

ロボット聴覚 —混合音の定位・分離—

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

1. 混合音からの3つの機能
 - 音源定位 (Sound source localization)
 - 音源分離 (Sound source separation)
 - 分離音の認識 (sound recognition)
2. 組み込みシステムの聴覚機能
3. 頭部音響伝達関数
4. 頭部音響伝達関数の近似
 - 聴覚エビポラ幾何
 - 散乱理論
5. モータ音のキャンセル

基本周波数とピッチの違い

1. 800, 1000, 1200Hz
FO = 200Hz



Demonstration 21. Shift of Virtual Pitch (1:08)

A tone having strong partials with frequencies of 800, 1000, and 1200 Hz will have a virtual pitch corresponding to the 200-Hz missing fundamental, as in Demonstration 20. If each of these partials is shifted upward by 30 Hz, however, they are no longer exact harmonics of any fundamental frequency around 200 Hz. The auditory system will accept them as being "nearly harmonic" and identify a virtual pitch slightly above 200 Hz (approximately $\lfloor \frac{1000}{4} + \frac{1000}{4} + \frac{1000}{4} \rfloor = 204$ Hz in this case). The auditory system appears to search for a "nearly common factor" in the frequencies of the partials.

Note that if the virtual pitch were created by some kind of distortion, the resulting difference tone would remain at 200 Hz when the partials were shifted upward by the same amount.

In this demonstration, the three partials in a complex tone, 0.5 s in duration, are shifted upward in ten 20-Hz steps while maintaining a 200-Hz spacing between partials. You will almost certainly hear a virtual pitch that rises from 200 to about $\lfloor \frac{1200}{4} + \frac{1200}{4} + \frac{1200}{4} \rfloor = 241$ Hz. At the same time, you may have noticed a second rising virtual pitch that ends up at $\lfloor \frac{1000}{4} + \frac{1000}{4} + \frac{1000}{4} \rfloor = 200$ Hz and possibly even a third one, as shown in Fig. 2 in Schouten et al. (1962).

In the second part of the demonstration it is shown that virtual pitches of a complex tone having partials of 800, 1000, and 1200 Hz and one having partials of 850, 1050, and 1250 Hz can be matched to harmonic complex tones with fundamentals of 200 and 210 Hz respectively.

Commentary

"You will hear a three-tone harmonic complex with its partials shifted upward in equal steps until the complex is harmonic again. The sequence is repeated once".

"Now you hear a three-tone complex of 800, 1000 and 1200 Hz, followed by a complex of 850, 1050 and 1250 Hz. As you can hear, their virtual pitches are well matched by the regular harmonic tones with fundamentals of 200 and 210 Hz. The sequence is repeated once".

2. 820, 1020, 1220Hz
FO = 210Hz



混合音処理への研究アプローチ

1. 信号処理からのモデル化

- マイクフォンアレイ
- ビームフォーマ(遅延加算型、死角生成型, 適応)
- 独立成分解析(ICA, independent component analysis)

2. 人の聴覚機能からのモデル化

- Computational Auditory Scene Analysis(CASA)
「音環境理解」
- Missing Feature Theory
音素修復(auditory induction)
- Sub-band analysis

5

混合音処理での注意

1. 研究室環境とは異なる実環境

- interfering sounds が非定常的、残響
- 複数の話者の同時発話
- 背景雑音としてTVやラジオからの音声・音楽
- 音源移動(移動話者、システムが移動)

2. 理論上の優越は必ずしも実世界での優越

- ICAかGSS(Geometrical Source Separation)か?
- 正確な測定がいつも役立つとは限らない(過剰適用)

3. 総合的な能力が重要

4. ロボットの耳に適用すると現実の問題が見えてくる

- 実時間処理
- 体の影響(空間伝達関数に加えて身体伝達関数が発生)
- モータ音の影響(ハーストノイズであり、毎回異なる現象)

5. 試行錯誤によるノウハウの確立を通じたより汎用な技術の確立

6

Kismet (MIT AIL) の顔の表情



play kismet-speech.mov
7

既存のロボットのマイクロフォンは

QRIO 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

- 耳はなかった。8/16本。



13

ロボット聴覚

- ロボット自身の耳で聞く研究は少ない
- 従来の研究
 - マイクは、人間の口元に装着。
 - 単一音源からの入力を想定。
 - モーターノイズが無視できるくらい対象音は大きい。
- “Stop-perceive-act” 戦略による処理の簡単化
- ロボット聴覚の機能は、組み込みシステムに音声認識を実現するための重要な一歩
- 情報家電の音声入力・音声コマンダー

14

混合音処理への研究アプローチ

1. 信号処理からのモデル化
 - マイクロフォンアレイ
 - ビームフォーマ(遅延加算型、死角生成型)
 - 独立成分解析(ICA, independent component analysis)
2. 人の聴覚機能からのモデル化
 - Computational Auditory Scene Analysis(CASA)
「音環境理解」
 - Missing Feature Theory
音素修復(auditory induction)
 - Sub-band analysis

15

音源定位の原理 (耳はいくつ必要)

1. マイクロフォンアレイによる方法

- ・ ビームフォーミング **N+1本**
- ・ 独立成分解析 (*Independent Component Analysis, ICA*) **N本**

2. 2本のマイクロフォンによる方法

- ・ 頭部伝達関数 (*Head-Related Transfer Function, HRTF*)
- ・ 方向通過型フィルタ (*Direction-Pass Filter, DPF*)

16

音場の区別

1. Near Field

- ・ 音源から球面波が届く
- ・ マイクロフォンの間隔で規定
- ・ 人の耳だと30cmまで

2. Far Field

- ・ 音源から平面波として届く
- ・ 数m程度

3. Reverberation/Reflect Field

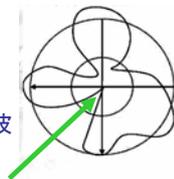
- ・ ambient noise
- ・ 数m程度以上 (アンテナの研究)

17

Beamformer (マイクロフォンアレイ)

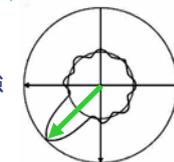
■ ナルフォーミング (null (beam) forming)

- 原理「N+1本のマイクロフォンでN個の音響的死角が構成できる」
- 特定の方向の音に対して逆相で波を重ね合わせて、notch作成
- 鋭い指向性の形成が可能



■ ビームフォーミング (beamforming)

- 特定の方向の指向性 (focus) を強調. 緩やかな指向性.
- 遅延型加算 (delay-and-sum)
- 適応型 (adaptive)



18

独立成分解析(ICA)

- Independent Component Analysis
- 原理「音源が情報論的に相互独立ならば、N個の音源はN本のマイクロフォンで分離できる」
- 時間領域ICA vs 周波数領域ICA
- Blind Source Separation
- 音源の性質について最小限の仮定
 - 出力の相互情報量を最小化
 - 非ガウス性の最大化(by 中心極限定理)
 - 尤度の最大化(by 最尤推定)
- Beamformer は方向情報が所与
- マイクロフォンに関する幾何学的情報(位置の測定)不用

19

混合音処理への研究アプローチ

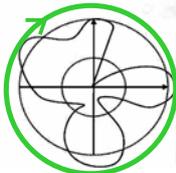
1. 信号処理からのモデル化

- マイクロフォンアレイ
- ビームフォーマ(遅延加算型、死角生成型)
- 独立成分解析(ICA, independent component analysis)
- 音源定位(sound source localization)
 - Steered Beamformer
 - MUSIC法

20

音源定位: Steered Beamformer

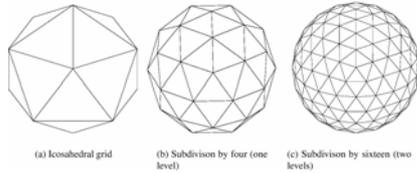
- 2D-Steered Beamformer
- Steered Delay-and-Sum Beamformer
- focus (beam) をscan
- Steered Null Beamformer
- (beam) をscan
- **MUSIC** (Multiple Signal Classification) 音源数が既知



21

Steered Beamformer

■ 3D-Steered Beamformer



■ 空間分割を粗いものから細かいものへ

22

音源分離 (sound source separation)

1. ビームフォーマ

- Delay-and-Sum (遅延加算)
- Null former (死角生成)
- Adaptive (適応)

2. Geometric Source Separation

3. Multi-channel Post-filter for GSS

4. 独立成分解析 (ICA, independent component analysis)

23

ICA(独立成分分析)とは

武田 龍君 (奥乃研B4)

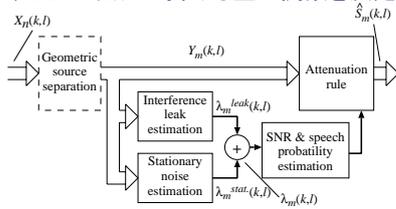
- 観測信号のみから信号を分離する
- 源信号の非ガウス性と独立性に着目して分離する
 - 独立性 = 結合密度がそれぞれの周辺分布の積に因数分解可能
 - $P_{xy}(X,Y) = P_x(X) P_y(Y)$
 - 音声信号は多くの場合非ガウス性(優ガウス分布)を持ち、ICAが適応できる
- 数学的モデルには瞬時混合モデルや畳み込み混合モデルなどがある

32

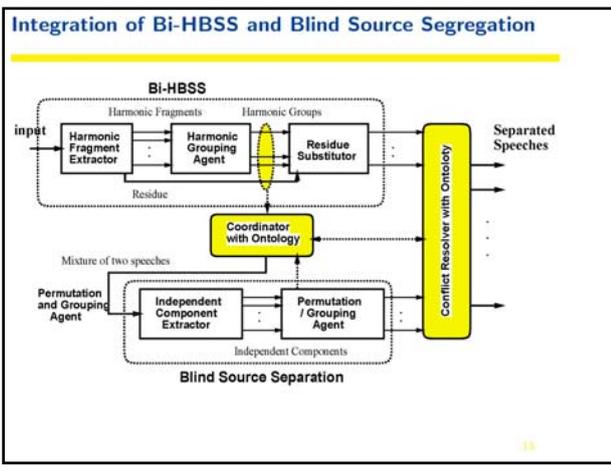
Multi-channel post-filter [Valin,Cohen]

■ 幾何学的音源分離 (Geometric Source Separation, GSS) によって分離された音を強調する手法

- 定常性の雑音推定と非定常性の雑音推定を行い, GSSの出力に掛ける重み関数を決定



36



38

混合音処理への研究アプローチ

1. 信号処理からのモデル化
 - マイクフォンアレイ
 - ビームフォーマ(遅延加算型、死角生成型)
 - 独立成分解析 (ICA, independent component analysis)
2. 人の聴覚機能からのモデル化
 - Computational Auditory Scene Analysis(CASA) 「音環境理解」
 - Missing Feature Theory 音素修復 (auditory induction)
 - Sub-band analysis

39

アクティブオーディションの課題

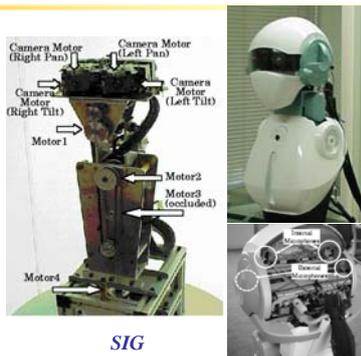
- active audition: active visionと同じく、マイクフォンの様々なパラメータを変更・適用して聞く。
- active sensing: 物体を認識するのに、物体を叩いて音を聞く。容器に中身が入っているかを振って調べる。
 1. 音声に限らない一般的な音の理解
 2. 定常雑音に限らない混合音の理解
 3. センサ情報統合
 - 信号レベル
 - シンボリックレベル
 4. ノイズキャンセル
 - 動作中のモータノイズは避けられない。
 - 一般にマイクはモータの近くにあるため、ノイズは比較的大きな音として収録されてしまう。

40

ヒューマノイド SIG

ソーシャルインタラクション用

- 4 DOFs
- 2組のマイク
- 1組のカメラ
- 機能的で美しい外装(デザインとしての研究テーマ)
- AIやセンサフュージョンの実世界への応用を目的



41

ロボットデザイン



- デザインされたロボットたち
 - J-Star99, SIG, Pino, Posy, SIG2
 - 人間との共生をテーマ
- デザイナー: 松井龍哉
 - Flower Robotics Inc. (Oct.)

42

我々のアプローチ

1. 2本のマイクロフォンを使用し、仮説推論により視覚と聴覚による定位の曖昧性を解消。
2. マルチモーダル情報の階層的な統合によるロバストな追跡を達成。
3. Gigabit Ethernetを介した4台PC群の分散処理による実時間処理の達成。
4. 受動的なソーシャルインターアクション。

43

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)

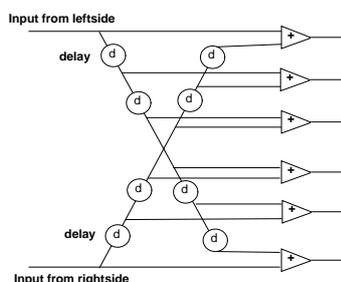
- ◆ 両耳間レベル差 (Interaural Level Differ.)
- ◆ 両耳間振幅差 (Interaural Amplitude Differ.)
- ◆ 両耳間強度差 (Interaural Intensity Differ.)

- これらの特徴と方向情報との対応は?
ITD, IPD & ILD, IAD, IID \Leftrightarrow
Azimuth & elevation

44

人間の音源定位モデル

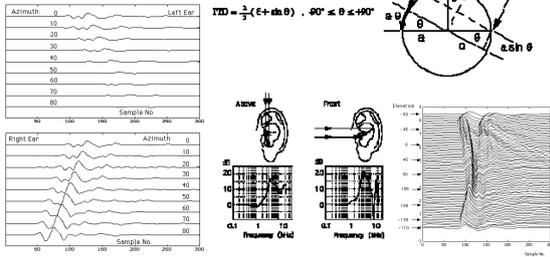
Jeffressモデル
時間差による
モデル化



45

頭部音響伝達関数(HRTF)

- Rayleigh卿の理論
- Head-shadow効果



46

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。



125Hz以上の周波数域では、反響が無い部屋



Anechoic room

47

無響室

- 272個のマイクロフォン(15度間隔) 直径4.6m、6.7m角
- 防音用耳カバの音源定位への影響
- 残響時間(60dB減衰時間) 0.01秒程度



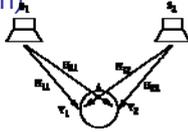
Auditory Localization Facility at Wright-Patterson AFB

48

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

$$\begin{bmatrix} r_1 \\ r_2 \end{bmatrix} = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} \begin{bmatrix} r_1 \\ r_2 \end{bmatrix} \quad \begin{bmatrix} g_1 \\ g_2 \end{bmatrix} = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix}^{-1} \begin{bmatrix} r_1 \\ r_2 \end{bmatrix}$$

49

聴覚エピソード幾何

HRTF (Head Related Transfer Function, 頭部伝達関数)

- バイノーラル (両耳聴) の研究でよく使われる
- 環境の変化に敏感 (通常は無響室で測定)
- 測定に時間がかかる
- 離散的な関数である

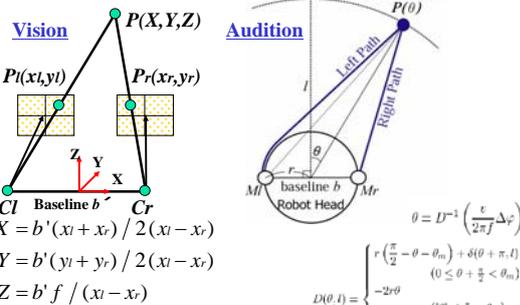


聴覚エピソード幾何

- ステレオビジョンで使われるエピソード幾何の聴覚への拡張
- 現状では水平方向の音源定位のみ
- 両耳間の位相差から、計算的に方向情報を算出
⇒ 測定不要、連続関数
- ステレオビジョンのエピソード幾何と情報統合が容易

50

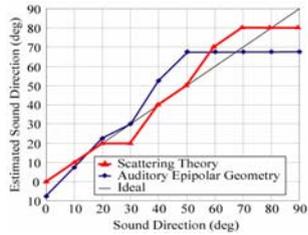
エピソード幾何 (視覚、聴覚)



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

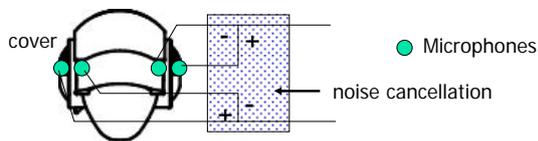


- 50度までは、同程度の精度
- 50度以上になると散乱理論の精度が高い。

59

外装によるノイズキャンセル

- 外装によってロボット内外を区別
- 1組の内部マイクをノイズ集音用に外装の内部に配置
- 1組の外部マイクを外装の音の集音用に外装の外部に配置
- 内部と外部のマイクの差を利用したノイズキャンセル



60

SIG ノイズの特徴

バーストノイズ

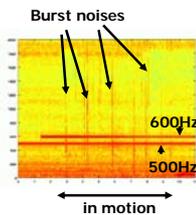
- 動作中にバーストノイズが発生。
- バーストノイズが特に悪影響を与えている。



- 少なくともバーストノイズのキャンセルは必須。

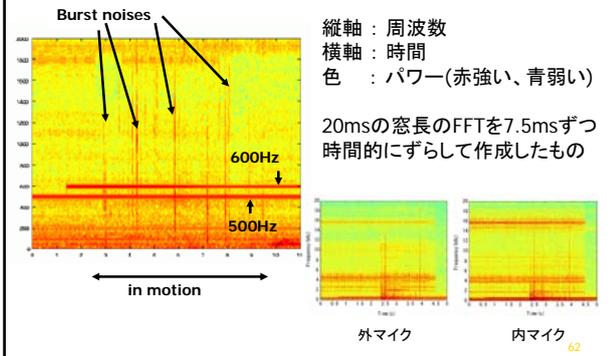
共鳴

- SIG の頭の直径は約 18 cm => 500Hzで $\lambda/4$ に相当
- 外装は、500Hzを中心周波数とした共鳴現象を持っているのでは？



61

スペクトログラム



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。



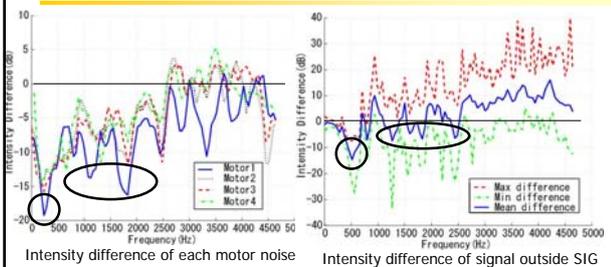
125Hz以上の周波数域では、反響が無い部屋



Anechoic room

63

外装の音響効果



- 500Hz 近辺での共鳴
- それ以外の周波数帯でも同様の現象あり
- 共鳴を考慮せずにノイズキャンセルをすることは困難

64

外装の音響効果を利用したノイズキャンセル

- Heuristics によるバーストノイズキャンセルフィルタ
- 音響測定結果をテンプレートとしてバーストノイズ判定に利用

Conditions:

- 内外のマイクの強度差がテンプレートのモータノイズの強度差に近い
- スペクトルの強度とパターンがテンプレートのモータノイズ周波数応答に近い.
- モータが動いている.

上記の3条件を満たした場合にバーストノイズと判定し、キャンセルする。

65

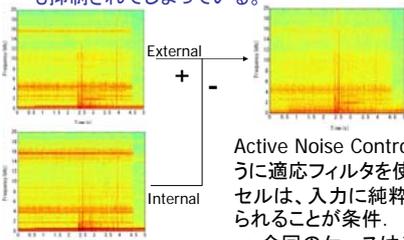
他の方法との比較

- FIR 適応フィルタによるノイズキャンセル(アクティブノイズコントロールなどでよく使われる)
- 外装の音響効果を考慮しない簡単なヒューリスティックによるバーストノイズを対象としたノイズキャンセル法

66

FIR 適応フィルタ

- 100次の FIR(Finite Impulse Response) フィルタ
- バーストノイズが残ってしまっている.
- 外部からの500Hz、600Hz のキャンセルされて欲しくない音も抑制されてしまっている。



Active Noise Controlに代表されるように適応フィルタを使ったノイズキャンセルは、入力が純粋なノイズだけが得られることが条件。
→ 今回のケースは当てはまらない。

67

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
- パーソナル空間の登場
 - TVの個人所有で変化が始まる

73

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)

74

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



75

Hearware – The Future of Hearing



Internet の登場

- 1986 Arpanet -> NSFNet
 - 大学間格差の解消
- Internet 民主化の始まりと期待
- 1989 Morris Worm 事件
 - MITのfree account からworm発生
- 新たな差別化の始まり
iam@alumni.princeton.ac.jp
- Acceptable Use Policyの登場
 - 学術用 vs AUP-freeの商用
- 新たな空間の誕生

77

Globalization vs. Localization

- Globalization によるcommunity の喪失
- Virtual community の誕生
- 新たな community rule の登場
 - ネット上の作品の著作権は？
 - Avex騒動: モナーと「のまネコ」



でも企業様は、「ネリソナリだ」といって、この2匹をお店の品物にしました。



78

本日の予定終了

- レポートは2題の回答
 - 1. 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(5ページ以上)
 - 2. 視聴覚情報統合について, ご自身の研究にどのように貢献するかについて述べよ(5ページ以上)。
-
- レポートの締切は12月20日(予定)
 - 提出先は10号館レポートボックス

81
