ロボット聴覚
ー混合音の分離・定位ー

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

既存のロボットのマイクロフォンは

QLIO SDR-4XII

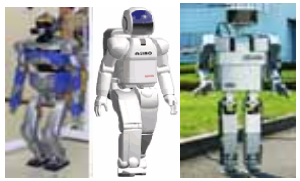
- 7本のマイクロフォン
  内1本は内部雑音除去用
- 音源定位は行う
- 音源分離は行わず

ASIMO

- 2本、音源定位のみ

HRP-2 (AIST・川田)

- 耳はない



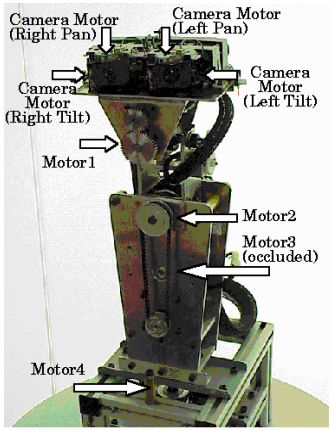
ロボット聴覚

- 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

ヒューマノイド　*SIG*

ソーシャルインターラクション用

- 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



SIG

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*）

マイクロフォンアレー

- ナルフォーミング(null 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」
- ビームフォーミング (beam forming) は、指向性を強調する。
  遅延型加算 (*delayed sum*) がよく使われる

独立成分解析（ICA）

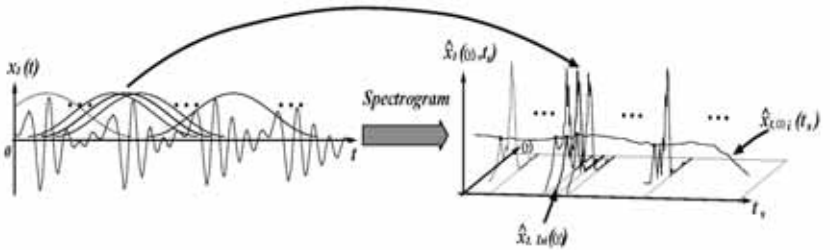
- 原理「音源が情報論的に相互独立ならば、N個の音源はN本のマイクロフォンで分離できる」
- Blind Source Separation
- 音源の性質について最小限の仮定
- 出力の相互情報量を最小にする

Blind Source Separation

1. Sound signal vector $s(t)$ of n components
   $s(t) = (s_1(t), \cdots, s_n(t))^T, \quad t = 0, 1, 2, \ldots$
2. Observed signal vector $x(t)$ by n microphones
   $x(t) = (x_1(t), \cdots, x_n(t))^T, \quad t = 0, 1, 2, \ldots$
3. Sources are mutually independent.
4. $x(t)$ is given by a linear operator A
   $$x(t) = As(t) = \left(\sum_k a_{ik} * s_k(t)\right) = \left(\sum_k \sum_{\tau=0}^{\tau_{max}} a_{ik}(\tau) * s_k(t-\tau)\right)$$
5. From $x(t)$, find a linear operator B s.t. $y(t) = Bx(t)$, mutually independent $y(t)$, without knowing operator A and the probability distribution of $s(t)$.

On-Line Algorithm Proposed by Murata and Ikeda

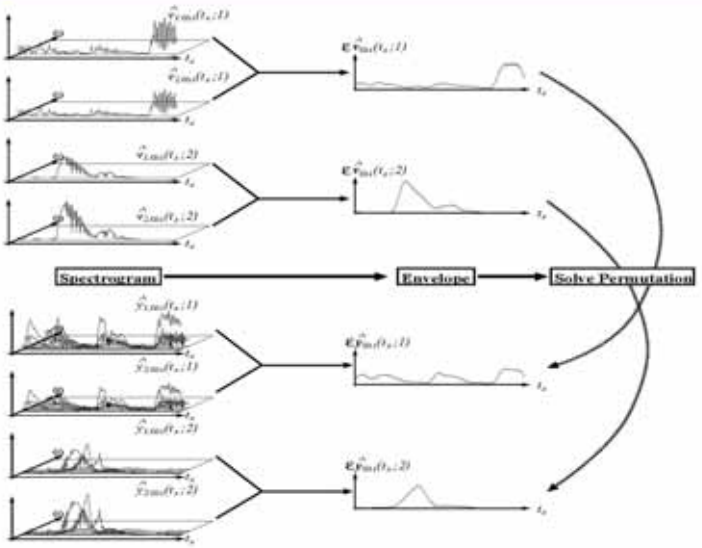
1. Human voice is stationary for a period $<$ 30∼40msec
2. Apply Windowed Fourier Transformation with Hamming window of 128 points to obtain spectrogram



3. Apply on-line Independent Component Analysis to each non-symmetric 65 points of frequency components
   $\hat{x}(\omega, t_s) = \widehat{A}(\omega)\hat{s}(\omega, t_s),$
   $\hat{u}(\omega, t_s) = \hat{x}(\omega, t_s) - B(\omega, t_s)\hat{u}(\omega, t_s)$
   $\hat{u}(\omega, t_s) = (B(\omega, t_s) + I)^{-1}\hat{x}(\omega, t_s)$
4. Learning rule:
   $B(\omega, t_s+\Delta T) = B(\omega, t_s) - \eta(B(\omega, t_s)+I)(diag(\phi(z)z^*) - \phi(z)z^*), \quad z = \hat{u}(\omega, t_s)$
   $\hat{v}_\omega(t_s; i) = (B(\omega, t_s) + I)(0, \ldots, \hat{u}_i(\omega, t_s), \ldots, 0)^T.$

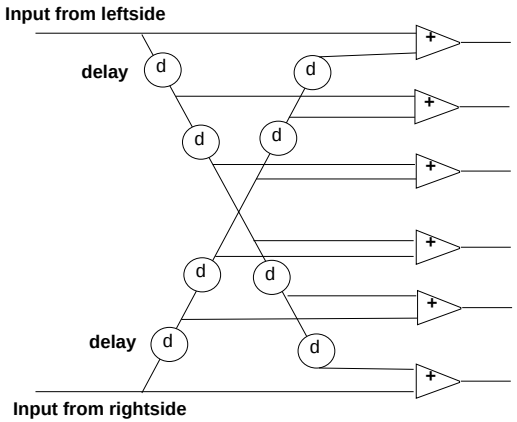
5. Reconstruct separated spectrogram based on the common temporal structure of original source signals. [Assumption] Common AM for the same sound source. Defining an envelope making operator by
   $$\mathcal{E}\hat{v}_\omega(t_s; i) = \frac{1}{M} \sum_{t'_s=t_s-M}^{t_s+M} |\hat{v}_w(t'_s; i)|,$$
   Solve permutation based on the correlation of envelopes between
   $\mathcal{E}\hat{v}_\omega(t_s; \sigma_w(i))$, and
   $\mathcal{E}\hat{y}_\omega(t_s; i) = \mathcal{E}\sum_{\omega'} \hat{v}_{\omega'}(t_s; \sigma_{\omega'}(i))$

Construct Separated Spectrogram



人間の音源定位モデル

Jeffressモデル
時間差による
モデル化



音源定位に関する特徴量

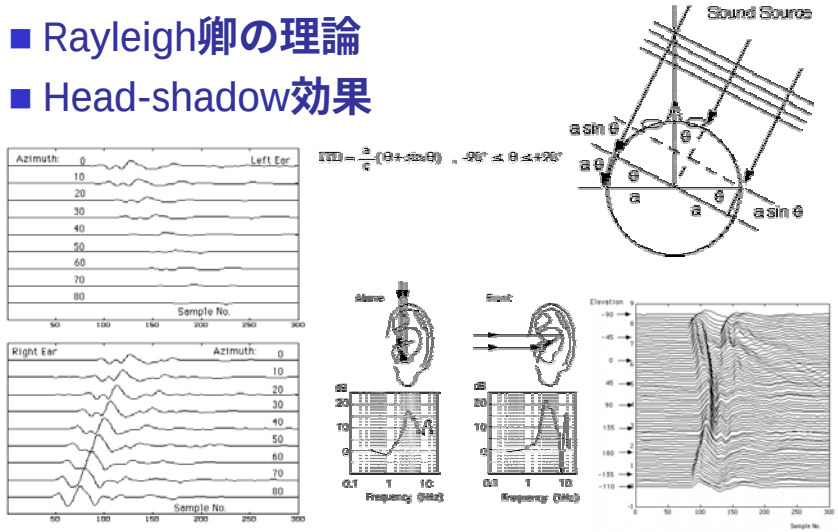
- **両耳間時間差** (Interaural Time Difference)
- **両耳間位相差** (Interaural Phase Difference)

- ◆**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

頭部音響伝達関数 (HRTF)

- Rayleigh卿の理論
- Head-shadow効果



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

無響室

■ 272個のマイクロフォン(15度間隔) 直径4.6m、6.7m角

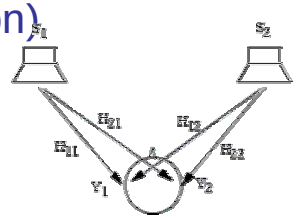
■ 防音用耳カバの音源定位への影響

■ 残響時間 (60dB減衰時間) 0.01秒程度

Auditory Localization Facility at Wright-Patterson AFB

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



$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} \begin{bmatrix} S_1 \\ S_2 \end{bmatrix} \quad \begin{bmatrix} S_1 \\ S_2 \end{bmatrix} = \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix}^{-1} \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}$$

聴覚エピポーラ幾何

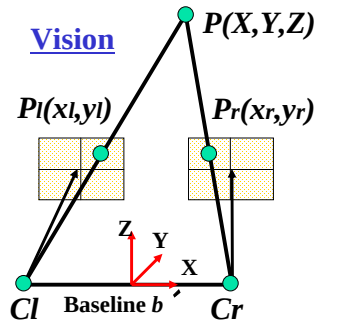
HRTF(Head Related Transfer Function、頭部伝達関数)

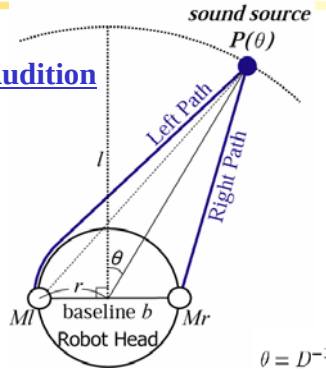
- バイノーラル(両耳聴)の研究でよく使われる
- 環境の変化に敏感(通常は無響室で測定)
- 測定に時間がかかる
- 離散的な関数である

聴覚エピポーラ幾何

- ステレオビジョンで使われるエピポーラ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ポーラ幾何と情報統合が容易

エピポーラ幾何(視覚、聴覚)





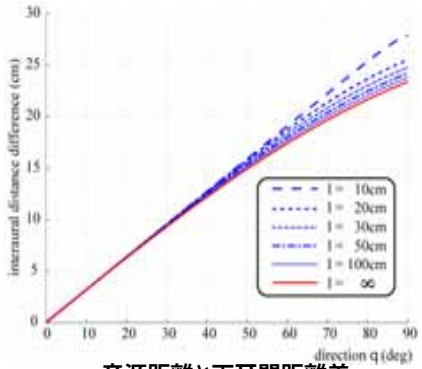
$$X = b'(x_l + x_r) / 2(x_l - x_r)$$
$$Y = b'(y_l + y_r) / 2(x_l - x_r)$$
$$Z = b' f / (x_l - x_r)$$

$$\theta = D^{-1}\left(\frac{v}{2\pi f}\Delta\varphi\right)$$

$$D(\theta, l) = \begin{cases} r\left(\frac{\pi}{2} - \theta - \theta_m\right) + \delta(\theta + \pi, l) & (0 \le \theta + \frac{\pi}{2} < \theta_m) \\ -2r\theta & (|\theta| \le \frac{\pi}{2} - \theta_m) \\ -r\left(\frac{\pi}{2} + \theta - \theta_m\right) - \delta(\theta, l) & (0 \le \frac{\pi}{2} - \theta < \theta_m) \end{cases}$$

音源距離との関係

- 50cm以上離れていれば、距離を無限と仮定することが可能
- 近接学(Proxemics) からも、インタラクションで50cm以上を仮定することは妥当[Hall 66]



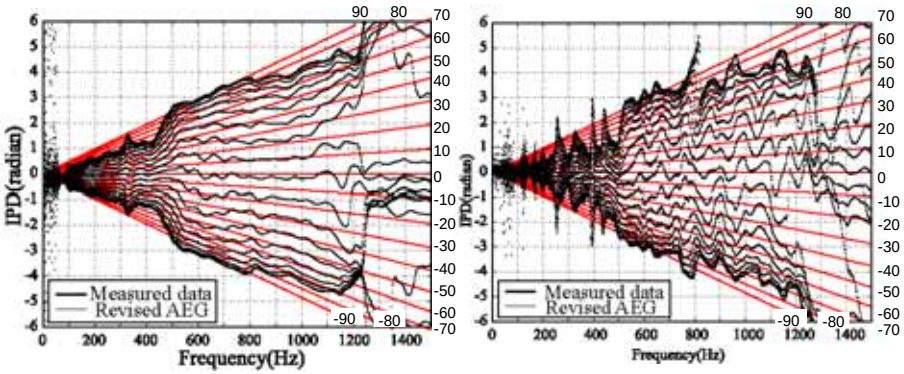
音源距離と両耳間距離差

$$D(\theta) = \lim_{l \to \infty} D(\theta, l) = r(\theta + \sin\theta)$$

Interpersonal distance in proxemics

50cm		1m	2m
intimate	personal	social	public

音響特性 (IPD)



Anechoic chamber (AEG)

Non-anechoic normal room (AEG)

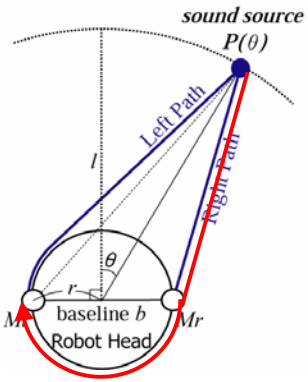
The AEG is efficient for sound source localization in an anechoic room.
In a non-anechoic room, it is not enough for robust localization.

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

■ **聴覚エピポーラ幾何**
- 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 :
$$\Delta\varphi = \frac{2\pi f}{v} \times r(\theta + \sin\theta)$$

散乱理論による頭部音響モデル

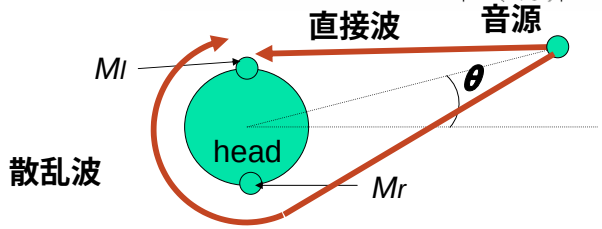
- 球体の頭部を仮定した場合、頭部表面の点でのポテンシャル

$$S(\theta, f) = -\left(\frac{v}{2\pi a f}\right)^2 \sum_{n=0}^{\infty} (2n+1) P_n(\cos\theta) \frac{h_n^{(1)}\left(\frac{2\pi r_0}{v} f\right)}{h_n^{(1)\prime}\left(\frac{2\pi a}{v} f\right)}$$

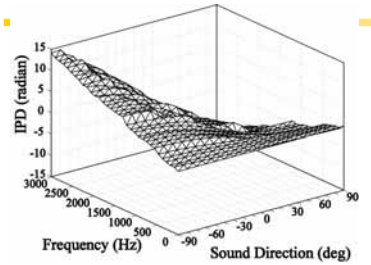
- IPD と IID の計算的な推定

$$\text{IPD:} \quad \Delta\varphi_s(\theta, f) = \arg\left(S_l(\theta, f)\right) - \arg\left(S_r(\theta, f)\right)$$

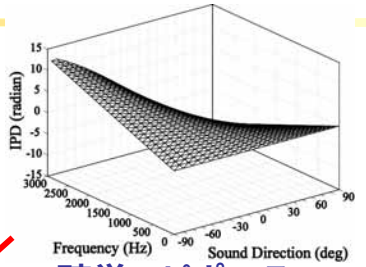
$$\text{IID:} \quad \Delta\rho_s(\theta, f) = 20\log_{10}\frac{|S_l(\theta, f)|}{|S_r(\theta, f)|}.$$



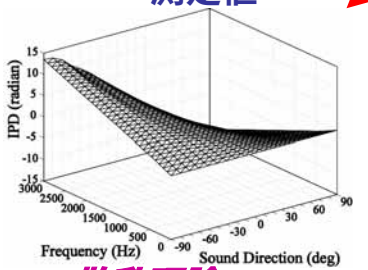
IPD推定の向上



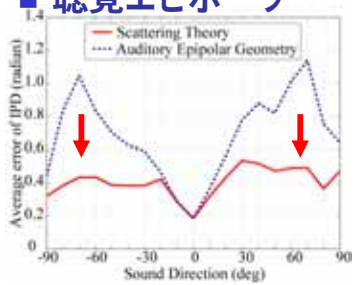
■ 測定値



■ 聴覚エピポーラ



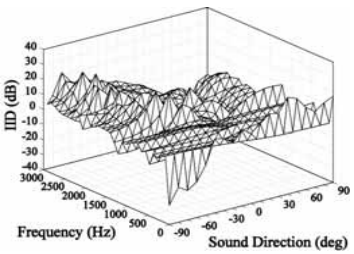
■ ***散乱理論***



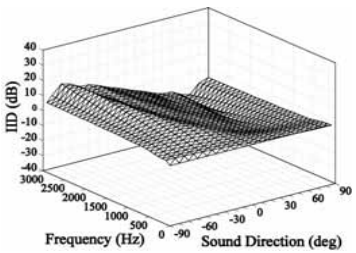
■ 誤差

IID 推定の向上

- 聴覚エピポーラ幾何
  - 大まかな3方向の推定: 正面、右、左
- 散乱理論
  - 方向ごとの計算的な推定



■ 測定値

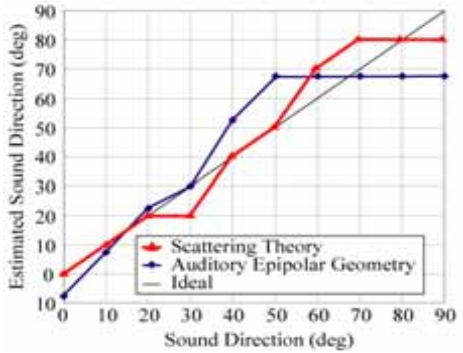


■ 散乱理論

実験1: 音源定位

- 100Hz の調波構造音 (100Hz – 3kHz) の定位

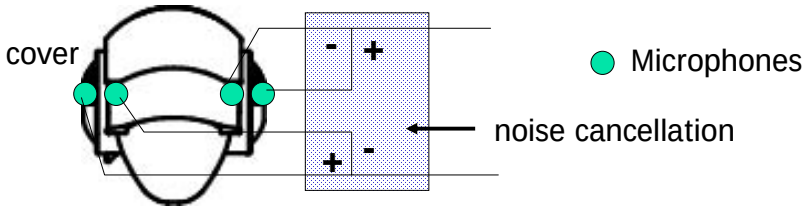
音源定位結果



- 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。

外装によるノイズキャンセル

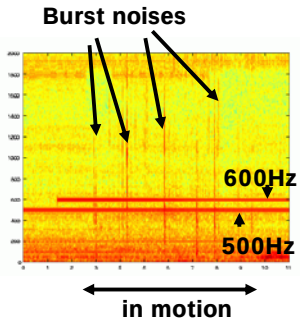
- 外装によってロボット内外を区別
- 1組の内部マイクをノイズ集音用に外装の内部に配置
- 1組の外部マイクを外部の音の集音用に外装の外部に配置
- 内部と外部のマイクの差を利用したノイズキャンセル



SIG ノイズの特徴

バーストノイズ

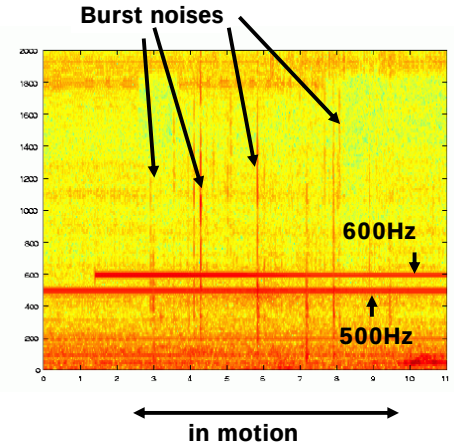
- 動作中にバーストノイズが発生.
- バーストノイズが特に悪影響を与えている.
- 少なくともバーストノイズのキャンセルは必須.



共鳴

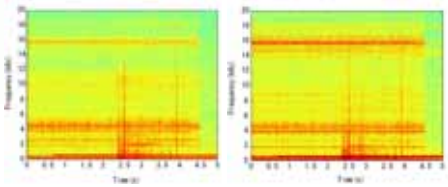
- SIG の頭の直径は約 18 cm => 500Hzで λ / 4 に相当
- 外装は、500Hzを中心周波数とした共鳴現象を持っているのでは？

スペクトログラム



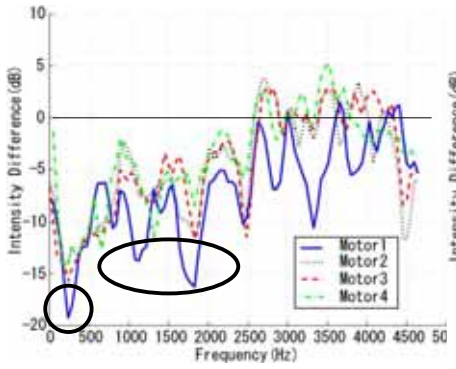
縦軸：周波数
横軸：時間
色　：パワー(赤強い、青弱い)

20msの窓長のFFTを7.5msずつ時間的にずらして作成したもの

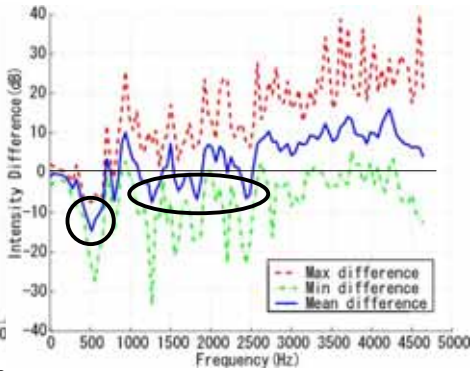


外マイク　　内マイク

外装の音響効果



Intensity difference of each motor noise



Intensity difference of signal outside SIG

- 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

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と近い
- 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.

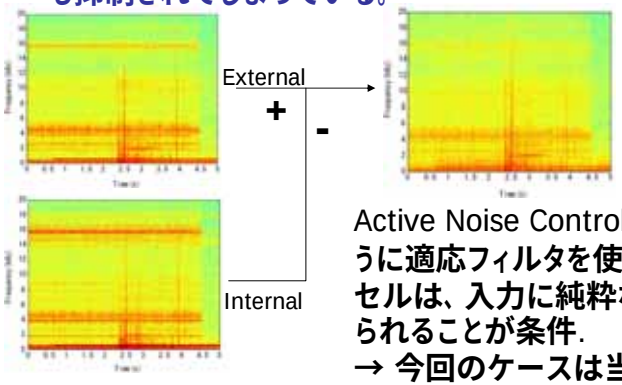
上記の3条件を満たした場合にバーストノイズと判定し、キャンセルする。

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

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、入力に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.

音響効果を考慮しないノイズキャンセル

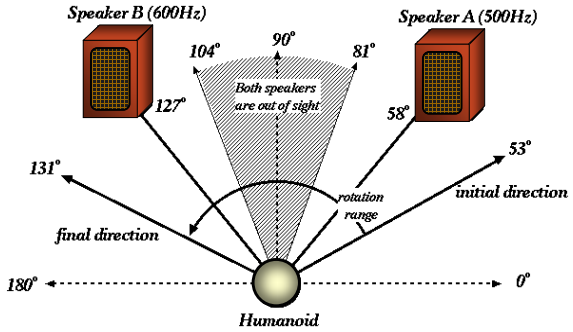
- 簡単な heuristics を使用したノイズキャンセル
- 仮定:
  - モータノイズは内部マイクの方が外部マイクよりも強く収音される。
  - 外部音は外部マイクの方が内部マイクよりも強く収音される。

conditions to judge burst noise

- 内部音のパワーは外部音よりも大きい
- 20以上の連続したサブバンドに渡ってパワーが大きい.
- モータが動作中である

実験

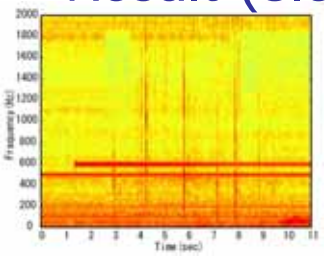
通常の部屋で、動作中に、2音源からの混合音の定位



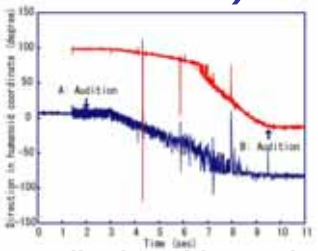
- スピーカ音源
  A(500Hz)、B(600Hz)
- Aの方向からBの方向へターン
- SIG はターンの間、両方の音を聞くことができる
- 斜線内では、両方のスピーカとも見えない。

1. 未知の環境での聴覚エピポーラ幾何による定位
2. 外装の音響効果を利用したノイズキャンセル
3. 聴覚と視覚の統合

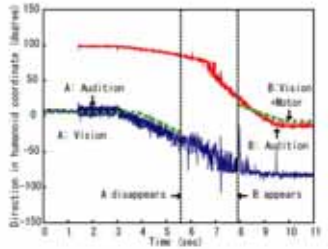
Result (Slow Rotation)



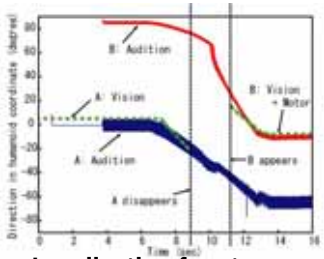
Spectrogram



Localization without noise cancellation

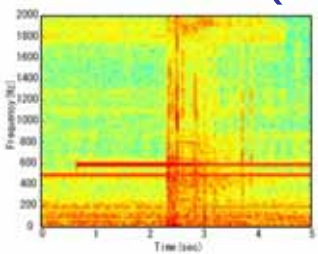


Localization with noise cancellation

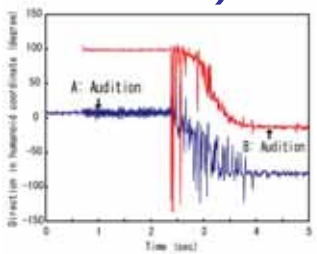


Localization for strong signal

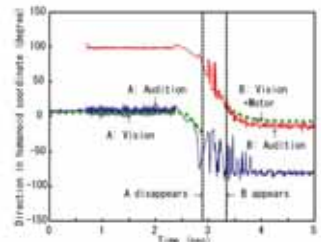
Result (Fast Rotation)



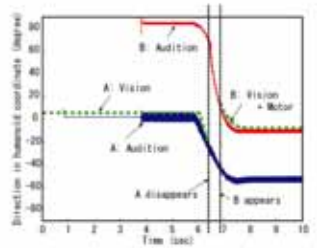
Spectrogram



Localization without noise cancellation



Localization with noise cancellation



Localization for strong signal