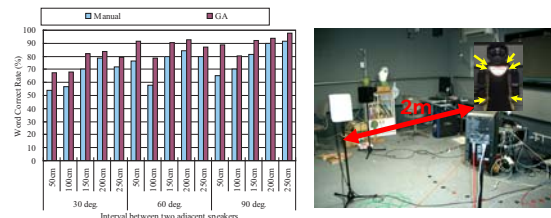


107

三話者同時発話認識の性能

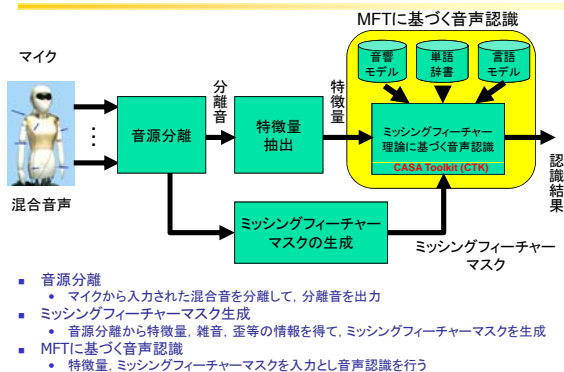
- 実験条件
 - 3本のスピーカから異なる単語を放送
 - Robovie-R2に8本のマイクロフォンを装着
 - 孤立単語認識 (200 語)
 - クリーン音声で作成した音響モデル

システムパラメータをGA(遺伝的プログラム)で最適化



123

音源分離と音声認識の統合



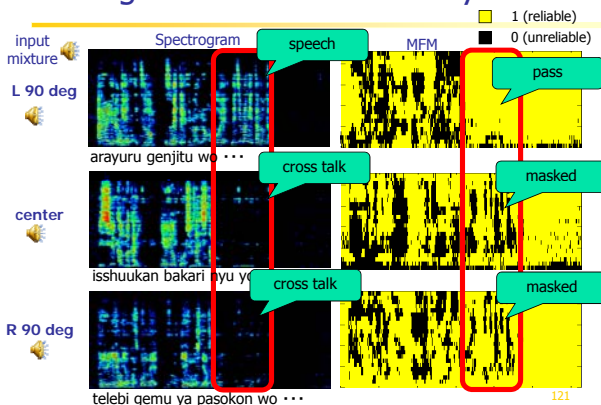
109

2005: 3 Actual Speakers in a larger room



124

MFM generated automatically



121

2006: 4x faster recognition



125

本日の話題

1. なぜ音がロボットやシステムでテキストや画像ほどには使われないのか。
2. 人と同じようにロボットやシステムが聞くことができるのか。
3. 音を聞くとときにどのような機能が必要か。
4. ロボットやシステムは現在どれだけのことが聞き分けられるのか。
5. 音楽は現在どの程度聞き分けられるのか。
6. 今後の展開

127

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
 - 無指向性マイクロフォン
 - 超指向性マイクロフォン(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

136

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？



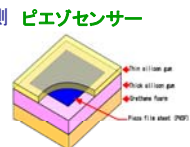
ゴヤ「1805年5月3日」

マネ「マクシミリアン皇帝の処刑」

131

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
- 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
- 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度へのみ音場を作成
→距離に関係なく、特定の人に話しかける



肌センサー



パラメトリックスピーカー

138

本日の話題

1. なぜ音がロボットやシステムでテキストや画像ほどには使われないのか。
2. 人と同じようにロボットやシステムが聞くことができるのか。
3. 音を聞くとときにどのような機能が必要か。
4. ロボットやシステムは現在どれだけのことが聞き分けられるのか。
5. 音楽は現在どの程度聞き分けられるのか。
6. 今後の展開

135

音の生成での課題

- プライバシー
- 聞く人に応じた生成



距離による挙動選択の例(動画)

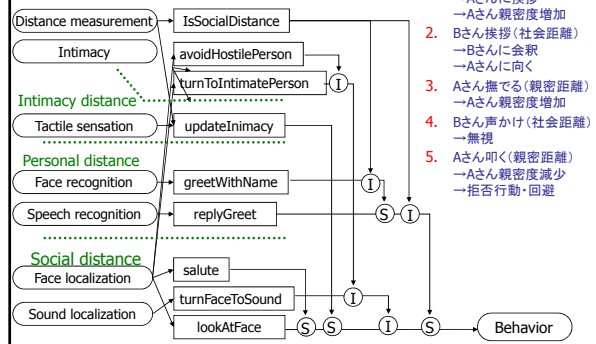


1. Aさん挨拶 (社会距離)
→Aさんに会釈
2. Bさん挨拶 (親密距離)
→Bさんに挨拶
3. Aさん呼びかけ (社会距離)
→接近を促す
4. Aさん接近
→声かける

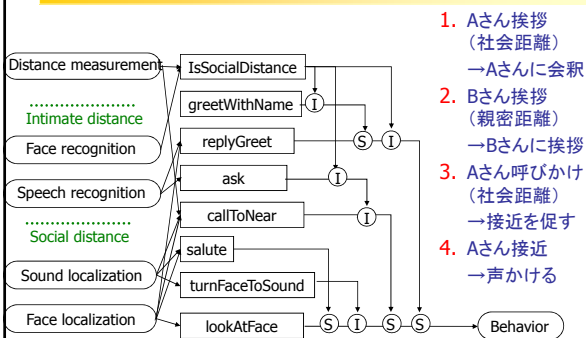
ステップ2と4は無指向性のスピーカーを利用
ステップ3は超指向性スピーカーを利用

142

Changing behavior by intimacy (implementation)



SA for Sensory modality selection



148

本日の話題

1. なぜ音がロボットやシステムでテキストや画像ほどには使われないのか.
2. 人と同じようにロボットやシステムが聞くことができるのか.
3. 音を聞かるときにどのような機能が必要か.
4. ロボットやシステムは現在どれだけのことが聞き分けられるのか.
5. 音楽は現在どの程度聞き分けられるのか.
6. 今後の展開

親密度による挙動選択の例(動画)



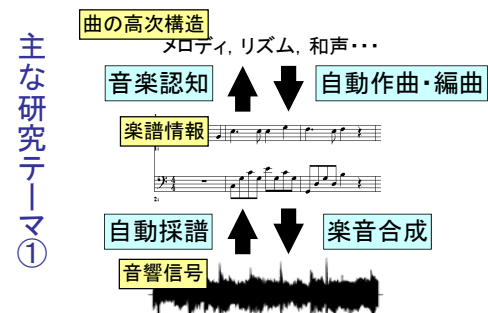
1. Aさん挨拶 (親密距離)
→Aさんに挨拶
→Aさん親密度増加
2. Bさん挨拶 (社会距離)
→Bさんに会釈
→Aさんに向く
3. Aさん撫でる (親密距離)
→Aさん親密度増加
4. Bさん声かけ (社会距離)
→無視
5. Aさん叩く (親密距離)
→Aさん親密度減少
→拒否行動・回避

ステップ3と5で皮膚センサを利用

145

音楽情報処理とは(その1)

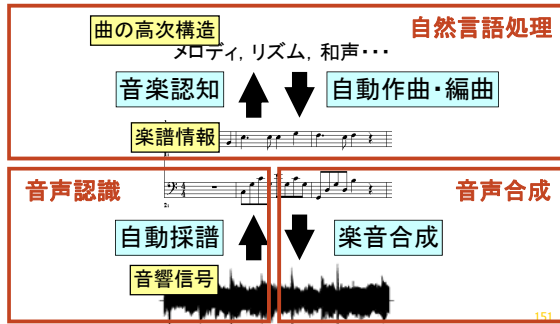
音楽情報処理 = 音楽を扱う情報処理技術全般



150

音楽情報処理の音声でのアナロジー

主な研究テーマ① 「音声」にたとえると・・・



研究事例紹介③ 未知の楽器を考慮した 楽器音の音源同定

音楽情報処理とは(その2)

主な研究テーマ②＝音楽情報検索

- デジタル音楽配信の普及
⇒入手可能な音楽データの急増
- 音は一覧性のないメディア
⇒目的の曲を探すには一曲一曲試聴が必要
- 代表的な3つのアプローチ
①鼻歌検索 ②印象語検索 ③類似楽曲検索
- 検索に適した高次アーカイブの構築
⇒音楽へのMPEG-7タグ付けが重要な課題
- <http://www.ismir.net/>

使用する楽器音の音響信号

1楽器あたり130～700個、計6,247個を使用

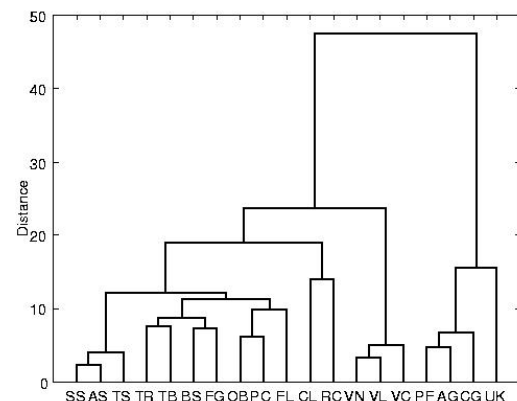
- 19種類のオーケストラ楽器の実楽器音データを「RWC-MDB-I-2001」から抜粋
- 半音ごとに全音域収録
- 各楽器、3楽器個体、3種類の音の強さ
- 通常の奏法のみ使用

↓
以上のデータから得られる各楽器の特徴空間上の分布を多次元正規分布で近似し、各楽器間のマハラノビス汎距離を使って階層的クラスタリング

音楽情報処理とは(その3)

主な研究テーマ③＝演奏の自動表情付け

- 入力: 楽譜or表情のないMIDIファイル
出力: 表情のついたMIDIファイル
- 2002年より、コンクール(Rencon)開始
- <http://shouchan.ei.tuat.ac.jp/~rencon/>
- その他の研究テーマ
- ジャムセッションシステム
- 音楽表現のための新インターフェース
- 感性情報処理、メディアアート、etc.
- 能動的音楽鑑賞



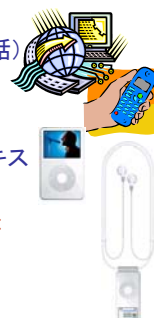
楽器カテゴリー設計結果

大分類	中分類	小分類	属する楽器
減衰系楽器	—	ウクレレ以外	PF, CG, AG
		ウクレレ	UK
持続系楽器	弦楽器	—	VN, VL, VC
	管楽器	サクソ	SS, AS, TS
		クラリネット	CL
		リコーダー	RC
		低音系+α	TR, TB, BS, FG
		高音系	OB, PC, FL

195

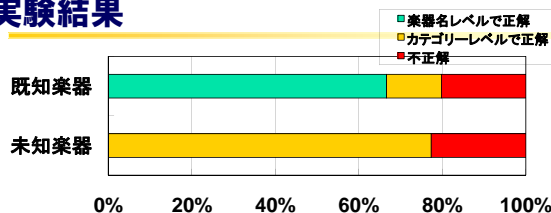
研究の背景

- 音楽データの爆発的な増加
 - デジタル音楽配信 (WEB・携帯電話)
 - 書誌情報を「知らない」楽曲の増加
- 書誌情報に基づく楽曲検索
 - タイトル・アーティストに着目したテキスト検索
 - 書誌情報を「知っている」ことが前提
 - データベース制作者サイド
 - ユーザサイド



書誌情報を「知らなくてもいい」検索システムが必要

実験結果



- 誤り率は、既知楽器で約20%、未知楽器で約23%。
- このような楽器音理解は、情報統合においても有用
e.g. 音から「楽器名はわからないが弦楽器」と同定
画像から「ある民族楽器」
⇒弦楽器に属する新たな楽器として再学習

199

イコライザの新しい概念

- The difference of equalizing targets.
 - 従来のグラフィックイコライザーは周波数帯域に基づく



e.g. graphic equalizer of Windows Media Player

- Drumix は楽器名に基づく



instrument names

230

ドラムパターン推定による ドラム音認識誤り補正手法

吉井 和佳† 後藤 真孝‡

駒谷 和範† 尾形 哲也† 奥乃 博†

†京都大学情報学研究科 知能情報学専攻

‡産業技術総合研究所

204

Drum Pattern Change

- You can edit drum patterns as you like!
 - MIDI-sequencer-like interface.
 - Automatic estimation of bar lines.
 - Beginner support by clustering drum patterns with a self-organizing map.



231

ドラム音認識とドラムパターン推定

- 楽曲推薦システムの性能向上
 - コンテンツ記述能力の強化が必要
 - 音楽の三大要素に着目
 - リズム・メロディー・ハーモニー
- リズム記述のためにドラムスに着目
 - ポピュラー音楽の多くで使用される楽器
 - 楽曲のリズムとドラムパートは密接な関係
 - 低次コンテンツ記述: ドラム音の発音時刻検出
 - 高次コンテンツ記述: ドラムパターン推定

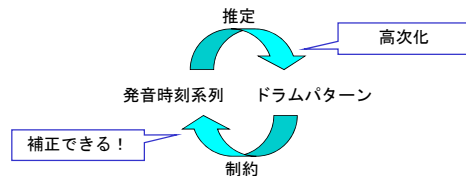
入力: 市販CDレベルの複雑な音響信号

とても難しい!

232

ドラムパターン推定

- ドラム音認識精度の改善
 - リズム階層間のインタラクションに着目
- 楽曲推薦システムへの応用
 - 高次のリズムコンテンツ記述が必要



235

ドラム音認識における課題と解決法

- 課題1: ドラム音の個体差
 - 対象楽曲中のドラム音の音響的特徴は未知
- 課題2: さまざまな楽器音の重畳
 - ドラム音の音響的特徴が変形する



ドラム音スペクトログラムをテンプレートとした

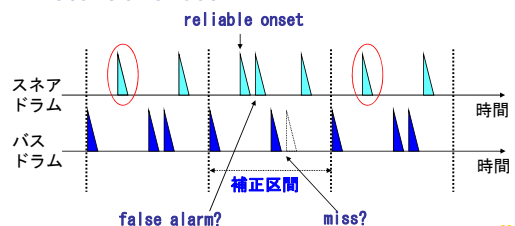
- テンプレート適応手法
 - 混合音中のドラム音スペクトログラムを推定
- テンプレートマッチング手法
 - 音の重畳に頑健な距離尺度を利用

233

アーキテクチャ

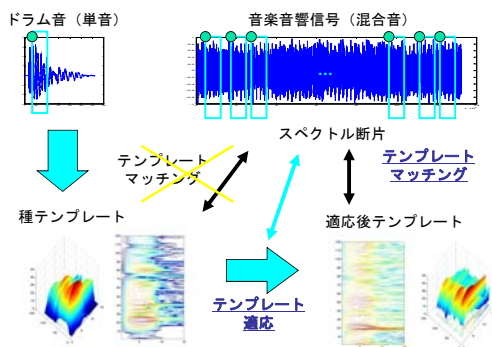
連続するドラムパターンに着目した信頼度付与

- Reliable onset → テンプレートの再適応
- Suspicious onset → 閾値を下げて再マッチング
- Potential onset → 閾値を上げて再マッチング



236

ドラム音認識のアーキテクチャ



34

実験結果

- 入力: ポピュラー音楽
 - 市販CDと同等の音楽音響信号50曲
- 評価: F値 (%) 再現率と適合率の調和平均

ドラム音認識コンテストで優勝!
(世界中から7エントリー)

	バスドラム	スネアドラム
TM	70.115	67.997
TM+TA	76.786	77.996
TM+TA+ECA	77.510	79.404
TM+TA+TCM	80.567	79.340
TM+TA+ECA+ECM	81.121	80.319

ベースライン
適応

適応+訂正

237

Drumix: ドラムパートのリアルタイム編集機能付きオーディオプレイヤー

- 能動的音楽鑑賞支援
 - 音楽にユーザ自ら手を加えて楽しむ
- 音響信号中のドラムパートを簡単にリアレンジ可
 - 音量調節機能
 - 音色置換機能
 - ドラムパターン編集機能
- ドラム音認識技術を用いて始めて実現
 - ドラム音の発音時刻
 - ドラム音スペクトログラム
 - ビート時刻

238

情報学ⅠA 2007年5月18日

マルチメディアコンテンツにおける音楽と映像の調和に関する分析

奥乃研究室 修士課程2回
西山正紘

242

Drumix: 能動的音楽鑑賞プレイヤー

Drumix: An audio player with
a function for re-arranging
drum parts in real time

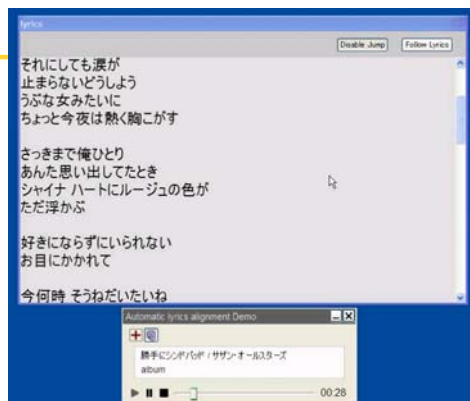
Kazuyoshi Yoshii
Masataka Goto
Kazunori Komatani
Tetsuya Ogata
Hiroshi G. Okuno

239



243

歌詞
を実
音楽
に対
応付
ける



241

音楽と映像の調和の研究目的

- 研究の課題点
 - 時間構造における調和しか扱っていない
 - 意味などの高次特徴量の自動抽出は困難
 - 時間構造と意味を統一的に扱うモデルは未提案
- 1. 時間構造と意味的調和に基づく音楽と映像の調和の計算モデルの提案
- 2. 提案するモデルに基づいた音楽と映像の調和要因の分析

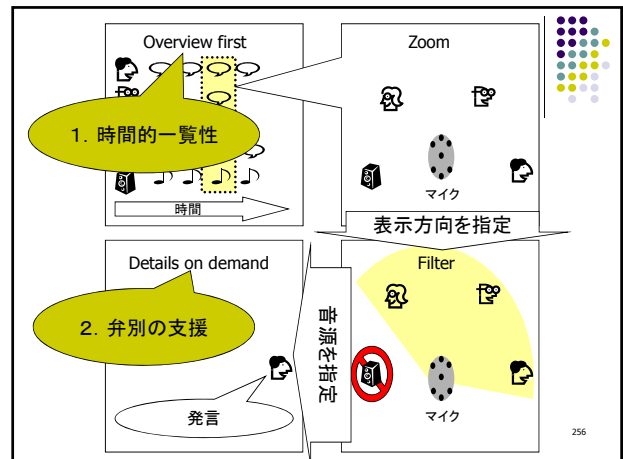


250

音の視覚化機能を有した 混合音録音再生システム

吉田雅敏† 海尻聡‡ 山本俊一‡ 中臺一博*
 駒谷和範‡ 尾形哲也‡ 奥乃博‡
 †京都大学工学部情報学科
 ‡京都大学大学院情報学研究科知能情報学専攻
 * (株) ホンダ・リサーチ・インスティテュート・ジャパン

253

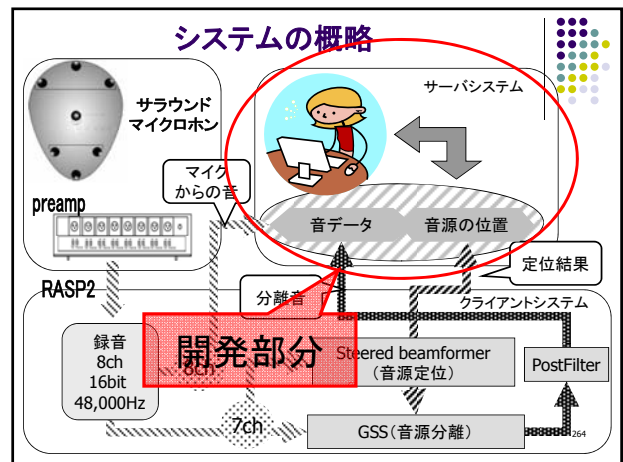


256

研究の目的

- 音情報を解り易く提示する
 - 音情報は、非常に有用である
 - ライフログ、監視、会議の議事録、etc.
- 音情報提示の問題点
 1. 時間的一覧性がない
 - 映像では
 - サムネイル表示により一覧性を持たせることができる
 - 早送り再生により全体の概略をつかめる
 2. 混合音の場合、弁別が困難である
 - 従来は単一音を別々に録音していた

254



インタフェースの設計思想

Overview first, zoom and filter, then details on demand.
 (Designing User Interface / Ben Shneiderman)

Overview first	Gain an overview of the entire collection.
Zoom	Zoom in items of interest.
Filter	Filter out uninteresting items.
Details on demand	Select an item or group and get details when needed.

IT技術による空間の変化

- 仮想現実 (virtual reality)、拡張現実 (augmented reality)
 - 新たな感覚の獲得
 - 遠隔操作、遠隔手術、マイクロデバイスを使用した手術等、人間生活を豊かに
 - 現実感のあるゲーム
 - 時空間・物理空間の現実からの遊離
- 能動的音楽鑑賞
- パーソナル空間の登場
 - TVの個人所有で変化が始まる

269

パーソナル空間の促進

- 高機能・大容量のパーソナルオーディオ機器の普及
- ノイズキャンセリングヘッドフォンの普及



- 公衆の場で自分の空間へ
- ワンセグ(地上波デジタルTV放送)



270

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



ご清聴ありがとうございます

275

Hearware – The Future of Hearing

- Victoria & Albert Museum (London)
June 2005-April 2006
- 補聴器を人の聴覚機能を拡張する機器に。
- Sound pollution 対策
- 騒音下でも静かな会話
- メガネのように補聴器も fashionableに



271

今後の展開

- ロボットを使った聴覚機能の開発
 - 音一般の分離・認識
 - 音響信号からシンボル、さらには、シンボルの実世界での意味づけ
- ロボットを使った人の聴覚機能の解明
 - 視聴覚情報統合(共感覚)
 - マルチモーダルなインタラクション
 - 福祉(聴覚障害者・高齢者へのサポート)
- 組み込みシステムとしてのロボット聴覚
 - 壁に耳あり
 - センサネットワーク

274