人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B002

ＡＩチャレンジ研究会 (第３２回)

Proceedings of the 32th Meeting of Special Interest Group on AI Challenges

CONTENTS

日　時　2010 年 11 月 26 日　　場　所　京都大学 百周年時計台記念館 国際交流ホールＩ
Kyoto University, Kyoto, Nov. 26, 2010

社団法人 人工知能学会
Japanese Society for Artificial Intelligence

社団法人　人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B002-1

歌声合成技術 VOCALOID とその組み込み機器への応用可能性

Singing Synthesizer “VOCALOID” and its possible application to embedded devices

剣持秀紀　吉岡靖雄 (ヤマハ(株) 研究開発センター)
Hideki KENMOCHI, Yasuo YOSHIOKA (Yamaha Corporation))

kenmochi@beat.yamaha.co.jp, yass@beat.yamaha.co.jp

Abstract—This paper describes overview of the commercial singing synthesis software “VOCALOID,” A prototype board where its synthesis engine is ported to a DSP is also shown. Its application possibility as an embedded device is discussed.

1. はじめに

最近，歌声合成ソフトウェア VOCALOID を用いて，多くのクリエイタが音楽制作を行っている．ニコニコ動画」などの動画サイトでは，「初音ミク」を筆頭とする歌声合成ソフトウェア VOCALOID を用いて作成された楽曲が数多く投稿され，クリエイタたちが楽曲制作を日夜競っている．VOCALOID は PC 上での音楽制作に特化した歌声生成ソフトウェアであるが，歌声を合成するという機能を考えると PC 以外の環境でも幅広い応用可能性も考えられる．また，歌声以外の音声についても，韻律を自由自在に操作できるという点はこれまでにない応用を生み出す可能性がある．

本稿では，VOCALOID 歌声合成システムを簡単に紹介し，その合成エンジンを汎用の DSP に移植し，ハードウェアとして実現した試作ボード“VOCALOID-board”について述べる．

2. VOCALOID 歌声合成システム

VOCALOID はヤマハが開発し，ライセンスを行っている歌声合成ソフトウェアである．人間の歌声から取り出した音声素片を連結することによって歌声を合成する．入力された楽譜情報をもとに素片を選択し，接続することで合成を行う．その構成を Fugure 1.に示す．

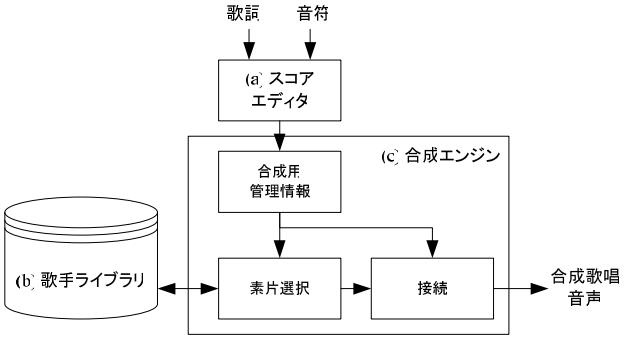


Figure 1 VOCALOID 歌声合成システム

ユーザはスコアエディタ(a)を用いて音符と歌詞を入力する(Figure 2.). 歌手ライブラリ(b)には実際の歌手の歌唱データから取り出した音声素片が含まれる．合成エンジン(c)は歌手ライブラリから取り出した必要な音声素片を連結して合成する．

Figure 2 スコアエディタ

合成の際の素片の連結時には，C-V(子音-母音)という素片の V(母音)の位置に音符開始タイミングが合うように素片の位置の調整が行われる．

素片連結時には，単に連結しただけではもちろん歌にならない．素片のピッチを所望のピッチに変換することと，素片接続位置での音色の合わせこみが必要となる．VOCALOID ではこの 2 つを周波数領域での信号処理にて行っている．すなわち，STFT により求められたスペクトルを周波数軸上でスケーリングすることでピッチを変換し，スペクトル包絡が時間的に滑らかになるように調整することで接続位置での音色の合わせこみを行っている．最後に IFFT(および Windowing & Overlapping)を行い合成波形を得る．

VOCALOID は歌声合成を第一の目的としているために，合成エンジンは音符と歌詞を入力としている．しかし一方で，歌声の合成から出発したという特質を活かし，韻律(イントネーションと音素継続長)を直接指定して合成することが可能なインタフェースも VOCALOID-flex という名称で提供されている．すなわち，韻律を与えれば歌声か話し声かを問わずそのまま合成することが可能である．

3. VOCALOID-board

VOCALOID は Windows PC 上で動作するソフトウェアなので，組み込み用途では制限が多い．そこで汎用の固定小数点 DSP に合成エンジンを移植し，小型のハードウェアとして実現したものが VOCALOID-board である．Figure 3.にその外観を示す．

VOCALOID-board の機能は以下の通りである．

* MIDI 入力により，歌声や話し声を合成
* 実時間にて合成を行い，出力する
* 最大負荷動作時 1W 以下の低消費電力
* 8cmx8cm のボード上に全機能を集約

歌声ライブラリは，PC ソフトウェア用のものがそのまま流用可能であり，SD カードにて提供される．

Figure 3. VOCALOID-board

VOCALOID-board は以下の各モードにて動作する．

(a) Playback **モード**

VOCALOID Editor にて作成したシーケンスデータ(VOCALOID-MIDI 形式)を再生するモードである．

(b) Realtime **モード**

あらかじめ歌詞を入力しておき，入力される MIDI の Note On/Off メッセージに従って発音するモードである．MIDI キーボード等を接続し，歌声を「演奏」することが可能．歌詞の入力は専用の MIDI メッセージを用いる．

(c) Voicesynth **モード**

音素長と各時刻でのピッチとダイナミクスを直接指定したものを入力として歌声や音声を出力することが可能なモードである．前述の VOCALOID-flex 機能に対応する．

4. VOCALOID-board の応用

VOCALOID-board は，用途として家電・業務用機器への歌声・音声合成機能の組み込みを想定している．特にエンタテインメントロボット分野で，インタラクティブな歌声の合成，表情豊かな話し声の合成機能が簡単に実現可能である．この分野で，単なる波形再生や TTS(Text-To-Speech)では不可能な価値を提供していきたい．その価値とは，リアルタイム性，インタラクティブ性，エンタテインメント性である．

リアルタイム性とは，メッセージを受け取ったら直ちに再生可能ということである．機器のメイン CPU に負担をかけることなく歌声や話し声の合成が可能である．インタラクティブ性とは，状況に応じて発話内容を変更可能ということである．エンタテインメント性とは，発話内容そのものが，親しみやすく楽しめるものに成り得るという点である．

5. VOCALOID-board の今後

今後さらに小型化を進めていきたい．また，試作ボードの評価使用を含むアライアンスプログラムの提供により，パートナー企業との協業を通じて用途開発および要求仕様の絞り込みを進めていきたい．

参考文献

1) H. *Kenmochi* and H. *Ohshita*, VOCALOID - commercial singing synthesizer based on sample concatenation, Proc. Interspeech, pp. 4009-4010. (2007.8).

低サイドローブ設計 64ch 球形マイクロホンアレイの開発

Design and Implementation of Omni-Directional Ball Microphone Array

○ **佐々木洋子** ⋆, **椛澤光隆** ⋆, **Simon THOMPSON**⋆, **加賀美聡** ⋆, **尾路京一** †

Yoko SASAKI⋆, **Mitsutaka KABASAWA**⋆, **Simon THOMPSON**⋆, **Satoshi KAGAMI**⋆, **Kyoichi ORO**†

⋆ 産業技術総合研究所 デジタルヒューマン工学研究センター

⋆**National Institute of Advanced Industrial Science and Technology**

† 関西電力（株）

†**Kansai Electric Power Co., Inc.**

y-sasaki@aist.go.jp

Abstract

This paper presents a microphone array design and the evaluation result of the developed microphone array. We propose an evaluation index of directional characteristic of Delay and Sum BeamForming to optimize microphone array design. Using beamforming simulation, we obtain a microphone arrangement which minimizes sidelobes, and improves the basic performance of beamforming. It has 64 microphones in a 350mm diameter ball designed to mount on a mobile robot and omni-directional directivity in azimuth and elevation. The performance of the proposed microphone array is verified in different real environments. Experimental results of sound localization show the effectiveness of the array in some challenging environment and its robustness for different pressure sound sources to cover larger areas.

1　はじめに

「どこから何の音がするか」周囲の音を捉える機能は，特に移動ロボットの環境知覚機能のひとつとして重要である．この機能を実現する方法として，複数のマイクロホンをロボットに搭載したマイクロホンアレイによるアプローチが一般的で，これまでに多くのシステムが提案されている [1,3,4,6]．一方，様々な条件が想定される実環境中では，未知の環境条件，音圧差・距離差の異なる複数音の扱いなど，まだ課題も多い．後段の信号処理部分の信頼性を増すためにも，マイクロホンアレイの基本性能の向上は重要な要素と言える．

遅延和ビームフォーミング (Delay and Sum BeamForming, DSBF) は，マイクロホンアレイによる音源定位・分離の最も簡単な手法である．環境の伝達関数など事前情報を必要とせず，計算も簡単なため移動ロボットに適した手法と言える．DSBF の性能は一般的にマイクロホン数に依存し，大規模なシステムほど高い性能が得られる．壁面状に配置した 1020ch アレイなど，大規模なシステムの有効性が示されている [7,9]．一方，DSBF はサイドローブが多く，鋭い指向性が得られない，という欠点も知られている．特に小規模なシステムでは，マイクロホン配置によってその特性が異なり，性能に大きく影響する．

DSBF 以外にも様々なアレイ信号処理手法がある [5]．Griffith-Jim 型に代表される適応型ビームフォーミングでは適応的に死角を形成することで高い SN 比が得られる．ただしステアリングベクトルの推定間違いが性能を劣化させるため，移動ロボットのような動的な条件では扱いが難しい．近年広く用いられている音源定位手法として，MUltiple SIgnal Classification (MUSIC) が挙げられる．事前に環境の伝達関数および音源数を与える必要があるが，鋭い指向性が得られ比較的少ないマイクロホン数でも高精度な音源定位が実現可能である．ただし，弱い音や距離の離れた音といった指向性の弱い信号の定位は難しい．

本稿では，ロボットに搭載可能な数十チャンネルの小規模なシステムを対象とした，DSBF に適したマイクロホンアレイの設計について述べる．ビームフォーミングの基本性能を向上させるため，ビームフォーミング時の指向特性を定量評価するための指標を定義し，方位角・仰角の全方位に高感度な特性を得られるマイクロホンアレイの設計を行う．後半では開発したマイクロホンアレイによる屋外での音源定位実験の結果について述べる．

2　遅延和ビームフォーミングの指向特性

本節では，まず DSBF の基礎式について整理し，マイクロホンアレイの性能を定量評価するための指向特性の評価指標を定義する．

2.1　遅延和ビームフォーミング

マイクロホンアレイの中心を原点とする極座標系で焦点 C の座標を (l, θ, ϕ) とおく．C から i 番目のマイクロホン $(i = 1, 2, \cdots, M)$ までの距離を L_i とすると，i 番目のマイクロホン入力に与える遅延は次式で表わされる．

$$\tau_i(l, \theta, \phi) = \frac{l - L_i(l, \theta, \phi)}{V_s} \tag{1}$$

ただし V_s は音速である．

各マイクロホン入力の位相を揃えて加算することで目的方向の信号が得られる．時刻 t における i 番目のマイクロホン入力を $x_i(t)$ とおくと，焦点 C に対する DSBF 出力は式 (2) となる．

$$s_c(t) = \frac{1}{M} \sum_{i=1}^{M} x_i(t + \tau_i) \tag{2}$$

焦点を全方位にスキャンさせることで,「空間スペクトル」と言われる各方向に対する音圧分布を示す曲線 (二次元定位の場合は曲面) が得られる．ひとつの点音源に対する空間スペクトルが指向特性を表わす．

2.2　指向特性の評価指標

対象音源の高精度な音源定位・分離を行うためには，メインローブは鋭く，サイドローブは低く抑えることが重要となる．そこで，DSBF 時の指向特性を定量評価するため，以下の 2 つの指標を定義する．

- IMs：球の表面積に対するメインローブエリアの面積の割合
- $ISmax$：サイドローブゲインの最大値

メインローブエリアとは焦点方向のゲインに対して-12dB までの範囲とし，サイドローブゲインとはそれ以外の範囲のピークゲインとする．Figure 1 に焦点方向を 0dB とした空間スペクトルの例を示す．中央の×が焦点，ハッチング部分がメインローブエリアとなる．

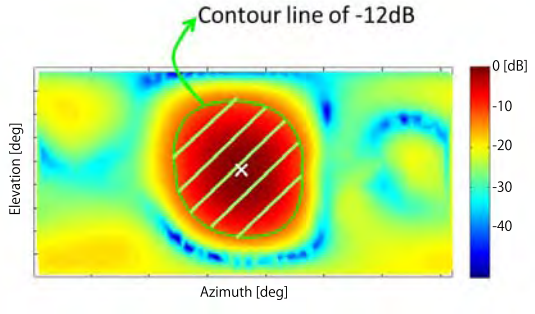


Figure 1: Directivity pattern of DSBF

上記指標をもとに，評価関数を次式のように定義する．

$$Eva = \alpha \frac{1}{a} \frac{1}{N} \sum_{\omega=\omega_L}^{\omega_H} IMs_\omega - \frac{1}{b} \frac{1}{N} \sum_{\omega=\omega_L}^{\omega_H} ISmax_\omega \tag{3}$$

ただし ω は周波数，ω_L, ω_H は対象とする下限・上限の周波数である．また N は ω_L から ω_H までの周波数ビンの数を表わす．α は重み付けのための係数である．a, b は各項の重みを等しくするための正規化係数である．本研究では，一様にランダムに生成したマイクロホンアレイに対する指標を用い，a を IMs の母平均，b を $ISmax$ の母平均とする．

3　マイクロホン配置の設計

3.1　低サイドローブ特性を得る配置

方位角・仰角の全方位に同一の指向性を持たせるため，マイクロホンの配置は同心球状とし，各球面上で偏りのないように分布させる．また低サイドローブ特性を得る方法として提案している，球の中心からの距離 r の関数としてマイクロホンの密度関数 [2] を用いる．この密度関数の基本的な考え方は，DSBF で各マイクロホン入力を同位相化することを仮想的にマイクロホンを音源から等距離の球面への射影することと捉えられるため，射影時に仮想球面上で均一に分布させるものである．つまり，マイクロホン密度 ρ は $\rho \propto 1/\cos(r)$ と定義できる．

式 (4) に 4 次関数として求めた密度関数を示す．

$$\rho(r) = \frac{0.328}{\cos\left(\frac{r}{R}\right)^4} + \frac{0.117}{\cos\left(\frac{r}{R}\right)^3} - \frac{0.496}{\cos\left(\frac{r}{R}\right)^2} + \frac{0.117}{\cos\left(\frac{r}{R}\right)} - 0.122 \tag{4}$$

各項の係数は射影後のマイクロホンが均一分布になるよう繰り返し計算により求めた．

式 (4) を基に以下の手順でマイクロホン配置を生成する．

1. アレイサイズおよびマイクロホン数を与える．
2. マイクロホンを配置する同心球の数を与え，各球の半径と配置するマイクロホン数のパターンを生成する．
3. 2. で生成したパターンごとに，各球面上で偏りなく分布させた配置を生成する．

有限個のマイクロホンに離散化すると，一つの条件 (アレイサイズ/マイクロホン数) に対して式 (4) を満たす配置が多数考えられる．ここで生成した複数の配置を式 (3) で定量評価し，最終的な配置を決定する．

3.2　マイクロホン数・アレイサイズの検討

DSBF では大規模なシステムほど性能がよく，マイクロホン数が多いほど高い SN 比が得られ，アレイサイズが大きいほどメインローブ幅が細くなる傾向にある．

まず必要なマイクロホン数およびアレイサイズを検討するために，シミュレーション上でマイクロホン数およびアレイサイズと性能の関係を評価する．

以下の条件を対象として，6×8 種類のそれぞれの条件について式 (4) を満たす 100 通りのマイクロホン配置を生成した．

- アレイ直径：0.2, 0.3, 0.4, 0.6, 0.8, 1.0 [m]
- マイクロホン数：30, 50, 70, 90, 110, 130, 150, 170

マイクロホン同士が干渉し合うものを除き，合計 3842 通りを用いて評価した．

なお，評価においては特定方向へ指向性が特化することを防ぐため，20 の異なる焦点に対して式 (3) の Eva を求め，平均値を評価値とした．次節以降のシミュレーションについても同様である．

Figure 2 に結果を示す．左から a) メインローブサイズ，b) サイドローブゲイン，c) 両者の等高線表示となっており，それぞれ縦軸がアレイサイズ，横軸がマイクロホン数となっている．グラフの右下の値がないのは，マイクロホン同士の干渉により十分な数の配置パターンが得られなかったためである．a) では上にいくほど値が小さくなっており，アレイサイズが大きいほど鋭いメインローブとなることが分かる．b) では左右に値が変化しており，マイクロホン数が SN 比に影響することが分かる．またマイクロホン間の距離が小さくなる右下ほどサイドローブゲインが下がっている．

Figure 2 の結果を踏まえ，我々の移動ロボットに搭載するマイクロホンアレイとして，本稿ではマイクロホン数：$M = 64$，アレイ直径を 360 mm と設定する．ロボットの詳細については 4.1 節で述べる．目標性能は，$(IMs, ISmax)$ = (15%, -14dB) とする．

3.3 配置シミュレーション

マイクロホン数を $M = 64$，最大アレイ直径を 360 mm とし，シミュレーション上で指向特性の評価を行った．周波数は 500〜3000Hz まで 100Hz おきに計算した．Figure 3 に 70000 通りのマイクロホン配置に対する評価値の分布を示す．縦軸がサイドローブゲイン，横軸がメインローブサイズとなっており，左下ほど性能が高いことを示している．横軸の下段の値は参考値であり，メインローブの形状が円であると仮定したときのメインローブ幅を表わしている．Figure 3 において，マゼンタ及び緑の点はそれぞれ $\alpha = 1.5, 1.7$ での評価値の上位 10 点を示している．

この結果から得られた最良配置の候補として，図中マゼンタで示した点では，すべて直径 350mm の球面上に 51 のマイクロホンを持つ配置であった．ここでシミュレーションから得られた配置をそのまま実装するのは構造が複雑になるため，以上の結果を踏まえ，以下のように実装のための制約条件を設定する．最外周のマイクロホンを 350mm の球面上に 50 個とし，高さを制限し均一に分布させるため，C60 フラーレンの頂点上に配置する．頂点は高さの異なる 8 層に分解できるため，下側 2 層を除いた 6 層上の計 50 点をマイクロホンの位置とする．その内側に配置する残りの 14 個のマイクロホンについては，同

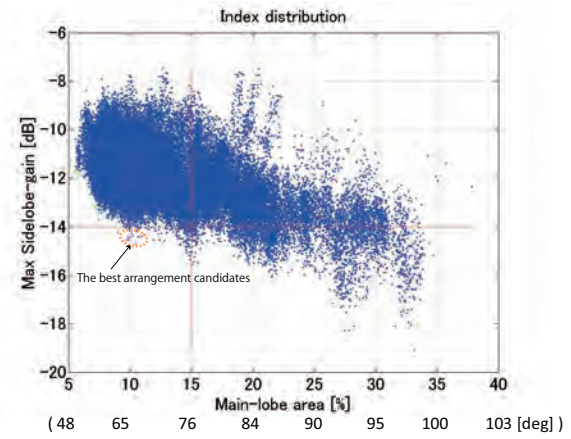


Figure 3: Evaluation Index Distribution (No Constraint)

様の手順でシミュレーションによる指向特性評価を行い，決定する．

Figure 4 に 3000 通りのマイクロホン配置に対する評価値の分布を示す．最良配置は，直径 350mm の球面上に 50 個，直径 150mm の球面上に 12 個，さらに内側に 2 個のマイクロホンがあり，評価値は $(IMs, ISmax)$ = (9.1%, -15.1dB) となった．配置の詳細については 4.1 節で述べる．

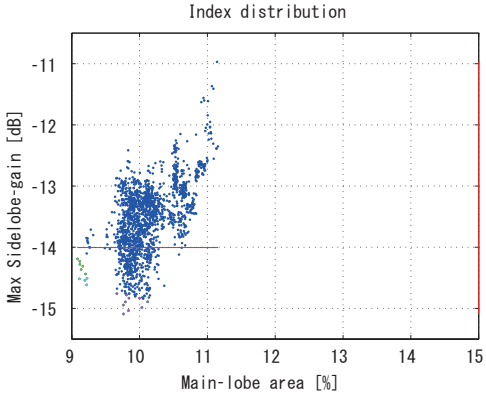
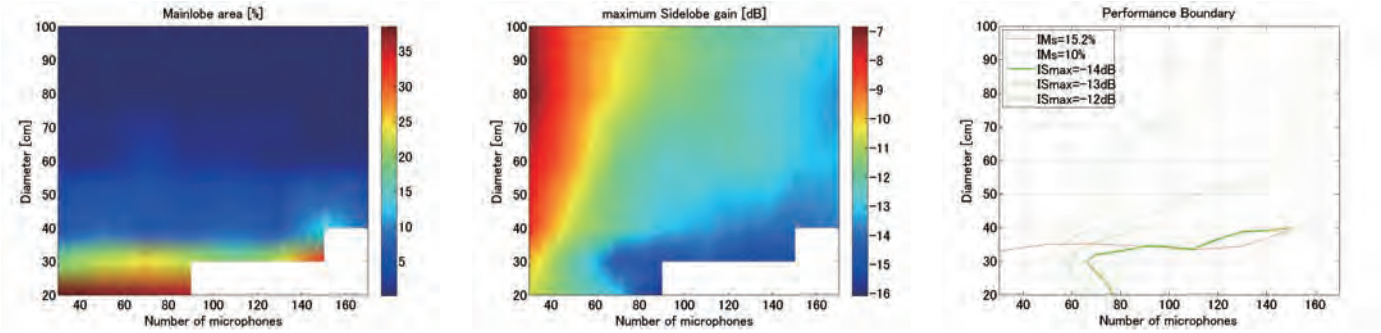


Figure 4: Evaluation Index Distribution (Fixed outside)

Figure 5 に設計したマイクロホンアレイの指向特性を示す．$(\theta, \phi) = (180°, 0°)$ が焦点である．それぞれの周波数における左右は同じデータを三次元表示したものと平面に展開したものである．平面展開では上下が広がっているが，全体的にサイドローブが低く抑えられていることが確認できる．

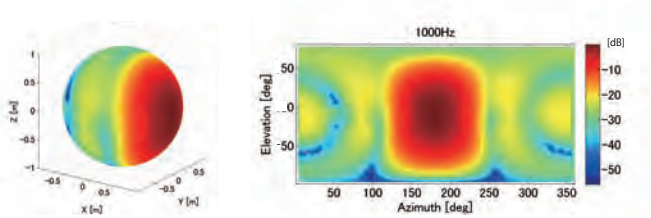
4 実装

前節で設計したマイクロホン配置をもとに，球形マイクロホンアレイを実装した．本節では，開発したマイクロホンアレイと，屋外での音源定位実験について述べる．

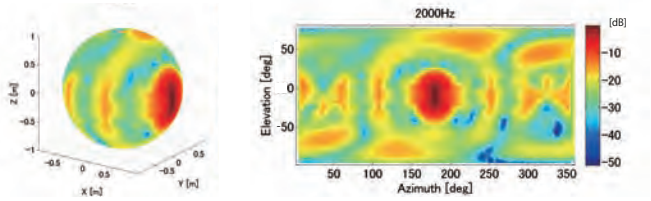


a) Mainlobe size　　b) Sidelobe gain　　c) Performance Boundary

Figure 2: Consideration of Array Size and Number of Microphones



a) 1000 Hz



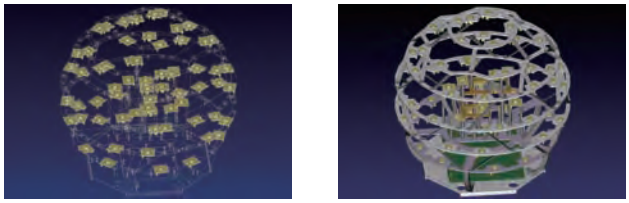
b) 2000 Hz

Figure 5: Directivity Pattern of Proposed Microphone Array

4.1　64 チャネル球形マイクロホンアレイ

Figure 6 に開発したマイクロホンアレイを示す．上段が CAD 図面のスナップショットで a) にマイクロホン基板，b) にマイクロホン基板とジグを表示している．ひとつのマイクロホン基板のサイズは 30×20 mm で，底部に設置したコントロール基板に接続されている．すべてのマイクロホンは上向きに設置している．ジグの設計においては構造物による音波の回り込みの影響を抑えられるよう考慮し，-20° から 90°(真上方向) までの範囲で直接波を捉えられるような設計になっている．c), d) はマイクロホンアレイおよび搭載したロボットの写真である．ロボットは Segway の RMP200 ATV をベースに天板等を改良したもので，2 輪駆動の倒立振子機構である．変電所内を自律走行し，各種機器を点検できるように設計されている．マイクロホンアレイはロボットの上面に設置した．

本節以降では，座標系として $\theta = 0°$ をロボットの正面，$\theta = 90°$ をロボットの左方向とし，仰角については，水平方向を $\phi = 0°$，真上方向を $\phi = 90°$ と設定する．



a) Microphones Drawing　　b) Fixture Drawing

c) Developed Microphone Array　　d) Mobile Robot

Figure 6: Developed Microphone Array

開発したマイクロホンアレイの仕様を Table 1 にまとめる．ゲインおよびサンプリング周波数はソフトウェア上で変更可能である．次節の実験は，すべて 16kHz サンプリングで行う．

Table 1: Spec. of The Microphone Array Board

Microphone	Primo EM100PT
Num. of Channels	64
Sampling Frequency	8, 16, 32, 48 [kHz]
Resolution	16 [bit]
Amplifier	AK4563A (Programmable Gain Amp.)
Interface	USB 2.0
Power Supply	+5 [v]

4.2 音源定位実験

ロボットに搭載したマイクロホンアレイを用いて，音源定位実験を行った．音源定位には DSBF の後段処理として FBS(Frequency Band Selection) を併用した手法 [8] を用いる．一回の計算に用いるデータ長は 1024 点 (64msec) とした．

4.2.1 静止時の音源定位精度評価

まずスピーカを音源とし定位角度の精度評価を行った．音源には，男声/女声の連続発話およびクラシック音楽を用い，スピーカ (YAMAHA101III) から再生した．背景雑音に対する音源の SNR は約 10dB である．音源までの距離を 3m, 9m とし，仰角を 0, 15, 30, 45deg と変化させた場合の平均角度誤差を Figure 7, 8 に示す．それぞれ，赤が角度誤差，緑/青が方位角/仰角成分を示している．最大でも 3.6deg と高精度に定位できている．誤差は主に仰角成分であり，仰角が大きいほど誤差が小さくなっている．これはマイクロホンを上向きに設置しているため，各素子の仰角方向の指向性が異なることと，ジグによる回り込みの影響が考えられる．

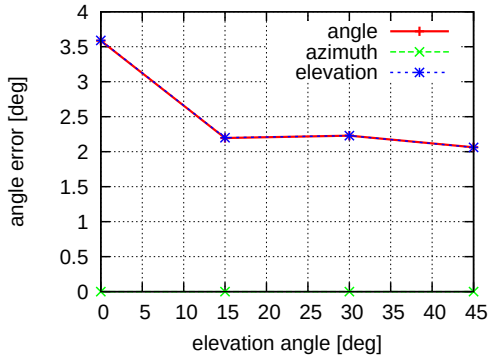


Figure 7: Average Sound Localization Error in Static Condition (distance=3m)

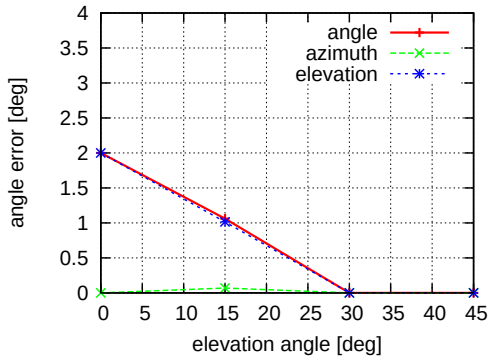


Figure 8: Average Sound Localization Error in Static Condition (distance=9m)

4.2.2 変電所内の音源定位

次に変電所の機械音を対象として音源定位実験を行った．定位計算の条件は前節と同様である．Figure 9 に実験を行った変電所内の配置図を示す．オレンジの丸で囲った 2 カ所が主な音源である．図の左方向が北となっており，南側の円は分路リアクトルの稼動音で直径約 2[m] の大口径ファンが主な音源，西側の楕円は変圧器の動作音で，人には敷地内全域で聞こえる低周波数帯域の音である．変圧器は 6 ブロックあり，各ブロックに 3 基ずつ並んでいる．これらの機械音を除くと比較的静かな環境である．水色の下向き矢印が静止時のマイクロホンアレイの位置を表わす．位置は GPS(Garmin Geko301) で取得した．マイクロホンアレイの向きは，図左方向が 0[deg]，反時計回りに方位角正方向となっている．また変圧器前の赤線はロボットの軌跡を表わす．位置は NTP でロボットと時刻を同期させた測量用のレーザ測距装置で測定した．ロボットの走行速度は約 1.3m/s であった。

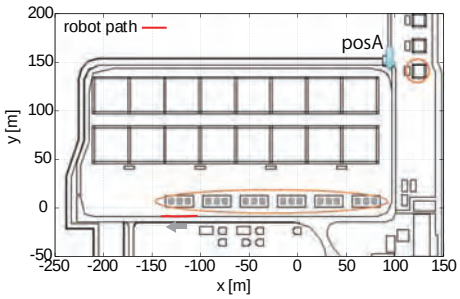


Figure 9: Robot Path in the Power Substation

Figure 10, 11 に音源定位の結果を示す．それぞれ上段が仰角，下段が方位角の結果となっている．Figure 10 は Figure 9 で示した posA で静止させた状態での結果である．$\theta = 180°$ 方向 10m の位置に分路リアクトルのファンがあり，3x6 基並んだ変圧器の中心が，およそ $\theta = 45°$ 方向に 180m 離れた位置となっている．グラフではそれぞれの音源を定位できていることが確認できる．

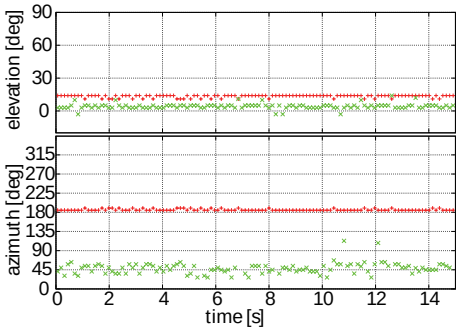


Figure 10: Sound Localization Result in the Power Substation: static condition at posA

Figure 11 は変圧器前を走行したときの結果である．通過

した3つの変圧器をそれぞれ定位できていることがわかる．また仰角の値を見ると，真横を通過する際に ($\theta = 270°$) 最も高くなっている．

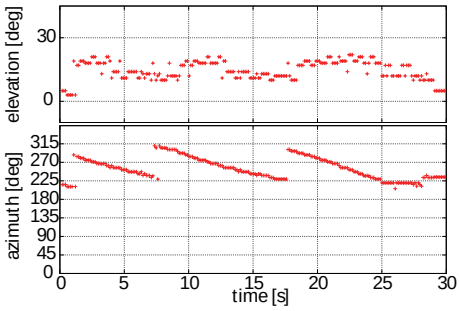


Figure 11: Sound Localization Result in the Power Substation: from moving robot

5 おわりに

本稿では，DSBF の基本性能を向上させるためのマイクロホンアレイの設計について述べた．指向特性を定量評価するための評価関数を定義し，様々な配置に対して DSBF の性能を定量評価することで，全方位に高感度な特性を持つ球形マイクロホンアレイを設計した．開発したマイクロホンアレイは，サイドローブを低減させたことで音圧差のある複数の音源を定位可能である．ロボットに搭載したアレイによる屋外での音源定位実験では，近くの音源とともに指向性の弱い離れた音源を定位可能で，ロボットの走行中も有効であることを確認した．

参考文献

[1] Hideki Asoh, Isao Hara, and Futoshi Asano. Tracking human speech events using a particle filter. In *Proceedings of IEEE International Conference on Acoustics, Speech, Signal Processing (ICASSP 2005)*, pp. MSP–P2.6, Philadelphia, USA, 2005.

[2] Tomoaki Fujihara, Yoko Sasaki, Satoshi Kagami, and Hiroshi Mizoguchi. Arrangement optimization for narrow directivity and high s/n ratio beam forming microphone array. In *Proceedings of the 7th Annual IEEE Conference on SENSORS (IEEE SENSORS 2008)*, pp. 450–453, Lecce, Italy, October 2008.

[3] Carlos Toshinori Ishi, Shigeki Matsuda, Takayuki Kanda, Takatoshi Jitsuhiro, Hiroshi Ishiguro, Satoshi Nakamura, and Norihiro Hagita. Robust speech recognition system for communication robots in real environments. In *Proceedings of IEEE-RAS International Conference on Humanoid Robots(HUMANOIDS2006)*, pp. 340–345, Genova, Italy, December 2006.

[4] Hyun-Don Kim, Jong-Suk Choi, and Munsang Kim. Speaker localization among multi-faces in noisy environment by audio-visual integration. In *Proceedings of IEEE-RAS International Conference on Robots and Automation (ICRA2006)*, pp. 1305–1310, Orlando, Florida, May 2006.

[5] Nikolaos Mitianoudis and Mike E. Davies. Audio source separation: Solutions and problems. *International Journal of Adaptive Control and Signal Processing*, Vol. 18, No. 3, pp. 299–314, March 2003.

[6] Kazuhiro Nakadai, Hirofumi Nakajima, Masamitsu Murase, Satoshi Kaijiri, Kentaro Yamada, Yuji Hasegawa, Hiroshi G. Okuno, and Hiroshi Tsujino. Real-time tracking of multiple sound sources by integration of in-room and robot-embedded microphone arrays. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-2006)*, pp. 852–859, Beijing, China, September 2006.

[7] Harvey F. Silverman, William R. Patterson III, and Joshua Sachar. Factors affectiong the performance of large-aperture microphone array. *The Journal of the Acoustical Society of America*, Vol. 111, No. 1, pp. 2144–2157, May 2002.

[8] Yuki Tamai, Yoko Sasaki, Satoshi Kagami, and Hiroshi Mizoguchi. Three ring microphone array for 3d sound localization and separation for mobile robot audition. In *Proceedings of 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS2005)*, pp. 903–908, Edmonton, Canada, August 2005.

[9] E. Weinstein, K. Steele, A. Agarwal, and J. Glass. Loud: A 1020-node modular microphone array and beamformer for intelligent computing spaces. Technical Report MIT-LCS-TM-642, MIT/LCS Technical Memo, April 2004.

社団法人　人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B002-3

言語・非言語情報を統合した指示パターンに対応するロボットの行動則獲得

Learning Robot Action Controller Corresponding to Direction by Verbal and Non-verbal Information

岡田　将吾，伊豆蔵　拓也，名渕　博人，高橋　徹，西田　豊明

Shogo Okada, Takuya Izukura, Hiroto Nabuchi, Toru Takahashi, Toyoaki Nishida

京都大学　情報学研究科　知能情報学専攻

Dept. of Intelligence Science and Technology, Kyoto University

okada_s@i.kyoto-u.ac.jp

Abstract

Human-Robot Interaction using free hand gestures and speaking word is more importance for humans which are operating robots in home or office environments. In this paper, we propose a novel technique for learning gesture command and spoken language command, action corresponding to these command by just observing interaction behavior of user with robot operated by a human operator. The main contribution of this paper is the introduction of a novel algorithm to segment and cluster patterns in its perceived signals. Proposed algorithm find gesture patterns and action patterns by using information of speech unit. The Experimental result shows that gesture patterns and action patterns are able to discovered with 85.0% ,88.0% respectively by using proposed pattern discovery algorithm.

1　はじめに

ユーザが自由なジェスチャ(非言語) と言語を用いてロボットに指示出来るインターフェイスはヒューマンロボットインタラクションにとって重要な機能の 1 つである．ロボットは聴覚とユーザの表出する自由なジェスチャを認識出来るセンサ双方を統合してユーザの指示行為を観測し，その観測結果の認識に基づき行動する必要がある．

本研究では，ユーザの言語による指示パターンと自由なジェスチャによる指示パターン，それらに対応するロボットの駆動パターンの対をインタラクションの履歴データからボトムアップに学習・獲得するシステムを提案する．まず提案システムでは人間とロボットの間で行われる言語および非言語による指示によりロボットをナビゲーションする，一連のインタラクション活動を観察する．観察から得られるユーザの発話・ジェスチャとロボットの駆動履歴に関する時系列データを，ロボットのための訓練データとして取得する．この訓練データから，ロボットに対する指示のパターンとそれに対応する適切な動作のパターンを学習により獲得する．

言語情報は，音声区間検出および音声単語認識により，発話区間および単語のシンボル情報として抽出する．ジェスチャは，光学式モーションキャプチャにより取得した位置座標データ系列として抽出する．この時系列データに対し，モチーフ発見アルゴリズムを用いることによりジェスチャパターンを抽出する．このアルゴリズムをロボットの駆動履歴系列データにも適用し，ロボットの駆動パターン（右に進む，左に回転する）を抽出し，抽出したパターン群に対しクラスタリングを行い，ジェスチャ・動作のシンボル化を行う．

本論文では，上記のシステムの内，ユーザの音声発話区間をジェスチャパターンおよびロボットの駆動パターン発見のための制約として用いる，制約付きパターン発見手法を提案し，その評価について報告する．音声認識を用いた言語パターンの獲得と，獲得したジェスチャパターン・駆動パターンの認識・生成モデルの構築については今後の課題とする．

提案するパターン発見手法の基盤には[Arita 02]で提案された方法を用いる．この手法では時系列データをその極値を与える点で区切ったのち，点同士のユークリッド距離に従ってノイズを除去した結果として得られるデータのセグメントを Motion unit として保持し，クラスタリングすることによってパターンを発見する．本研究では，Motion unit を抽出する手法にジェスチャの連続性を加味したルールを加えることで拡張し、ノイズに対して頑健なパターン発見手法を提案する．またクラスタリング部には，HMM 同士の kullback leibler 距離を用いた階層的クラスタリングを用いた．評価実験では，狭路を含む迷路をユーザが発話とジェスチャを用いてナビゲーションするタスクを行い，ジェスチャ・駆動パターンの抽出を行った．

2　関連研究

提案システムでは，一連のインタラクション活動を観測し，そのインタラクションデータを訓練データとして利用してロボットの行動制御則を獲得するという観点から，提案システムにおけるロボットの学習方法は，事例からの学習 (Robot learning from demonstration) [Argall 09]の一種と見なすことが出来る．

一連のインタラクション活動はセンサーを通じて連続的な時系列データに変換される．システム側はこの連続的なインタラクションデータから学習対象動作（ロボットの駆動パターンや，人間のジェスチャ）を分節化し，これをモデル化する必要がある．しかし従来多くの動作・運動学習に関する研究では，学習対象の動作が人間の手によって予め分節化されており，学習対象のカテゴリ数も予め与えられていることが指摘されている[Breazeal 02]．

これらの研究に対し[Kulic 09]では, 動作のモデル化にモデルの複雑度を入力データによって変化可能な Factorial-HMM を新規に提案し，階層的クラスタリングと HMM に基づく分節化手法と併用することで，連続動作データから動作の認識・生成モデルを教師無し学習により獲得する手法を提案している．この研究では動作の階層的構造も同時に獲得可能である．

[Kulic 09]の研究では，連続時系列データからの動作パターンの学習・獲得を実現しているが，本研究の目的とするユーザの指示とそれに対するロボットの動作の対の学習・獲得には着目していない．これに対し[Mohammad 09]では，変化点検知アルゴリズムに基づく制約付きパターン発見アルゴリズムを新規に提案し，さらに発見されたユーザのジェスチャ指示パターンとロボットの駆動パターンの時間的因果関係を grager causality analysis を用いて発見する手法を提案した．上記の時系列マイニング手法を，ユーザの自由なジェスチャを用いたナビゲーションタスクに利用し，ユーザ個人に依存するジェスチャによる指示パターンとロボットの駆動パターンの組み合わせを教師無し学習により獲得した．[Mohammad 09]では，インタラクションにおける指示をジェスチャに限定しており，発話による言語指示を用いていない．これに対し本研究では，言語指示パターン，非言語パターンと，それに対するロボットの駆動パターンをインタラクションデータから獲得する手法を提案する．本研究ではジェスチャパターンとロボットの駆動パターン，それらの組み合わせパターンをボトムアップに獲得することを目指している．言語パターンについての情報 (辞書・文法) は，人手により与えるものとする．

3　問題設定

本論文ではロボットのナビゲーションタスクを想定し，以下のように問題を設定した．

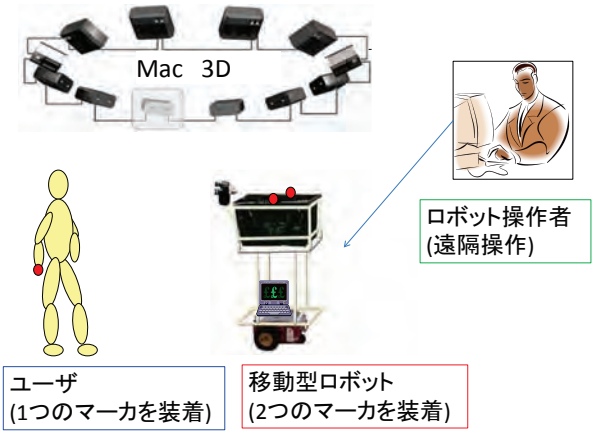


Figure 1: Environment for experimentation

3.1　環境設定

本研究ではユーザのジェスチャを観測するため，Motion Analysis 社のリアルタイム光学式モーションキャプチャである,「MAC3D System」を用いた．このシステムではマーカーが反射する光を複数のカメラで測定することにより，マーカーの三次元座標を 1 秒間に 120 フレーム取得することができる．このマーカーを指示者であるユーザの右腕に 1 つと動作を行うロボットに 2 つ取り付ける．また，指示者に従って動作するロボットとして，Mobilerobots 社の Pioneer3 を基盤とした移動型ロボットを使用する．ロボットが駆動可能な自由度は，向きの変更（左回転と右回転）と前進，後退の三つである．

3.2　インタラクションデータの取得

データ取得環境を Figure 1 に示す．本実験環境ではロボットに指示をするユーザ (以下ユーザと呼称する．) とロボット，またユーザに見えない場所でロボットを操作する操作者の三者によりデータ取得が行われる．ユーザは自由にジェスチャと言語を用いて指示を行い，ロボットを誘導する．ユーザの指示に対して，操作者は遠隔からロボットを操作する．ロボットのナビゲーションタスクを通じて，訓練データはユーザとロボットに取り付けられたマーカの位置座標の多次元時系列データとして収集される．この連続的な時系列データ (インタラクションの履歴データ) から，時系列マイニングを用いてジェスチャ・ロボットの駆動パターンをそれぞれ抽出する．

3.3　本研究で行う音声処理と言語処理

如何なる環境でも音声認識が行えるよう，ロボットにマイクなどの聴覚モジュールを保持させることが最終的に重要であるが，現段階ではシステム内部のアルゴリズムの評価に重点を置くため，出来る限りユーザのクリアな音声を取得することを目指しユーザに接話マイクを装着した．振幅と零交差に基づく入力検知を用いて音声区間の

Table 1: The word set used in the experiment

目的語句	副詞句	動詞句
右に	はやく	いって
左に	ゆっくり	きて
前に	たくさん	まわって
後ろに	すこし	とまって
こっちに	すぐ	さがって

検出を行った．発話された音声認識に関しては今後の課題とし，手動で音声データにアノテーションを付加する．以下の Table 1 に，本実験において使用される単語の一覧を示す．各品詞の単語数種類の組み合わせで数十程度のオーダーとなる．

4 提案システム

本節では観測された連続時系列データよりジェスチャおよび駆動パターンを発見するための方法を述べる．提案するパターン発見手法では，検出された発話区間に基づきセグメンテーション，およびクラスタリングを行いパターンをシンボル化する．

4.1 システムの概要

システムは，連続時系列データからジェスチャおよびロボットの駆動パターンを抽出する．セグメンテーション部分と，抽出されたパターンをクラスタリングする部分とから構成されている．Figure 2 にシステムの流れを示す．

4.2 セグメンテーション部分の実装

まずはシステムのセグメンテーション部について，入力データ，手法，そして全体での処理の流れの順で説明する．

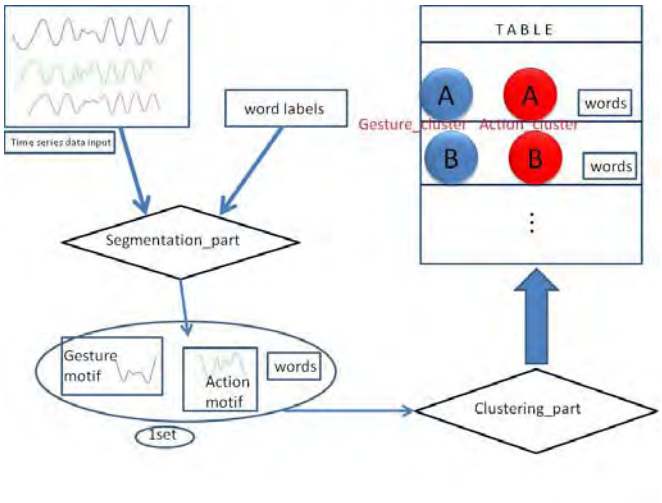


Figure 2: System flow

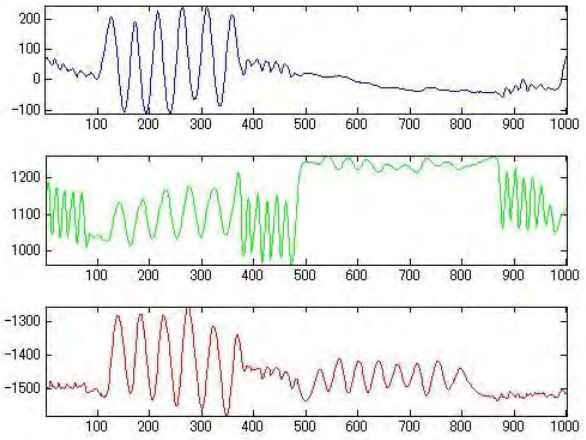


Figure 3: An example of input data

4.2.1 入力データ

システムへの入力は，MAC3D によって取得したマーカの三次元位置座標の時系列データと，言語ラベルである．ここで，ロボットの向きとはロボットに取り付けられたマーカから算出した，ロボットの中心から向いている方向へのベクトルの値のことである．また，言語ラベルとは，言語アノテーションと音声データから自動検出した発話区間である．例として，ジェスチャの三次元時系列データを Figure 3 に示す．上から順に，マーカの x 座標，y 座標，z 座標の軌跡である．

4.2.2 パターンの抽出

本研究では[Arita 02]で提案された手法の一部を基盤として新規の手法を提案した．基盤手法では，時系列データをその極値を与える点で区切ったのち，点同士の距離や値の変化量などの尺度に従って適度にノイズを除去した結果として得られるデータのセグメントを Motion unit と呼称し，これらを Nearest Neighbor 法によってクラスタリングしている．また，Motion unit 同士の類似度は DP マッチングによって計算される．本研究ではセグメンテーションの第一段階として，有田らの手法の一部分である，Motion unit を得る手法を改良したものを用いる．与えられた入力時系列に対して，まずその極値を与える点を求めたのちノイズや微小な振動などを取り除く作業を行う．このとき点同士の距離が近い点と値の変化量の小さい点を単純に取り除いていくだけではノイズが増えたときに対処しきれず，除去すべき点が残ってしまったり除去すべきでない点が除去されてしまったりすることがある．Figure 4 に具体的な事例を示す．

したがって基盤手法を用いた場合，後に Motif の抽出を行う際に影響が出てしまうため，本研究では，ノイズ除去の手法を以下のように改良した．

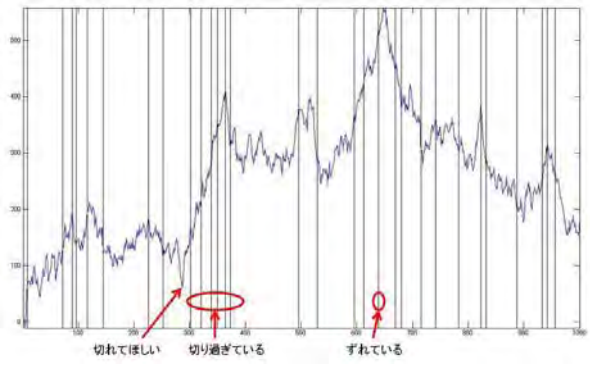


Figure 4: An example of failure case of segmentation

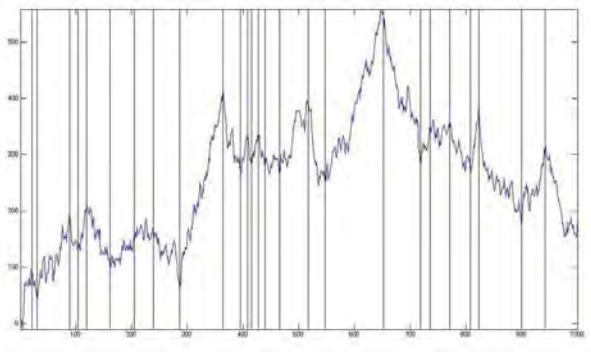


Figure 5: A segmentation result by improved method

- セグメント候補点（すなわち時系列データの極値を与える点）を端から順次見てその場ですぐに破棄するか残すかを判断するのではなく，バッファを用いて候補点を一定量保持する．
- ある点から見て次のセグメント候補点における値が十分に大きく変化しているかどうかの判断を，バッファ内に保持してあるデータでその点に十分近いものと比較する．
- 値の変化量の大きい点を探索の結果，見つけた場合，変化が正方向か負方向かを記憶しておき，さらにその次の大きな値の変化が先の変化とは逆方向であるときのみ，バッファから最小値かあるいは最大値を与える点をセグメント点として選択する．
- 一定時間以上値の大きな変化が観測されない状態が続いたのちに大きな変化に行き当たった場合は，前回の大きな変化点（変化後の点）と今回の変化点（変化前の点）とをセグメント点として選択する．
- セグメント点を選択した時点でバッファを破棄し，再びセグメント点から観測を行う．

以上のジェスチャおよび駆動パターンの抽出に特化した改良を加えることにより，時系列データの分割をさらに適切な点で得ることができる．改良した手法によって Figure 4 で用いたデータを分割した結果を Figure 5 に示す．以下では，この手法を用いたパターン抽出方法を述べる．

まず，上述した手法によって本研究においての入力となる時系列データを分割した様子を，Figure 6 に示す．

ここで，Figure 6 において赤い矢印で示された区間は，その始点と終点との差が閾値以上となっているセグメントの列からなる区間である．連続時系列データの中に発見すべきパターンが埋もれている場合，その周辺では必ず何らかの形でデータの動的傾向に変化が生じているはずであり，また逆に，ある次元に閾値以上の変化があり動的傾向に変化が生じているならばその周辺には何かしらの

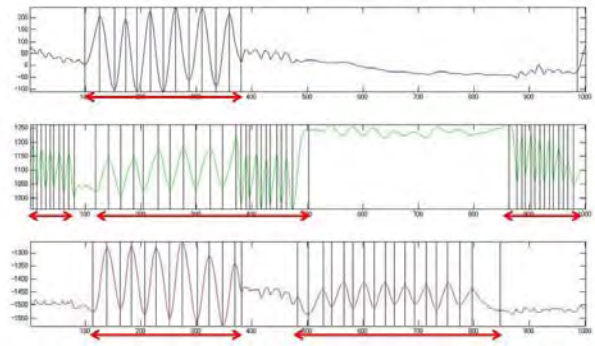


Figure 6: The approach for decision of segmentation unit

モチーフがあると考える．そこで本研究では，Figure 6 の例でいえば赤い矢印で示された区間とそうでない区間で，時系列データを変化の有無に応じて 1 か 0 かの二通りに分け，それらを全次元についてマージした列を考え，それらの共通する区間を一つの Motif であるとして抽出する．Figure 7 に，抽出される区間を状態ごとに分けて示した．

4.2.3 セグメンテーション処理の流れ

一段階目に，本システムでは言語ラベルを受け取るとラベル区間の周辺に何らかのジェスチャパターンが存在すると仮定してラベルの近辺に探索範囲を絞り込んでパ

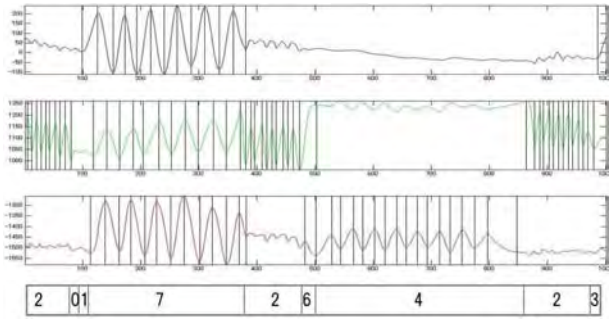


Figure 7: An example of extracted segments

ターンの発見を行う．パターン発見のアルゴリズムの手順を以下に示す．

Step1. 言語ラベルのある区間についてデータの分割を行い，その動的傾向を確認する．

Step2. 傾向に変化がなければデータの分割を行う範囲を左右に広げて Step1 へ．

Step3. 傾向に変化があり，最も長い系列がそうでない系列に挟まれている場合，その系列を抽出する．

Step4. そうでなければ，最も長い系列が伸びている方向へ探索範囲を広げて Step1 へ．

二段階目に，一段階目でジェスチャが抽出されなかった系列データに対して再度上記のアルゴリズムを適用する．

以上のステップを，探索範囲が隣り合う言語ラベルの区間に及ぶまで繰り返す．そうして全ての言語ラベルについてジェスチャのパターンが発見されたら，次にアクションのパターン発見に移る．ジェスチャは言語ラベルのある近辺から探索範囲を左右に少しずつ伸ばしていく方針をとるが，アクションではジェスチャにとっての言語ラベルの役割をジェスチャのある区間に担わせる．また，アクションは初期探索区間から右方向（時間の進む方）へしか探索範囲を広げない．これは指示を受けるよりも早くロボットが適切な動作を取ることは仮定していないためである．構成されるアルゴリズムを以下に示す．

Step1. ジェスチャのある区間についてデータの分割を行い，その動的傾向を確認する．

Step2. 傾向に変化がなければデータの分割を行う範囲を右に広げて Step1 へ．

Step3. 傾向に変化があり，最も長い系列がそうでない系列に挟まれている場合，その系列を抽出する．

Step4. そうでないなら，探索範囲を右へ広げて Step1 へ．

以上のステップを，繰り返し，全てのジェスチャに対し駆動パターンが発見できたら終了する．これを，以降のクラスタリング部分の入力とする．

4.3 クラスタリング部分の実装

4.2.3 で述べたセグメンテーション処理後に得られる，ジェスチャおよび動作の候補である時系列パターン群をクラスタリングする．今回行ったロボットナビゲーションの実験では，上下，左右に手を小刻みに動かすような，ビートジェスチャ[西田 豊明 角 康之 松村 09]が多用された．ビートジェスチャから得られる時系列パターンは繰り返し構造を持つため，セグメントされたジェスチャパターン（時系列パターン）は例えば，五回小刻みに動かしたジェスチャや一回だけ小刻みに動いたジェスチャなど任意の繰り返し構造を持っている．

今回「左に行く」という概念をクラスタとして抽出したいため，これらの繰り返し構造を持つジェスチャパターン間の類似度が高くなるように距離関数を定義したい．本研究ではエルゴード型（全結合型）の HMM をこの問題に対して利用する．エルゴード型の HMM では全状態間の遷移を許すため，上記のような繰り返し構造を持つジェスチャの学習に有用である．

クラスタリングの手法には Ward 法をクラスタ併合の基準とする凝集型階層的クラスタリング法を用いた．以下にセグメントされたジェスチャパターンのクラスタリング手法を述べる．

1. セグメントされた時系列パターン群の数 L だけエルゴード型の HMM を用意する．
2. 各時系列データを各 HMM でそれぞれ学習する．HMM のパラメータ推定には EM アルゴリズムを用いる．
3. L 個の HMM 同士の Kulbuck-Leibler 距離[Rabiner 89]を算出し，$L \times L$ の距離行列 D を作成する．
4. D に基づき階層的クラスタリングを行う．ここでクラス数はパラメータとして事前に設定する．

ロボットの動作パターンに関しても，上記の手法を適用した結果，ジェスチャ同様クラスタリング精度が良好であったため，上記の手法を用いた．階層的クラスタリングにおけるクラスタ数の推定は以下の要領で行う．クラス数 K を 2 から 30 まで変化させた各場合において Ward 法を適用して，Ward 法による距離基準である，クラスタをマージした場合のクラス内分散とマージする前の 2 つのクラスタのクラス内分散の和の差 $\triangle E$ をプロットする．各プロット点を境として対象となるプロット点から前のプロット点集合と後のプロット点集合をそれぞれ線形近似した後，その傾きを求める．この傾きのなす角度が最小となる点のクラス数をクラスタ数として推定する．

5 評価実験

本研究で提案した手法を，実際の実験データに対して用い評価を行う．実験では 1 人のユーザがナビゲーションタスクを 8 分 50 秒間行い，計 63600 フレームの多次元時系列データが得られた．セグメンテーションの結果を Recall, Precision として算出し，Table 2 に示す．ここでセグメンテーションが正解したかどうかの判定について，ビデオ分析とモーションキャプチャのデータから正解区間を決定

し，この区間と抽出したジェスチャ区間が 80% 以上共通していれば正解とした．

Table 2 より，ジェスチャパターンのアクションパターン共に良好な recall, precision の値が得られた．しかしながら発話区間の推定で，息継ぎなどの部分も発話区間として抽出した場合に，その付近にある無意味なパターンを抽出してしまう場合があり，これらを抽出してしまった．

5.1 クラスタリングの評価方法

まず，クラスタリングが正しく行われたかどうかを判断する基準を述べる．

クラスタリング精度の評価には Purity [Manning 07] を用いる．まず Purity は以下の式で算出される，

$$Purity(\Omega, \mathcal{C}) = \frac{1}{N} \sum_{l} \max_{k} |\omega_l \cap c_k| \qquad (1)$$

式 (1) で N は学習データの総数であり，$\Omega = \{\omega_1, \omega_2, ..., \omega_L\}$ はクラスタリング後のクラスタの集合を表し，$\mathcal{C} = \{c_1, c_2, ..., c_K\}$ はジェスチャの正解カテゴリ (クラス) の集合を表す．$|\omega_l \cap c_k|$ は集合 $\omega_l \cap c_k$ に属するデータ数を示す．

5.2 クラスタリング結果

実験によって得られたジェスチャパターンは Come, Forward, Forward2, To left, To right, Stop, Turn left, Turn right の計 7 種類のジェスチャと Forward, Backward, To left, To right, Round clockwise, Round counterclockwise の計 6 種類であった．Figure 2 より，インタラクションデータより無意味なパターンが含まれていた．本システムにおけるクラスタリングの目的は，同じカテゴリのジェスチャ・駆動パターンを異なるクラスタとしてはじくことである．無意味なパターンを実際のジェスチャや駆動パターンと異なるクラスタとしてはじくことである．

以下に階層的クラスタリングした結果を Table 3 に示す．ジェスチャ12 のクラスタの内訳として 7 種類のジェスチャにそれぞれ対応するクラスタが出力された．この内 3 クラスは無意味なパターン集合によるものであったが，一部無意味なパターンが Backward, To left とチャンキングする場面が見られた．これは Backward, To left 内のパターンにセグメント境界が上手く検出できず，無意味な動きとチャンキングしたまま 1 つのパターンとして抽出されたためである．

Table 2: The result of segmentation

	Recall	*Precision*
ジェスチャパターン	0.86	0.90
駆動パターン	0.90	0.95

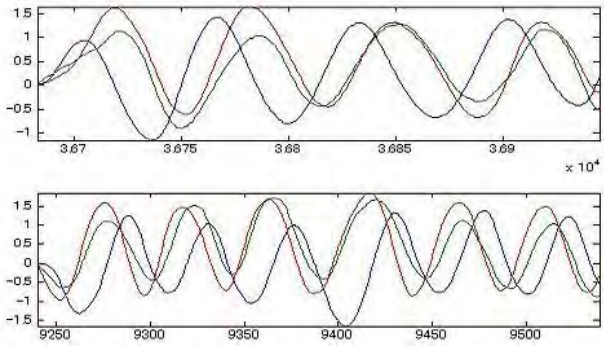


Figure 8: A failure case of clustering (upper shows the gesture pattern which means Turn right, lower shows the gesture pattern which means Turn left)

また Turn left，Turn right に関するクラスタが 2 クラスずつに分かれた上，お互いのクラスタに属するはずのパターンが誤ってチャンキングしてしまうといった場合があった．Turn right クラスのジェスチャと Turn left クラスのジェスチャは似通った傾向を持っている（Figure 8）．Turn right は手を床に平行に時計回りに動かすジェスチャで，turnleft は反時計回りに動かすジェスチャである．このようにセグメントする位置が少しずれるとこれらのパターンについては区別がつかなくなることがわかる．上記の問題に対処するためセグメンテーションの精度およびクラスタリングの精度を改善する必要があるものの，連続動作から発話区間を制約としてパターンを発見する本手法の有効性を示した．

6 結論

本研究ではユーザの言語・非言語指示とロボットの駆動パターンの組み合わせをボトムアップに獲得するため，ユーザの音声発話区間をジェスチャパターンおよびロボットの駆動パターン発見のための制約として用いる，制約付きパターン発見手法を提案した．実験の結果，7 種類のジェスチャを 85% の精度 (purity) で，6 種類の駆動パターンを 88% の精度 (purity) で抽出可能であることを示した．

現段階ではユーザの用いた言語による指示を，人手によりテキストにして書き出しているが，今後この部分を音声認識による自動獲得に移行していく予定である．

Table 3: The result of clustering

ジェスチャ		駆動パターン	
クラスタ数	Purity	クラスタ数	Purity
12	0.85	6	0.88

参考文献

[Argall 09] Argall, B. D., Chernova, S., Veloso, M., and Browning, B.: A survey of robot learning from demonstration, *Robotics and Autonomous Systems*, Vol. 57, No. 5, pp. 469–483 (2009)

[Arita 02] Arita, D., Yoshimatsu, H., and Taniguchi, R.: Frequent motion pattern extraction for motion recognition in real-time human proxy, in *Proceedings of JSAI Workshop on Conversational Informatics*, pp. 25–30 (2002)

[Breazeal 02] Breazeal, C. and Scassellati, B.: Robots that imitate humans, *Trends in Cognitive Sciences*, Vol. 6, No. 11, pp. 481–487 (2002)

[Kulic 09] Kulic, D., Takano, W., and Nakamura, Y.: Online Segmentation and Clustering From Continuous Observation of Whole Body Motions, *IEEE Transactions on Robotics*, Vol. 25, No. 5, pp. 1158–1166 (2009)

[Manning 07] Manning,C.D., Raghavan, P., and Schutze, H.: *Introduction to Information Retrieval*, Cambridge University Press (2007)

[Mohammad 09] Mohammad, Y. F. O., Nishida, T., and Okada, S.: Unsupervised simultaneous learning of gestures, actions and their associations for Human-Robot Interaction, in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2537–2544 (2009)

[Rabiner 89] Rabiner, L. R.: A tutorial on hidden markov models and selected applications in speech recognition, in *Proc. IEEE*, pp. 257–286 (1989)

[西田 豊明 角 康之 松村 09] 西田 豊明 角 康之 松村 真宏：社会知デザイン, オーム社 (2009)

社団法人　人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B002-4

Robust Speech Recognition Using Optimized Wavelet Filtering in Reverberant Conditions

Randy Gomez and Tatsuya Kawahara
Kyoto University,
Academic Center for Computing and Media Studies (ACCMS),
Sakyo-ku, Kyoto 606-8501, JAPAN

Abstract

Speech recognition in reverberant environments is a difficult task. Reverberation has the effect of degradation of recognition performance due to acoustic mismatch. We present an optimization method of the wavelet parameters for dereverberation in automatic speech recognition (ASR). By tuning the wavelet parameters to improve the acoustic model likelihood, wavelet-based dereverberation methods become more effective in the ASR application. We evaluate several existing wavelet-based methods and optimize them, based on our proposed scheme. Experimental evaluations through ASR experiments demonstrate significant improvement for all methods with the proposed optimization.

Index Terms: Robustness, Speech recognition, Dereverberation

1 Introduction

Reverberation is a phenomenon caused by the reflection of the speech signal in an enclosed environment. When analyzing in short time fourier transform (STFT), the current observed speech frame is smeared with the speech energy of the preceding frames. This degrades the acoustic quality of the speech signal and is detrimental to the ASR system. The reverberant speech model $X(f,t)$ we adopt is based on the additive effects of the early $X_E(f,t)$ and late $X_L(f,t)$ reflection,

$$\begin{aligned} X(f,t) &\approx X_E(f,t) + X_L(f,t) \\ &\approx S(f,t)H(f,0) + \sum_{d=1}^{D} S(f,t-d)H(f,d) \end{aligned} \quad (1)$$

where $S(f,t)$ and $H(f,t)$ are the frequency response of the clean speech and the room impulse response (RIR), respectively. D is the number of frames, over which the reverberation (smearing) has an effect. The early reflection is due to the direct signal and some reflections that occur at earlier time, while the late reflection, whose effect spans over frames, can be treated as long-period noise [1]-[4]. The former is mostly addressed through Cepstral Mean Normalization (CMN) in the ASR system as it falls within the frame. In our application, dereverberation is defined as suppressing the effects of the late reflection. Since the late reflection can be treated as noise, we can apply existing wavelet-based denoising techniques to dereverberation problems based on the context of our reverberant speech model.

Most of the speech enhancement algorithms are applied in the frequency domain, using short-time Fourier transform (STFT) where the time resolution is the same for all frequency components. Some enhancement methods are applied in wavelet domain which provides more flexible time-frequency representation of speech. There have been a lot of research involving wavelet-based speech enhancement primarily in denoising [5]-[8]. Originally, wavelet-based enhancement methods were proposed to address denoising problems. Most recently, it is expanded to address the effects of reverberation.

Existing wavelet-based methods are generally designed to enhance the speech waveform, but this does not guarantee an improvement in performance for ASR application. In this paper, we present a method of optimizing the wavelet parameters for dereverberation in ASR. In our proposed scheme, prior to wavelet-based dereverberation, the wavelet parameters are optimized to improve the likelihood of the acoustic model. We expand existing wavelet-based speech enhancement methods for the dereverberation application. Then, we incorporate the proposed scheme of optimizing the wavelet parameters for effective dereverberation in the ASR application. In this paper, noise and late reflection are jointly referred to as "contaminant signal".

The paper is organized as follows; Section 2 gives the background of the different wavelet-based methods which we will evaluate and optimize. In Section 3, we present the optimization method of wavelet parameters. Experimental set-up and ASR evaluation results are pre-

Randy Gomez is a research fellow of the Japan Society for the Promotion of Science (JSPS).

sented in Section 4. Finally, we conclude this paper in Section V.

2 Dereverberation Methods using Wavelets

In this section, we will discuss existing wavelet-based methods. Specifically in this paper, we consider five wavelet-based methods. The last method was previously proposed by the authors [9].

2.1 WaveShrink

The basic wavelet enhancement approach [10] is based on the idea that real-world signals do not necessarily require full resolution treatment. In speech application, a limited number of wavelet coefficients in the lower band are deemed sufficient to reconstruct the speech signal. These coefficients are characterized by higher values compared to the contaminant signals (i.e. noise or late reflections). Thus, by shrinking the contaminant wavelet coefficients, its effects are removed. In general, the waveshrink approach is applicable when the contaminant signal is homogeneously concentrated on the other side of the spectrum (e.g. higher frequencies). Problems may arise in ASR applications, because some parts of speech have important information in the higher frequencies (i.e consonants and unvoiced regions).

2.2 Thresholding

An improved version of the waveshrink approach is implemented by means of a thresholding algorithm. Unlike its predecessor, the thresholding approach is more flexible in dealing with the wavelet coefficients by defining a threshold criterion. A particular wavelet coefficient of interest may be shrunk or scaled based on this criterion. An example based on soft thresholding [11] is defined as

$$\bar{x} = \begin{cases} 0 & ,\mid x \mid \leq \ thr \\ sign(x)(\mid x \mid - thr) & ,\mid x \mid > \ thr \end{cases} \tag{2}$$

Based on the threshold *thr*, Eq. (2) can be interpreted as setting the contaminant subspace to zero, and implementing a magnitude subtraction in the speech plus contaminant subspace. The threshold that defines the subspace of the contaminant signal can be calculated [11] as

$$thr = \sigma\sqrt{2 \ log(L)}, \tag{3}$$

where L is the length of the contaminant signal with variance σ^2. Other thresholding criteria are *Hard*, *Firm*, *Garrote* and $Step - garrote$. The thresholding technique has some known problems; If the spectrum of the contaminant signal is not uniform, the method has difficulty in distinguishing the desired subspace from the contaminant subspace. Since thresholding is directly applied to the wavelet coefficients, the quality of the reconstructed signal is sensitive to the threshold.

2.3 Improved Wavelet-based Speech Enhancement System

To address the problems in both the waveshrink and thresholding methods, a more advanced method is proposed [12]. This system employs an automatic pause detection algorithm using a voice activity detection (VAD) and introduces several threshold profiles for different types of contaminant signals. With the VAD, a more accurate estimation of noise power is achieved. In addition to the VAD, it incorporates speech signal features in the system. It also implements a mechanism that efficiently selects suitable parameters for voiced, unvoiced and silence regions, separately. The use of several threshold profiles enables switching several threshold criteria according to the contaminant signal. Consequently, the system can cope with colored and non-stationary contaminant signals.

2.4 Wavelet Extrema Clustering

Another method based on the adoption of the speech production model is the wavelet extrema clustering. It assumes that the detrimental effects of the contaminant signal introduce zeros into the overall system and only affects the speech excitation sequence (not the all-pole filter) [13]. A class of wavelets are employed to decompose the LPC residuals to calculate the wavelet extrema. The underlying impulsive structure of the desired speech (non-reverberant) are captured by locating the extrema which has the characteristics of being well clustered. The extrema at each wavelet scale are effective indicators of the impulses (clean speech) in the contaminated signal. These are used to reconstruct the non-reverberant speech.

2.5 Wavelet Filtering with Wiener Gain

We have previously expanded the multi-band wavelet domain filtering [9] to address the dereverberation problem [14]. The general expression of the Wiener gain at band m [14] is expressed as

$$\kappa_m = \frac{S(\upsilon,\tau)_m^2}{S(\upsilon,\tau)_m^2 + X_L(\upsilon,\tau)_m^2}, \tag{4}$$

where $S(\upsilon,\tau)_m^2$ and $X_L(\upsilon,\tau)_m^2$ are wavelet power estimates for the clean speech and the late reflection, respectively. And υ and τ are the wavelet parameters scale and shift, which will be explained in Section 3. Wavelet filtering is carried out by weighting the reverberant wavelet coefficients $X(\upsilon,\tau)$ with the Wiener gains as,

$$X(\upsilon,\tau)_m(enhanced) = X(\upsilon,\tau)_m \ . \ \kappa_m. \tag{5}$$

In Eq. (5), the Wiener weighting κ_m dictates the degree of suppression of the late reflection to the observed signal. If the late reflection power estimate is greater than the estimate of the speech power, then κ_m for that band may be set to zero or a small value. This attenuates the effect of the late reflection. Moreover, if the power of the clean speech estimate is greater, the Wiener gain will emphasize its effect. The enhanced wavelet coefficients are converted back to the time domain through

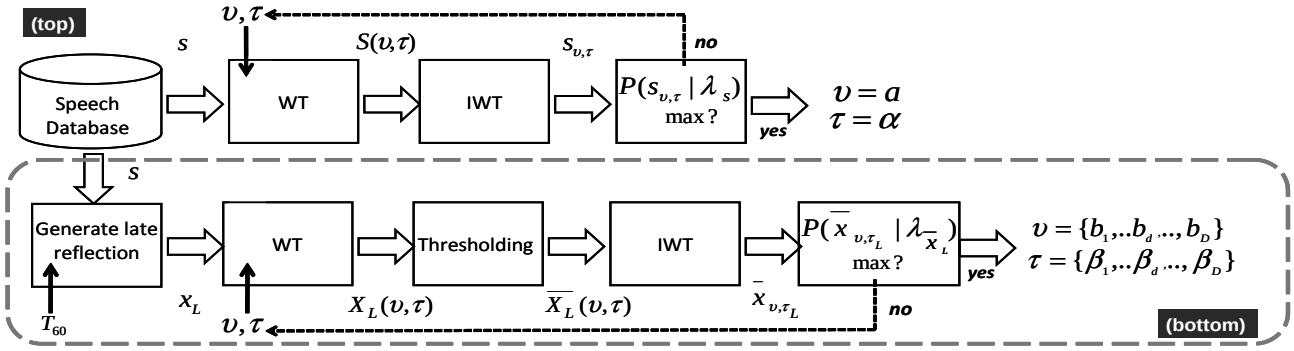


Figure 1: Block diagram of the wavelet optimization scheme.

inverse wavelet transform (IWT). In our previous work [14], the wavelet parameters are not optimized to track the clean speech and the late reflection given a reverberant observation.

3 Optimizing wavelet Parameters υ and τ based on Acoutsic Model Likelihood

A wavelet is generally expressed as

$$\Psi(\upsilon,\tau,t) = \frac{1}{\sqrt{\upsilon}}\Psi\left(\frac{t-\tau}{\upsilon}\right), \tag{6}$$

where t denotes time, υ and τ are the scaling and shifting parameters respectively. $\Psi\left(\frac{t-\tau}{\upsilon}\right)$ is often referred to as the mother wavelet. Assuming that we deal with real-valued signal, the wavelet transform (WT) is defined as

$$F(\upsilon,\tau) = \int f(t)\Psi(\upsilon,\tau,t)dt, \tag{7}$$

where $F(\upsilon,\tau)$ is the wavelet coefficients and $f(t)$ is the time-domain function. With an appropriate training algorithm we can optimize τ and υ so that the wavelet captures specific characteristics of a certain signal of interest. The resulting wavelet is sensitive in detecting the presence of this signal given any arbitrary signal.

For illustration purpose, we will only show the optimization of the wavelet parameters υ and τ for the wavelet filtering method discussed in Section 2.5. In the wavelet filtering method, we are interested in detecting the power of clean speech and late reflection given a reverberant signal.

We optimize the wavelet to detect clean speech and late reflection separately based on the acoustic model likelihood as shown in Fig. 1. In ASR, we assume that the speech does not vary for a certain time-frame. Thus, optimizing a single wavelet template for speech will be sufficient. In Fig. 1 (top) we illustrate the optimization of the wavelet for clean speech. Wavelet coefficients $S(\upsilon,\tau)$, extracted through Eq. (7), are converted back to time domain $s_{\upsilon,\tau}$. Likelihood scores are computed using the clean speech acoustic model λ_s. The process is iterated, adjusting υ and τ. The corresponding υ=a and τ=α that result to the highest score are selected. In the case of the late reflection in Fig. 1 (bottom), D templates are to be optimized for both scale $(\upsilon_1,...\upsilon_D)$ and shift $(\tau_1,...,\tau_D)$. These correspond to D preceding frames that cause smearing to the current frame of interest. We note that the effect of smearing is not constant, thus D templates are created. By estimating the reverberation time T_{60}, we can generate the impulse response and its corresponding late reflection coefficients h_L. Both T_{60} estimation and impulse response generation are discussed in [15]. Then, late reflection observations x_L are generated by convolving the clean speech with h_L. Next, wavelet coefficients $X_L(\upsilon,\tau)$ are extracted through WT (Eq. (7)). To make sure that $X_L(\upsilon,\tau)$ is void of speech characteristics, thresholding is applied to $X_L(\upsilon,\tau)$. Speech energy is characterized with high coefficient values [11] [12] and thresholding sets these coefficients to zero,

$$\bar{X}_L = \begin{cases} 0 & ,\mid X_L \mid > \ thr \\ X_L & ,\mid X_L \mid < \ thr \end{cases} \tag{8}$$

thr is calculated similar to that in Eq. (3). The thresholded signal is converted back to time domain $\bar{x}_{\upsilon,\tau_L}$ and evaluated against a late reflection model $\lambda_{\bar{x}_L}$. The parameters υ and τ are adjusted and the corresponding υ=$\{b_1,...b_D\}$ and τ=$\{\beta_1,...\beta_D\}$ that result to the highest likelihood score are selected. We note that the acoustic model λ_s is trained with clean speech data, while $\lambda_{\bar{x}_L}$ uses the synthetically generated late reflection data with thresholding applied.

By using these optimized wavelet parameters, we can estimate both the clean speech and late reflection power directly from the observed reverberant signal $X(\upsilon,\tau)$ and use these to estimate the Wiener gain in Eq. (4). Thus, the speech power estimate becomes

$$S(\upsilon,\tau)_m^2 \approx X(a,\alpha)_m^2, \tag{9}$$

and the late reflection power $X_L(\upsilon,\tau)_m^2$ estimate

$$X_L(b_d,\beta_d)_m^2 \approx \begin{cases} X(b_1,\beta_1)^2 \ , \ d=\ 1 \\ \dfrac{\sum_{k=1}^{d-1} X(b_k,\beta_k)^2}{d-1} + X(b_d,\beta_d)_m^2 \ , \\ \text{otherwise} \end{cases} \tag{10}$$

where d (smearing effect) is the d-th frame template (for k:1,...,D).

Table 1: System specification used in evaluating the system

Sampling frequency	16 kHz
Frame length	25 ms
Frame period	10 ms
Pre-emphasis	$1 - 0.97z^{-1}$
Feature vectors	12-order MFCC, 12-order ΔMFCCs 1-order ΔE
HMM	8256 Gaussian pdfs
Training data	Adult by JNAS
Test data	Adult by JNAS

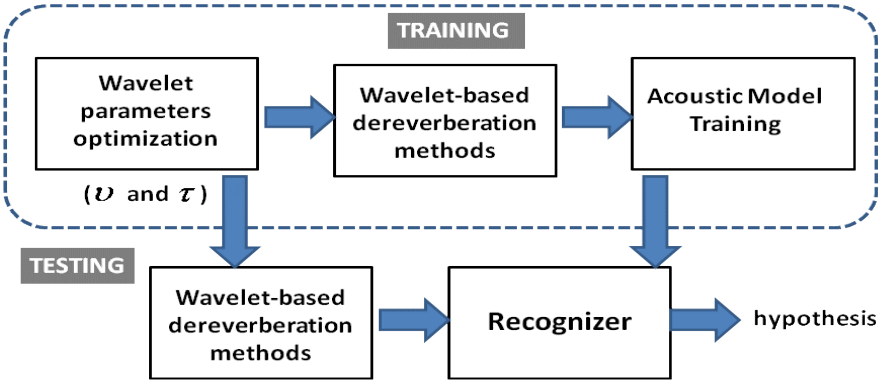


Figure 2: Overall system diagram (Training and Testing).

Table 2: Recognition performance for different wavelet-based methods (No adaptation).

	200 ms	**400 ms**	**600 ms**	**average**
No processing; clean model	68.6 %	41.3 %	21.4 %	43.8 %
No processing; reverberant model	75.4 %	61.2 %	32.1 %	56.2 %
(1) WaveShrink (Sec. 2.1)	75.9 %	63.3 %	40.6 %	60.0 %
(1+) WaveShrink + wavelet optimization	76.7 %	65.4 %	44.9 %	62.3 %
(2) Soft thresholding (Sec. 2.2)	76.5 %	65.8 %	46.7 %	63.0 %
(2+) Soft thresholding + wavelet optimization	78.1 %	67.1 %	49.2 %	64.8 %
(3) Improved wavelet-based speech enhancement (Sec. 2.3)	77.3 %	66.7 %	50.6 %	64.8 %
(3+) Improved wavelet-based speech enhancement + wav. opt.	79.1 % %	68.5 %	54.0 %	67.2 %
(4) Extrema clustering (Sec. 2.4)	78.4 %	67.1 %	59.7 %	68.4 %
(4+) Extrema clustering + wavelet optimization	80.8 %	69.8 %	62.9 %	71.1 %
(5) Wavelet filtering (Sec. 2.5)	81.5 %	71.4 %	64.5 %	72.5 %
(5+) Wavelet filtering + wavelet optimization	83.2 %	74.6 %	68.6 %	75.5 %

4 Experimental Evaluations

We have evaluated the proposed scheme and the five wavelet-based methods described in Section 2. Evaluation is carried out in large vocabulary continuous speech recognition (LVCSR). The training database is from the Japanese Newspaper Article Sentence (JNAS) corpus with a total of approximately 60 hours of speech. The open test set is composed of 200 utterances uttered by 50 speakers. ASR experiments are carried out on the Japanese dictation task with a 20K vocabulary. The language model is a standard word trigram model. The acoustic model is a phonetically tied-mixture (PTM) HMMs with 8256 Gaussians in total. System specification is summarized in Table 1.

We experimented in the condition of reverberation time: T_{60}=200 ms, 400 ms and 600 ms. Reverberant training data are synthetically produced with the automatically generated RIR as discussed in [15]. Test performance is evaluated using real data recorded in a room with known reverberation time: T_{60}=200 ms, 400 ms and 600 ms. In the experiments, we used a total number of bands $M = 5$ which was found to be effective [1][3]. The wavelet used here is the Daubechies wavelet which was also used in [14].

Table 3: Recognition performance for different wavelet-based methods (MLLR adaptation).

	200 ms	**400 ms**	**600 ms**	**average**
No processing; clean model	70.3 %	43.2 %	24.8 %	46.1 %
No processing; reverberant model	76.5 %	63.2 %	35.1 %	58.2 %
(1) WaveShrink (Sec. 2.1)	76.4 %	64.8 %	41.1 %	60.8 %
(1+) WaveShrink + wavelet optimization	77.9 %	67.2 %	46.4 %	63.8 %
(2) Soft thresholding (Sec. 2.2)	77.8 %	67.5 %	47.1 %	64.1 %
(2+) Soft thresholding + wavelet optimization	79.0 %	68.6 %	51.4 %	66.3 %
(3) Improved wavelet-based speech enhancement (Sec. 2.3)	78.5 %	67.9 %	52.1 %	65.1 %
(3+) Improved wavelet-based speech enhancement + wav. opt.	80.0 % %	69.5 %	56.2 %	68.5 %
(4) Extrema clustering (Sec. 2.4)	79.6 %	68.2 %	61.5 %	69.7 %
(4+) Extrema clustering + wavelet optimization	81.5 %	70.7 %	64.1 %	72.1 %
(5) Wavelet filtering (Sec. 2.5)	82.7 %	72.7 %	66.9 %	74.1 %
(5+) Wavelet filtering + wavelet optimization	84.2 %	76.3 %	69.5 %	76.6 %

The process flow of the experiment is shown in Fig. 2. During training, we optimize the wavelet parameters. Using the optimized wavelet parameters, we implemented the wavelet-based dereverberation methods discussed in Section 2, then trained individual acoustic models. During testing, the optimized wavelet parameters were used together with the wavelet-based dereverberation methods to process the reverberant test data. Then, processed data were evaluated in ASR. In our experiments, the actual optimization of the wavelet parameters may vary for each of the different wavelet-based dereverberation methods, depending on individual unique requirements. Nevertheless, the criterion of maximizing the likelihood for the ASR application is maintained for all the methods.

We also implemented a model adaptation based on Maximum Likelihood Linear Regression (MLLR) [16][17]. Model adaptation is used to minimize the mismatch between training and testing conditions. The MLLR adaptation estimates linear transformations for groups of model parameters to maximize the likelihood of the adaptation data. In our adaptation experiment, we used 50 adaptation utterances.

We show the ASR performance in word accuracy for all methods in Tables 2-3. The conventional acoustic model training based on Baum-Welch is used in Table 1 (No adaptation). In Table 2, acoustic model adaptation was implemented using MLLR. In the case of the MLLR, the adaptation data is limited to using only 10 adaptation utterances. In usual case, several adaptation utterances are used (more than 10) for improved performance. In this experiment, we only wanted to verify whether adaptation works in our proposed method.

For reference, we show on the top the results when the reverberant data are not processed and matched against clean and reverberant acoustic models, respectively. We show the results based on waveshrink and thresholding (Sections 2.1 and 2.2) in (1) and (2), respectively. The improvement in (1+) and (2+) from (1) and (2) are the results when the wavelet parameters are optimized. The improved wavelet-based enhancement system that incorporates VAD and threshold profiles (Section 2.3) is shown in (3). In (3+), an improvement in performance is attained when wavelets are optimized as compared to (3). Another method based on extrema clustering (Section 2.4) is provided in (4) together with the optimized wavelet version in (4+). The result of our previous dereverberation approach (Section 2.5) [14] is shown in (5), while the result of incorporating wavelet optimization discussed in Section 3 is given in (5+).

The results in Tables 2-3 show that all the methods (1-5) benefit from the proposed method. By optimizing the wavelet parameters, the dereverberation process is more tuned to improving the acoustic model likelihood. As a result, it becomes more effective in the ASR application. Moreover, we observe a consistent improvement in recognition performance when the model adaptation was conducted. Thus, the proposed optimized dereverberation method also works in the context of adaptation.

We note that in (1),(2) and (3), dereverberation is implemented by means of directly thresholding the wavelet parameters. This may have detrimental effects to the speech recognition performance due to the non-smooth nature of the thresholding function. In our method, thresholding is only used to select the the optimal wavelet parameters and not directly applied to the wavelet coefficients. The actual weighting of the wavelet coefficients is through Wiener filtering, which is a smoother weighting function based on the power ratio of the estimated clean speech and late reflection. Moreover, (1),(2),(3),(4) and (5) are originally based on improving the speech quality (hearing) of the dereverberated signal. However, improving the speech quality may not necessarily translate to improvement in ASR performance. Thus, when we optimized the system for ASR, we have achieved improvement in the recognition performance.

5 Conclusion

Wavelet-based speech enhancement approach has been successfully used in addressing denoising problems. Its application has been extended to reverberant scenarios. Although satisfactory improvement in signal-to-noise ra-

tio has been reported, the existing approach is primarily optimized for improved human perception. In our method, we are interested in optimizing the wavelet-based dereverberation for ASR.

We proposed to optimize the wavelet parameters used in dereverberation in ASR. This scheme guarantees that the optimized parameters improve the model likelihood used in ASR. We have evaluated existing wavelet-based methods. Moreover, we have shown that our approach is effective in improving the ASR performance when applied to different wavelet-based dereverberation methods. In the future, we investigate the effects of contaminated noise and extend this work to deal with both noisy and reverberant environment conditions.

References

[1] R. Gomez, J. Even, H. Saruwatari and K. Shikano, "Distant-talking Robust Speech Recognition Using Late Reflection Components of Room Impulse Response" *In Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing ICASSP*, 2008.

[2] R. Gomez, J. Even, H. Saruwatari, and K. Shikano, "Fast Dereverberation for Hands-Free Speech Recognition" *IEEE Workshop on Hands-free Speech Communication and Microphone Arrays HSCMA*, 2008

[3] R. Gomez, T. Kawahara, "Optimization of Dereverberation Parameters based on Likelihood of Speech Recognizer" *In Proceedings of Interspeech*, 2009.

[4] R. Gomez and T. Kawahara, "Robust Speech Recognition Based on Dereverberation Parameter Optimization Using Acoustic Model Likelihood" *In Proceedings of the IEEE Transactions on Audio, Speech, and Language Processing*, 2010.

[5] Vaseghi SV., "Advanced Digital Signal Processing and Noise reduction" *2nd ed. Wileys*, 2005.

[6] Q. Fu and EA. Wan, "Perceptual Wavelet Adaptive Denoising of Speech" *In Proceedings of EUROSPEECH*, 2003.

[7] JW. Seok and KS. Bae, "Speech Enhancement with Reduction of Noise in the Wavelet Domain" *In Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing ICASSP*, 1997.

[8] M. Jansen, "Noise Reduction by Wavelet Thresholding" *In Proceedings of the IEEE International Symposium on Industrial Electronics, ISIE*, 2001.

[9] E. Ambikairajah et. al., "Wavelet Transform-based Speech Enhancement" *In Proceedings of International Conference on Speech and Language Processing ICSLP*, 1998.

[10] H.Y. Gao, "wavelet Shrinkage Denoising", *Computational Graphical Statistics* 1998.

[11] D.L. Donoho, "Denoising by soft thresholding", *IEEE Trans. Info. Theory* 1995.

[12] H. Sheikhzadeh and Hamid. Abutalebi, "An Improved Wavelet-based Speech Enhancement System" *In Proceedings Eurospeech*, 2001.

[13] S. Griebel and M. Brandstein, "Wavelet Transform Extrema Clustering for Multi-channel Speech Dereverberation" *In Proceedings of the IEEE Workshop on Acoustic Echo and Noise Control*, 1999

[14] R. Gomez and T. Kawahara, "Optimizing Spectral Subtraction and Wiener Filtering for Robust Speech Recognition in Reverberant and Noisy Conditions" *In Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing ICASSP*, 2010.

[15] R. Gomez, T. Kawahara, "Tight Integration of Dereverbeartion and Automatic Speech Recognition" *In proceedings of the Asia Pacific Signal and Information Processing Association APSIPA*, 2009.

[16] M.J.F. Gales and P.C. Woodland, "Mean and variance adaptation within the MLLR framework" *In Proceedings of Computer Speech and Language*, 1996.

[17] Leggeter, C.J., Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models" *In Proceedings of Computer Speech and Language*, 1995.

社団法人　人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B002-5

Programming by Playing and Approaches for Expressive Robot Performances

Angelica Lim, Takeshi Mizumoto, Toru Takahashi, Tetsuya Ogata, and Hiroshi G. Okuno
Graduate School of Informatics, Kyoto University
36-1 Yoshida-Honmachi
Sakyo-ku, Kyoto, 606-8501 JAPAN
{angelica, mizumoto, tall, ogata, okuno}@kuis.kyoto-u.ac.jp

Abstract

This paper extends our work with a theremin-playing robot accompanist. Here, we consider that a good accompanist should play with "expression": small deviations in volume, pitch and timing. We propose a Programming by Playing approach that allows a human flutist to transfer a performance to a robot theremin-ist, keeping these expressive changes intact. We also examine precisely what makes music robots play more or less "robotically, and survey the eld of musical expression in search of a good model to make robots play more like humans.

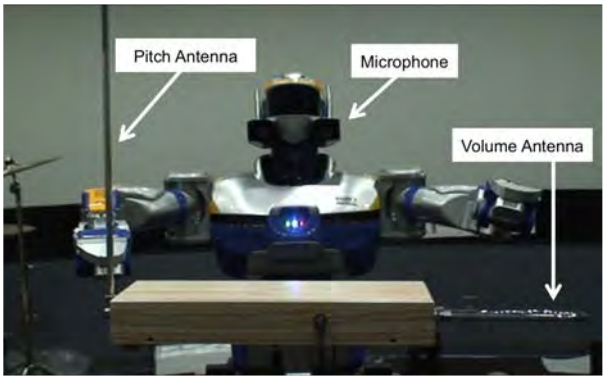


Figure 1: HRP-2 robot listens to a performance with its microphone, then replays it on the theremin by varying pitch and volume.

1 Introduction

A major challenge in human-robot interaction is the current lack of "humanness" in robot communication. Whereas humans express emotions using vocal inflection, expressive gestures and facial expression, robots have difficulty detecting these implicit emotions. Conversely, robot speech and movements remain dry, flat and unnatural. How can we make robots both detect these inexplicit emotions, and respond in emotionally empathetic, expressive ways? In the field of computer music, adding expression to synthesized music has already been a major goal since the 1980's [Todd, 1985a]. Musical expression is the result of adding variations [Sundberg, 1993] to a neutral ("robotic") performance, giving pleasing, natural renditions, sometimes even evoking emotions from listeners. Furthermore, there is evidence that communication of emotions in music follow the same patterns as speech [Juslin and Laukka, 2003]. Thus, we pursue the possibility that by giving robots musical expression detection and production abilities, we are one step closer to natural human-robot interaction.

We first propose a method called *Programming by Playing*: our anthropomorphic robot [Mizumoto *et al.*, 2009] listens to a flutist's performance with its own microphone, then replays the piece on the theremin with the same timing and dynamics as the human (Fig. 1). In the field of music robots, Solis et al. [Solis *et al.*, 2007] have already achieved an impressive increase in expressiveness by training an artificial neural network (ANN) to reproduce a human flutist's vibrato and note length. However, expression is a multifaceted problem that we can attack from many angles; for example, many musicians are able to play a given piece in a "sad" or "happy" manner on demand [Gabrielsson and Juslin, 1996]. How could we make robots play with emotion, too?

In the second part of this paper, we survey musical expression research not only from a computational music perspective, but also a psychological perspective. We first review some factors which make a performance expressive or not, then describe a 5-dimensional musical expression model [Juslin, 2003] suggested by music psychologist Juslin for the case of human musicians. We suggest that by extending the Programming by Playing approach to consider such a model, music robots could both perceive human musician's emotional intentions, and produce these emotions in their own playing as well.

2 A programming by playing approach

Let us begin by considering the simplest method for giving robots the appearance of human expressiveness: mimicry. At first sight, translating a human performance to a robot performance seems like a simple problem of music transcription. The naive approach would be to segment the performance into notes (using note onset detection, for example), extract each note's pitch and volume, and create a robot-playable MIDI file that contains each discretized note. This technique has worked well for piano because a piece can be represented simply by 3 parameters for each note: note length, pitch, and key-strike velocity [Raphael, 2009].

We claim that, while MIDI transcription may work well for piano, this note-level representation is an oversimplification for continuous instruments such as flute, voice and violin. Here are some concrete examples:

- *Intra-note volume changes* over the course of a note (e.g. *crescendo* or *diminuendo*) add fullness and expression for many continuous instruments. This is often overlooked because single piano notes cannot change volume in a controlled manner over time.
- *Intra-note pitch variation* known as vibrato can vary in speed and depth within a note. In most MIDI representations, vibrato speed and depth are set to constant values, if present at all.
- *Pitch bends*, or purposely playing slightly flat or sharp for expressive effect may be discretized to the nearest semi-tone.
- *Articulation* such as legato, attacked, staccato is produced by musicians using carefully composed note volume envelopes. In MIDI, this is often abstracted into a single average volume per note.
- *Timbre.* For instruments with timbral characteristics, tones can be "bright" or "dull" depending on their spectral composition; this information may be lost, too.

In summary, many critical details that may make a performance expressive can be lost when representing a piece symbolically! Thus, we must take care to represent our score in as rich a way as possible.

2.1 An Intermediate Representation: The Theremin Model

Raphael [Raphael, 2009] has proposed that the essence of an expressive melodic performance can be represented using a simple, but capable "theremin model". The theremin model takes after the electronic instrument of the same name that produces a pure sinusoidal pitch. Players can modulate the theremin's pitch frequency and volume independently, by moving their hands closer or farther from the respective pitch or volume antennas. We therefore represent a performance as a pitch trajectory and volume trajectory that continuously varies over time. Equation 1 represents the discrete sound signal s at time t:

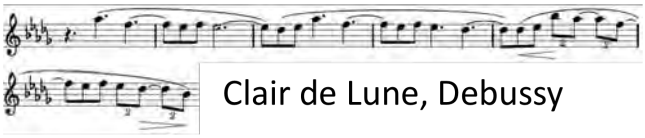


Figure 2: Example piece played by human flutist

$$s(t) = a(t) * \sin(f(t) * 2\pi * t), \quad (1)$$

where:

- $a(t)$ is the amplitude (a.k.a. power)
- $f(t)$ is the fundamental frequency (a.k.a pitch)

With a sufficient number of samples per second, this representation can capture almost all of the subtle information described in the previous section. For example, an attacked note would be equivalent to a sharp increase and quick drop in $a(t)$. Vibrato and note changes are captured in modulations over time in $f(t)$. Unfortunately, timbral characteristics, otherwise known as tone color, are not representable here, as a theremin's sound is characteristically composed of only a pure sine wave. See [Raphael, 2009] for a modified theremin model which adds timbre as a function of amplitude using hand-designed functions.

This simple representation captures the essential details of a performance while allowing for inter-instrument transfer. As noted in [Williamon, 2004], "The communication of emotion in music is generally successful despite individual differences in the use of acoustic features among performers... and different musical instruments." In more concrete terms, we can take as input a recording of a human's performance on flute, and output a performance by our robot thereminist.

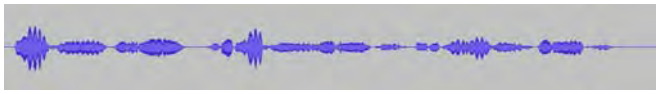
2.2 Acoustic Processing

The input to our system is a wave file recording of a piece played by an intermediate flute player. It is recorded using the robot's own microphone, sampled at 44.1 kHz. As an example, consider the excerpt from Clair de Lune as shown in Fig. 2. Processing of the flute recording is composed of three parts: robot noise removal, continuous power extraction, and continuous fundamental frequency extraction.

2.2.1 Noise Reduction

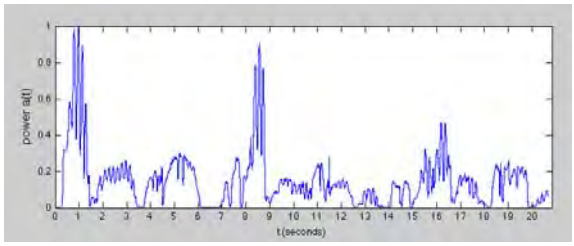
To increase robustness in our next steps, we first remove the robot fan noise also captured during recording. We use a filter called a spectral noise gate, which is likened to "background subtraction". By analyzing the frequency spectrum of a "silent" part of the recording (ie. when the flutist is not playing) we can reduce the fan noise by 24 dB from the entire recording (see Fig. 3). An FFT size of 2048 is used, resulting in 1024 frequency bands.

(a)

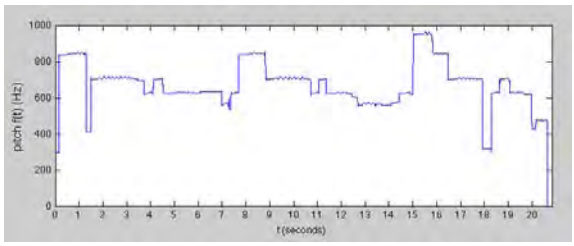


(b)

Figure 3: Clair de Lune original recording before (a) and after (b) fan noise reduction.



(a)



(b)

Figure 4: Continuous power a(t) and pitch f(t) extracted from flutist's Clair de Lune recording.

2.2.2 Continuous Power Estimation

We now have the filtered recorded signal $x(t)$. To extract the power $a(t)$, we use window sizes of 512 and sum the values of $x(t)^2$ of each bin. We then normalize the result to values between 0 and 1. The resulting power is plotted in Fig. 4(a).

2.2.3 Continuous Fundamental Frequency Estimation

Using the same input signal $x(t)$, we estimate the fundamental frequency at windows of 2048 using multi-comb spectral filtering and a hopsize of 1024. Instead of discretizing to the nearest semi-tone on the melodic scale, we measure to the nearest frequency in Hz. We can visualize the pitch estimation in Fig. 4(b).

2.3 From Representation to Performance

To convert the theremin model representation to a performance, we must first consider two constraints: instrument-related constraints and player (robot) constraints. Finally, we can convert our intermediate representation to a score playable by our robot thereminist.

2.3.1 Instrument-related Constraints

In this step we modify our performance representation depending on our target instrument. Consider that during silent sections of the recording where $a(t)$ is 0, the detected frequency $f(t)$ could have an arbitrary number of possible settings. To relate this situation to other instruments, a marimba player, for example, may return to "home position" during silent rests, and perhaps a flute player may hold the flute neutral with no keys depressed. In the case of our target instrument, the theremin, we assume that a theremin player would anticipate the next note during rests. Concretely, where $a(t)$ is 0, we set $f(t)$ to the next non-zero value of $f(t+k)$ where k is positive. Other possible modifications that may fall under Instrument-related constraints may include changing register (in case the human's instrument is, for example, a bass instrument, and the robot's instrument is soprano).

2.3.2 Player-related Constraints

Beginner and expert musicians have very different capacities. In our case, our player is an HRP-2 robot produced by Kawada Industries. However, in [Mizumoto *et al.*, 2009] Mizumoto et al. showed that the theremin-playing capabilities can be easily transfered to other robots, including a humanoid robot developed by Honda. In tests with another Kawada Industries robot, Hiro, we found that Hiro can change notes faster than HRP-2, due to a difference in arm weight. Thus, we must either modify our representation to be "easy" enough for our particular robot to play, or program these constraints into the motor module directly. For now, we scan our representation for any changes in frequency or volume that would violate the maximum acceleration of our robot arm, and remove them.

2.3.3 Generating a Robot-Playable Score

In this final step, we convert our intermediate representation score to a robot playable score. In preliminary experiments, we found that our system could handle a score with 3 pitch/volume targets per second (i.e., an update rate of 3 Hz) and still play in real-time using feedforward control. Using our Programming by Playing method, we thus update our robot's target note and volume multiple times per note, achieving more subtle tone and volume variations.

2.4 Preliminary Results and Improvements

We implemented Programming by Playing coupled with the theremin volume/pitch model to transfer the performance of "Clair de Lune" by a human flutist to a robot thereminist. In informal listening tests, the resulting performance does indeed sound more natural than our score-based method, but the reader is encouraged to evaluate the performance for themselves at

`http://winnie.kuis.kyoto-u.ac.jp/members/angelica/pbp`.

Although vibrato could be heard slightly, our maximum update rate of 3 Hz may have been too little to fully define vibrato (which previously had been hand defined

at 5-10 Hz). It also remains to be seen whether using the theremin model representation could be applied to instrument pairs other than flute-theremin. In particular, we have not implemented timbre into our performance representation, though this could be implemented with a third continuous parameter containing the extracted spectral centroid of the original recording.

An immediate use for Programming by Playing is allowing a human ensemble player to program the robot with his own style. That is, it is much easier to synchronize with a duet player that plays with natural timings, pauses, and articulations similar to one's own. Other uses for this version of Programming by Playing could include embodying famous musicians in a music robot based on their music recording.

Up until now, we have taken a relaxed approach to music expressiveness. As previously conjectured, intra-note volume variation, vibrato, pitch bends, articulation, and potentially timbre all contribute to making a performance more expressive. In the next section, we will see why these minute details are so important, and examine how we can exploit them to generate expressive performances "from scratch".

3 Expressive performances

3.1 Definitions

Expression is the most important aspect of a musician's performance skills, reports a nationwide survey of music teachers [Laukka, 2004]. But what is expression exactly? According to the survey, most teachers define expressivity as the communication of the emotional content of a piece, such as joy, sadness, tenderness or anger. Occasionally an expressive performance can even evoke these emotions in the listener ('being moved'), though it is not obligatory for music to be expressive [Davies, 1994]. What else makes human performers sound so different from the "dead-pan" rendition of a piece by a computer?

Another typical definition of expressiveness is "deviation from the score". Although scores may be marked with dynamic markings such as *decrescendo* or *accelerando*, expert performers contribute other expressive changes to the score [Palmer, 1997]. Typical examples include [Kirke and Miranda, 2009]:

- unmarked changes in tempo (such as playing faster in upward progressions of notes)
- loudness (high notes played slightly louder)
- modifications in articulation (staccato or legato)
- changes in intonation (making notes slightly flatter or sharper)
- adding vibrato at varying frequencies
- changing the timbre, if applicable to the instrument

The regularity of these deviations suggest that performances may be either subject to a set of grammar-like rules, or learned to some extent, and has thus spawned a vast number of attempts to reproduce these human-like qualities using computational methods.

3.2 A Need for Psychological and Physical Models

Automated computer systems for expressive music performance (CSEMPs) are programs which take a score as an input and attempt to output expressive, aesthetically pleasing, and/or human-like performances of the score. A recent survey of CSEMPs [Kirke and Miranda, 2009] outlined the various approaches including rule-based, linear regression, artificial neural network, case-based and others. There are too many approaches to outline here, but it is the conclusion of the survey that sparks the most interest.

According to the review, "Neurological and physical modeling of performance should go beyond ANNs and instrument physical modeling. The human/instrument performance process is a complex dynamical system for which there have been some deeper psychological and physical studies. However, attempts to use these hypotheses to develop computer performance systems have been rare." [Kirke and Miranda, 2009] They cite an attempt to virtually model a pianist's physical attributes and constraints [Parncutt, 1997] as one of these rare cases. Thus, in the following sections, we delve deeper into the phenomenon of expression, in order to better understand this challenge.

3.3 Factors

What factors can make a performance expressive or not? Though researchers typically focus on how the *performer* is expressive, the phenomenon can involve environmental factors, too. We briefly overview these factors from [Juslin, 2003], to better understand the variables involved.

3.3.1 The Piece

The musical composition itself may invoke a particular emotion. For example, Sloboda [Sloboda, 1991] found that certain scores consistently produced tears in listeners: scores containing a musical construct called melodic appogiaturas. Shivers were found in participants during points of unprepared harmonies or sudden dynamic change in the score. Score-based emotions have been well-studied, and in a recent review of 102 studies by Livingstone et al. [Livingstone *et al.*, 2010], it was found that happy emotions are most correlated with pieces in major keys, containing simple harmonies, high pitch heights, and fast written tempos. Loud pieces with complex harmonies, in a minor key with fast tempos were considered "angry", and so on. Though we choose not to treat this score-based emotion in the present paper, this is useful to know so we do not confuse emotion evoked by a written score with emotion projected by a performer.

3.3.2 The Listener

The musical background and preferences of the listener may have an effect on the perceived expressiveness of a piece. For example, listeners with less musical education appear to rely more heavily on visual cues (such as gestures or facial expression) rather aural cues when deciding on an affective meaning of a musical perfor-

mance [Thompson *et al.*, 2005]. However, even children at the age of 5 years are able to differentiate happy and sad pieces based on whether the tempo is fast or slow, and six-year-olds can classify additionally based on major versus minor mode [Dalla Bella *et al.*, 2001]. Interestingly, detection of basic emotions such as joy, sadness, and angry even appear to be cross-cultural: Western and Japanese listeners are able to distinguish these emotions in Hindustani ragas [Balkwill and Thompson, 1999]. Thus, though we should take care during evaluations of expressiveness, we should know that detection of emotion in music is not as elusive as it may seem.

3.3.3 The Context

The performance environment, acoustics or influence from other individuals present can also affect the expression perceived [Juslin, 2003]. For example, music at a patriotic event may evoke more emotion in that context than in another. Another example is Vocaloid's virtual singer Hatsune Miku, who performs at concerts to a large fanbase despite being a synthetic voice and personality. In these cases, perceived expressiveness may also depend on factors such as visual and cultural context.

3.3.4 The Instrument

Whereas percussion instruments such as piano can only vary timing, pitch and volume, continuously controlled instruments such as flute and violin have many more expressive features. They can change timbre to obtain "bright" versus "dull" tones [Raphael, 2009], have finer control over intensity and pitch, and can produce vibrato. Interestingly, human voice is also in this set of continuously controlled instruments. Since many studies find that timbre, pitch variations and vibrato [Livingstone *et al.*, 2010] can have an effect on the perceived expressiveness, the choice of instrument can limit or extend the ability to convey a particular emotion.

3.3.5 The Performer

Clearly the most important factor of expression lies in the performer, which is why this factor has been so extensively studied. The musician's structural interpretation, mood interpretation, technical skill and motor precision can all affect the perceived expressiveness. We explore the expressive aspects of a performer in detail in the next section.

3.4 A Model for Performer Expressiveness

Up until now, performer expressiveness has been informally described by a large number of performance features, such as playing faster and louder, and with more or less vibrato. Are there any models that can bring order and sense to these empirically derived findings?

Four computational models for expressive music performance were considered in [Widmer and Goebl, 2004]: KTH's rule-based model [Bresin *et al.*, 2002], Todd's model based on score structure [Todd, 1985b], Mazzola's mathematical model [Mazzola, 2003], and Widmer's machine learning model [Widmer and Goebl, 2004]. However, according to the CSEMP review, they are still not sufficient. As the review points out, we should search for a model that adheres to certain requirements: it should take into account psychological and neurological factors, as well as physical studies.

Music psychologist Juslin proposed a 5-faceted model [Juslin, 2003] [Juslin *et al.*, 2002] that separates expressive performance into a manageable, but all-encompassing space: Generative rules, Emotion patterns, Random variance, Motion-inspired patterns, and Stylistic unexpectedness (called GERMS). Details of each element are described shortly. Juslin et al. implemented the first 4 parts of the model in 2002 using synthesis [Juslin *et al.*, 2002], and tested each facet in a factorial manner. Their results, along with evidence that each of these facets corresponds to specific parts of the brain [Juslin and Sloboda, 2010], make this model promising. Even if Juslin's model is not quite correct, we claim that it is still very useful for designing factorized modules for robot expression.

3.4.1 Generative rules for musical structure

Similar to speech prosody, musicians add beauty and order to their playing by adding emphasis to remarkable events [Juslin and Sloboda, 2010]. By adding the following features, the musician makes their structural interpretation of a piece clear:

- Slow at phrase boundaries [Clarke, 1988]
- Play faster and louder in the center of a phrase [Todd, 1985b]
- Micropause after phrase and subphrase boundaries [Friberg, A. And Sundberg, J. And Fryden, 1987]
- Strong beats louder, longer, and more legato [Palmer and Kelly, 1992]

A complete and slightly different ruleset is listed in Juslin's experiments [Juslin *et al.*, 2002]. Listeners rated synthesized pieces with this component as particularly "clear" and "musical".

3.4.2 Emotion

We previously defined musical expression partly as the ability to communication emotion. Particular sets of musical features can evoke emotions, such as happiness, sadness, and anger. Livingstone et al. recently surveyed 46 independent studies and summarized the main acoustic features corresponding to each of 4 basic emotions [Livingstone *et al.*, 2010]. We reproduce here the most notable of each group. Note that the order may matter (i.e., first features characterizing the emotion more strongly). In the case of conflicting reports, we removed the one with less experimental backing.

1. **Happy**: Tempo fast, Articulation staccato, Loudness medium, Timbre medium bright, Articulation variability large, Note onset fast, Timing variation small, Loudness variability low, Pitch contour up, Microstructure regularity regular, F0 sharp
2. **Angry**: Loudness loud, Tempo fast, Articulation staccato, Note onset fast, Timbre bright, Vibrato

large, Loudness variability high, Microstructural regularity irregular, Articulation Variability large, Duration contrasts sharp

3. **Sad**: Tempo slow, Loudness low, Articulation legato, F0 flat, Note onset slow, Timbre dull, Articulation variability small, Vibrato slow, Vibrato small, Timing variation medium, Pitch variation small, Duration contrasts soft

4. **Tender**: Loudness low, Tempo slow, Articulation legato, Note onset slow, timbre dull, Microstructural regularity regular, Duration contrasts soft

In the evaluation of this factor, happiness versus sadness were implemented by varying tempo, loudness, and articulation. Upon adding emotional cues, listeners judged the piece as "expressive" and "human" by a large factor.

3.4.3 Randomness

Humans, unlike computers, cannot reproduce the exact same performance twice. In studies on finger tapping [Madison, 2000], even professional musicians varied 3-6% (of the inter-onset interval) in tapping precision. It is thus why some software programs such as Sibelius add some random fluctuation to make MIDI playback sound more human [Kirke and Miranda, 2009]. Interestingly, these fluctuations are not completely random; the variation can be simulated by a combination of 1/f noise and white noise [Gilden *et al.*, 1995]. Motor delay noise was simulated in [Juslin *et al.*, 2002] by adding white noise to each note onset time and sound level. Internal timekeeper lag was added by white noise as a function of the note length, filtered to obtain 1/f pink noise.

Although the idea of making robots purposely less precise sounds intriguing, it remains to be seen whether music robots do actually play as perfectly as the computer clocks that control them. Do they achieve perfect timings despite variations in environment such as network lag and motor delay? In computer synthesis tests this randomness factor made performances more "human" over the neutral versions.

3.4.4 Motion constraints

The fourth component refers to two kinds of motion constraints. One pertains to voluntary patterns of human biological motion. Mainly, the final ritardandos of musical performances has been found to follow a function similar to that of runners' decelerations [Friberg and Sundberg, 1999], but more examples can be found in [Juslin *et al.*, 2002]. The other kind of motion constraint is information that specifies that the performer is human. For example, a pianist could not physically play two distant notes faster than two notes side-by-side. This is an involuntary motion constraint.

In terms of robot implementation, safety mechanisms are probably already programmed into lower level motor controls of our music robots. This corresponds to the latter, involuntary constraint. However, similar to the Player-related constraints described in our Programming by Playing approach, it could be possible to add additional motor constraints that mimic natural human movement curves. For example, our pitch or volume trajectories could be smoothed or interpolated with splines. As for the effect of adding the biological motion constraint: listeners rated synthesized pieces more "human".

3.4.5 Stylistic unexpectedness

Despite the systematic discovery of many common expressive features among musicians, humans of course have the freedom to change their style on a whim. For examples, some performers may intentionally play the repeat of a same phrase differently the second time, or a musician may pause longer than usual for dramatic effect. Indeed, in a study on pianists playing the same piece, it was found that graduate students had rather homogenous timing patterns, whereas experts showed more originality and deviations [Repp, 1997].

This element was not included in Juslin's tests due to the difficulty in implementation. Indeed, this could be the crux of what gives originality to a robot's performance. Could we use Programming by Playing to learn the probabilistic tendency of one or many human artists? Could we shape a music robot's "personality" based on this factor (more or less showmanship, or extroversion)? How exactly to approach this module is an open area for research, and perhaps AI in general.

3.5 Towards an Expressive Music Robot

It seems clear that an expressive music robot should thus have 5 modules:

1. **Prosody controller**: to clarify music structure

2. **Emotion controller**: to store and produce an intended emotion

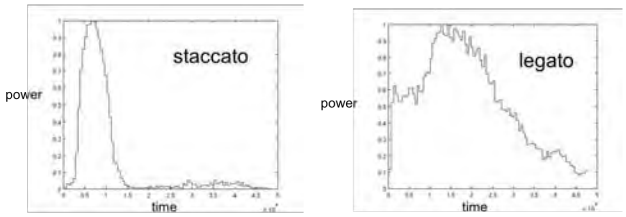
3. **Humanness controller**: to add randomness to imitate human imprecision

4. **Motor smoothness controller**: to mimic human biological movement

5. **Originality controller**: to add unexpected deviations for originality

Although we are still far from implementing this model in full, we have started by implementing the Prosody and Emotion controller. We start with a hand-entered score of the traditional folk song, Greensleeves. Then, it is modified using the generative rules for musical structure mentioned previously. We then address Emotion using Programming by Playing. Focusing on the articulation feature, we record a flutist playing notes in each of the Happy (staccato) and Sad (legato) styles.

We extract volume envelopes for each type as shown in Fig. 5, and apply the volume envelopes to all notes in the continuous volume representation. Our result is two different performances, one to convey sad emotion and the other conveying happiness. It is unclear whether the robot performances effectively convey the emotions as desired, but expressiveness again seems improved over the neutral version. In addition, we have achieved expressiveness without resorting to mimicry.



(a) (b)

Figure 5: Volume envelopes for staccato and legato articulations.

In an ideal version of Programming by Playing, more features (not only articulation) should be extracted. By extracting these acoustic features automatically, perhaps similar to [Mion and De Poli, 2008], we could recognize the emotional content of the human musician.

4 Conclusion and future work

In this paper, we introduced a paradigm called Programming by Playing. We showed how it could be used for expressive robot performance through both mimicry and generation. A key point of the approach was that small details in performance can have a great impact on a performance's expressive content; thus, a good symbolic representation is important.

We also tried to demystify the phenomenon called expression – by applying a 5-facet model to music robot design, we realize that features for structural clarity and emotion are distinct. Another interesting find was that in order to sound more human, we may need to add slight human imprecision. This may be contrary to our current efforts to make "virtuoso" music robots that play faster, but more unrealistically. And finally, the key ingredient missing before music robots will be accepted is a kind of originality or "personality", giving the element of surprise to performances.

All of these factors may be applicable to robot design in general, for example making synthetic voice and movement less "robotic". Yet, what is the goal for music robots? Do we want them to sound more realistic, more human? If that is the case, this complex phenomenon called expression may be the missing ingredient.

Acknowledgments

This work was partially supported by a Grant-in-Aid for Scientific Research (S) No.1910003 and the Global COE Program from JSPS, Japan.

References

[Balkwill and Thompson, 1999] Laura-Lee Balkwill and William Forde Thompson. A cross-cultural investigation of the perception of emotion in music: Psychophysical and cultural cues. *Music Perception*, 1999.

[Bresin *et al.*, 2002] R. Bresin, A. Friberg, and J. Sundberg. Director musices: The KTH performance rules system. *SIGMUS*, pages 43–48, 2002.

[Clarke, 1988] E.F. Clarke. Generative principles in music performance. *Generative processes in music: The psychology of performance, improvisation, and composition*, pages 1–26, 1988.

[Dalla Bella *et al.*, 2001] S Dalla Bella, I Peretz, L Rousseau, and N Gosselin. A developmental study of the affective value of tempo and mode in music. *Cognition*, 80(3):B1–10, July 2001.

[Davies, 1994] Stephen Davies. *Musical meaning and expression*. Cornell University Press, 1994.

[Friberg, A. And Sundberg, J. And Fryden, 1987] L. Friberg, A. And Sundberg, J. And Fryden. How to terminate a phrase. An analysis-by-synthesis experiment on a perceptual aspect of music performance. *Action and perception in rhythm and music*, 55:49–55, 1987.

[Friberg and Sundberg, 1999] Anders Friberg and Johan Sundberg. Does music performance allude to locomotion? A model of final ritardandi derived from measurements of stopping runners. *The Journal of the Acoustical Society of America*, 105(3):1469, 1999.

[Gabrielsson and Juslin, 1996] Alf Gabrielsson and Patrik N. Juslin. Emotional Expression in Music Performance: Between the Performer's Intention and the Listener's Experience. *Psychology of Music*, 24(1):68–91, April 1996.

[Gilden *et al.*, 1995] D. L. Gilden, T. Thornton, and M. W. Mallon. 1/f noise in human cognition. *Science*, 267(5205):1837, 1995.

[Juslin and Laukka, 2003] PN Juslin and P. Laukka. Communication of emotions in vocal expression and music performance: Different channels, same code?. *Psychological Bulletin*, 129(5):770–814, 2003.

[Juslin and Sloboda, 2010] Patrik N. Juslin and John Sloboda. *Handbook of Music and Emotion*. Oxford University Press, USA, 1 edition, February 2010.

[Juslin *et al.*, 2002] PN Juslin, A Friberg, and R Bresin. Toward a computational model of expression in music performance: The GERM model. *Musicae Scientiae*, 6(1; SPI):63–122, 2002.

[Juslin, 2003] PN Juslin. Five facets of musical expression: A psychologist's perspective on music performance. *Psychology of Music*, 31(3), 2003.

[Kirke and Miranda, 2009] A Kirke and ER Miranda. A Survey of Computer Systems for Expressive Music Performance. *ACM Computing Surveys*, 2009.

[Laukka, 2004] P Laukka. Instrumental music teachers' views on expressivity: a report from music conservatoires. *Music Education Research*, 2004.

[Livingstone *et al.*, 2010] Steven R Livingstone, Andrew R Brown, Ralf Muhlberger, and William F Thompson. Modifying Score and Performance Changing Musical Emotion : A Computational Rule System for Modifying Score and Performance. *Computer Music Journal*, 34(1):41–65, 2010.

[Madison, 2000] G Madison. Properties of Expressive Variability Patterns in Music Performances. *Journal of New Music Research*, 2000.

[Mazzola, 2003] Guerino Mazzola. *The Topos of Music: Geometric Logic of Concepts, Theory, and Performance*. Birkhäuser Basel, 1 edition, January 2003.

[Mion and De Poli, 2008] Luca Mion and Giovanni De Poli. Score-Independent Audio Features for Description of Music Expression. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):458–466, 2008.

[Mizumoto *et al.*, 2009] Takeshi Mizumoto, Hiroshi Tsujino, Toru Takahashi, Tetsuya Ogata, and Hiroshi G Okuno. Thereminist robot : development of a robot theremin player with feedforward and feedback arm control based on a theremin's pitch model. In *IROS*, pages 2297–2302, 2009.

[Palmer and Kelly, 1992] C Palmer and MH Kelly. Linguistic Prosody and Musical Meter in Song. *Journal of Memory and Language*, pages 525–542, 1992.

[Palmer, 1997] C. Palmer. Music performance. *Annual Review of Psychology*, 48(1):115–138, 1997.

[Parncutt, 1997] R. Parncutt. Modeling piano performance: Physics and cognition of a virtual pianist. In *ICMC*, pages 15–18, 1997.

[Raphael, 2009] Christopher Raphael. Symbolic and Structrual Representation of Melodic Expression. In *ISMIR*, pages 555–560, 2009.

[Repp, 1997] BH Repp. The aesthetic quality of a quantitatively average music performance: Two preliminary experiments. *Music Perception*, 1997.

[Sloboda, 1991] JA Sloboda. Music Structure and Emotional Response: Some Empirical Findings. *Psychology of music*, 1991.

[Solis *et al.*, 2007] Jorge Solis, Kei Suefuji, Koichi Taniguchi, Takeshi Ninomiya, and Maki Maeda. Implementation of Expressive Performance Rules on the WF-4RIII by modeling a professional flutist performance using NN. In *ICRA*, pages 2552–2557, 2007.

[Sundberg, 1993] J. Sundberg. How can music be expressive? *Speech communication*, 13(1-2):239–253, 1993.

[Thompson *et al.*, 2005] W.F. Thompson, Paul Graham, and F.A. Russo. Seeing music performance: Visual influences on perception and experience. *Semiotica*, 156(1/4):203–227, 2005.

[Todd, 1985a] Neil Todd. A model of expressive timing in tonal music. *Music Perception*, 3(1):33–57, 1985.

[Todd, 1985b] Neil Todd. A Model of Expressive Timing in Tonal Music. *Music Perception: An Interdisciplinary Journal*, 3(1):33–57, 1985.

[Widmer and Goebl, 2004] Gerhard Widmer and Werner Goebl. Computational Models of Expressive Music Performance: The State of the Art. *Journal of New Music Research*, 33(3):203–216, September 2004.

[Williamon, 2004] Aaron Williamon. *Musical excellence: strategies and techniques to enhance performance*. Oxford University Press, 2004.

社団法人　人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B002-6

JOINT USE OF DISTRIBUTED MICROPHONE ARRAY AND LASER RANGE FINDERS FOR SPEAKER IDENTIFICATION IN MEETING

Jani Even, Panikos Heracleous, Carlos Ishi and Norihiro Nogita

ATR Intelligent Robotics and Communication Laboratories
2-2-2 Hikaridai, Kyoto 619-0288, Japan
even@atr.jp

ABSTRACT

This paper presents a text-independent speaker identification system for meetings. During the meeting, all of the meeting participants carry a microphone while a human tracker monitors their movements. The human tracker is based on scanning laser range finder and gives the positions of all the participants at any time. The position information is used to track the geometry of the distributed microphone array formed by all of the microphones. Using the geometry of the distributed array it is possible to cancel interfering speeches and noises from the audio stream assigned to each of the participants. Then, using these processed audio streams, the participants are identified by means of Gaussian mixture models (GMM) that were trained before hand. The proposed system is able to perform identification of simultaneously speaking participants and is thus a good candidate system for meeting diarization task. In particular, the use of laser range finders is a novel approach that makes the position estimation immune to acoustic noise and reverberation. An experiment conducted with three subjects reproducing a meeting configuration demonstrates the performance of the system for identification.

1. INTRODUCTION

These last years, the speech recognition community has been intensively working on the transcription of meetings [1, 2, 3]. An important task in meeting transcription is speaker diarization (i.e. to find "Who spoke when").

In a meeting, it is desirable to impose the least constraints to the participants. For example participants should be allowed to seat freely. Thus a convenient speaker diarization system should be flexible relatively to the positioning of the participants. For hands-free diarization, single microphone [4] or multiple microphones [5] approaches were proposed. Using multiple microphones, it is possible to estimate the position of the speakers using the time differences of arrival [6]. However, in a real environment, the accuracy of the position estimation is reduced because of reverberation and noise. Moreover, prior to diarization, separating the audio streams captured by a distant microphone array requires heavy computation.

For a meeting rooms equipped with a microphone array or with distributed microphones, the observed audio streams are usually processed to obtain one stream for each active participant (for example with audio beamforming in [7]).

Nowadays, with the proliferation of portable devices (laptop computers, PDAs and smart phones), it is not rare that in a meeting situation, each of the participants may be carrying a device having a microphone. Thus the speaker localization and the acquisition of the data streams may be performed using these microphones [8, 9]. Such a set of microphones is referred to as a distributed array. These approaches usually require the different devices to communicate together in order to acquire all the audio streams, then distributed or centralized processing may be applied to perform localization, diarization or other tasks.

As the first step in developing a multi-modal front-end for speaker diarization exploiting a distributed array, this paper discusses the signal processing involved in the speaker identification task (at this first step networking problems are not treated yet). The proposed front-end exploits audio data from the tie microphones and position information given by a human tracker system based on laser range finders (LRF) [10]. During the meeting, the positions of the different participants are tracked using the LRF and one audio stream is obtained for each of the participants using a tie microphone (a microphone fixed in front of the torso). Then speaker identification is performed by using Gaussian mixture models (GMM) of the mel-frequency cepstral coefficients (MFCCs) extracted from theses audio streams [11]. For each participant, the tie microphone fixed on their torso is dominated by their voice when talking. But the speech signal from the tie microphones also contains environmental noises and interferences from the other participant voices if they are talking. If a participant is silent, the interfering voice of the person, or persons, talking at that moment is likely to be the dominant signal. Thus it is necessary to implement a accept/reject system to detect the active channels.

With this system, the participants are able to seat freely,

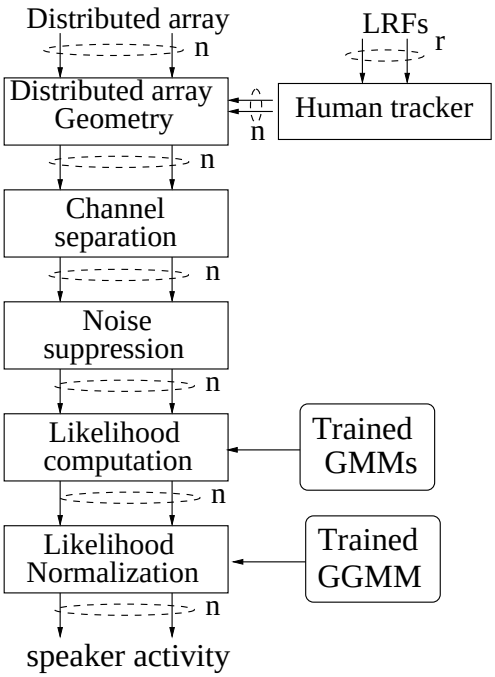


Fig. 1. Outline of the speaker identification system.

to stand and even move during the meeting because the LRFs tracks their positions. Moreover using LRF based tracking is preferable to using audio data for tracking as it insensitive to the acoustic noise and to the reverberation. It is also an interesting alternative to camera based tracking as it is very precise. Note that the number of participants is estimated by the human tracker independently of the fact that they are talking or not.

Experiments were conducted in a realistic meeting situation to demonstrate the capacity of the proposed front-end to identify active participants.

2. METHOD

Fig. 1 gives an outline of the proposed front-end for speaker identification. A presentation of the different modules follows.

2.1. Laser range finder

The motion of the participants in the meeting area is monitored using r LRFs mounted on pole around the meeting area's perimeter (represented by the circles in Fig. 5). The scanning laser range finders are mounted above the obstacles, like the table and chairs, to a height where the torso of the participants (sitting or standing) could be easily observed. To reduce the errors due to noise and occlusion, each person is tracked with a particle filter using a linear motion model with random perturbations. The likelihood is evaluated based on the potential occupancy of each particle's position. By computing a weighted average across all particles, the $\{x, y\}$ position is calculated at a frequency of approximately 37 Hz. Details of the algorithm are presented in [10].

At a given time t, the estimated number of participants in the meeting is $n(t)$ and their estimated positions are $\{x_i(t), y_i(t)\}_{i\in[1,n(t)]}$.

2.2. Noise cancellation

Each of the participants is wearing a tie microphone attached in the front of their torso. The position of these microphones are given by the LRF based tracking system that tracks the position of the torso of all the participants. In this paper, we assume for simplicity that the correspondence between a microphone and a given position is known. Thus the set of tie microphones defines a distributed microphone array whose geometry is known.

The goal of the noise cancellation module is to provide an audio stream for each of the $n(t)$ detected participants that contains less interference from the other participants and fewer environmental noise than the unprocessed streams from the tie microphones (the observed signals). These streams are obtained by filtering the observed signals in the frequency domain. After performing a F bins short time Fourier transform (STFT), the vector of observation in the fth frequency bin is

$$\mathbf{X}(f,k) = \begin{bmatrix} X_1(f,k) \\ X_2(f,k) \\ \vdots \\ X_n(f,k) \end{bmatrix}$$

where k denotes the frame index.

Let us define

$$\mathbf{S}(f,k) = \begin{bmatrix} S_1(f,k) \\ S_2(f,k) \\ \vdots \\ S_n(f,k) \end{bmatrix}$$

the vector containing the speech of all participants at frame index k and frequency bin f. Considering only direct path propagation we can write the mixing process as

$$\mathbf{X}(f,k) = \widehat{\mathbf{A}}(f,k)\mathbf{S}(f,k)$$

where $\widehat{\mathbf{A}}(f,k)$ is the matrix of general term

$$A_{ij}(f,k) = \frac{1}{4\pi r_{ij}(k)} e^{-j2\pi f r_{ij}(k)/c}$$

with c is the celerity of sound and $r_{ij}(k)$ the distance between the jth speech source (the mouth of the jth participant) and the ith microphone (fixed to the ith participant).

The distance $r_{ij}(k)$ is decomposed in two terms d_i, the distance between the mouth of the ith participant and the

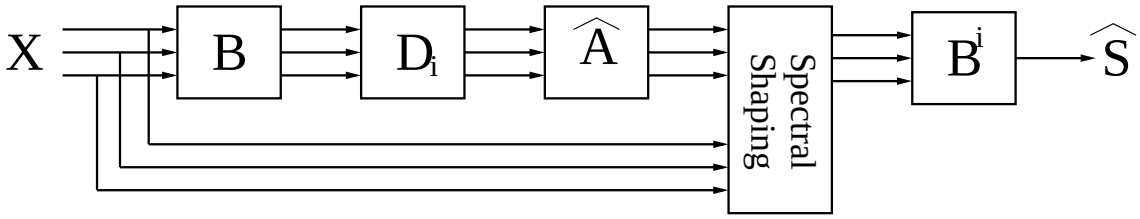


Fig. 2. Noise suppression.

microphone fixed to his/her torso (assumed constant), and $d_{ij}(k)$, the distance between the microphones i and j. We have

$$r_{ij}(k) = \sqrt{d_i^2 + d_{ij}^2(k)}.$$

The distances $d_{ij}(k)$ are obtained using the positions given by the human tracker whereas the distances d_i are assumed to be known.

A separation matrix is obtained by taking the inverse of the mixing matrix

$$\mathbf{B}(f,k) = \widehat{\mathbf{A}}^{-1}(f,k).$$

Then the separated audio streams for the n participants are

$$\mathbf{Y}(f,k) = \mathbf{B}(f,k)\mathbf{X}(f,k).$$

Rather than using these separated streams, better results were obtained by applying a post-filter approach as the ones in [12, 13] (see Fig. 2 where $n = 3$).

Let us define the noise estimate

$$\mathbf{N}_i(f,k) = \widehat{\mathbf{A}}(f,k)\mathbf{D}_i\mathbf{Y}(f,k)$$

where $\mathbf{D}_i$ is a diagonal matrix with all entries set to one except the ith entry which is null. $\mathbf{N}_i(f,k)$ is the estimate of the contribution in the observed signals of all the signals except the ith participant speech. Then an estimate of the contribution of the ith participant speech is obtained by using spectral shaping (We use a post-filter similar to the one used in [12, 13]) The gain for the ith signal is

$$G_i^{(j)}(f,k) \quad = \quad \frac{|X^{(j)}(f,k)|^2}{|X^{(j)}(f,k)|^2 + \alpha|N_i^{(j)}(f,k)|^2}$$

where the superscript (j) denotes the jth component and α is a parameter controlling the noise reduction. The ith component of the filtered target speech is

$$\widehat{Z}_i^{(j)}(f,k) \quad = \quad \sqrt{G_i^{(j)}(f,k)|X^{(j)}(f,k)|^2}\frac{X^{(j)}(f,k)}{|X^{(j)}(f,k)|}.$$

Finally the speech estimate $\widehat{S}_i(f,k)$ is obtained by taking

$$\widehat{S}_i(f,k) = \mathbf{B}^i(f,k)\widehat{\mathbf{Z}}_i(f,k)$$

where $\mathbf{B}^i(f,k)$ is the ith row of the matrix $\mathbf{B}(f,k)$ (the row corresponding to the ith participant).

2.3. Corpus and GMM

Text-independent speaker identification is performed by scoring the MFCCs (12 MFCCs and the energy, their derivatives and their accelerations) extracted from the audio streams of each participant by means of GMMs corresponding to the target speakers [11]. In this experiment, nine speakers were considered (5 females and 4 males). In the remainder of the paper, the speakers are designated by the letters $\{a, b, c, \cdots, i\}$. For each speaker a common training set of 100 Japanese sentences from the JNAS database [14] was recorded using a tie microphone while sitting at the table in the experiment room. Then a GMM was trained for each of the speakers using the 100 utterances. The GMM for all the speakers are designated by $\{\lambda_a, \lambda_b, \cdots, \lambda_i\}$. A general GMM was also trained using the 900 utterances (referred to as GGMM in Fig. 1). The general GMM is designated by λ_G.

The test set was recorded in the same room while monitoring the speaker movement with the LRF based human tracker system. Only three $\{a, b, c\}$ of the nine speakers were sitting around the table and were not constrained of any manner (see Fig. 5). Three different sets of 50 sentences from the JNAS database were prepared and each speaker was assigned one of these sets. Using these sets, 350 test utterances were recorded. First each of the speaker was reading alone its test set (the two other persons are sitting around the table but are remaining silent). These are the test sets T_a, T_b and T_c. Then the three combinations of two speakers simultaneously reading were recorded (test sets T_{ab}, T_{ac} and T_{bc}). Finally, the three speakers were reading simultaneously (test set T_{abc}).

Training and scoring were performed with Htk 3.41 [15] using the whole test utterances.

2.4. Activity detection

The GMMs are used to determine for each utterance which of the participants are active. For decision based on likelihood, it is usual to apply some sort of normalization [16, 17]. In this paper, for a given stream $\widehat{S}_k$ of a given test utterance the likelihood given by the GMMs are normalized using the following likelihood ration

$$\bar{p}(\widehat{S}_k|\lambda_i) = \log p(\widehat{S}_k|\lambda_i) - \log p(\widehat{S}_k|\lambda_G).$$

Then the accept/reject procedure is conducted by comparing the largest normalized likelihood

$$\bar{p}(\widehat{S}_k|\lambda_j) \quad = \quad \max_i \bar{p}(\widehat{S}_k|\lambda_i)$$

to a threshold ϵ

- if $\bar{p}(\widehat{S}_k|\lambda_j) \geq \epsilon$ then speaker j is active in stream $\widehat{S}_k$.
- if $\bar{p}(\widehat{S}_k|\lambda_j) < \epsilon$ then no speaker is active in stream $\widehat{S}_k$.

For each utterance, this test is conducted for all the audio streams.

3. EXPERIMENTS

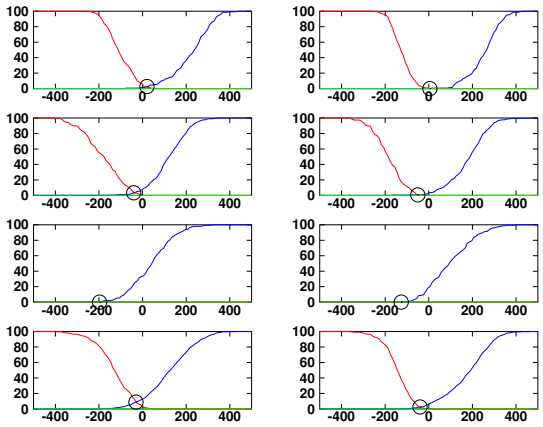


Fig. 3. Deletion (blue), insertion (red) and substitution (green) versus threshold ϵ for single speaker (top row), two speakers (second row), three speakers (third row) and all cases (bottom) for unprocessed (left) and processed audio streams (right).

The experiment setup is described in Fig. 5. The four circles in the corners represent the pole mounted LRFs used by the human tracker, the cross gives the position origin and the probability densities of the positions of the three speakers during the experiment also appear. In this first step, the tie microphones are still wired microphones connected to the same computer.

Note that for all test sets except the test set T_{abc}, at least one of the speakers is silent. The results of the speaker identification experiment are given in terms of deletion, insertion and substitution errors:

- An insertion error occurs when the largest normalized likelihood associated to the audio stream of a silent speaker is larger than the threshold ϵ,

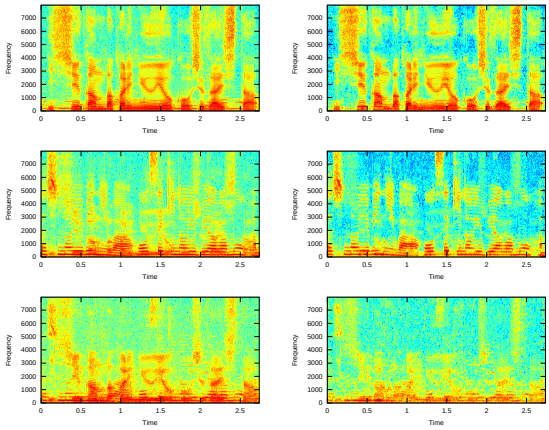


Fig. 4. Spectra of unprocessed audio streams (left) and processed audio streams (right) when speakers a and b are talking.

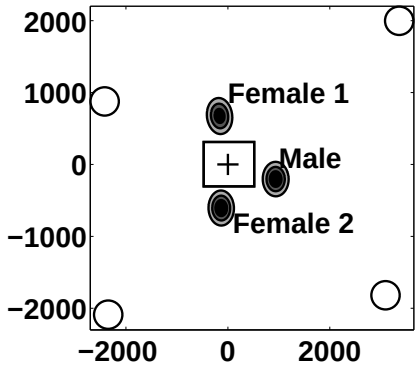


Fig. 5. Pole mounted LRFs (circles), table (rectangle), position origin (cross) and probability densities of the three speakers position (distances are in mm)

- A deletion error occurs when the largest normalized likelihood associated to the audio stream of an active speaker is smaller than the threshold ϵ,
- A substitution error occurs when the largest normalized likelihood associated to the audio stream of an active speaker is larger than the threshold ϵ but is not the correct one (for example $\bar{p}(\widehat{S}_k|\lambda_a)$ is the largest normalized likelihood but the audio $\widehat{S}_k$ is associated to the speaker b).

Two different cases were compared where the audio stream of each speaker is obtained by

- her or his own tie microphone (unprocessed),
- the processed stream $\widehat{S}_k$ she or he is assigned (processed with $\alpha = 17$).

Table 1. Deletion percentage for selected threshold.

	unprocessed	processed
one speaker	2.67	0
two speakers	3.33	0.67
three speakers	0	0
all	8.83	2.5

Table 2. Insertion percentage for selected threshold.

	unprocessed	processed
one speaker	2	0
two speakers	4	0.67
three speakers	0	0
all	10	2

The insertion, deletion and substitution percentages are plotted for different values of the threshold ϵ in Fig. 3. First, we can see that no substitution occurred for this experiment. In all figures, the black circle represent the threshold for which a trade off between insertion and deletion errors is obtained. The percentages for these thresholds are given in Tables 1 and 2. In particular, the processed audio streams give a better performance when considering one unique threshold for all the test sets (bottom of fig 3 and last rows of tables 1 and 2), which is the operating condition.

The spectra of the audio streams are given in Fig. 4 for the test set T_{ab}. The channel assigned to speaker a (top) and b (middle) contain less noise and fewer interference after processing. The processing also reduces the amount of speech that leaks in the bottom channel assigned to the silent speaker c.

4. DISCUSSION

The human tracker is a fast and accurate way of obtain the position of the speakers in the $\{x, y\}$ plane but we have no access to the z coordinate of the mouth or the tie microphone. In this paper, we assumed that all the tie microphones were at the same height and also used an approximation of the distance between a speaker mouth and tie microphone. But during the test recording, the three subjects fixed the tie microphone as they desired. Despite this mismatch, the proposed approach was able to improve significantly the performance for the considered task.

The experiment in this paper was conducted using tie microphones connected to the same computer in order to deal with the signal processing part only. In a real situation, the participants are likely to use microphones connected to devices that communicate using a wireless network. Then for usual approaches one of the most important problem is to synchronize the audio data in order to perform collaborative array processing (like beamforming for estimating the positions)[9]. But with the proposed approach, the localization is performed by the human tracker thus synchronization may be a less sensitive issue.

5. CONCLUSION

In this paper, we proposed an experiment to test the use of LRF based human tracker in a multi-modal front-end for speaker diarization in a meeting situation. Since the positions of all the participants are known at each instant, it is possible to use this information for monitoring a set of tie microphones worn by the participants. Then applying appropriate array processing techniques to this distributed microphone array, it was possible to improve the accuracy in a speaker detection and identification task.

6. ACKNOWLEDGEMENTS

This work was supported by the Ministry of Internal Affairs and Communication.

7. REFERENCES

[1] A. Waibel, H. Yu, M. Westphal, H. Soltau, T. Schultz, T. Schaaf, Y. Pan, F. Metze, and M. Bett, "Advances in meeting recognition," *HLT '01: Proceedings of the first international conference on Human language technology research*, pp. 1–3, 2001.

[2] J. Carletta, "Unleashing the killer corpus: experiences in creating a multi-everything ami meeting corpus," *Language ressource and evaluation journal*, vol. 41, no. 2, pp. 181–190, 2007.

[3] J.G. Fiscus, J. Ajot, and J.S. Garofolo, "The rich transcription 2007 meeting recognition evaluation," *Lecture note in computer science*, vol. 4625, pp. 373–389, 2008.

[4] H. Sun et al., "Speaker diarization system for rt07 and rt09 meeting room audio," *ICASSP 2010*, pp. 4982–4985, 2010.

[5] H. Sun et al., "Speaker diarization for meeting room audio," *INTERSPEECH 2009*, pp. 900–903, 2009.

[6] M.S. Brandstein and H.F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," *ICASSP 1997*, pp. 375–378, 1997.

[7] A. Stolcke, G. Friedland, and D. Imseng, "Leveraging speaker diarization for meeting recognition from distant microphones," *ICASSP 2010*, pp. 4390–4393, 2010.

[8] T.S. Wada, E. Robledo-arnuncio, G. Yue, and B.H. Juang, "Immersive acoustic signal processing for intelligent collaboration," *Proc. 9th Western Pacific Acoustics Conference*, p. 653, 2006.

[9] Y. Jia, Y. Luo, Y Lin, and I. Kozintsev, "Distributed microphones arrays for digital home and office," *ICASSP 2006*, pp. 1065–1068, 2006.

[10] D.F. Glas et al., "Laser tracking of human body motion using adaptive shape modeling," *Proceedings of 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 602–608, 2007.

[11] D.A. Reynolds and R.C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE transaction on speech and audio processing*, vol. 3, no. 1, pp. 72–82, 1995.

[12] Y. Takahashi, Y. Uemura, H. Saruwatari, K. Shikano, and K. Kondo, "Structure selection algorithm for less musical-noise generation in integration systems of beamforming and spectral subtraction," *2009 IEEE Workshop on Statistical Signal Processing SSP2009, Cardiff, Wales, UK*, pp. 701–704, 2009.

[13] J. Even, H. Saruwatari, K. Shikano, and T. Takatani, "Speech enhancement in presence of diffuse background noise: Why using blind signal extraction?," *International Conference on Acoustics, Speech, and Signal Processing ICASSP 2010, Dallas, USA*, pp. 4770–4773, 2010.

[14] K. Ito, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "Jnas: Japanese speech corpus for large vocabulary continuous speech recognition research," *The Journal of Acoustical Society of Japan*, vol. 20, pp. 196–206, 1999.

[15] S.J. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book Version 3.4*, Cambridge University Press, 2006.

[16] A. Rosenberg, J. DeLong, C. Lee, B.H. Juang, and F. Soong, "The use of cohort normalized scores for speaker verification," *Proc. ICSLP*, pp. 599–602, 1992.

[17] T. Matsui and S. Furui, "Likelihood normalization for speaker verification using a phoneme- and speaker-independent model," *Speech communication*, vol. 17, no. 1-2, pp. 109–116, 1995.

社団法人　人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B002-7

ロボットの実環境におけるピッチ抽出に関する考察

Considerations on pitch extraction for robots in real noisy environments

○石井カルロス寿憲（ＡＴＲ知能ロボティクス研究所）
梁棟（大阪大学工学部，ＡＴＲ知能ロボティクス研究所）
石黒浩（大阪大学工学部，ＡＴＲ知能ロボティクス研究所）
萩田紀博（ＡＴＲ知能ロボティクス研究所）

* Carlos Toshinori ISHI, Liang DONG, Hiroshi ISHIGURO, Norihiro HAGITA (Intelligent Robotics and Communication Labs., ATR)

carlos@atr.jp, liang@atr.jp, ishiguro@ams.eng.osaka-u.ac.jp, hagita@atr.jp

Abstract - Pitch extraction is important for communication robots, since pitch may carry information about intention, attitude or emotion expression from the user's speech. However, current pitch extraction methods are not robust enough in real noisy environments. In the present work, we make use of microphone-array technology, and evaluate pitch extraction of multiple speakers in real noisy environments. The MUSIC method for sound source localization, adaptive beamformer for source separation and SACF method for pitch extraction have been used.

1. はじめに

音声に含まれるピッチ情報は、アクセントやイントネーションのみならず、発話者の意図・態度・感情などの表現に大きな役割を果たす[1]。従って、ロボットと人との音声コミュニケーションにおいて、発話者のピッチ抽出はコミュニケーションをより円滑に進めるため、重要である。

ロボットに取り付けたマイクロホンは通常離れた位置（１ｍ以上）にあり、例えば電話音声のようにマイクと口との距離が数センチの場合と比べて、信号と雑音の比（SNR）は低くなる。このため、傍にいる他人の声や環境の雑音が妨害音となり、ロボットによる目的音声の認識を始め、ピッチ情報の抽出も難しくなる。

「ピッチ」とは、知覚される声の高さを表現する用語であるが、声の高さの生成に関する声帯振動の基本周波数（F0）と大きく関連しているため、「ピッチ抽出」と「F0 抽出」を同等に扱う場合が多い。厳密には、観測される F0 は、発声様式によって知覚されるピッチと必ずしも対応するとは限らないが、通常発声の場合は、同等扱いが可能である。

F0 抽出に関しては過去にさまざまな研究がされている[2]-[4]。しかしながら、その大半ではクリーンな発話あるいは適度な雑音が伴う単一なピッチトラックしか対応できなく、ロボットが動作する実環境のデータを評価するものも少ない。

以上の実状を踏まえ、本研究では、ロボット聴覚におけるマイクロホンアレイ技術を利用し、雑音環境でのピッチ抽出の実現を試みた。我々の研究室の人型コミュニケーションロボット「ロボビー」を使って、実環境の雑音環境で収録したデータを用いて評価を行った。

本研究では、分解能が高いMUSIC法 (MUltiple SIgnal Classification)に基づく音源定位法、指向的雑音除去の効果が優れたAdaptive-Beamfomerに基づく音源分離法、および音の歪みに強いSACF（Summary Autocorrelation Function）に基づいたピッチ抽出法を組み合わせ、ピッチ抽出を評価した。

2. ハードウェアおよび収録データ

2.1 マイクロホンアレイ

14 個のマイクロホンによるアレイを、図１に示すようにロボビーの胸部にフィットするよう作成した。著者の過去の研究[5]に用いたものと同様である。

マイクロホンアレイのオーディオ信号のキャプチャには、Tokyo Electron Device Limited の TD-BD-16ADUSB という 16 チャンネルの A/D 変換機を用いた。マイクロホンには、Sony の無指向性のコンデンサーマイク ECM-C10 を用いた。オーディオ信号は、音声認識で一般的に使用される 16 kHz/16 bit でキャプチャした。

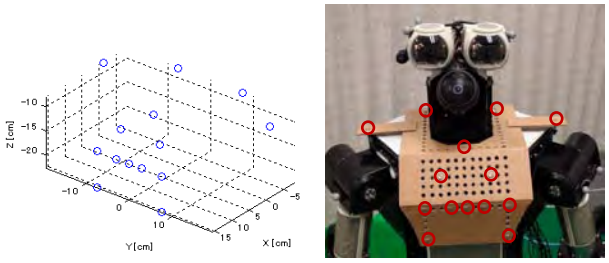


Fig. 1. (a) The geometry of the 14-element microphone array. (b) Robovie wearing the microphone array.

2.2 実験のセットアップ

マイクロホンアレイをロボビーの胸部にフィットさせた。ロボットの内部雑音も考慮させるため、

ロボットの電源は入れた状態にした。音源となる話者はロボットの周りのさまざまな方位に配置し、ロボットに向かって自然に発話するよう指示した。各音源のレファレンスとなる信号を求めるため、各話者には追加のピンマイクロホンを持たせた。これらの追加のマイクロホンから得られた信号を本稿で「音源信号」と呼ぶ。なお、これらの音源信号は、分析と評価に用いるためであり、最終的な実装には不要である。

2.3 データ収集および環境の条件

マイクロホンアレイによるデータ収録環境は、ロボビーの実証実験を行った「ユニバーサル・シティ・ウォーク大阪」という野外のショッピングモールの通路(UCW)である。UCW での主な雑音源は、天井に設置されているスピーカーから流れてくるポップ・ロックミュージックとなる。通路内のさまざまな位置およびさまざまな向きで収録を行った。30 秒のトライアルを 13 個(UCW1~UCW13 と呼ぶ)収録した。図 2 にロボットの位置とスピーカーの位置関係を示している。4 個のトライアル(UCW1~4 = “UCW-a”)で、ロボットは天井のスピーカーから(およそ 7 メートル)離れている。5 個のトライアル(UCW5~9 = “UCW-b”)で、ロボットは 1 個のスピーカーに比較的近い(およそ 4 メートル)。残り 4 個のトライアル(UCW10~13 = “UCW-c”)では、ロボットは 1 個のスピーカーの真下に位置している。

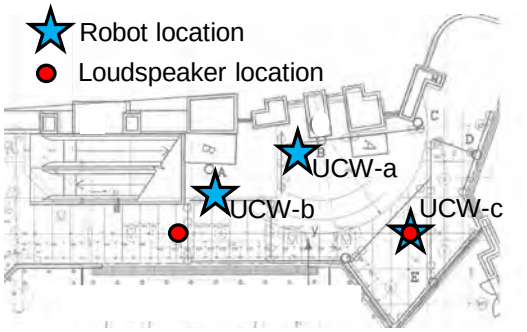


Fig. 2. A map of the UCW hallway, with locations of the robot and the ceiling loudspeakers.

ターゲット音源は 2 つ(男性話者 2 名)で、ロボットの周りにおよそ 1m 離れた位置に配置した。各トライアルにおいて、概ね最初の 10 秒間に 1 人目の話者、次の 10 秒間に 2 人目の話者、最後の 10 秒間に同時に発話するようにした。13 個のトライアルのうち、UCW7 と UCW8 では、1 個の音源がしゃべりながらロボットの周りを動いている。

2.4 ピッチの正解データの作成

話者の口元に設置したレファレンス・マイクの音を利用して、各音源のピッチの正解データを作成した。図 3 にその概要を示す。これらのマイクの音声は、SN 比が比較的高いもので、ピッチ抽出法として一般的に用いられる LPC 残差波形の自己相関関数のピーク探索による手法で、正解データを求めた。ただし、SN 比は高いとはいえ、話者が同時に発話する場合、leakage が起きてしまうため、図 3 に示すように、前処理として、時間周波数領域で、バイナリマスクにより、妨害音を抑圧した。各音源において得られた F0 の軌道を確認し、手直し後、正解データとして用いた。

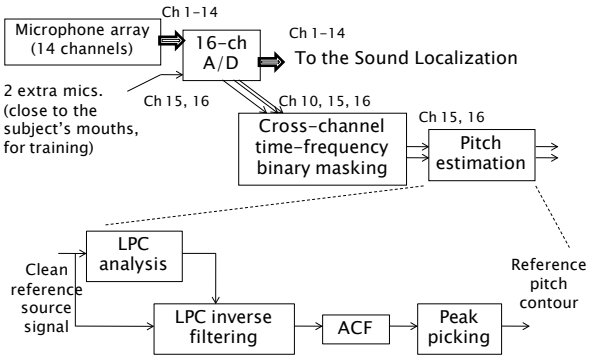


Fig. 3. Obtaining the reference pitch contours from the reference microphones.

3. 手法

図 4 に手法の概要を示す。MUSIC 法による音源定位の結果を用いて、各音源を Adaptive-Beamfomer により分離し、SACF 法によりピッチ抽出を行う。それぞれのブロックについて本節で説明する。

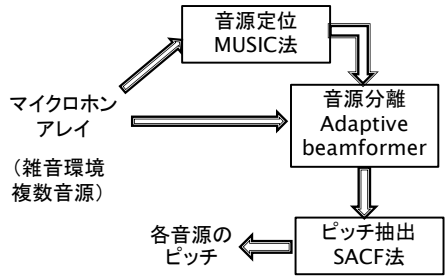


Fig. 4. Overall block diagram of the evaluated pitch extraction.

3.1 音源定位

複数の音源が存在する環境で、各音源の位置情報を得るため、定位精度の高い MUSIC (Multiple signal Classification) 法を使った[5,6]。14 チャンネルのマイクホーンアレーの入力から MUSIC spectrum を計算し、各音源の DOA(Direction Of Arrival)を推定する。

図 5 に MUSIC 法による音源定位法を示す。通常の手法との違いとして、リアルタイム処理を可能にするため、フレーム長を 4ms (FFT 点数=64) にし、雑音空間の固有ベクトルの次元を決定するため必要な音源数を固定し、MUSIC スペクトルのピーク検索に MUSIC パワーの閾値を用いている。

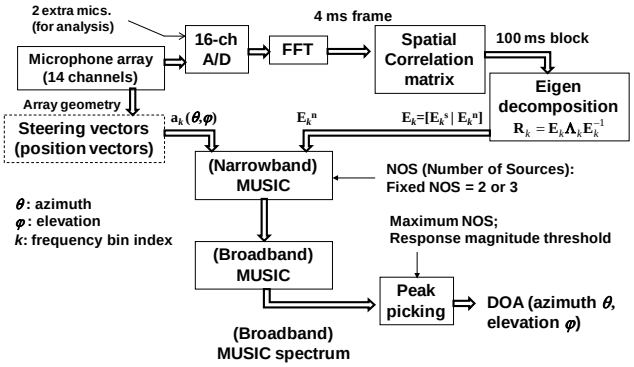


Fig. 5. The MUSIC-based sound localization algorithm, and related parameters.

3. 2 音源分離

図 2 に音源分離に用いた適応ビームフォーマーの流れを示している。MUSIC spectrum から各音源の推定 DOA 情報を利用し、空間フィルタを形成する。ターゲット音源方向にフォーカスを形成し、雑音方向にヌルを形成する[7]。フィルタを多入力にかけて、ターゲット音源の音声を分離する。

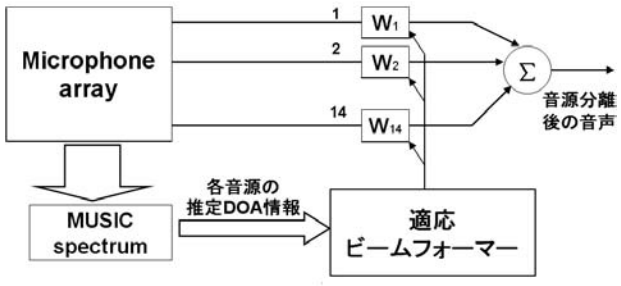


Fig. 6 Speech separation using adaptive beamformer

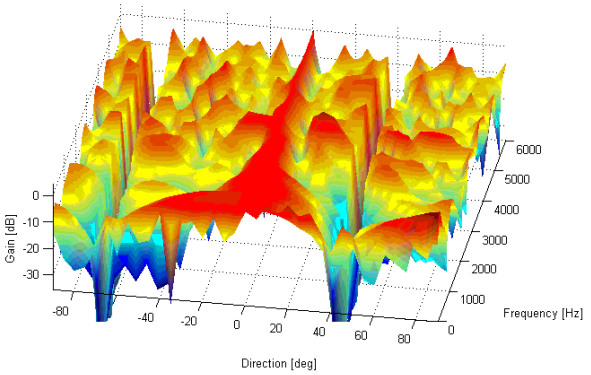


Fig. 7 Example of beamformer gain for a target source close to 0 degrees, and interference sources at 50 and -60 degrees.

3. 3 ピッチ抽出

雑音に埋もれた音声信号は、音源分離を行っても、劣化により基本周波数成分が抑圧される場合が多い。特にマイクロホンアレイの大きさが小さい程、低周波数の成分で劣化が起きてしまう。しかし、音声の有声区間（声帯が振動して発声される区間）は、声帯振動の基本周波数（F0 またはピッチ）の成分と複数の倍音（N*F0,N∈2,3,4…）から成る。本研究では、この特徴を生かした聴覚モデルに基づいた SACF 法を用いてピッチ抽出を試みた。

SACF（Summary autocorrelation function）は、図 8 に示すように、音声信号に内耳フィルターバンク(cochlear filterbank)を通し、各フィルターチャンネル出力の自己相関関数(ACF)を求め、全フィルターチャンネルの ACF を足し合わせて求める[7]。

$$\mathrm{acf}(n,c,\tau)=\sum_{k=0}^{K-1}x(n-k,c)x(n-k-\tau,c)w(k) \quad (1)$$

$$\mathrm{sacf}(n,\tau)=\sum_{c}\mathrm{acf}(n,c,\tau) \quad (2)$$

Cochlear filterbank としては、Matlab 用の Auditory Toolbox [9] の Gammatone filter を用いている。Gammatone とは、gamma 関数と tone の積から成るインパルス応答$g_{f_c}(t)$を持つ帯域通過フィルタである。

$$g_{f_c}(t)=t^{N-1}\exp[-2\pi tb(f_c)]\cos(2\pi f_c t+\phi)u(t) \quad (3)$$

ただし、高周波数に対応するチャンネルでは、チャンネル出力の振幅包絡を Hilbert transform により求めた後、自己相関関数を計算する。

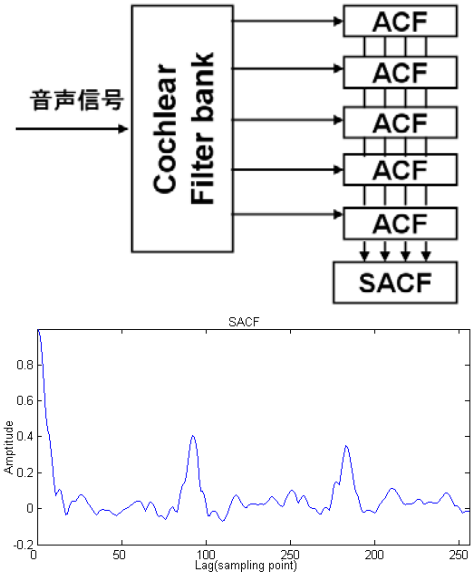


Fig. 8　Pitch extraction method based on SACF.

図 8b に SACF の例を示す。SACF が（遅延 0 を除いた）最大のピークを有する遅延が基本周期に対応し、その逆数をサンプリングレートで掛けることにより、信号のピッチ（基本周波数；F0）が推定される。

周期性を持つ信号に対して自己相関関数を取ると、周期の倍数のところにもピークが現れるため、SACF から正確に F0 を検出するため、peak pruning 手法[10]を使用した。Peak pruning の過程の例を図 9 に示す。処理としては、SACF から SACF の遅延軸で 2 倍に伸ばしたものを差し引いて PSACF が得られる。PSACF では、真の F0 に対応するラグにピークが残ることが分かる。

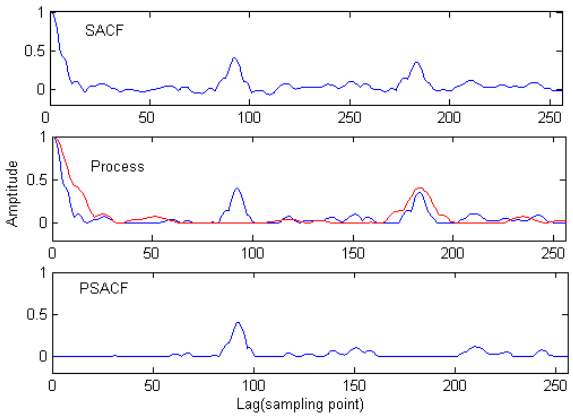


Fig. 9　Example of SACF peak pruning.

4.　実験結果と分析

4.1 評価のセットアップ

ピッチ抽出の効果を測るため、4 つの尺度を用いた。1 つ目はピッチ抽出の正解率で、正解ピッチの何パーセントを検出したかを表す。2 つ目はグロスエラー率で、正解ピッチとのずれが大きい（半音以上の）誤り率である。3 つ目は挿入誤り率で、ピッチが存在しない区間で検出した誤り率である。4 つ目は脱落誤り率で、ピッチが存在する区間で検出しなかった誤り率である。

ピッチ抽出に関しては、4 種類の手法を比較した。

a)　Raw-SACF : 音源分離なしで、シングルマイクで採ったデータに対して SACF でピッチを抽出（ベースライン）；

b)　DS-SACF : DS(Delay Sum)ビームフォーマーを用いた音源分離を施し、SACF でピッチを抽出；

c)　NULL-SACF : 妨害音に NULL を形成した適応ビームフォーマーを用いた音源分離を施し、SACF でピッチを抽出；

d)　NULL-PSACF：c)と同様の適応ビームフォーマーを用いた音源分離を施し、Peak pruning を行った SACF（PSACF）でピッチを抽出[6]。

ピッチを正解データは、2. 4 節に記述した通り、マイクロホンアレイとは別に、話者の口元で採ったリファレンスマイクのデータから求めた。

4.2 ピッチ抽出の分析

図 10 に UCW で採った 13 個の異なった収録環境において、各ピッチ抽出法のパフォーマンス（正解率、グロス誤り、挿入誤り、脱落誤り）を示している。

図 10 に示す結果より、音源分離無しの a)のピッチ抽出法の正解率と脱落誤り率が、音源分離を行った b), c), d) と比較して明らかに劣っている。b)の DS ビームフォーマーを　また、b), c), d) のうち、d)の適応ビームフォーマー+PSACF のピッチ抽出法で、最も良い正解率と低い誤り率が得られた。

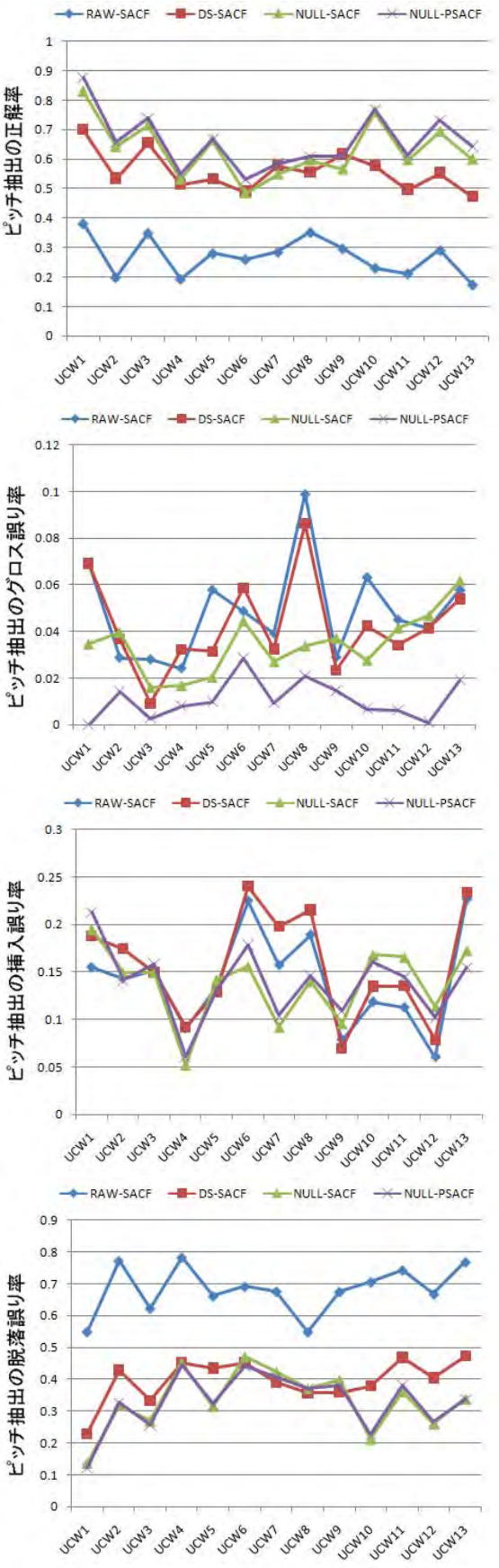


Fig. 10　Pitch extraction performance results for each trial in UCW.

5. まとめ

本研究では、マイクロホンアレイ技術を利用して、雑音環境で複数話者のピッチ抽出を試みた。

評価結果より、適応ビームフォーマーを使った音源分離は、ターゲット音源に集中する一方、雑音源の影響を抑えるため、ピッチ抽出の効果を向上した。Peak Pruning 法を使った SACF で、最も良い正解率と、低い誤り率が得られた。しかし、脱落誤りと挿入誤りは、まだ高いので、今後は、その改善に向けて誤りの原因の詳細な分析を進める予定である。

謝辞

本研究は総務省の研究委託により実施したものである

参 考 文 献

1) Ishi, C.T., Ishiguro, H., Hagita, N. (2008). Automatic extraction of paralinguistic information using prosodic features related to F0, duration and voice quality. Speech Communication 50(6), 531-543, June 2008.
2) Alain de Cheveign´e and Hideki Kawahara. Yin, a fundamental frequency estimator for speech and music. Journal of the Acoustical Society of America, 111(4), 2002.
3) Boris Doval and Xavier Rodet. Estimation of fundamental frequency of musical sound signals. In International Conference on Acoustics, Speech and Signal Processing, pages 3657-3660. IEEE, 1991.
4) Boris Doval and Xavier Rodet. Fundamental frequency estimation and tracking using maximum likelihood harmonic matching and HMMs. In International Conference on Acoustics, Speech and Signal Processing, volume I, pages 221-224. IEEE, 1993
5) Ishi, C.T., Chatot, O., Ishiguro, H., and Hagita, N. (2009). “Evaluation of a MUSIC-based real-time sound localization of multiple sound sources in real noisy environments,” Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2009), 2027-2032.
6) F. Asano, M. Goto, K. Itou, and H. Asoh, “Real-time sound source localization and separation system and its application on automatic speech recognition,” in Eurospeech 2001, Aalborg, Denmark, 2001, pp. 1013-1016.
7) F. Asano et al., “Detection and separation of speech events in meeting recordings using a microphone array,” EURASIP Journal on Audio, Speech, and Music Processing, Volume 2007, Article ID 27616, 8 pages
8) Wang, D. L. and Brown, G. J. (Eds.) (2006) Computational auditory scene analysis: Principles, algorithms and applications. IEEE Press/Wiley-Interscience
9) Webpage of Matlab auditory toolbox http://cobweb.ecn.purdue.edu/~malcolm/interval/1998-010/
10) D. P. W. Ellis, Prediction-driven computational auditory scene analysis. PhD thesis, MIT, Cambridge, Massachusetts, ISA, June 1996

社団法人　人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B002-8

ICA に基づく音声対話ロボット雑音抑圧における確率統計モデルを用いたパーミュテーション解決法

Permutation solver using probability statistics model for ICA-based noise reduction in spoken dialogue robot

† 平田将久，† 脇坂龍，†† 八田俊之，† 猿渡洋，† 鹿野清宏，††† 高谷智哉

†Nobuhisa Hirata, †Ryo Wakisaka, ††Toshiyuki Hatta, †Hiroshi Saruwatari
†Kiyohiro Shikano and †††Tomoya Takatani

† 奈良先端科学技術大学院大学　†Nara Institute of Science and Technology
†† 大阪府立工業高等専門学校　††Osaka Prefectural College of Technology
††† トヨタ自動車株式会社　†††TOYOTA MOTOR CORPORATION

nobuhisa-h@is.naist.jp

Abstract

In this paper, first, a new permutation solving method using probability statistics model is proposed for realizing high performance ICA-based noise reduction used in a spoken dialogue robot. In this method, a shape difference between probability density functions of sources can cope with the permutation problem in realistic sound mixtures consisting of point-source speech and diffuse noise. Next, to achieve high recognition accuracy for the early utterance of the target speaker, we introduce a new rapid ICA initialization method combining image information and a pre-stored initial separation filter bank. The experimental results show that the proposed approaches can remarkably improve the word recognition accuracy in the real-time ICA-based noise reduction developed in the robot dialogue system.

1　はじめに

近年，人と音声コミュニケーションができる音声対話ロボットの研究が盛んに行われている．しかし，実環境下においてロボットから離れた位置から対話ができるハンズフリー音声対話システムを実現する際，環境雑音によって音声認識率が低下するという問題点がある．従来の雑音抑圧技術として独立成分分析 (independent component analysis: ICA) [1]があるが，ICA は音声と環境雑音が混合した信号から環境雑音を推定する能力が高いことがわかっている [2]．そこで，Takahashi らはブラインド空間的サブトラクションアレー (blind spatial subtraction array: BSSA) [3]という雑音抑圧手法を提案している．BSSA は，環境雑音を含んだ観測信号から，ICA によって推定した環境雑音をスペクトル減算 (spectral subtraction: SS) [4]することで目的音抽出を行う手法であり，リアルタイム化も行われている [5]．しかしリアルタイム BSSA では，ユーザ方位情報が未知であるため，いかなる方位のユーザに対しても，ICA における分離フィルタ初期値として正面方位の死角ビームフォーマ (null beamformer: NBF) [6]を使用せざるを得ない．更に ICA にて精度良く雑音推定するには，ある程度分離フィルタの学習時間が必要である．また，ICA は信号間の独立性のみを用いて分離を行うため，分離信号における順序の不定性の問題（パーミュテーション問題）が生じる．従って，異なる周波数毎に ICA を行う周波数領域 ICA (FDICA) では，この問題が生じ，分離信号を大きく歪ませてしまう可能性がある．従って，フィルタの学習が収束するまでに入力される信号に対しては雑音抑圧性能が低く，ロボット音声対話におけるユーザの第一発話目の音声認識率が極めて低い．

上記を解決するため本研究では，ロボットにはカメラが搭載されていて，そのカメラの画像情報からユーザ方位情報を瞬時に推定できると仮定し，予め過去に学習した ICA フィルタを得られたユーザ方位情報にタグ付けをして保存することでフィルタバンクを作成し，そのフィルタバンクに存在する話者方位の ICA フィルタを初期値として使うことで，ロボット音声対話におけるユーザの第一発話目の音声認識率の向上を目指す．また，ICA におけるパーミュテーション問題解決として，音声と拡散性雑音の分離問題に対応させるため，ガンマ分布に分離信号をフィッティングさせる方法を提案する．

2　ICA を用いた目的音声抽出

2.1　ICA による雑音推定

本稿では，点音源で近似される一つの目的信号と，点音源で近似されない雑音信号がある環境を想定する．このような環境の場合，ICA は目的信号を推定するよりも，雑音信号を推定する精度のほうが高いということが明らかになっている [2]．マイクロホン数を J とすると，時間

周波数領域における観測信号は以下のように表現できる．

$$\boldsymbol{x}(f,\tau) = \boldsymbol{h}(f,\theta)s(f,\tau,\theta) + \boldsymbol{n}(f,\tau) \tag{1}$$

ここで，$\boldsymbol{x}(f,\tau) = [x_1(f,\tau),\ldots,x_J(f,\tau)]^{\mathrm{T}}$ は観測信号ベクトル，$\boldsymbol{h}(f,\theta) = [h_1(f,\theta),\ldots,h_J(f,\theta)]^{\mathrm{T}}$ は，目的音源から各マイクロホンへの伝達関数ベクトル，$s(f,\tau,\theta)$ は目的信号，$\boldsymbol{n}(f,\tau) = [n_1(f,\tau),\ldots,n_J(f,\tau)]^{\mathrm{T}}$ は加法性の雑音信号ベクトルを示す．ただし，f は周波数領域番号，τ は分析フレーム番号，θ は画像情報に基づいて推定された目的信号方位を表す．FDICA では，観測信号を以下の式に基づいて分離を行う．

$$\boldsymbol{o}(f,\tau,\theta) = \boldsymbol{W}_{\mathrm{ICA}}(f,\theta)\boldsymbol{x}(f,\tau) \tag{2}$$

$$\boldsymbol{o}(f,\tau,\theta) = [o_1(f,\tau,\theta),\ldots,o_K(f,\tau,\theta)]^{\mathrm{T}} \tag{3}$$

ここで $\boldsymbol{o}(f,\tau,\theta)$ は分離信号ベクトル，K は出力音源数，$\boldsymbol{W}_{\mathrm{ICA}}(f,\theta)$ は θ 方位の信号をキャンセルするための分離行列である．分離行列は以下の更新式に基づいて反復的に求められる．

$$\boldsymbol{W}_{\mathrm{ICA}}^{[p+1]}(f,\theta) = \mu[\boldsymbol{I} - \langle\boldsymbol{\varphi}(\boldsymbol{o}(f,\tau,\theta))\boldsymbol{o}^{\mathrm{H}}(f,\tau,\theta)\rangle_\tau]\boldsymbol{W}_{\mathrm{ICA}}^{[p]}(f,\theta) + \boldsymbol{W}_{\mathrm{ICA}}^{[p]}(f,\theta) \tag{4}$$

ここで p は反復回数，μ はステップサイズ，$\boldsymbol{M}^{\mathrm{H}}$ は行列 $\boldsymbol{M}$ の複素共役転置，$\langle\rangle_\tau$ は時間平均，$\phi(\cdot)$ は非線形関数ベクトルを表す．雑音推定を行うため，分離信号ベクトルから，目的音推定信号 $o_U(f,\tau,\theta)$ を以下のように取り除いた信号ベクトル $\boldsymbol{q}(f,\tau,\theta)$ を得る．

$$\boldsymbol{q}(f,\tau,\theta) = [o_1(f,\tau,\theta),\ldots,o_{U-1}(f,\tau,\theta),0, o_{U+1}(f,\tau,\theta),\ldots,o_K(f,\tau,\theta)]^{\mathrm{T}} \tag{5}$$

次に射影法によって，利得の正規化を行う．この処理は以下の式によって与えられる．

$$\hat{\boldsymbol{q}}(f,\tau,\theta) = [\hat{q}_1(f,\tau,\theta),\ldots,\hat{q}_J(f,\tau,\theta)]^{\mathrm{T}} \tag{6}$$

$$= \boldsymbol{W}_{\mathrm{ICA}}^{+}(f,\theta)\boldsymbol{q}(f,\tau,\theta) \tag{7}$$

ここで，$\boldsymbol{M}^{+}$ は行列 $\boldsymbol{M}$ の Moore-Penrose 型一般逆行列を表す．ICA では，信号間の独立性のみを用いて分離を行うため，分離信号における順序の不定性の問題（パーミュテーション問題）が生じる．従って，異なる周波数毎に ICA を行う FDICA では，この問題が生じ，分離信号を大きく歪ませてしまう可能性がある．

2.2 目的音声抽出

目的音声抽出におけるポスト処理として，本研究では Wiener filter (WF) [10]を使用する．ICA によって推定した雑音信号を用いて，以下のように各チャネル毎に WF のゲイン係数を得る．

$$g_j(f,\tau,\theta) = \frac{|x_j(f,\tau,\theta)|^2}{|x_j(f,\tau,\theta)|^2 + \beta|\hat{q}_j(f,\tau,\theta)|^2} \tag{8}$$

ここで $g_j(f,\tau,\theta)$ は j チャネルにおけるゲイン係数，β は雑音抑圧の処理強度パラメータを表す．最終的に，各チャネル毎にゲイン係数 $g_j(f,\tau,\theta)$ をマイクロホンの観測信号に適用することで，以下のように推定目的信号を得る．

$$s_j^{(\mathrm{WF})}(f,\tau,\theta) = \sqrt{g_j(f,\tau,\theta)|x_j(f,\tau,\theta)|^2}\,\frac{x_j(f,\tau,\theta)}{|x_j(f,\tau,\theta)|} \tag{9}$$

ここで，$s_j^{(\mathrm{WF})}(f,\tau,\theta)$ は j チャネルにおける推定目的音声信号を表す．最後に，WF によって得られた各チャネル毎の推定目的音声信号に対して，遅延話法 (delay and sum: DS) により目的音声強調を行い，最終出力音声信号を得る．

$$s_{\mathrm{DS}}(f,\tau) = \boldsymbol{w}_{\mathrm{DS}}(f,\theta)^{\mathrm{T}}[s_1^{(\mathrm{WF})}(f,\tau,\theta),\ldots,s_J^{(\mathrm{WF})}(f,\tau,\theta)]^{\mathrm{T}} \tag{10}$$

$$\boldsymbol{w}_{\mathrm{DS}}(f,\theta) = [w_1^{(\mathrm{DS})}(f,\theta),\ldots,w_J^{(\mathrm{DS})}(f,\theta)]^{\mathrm{T}} \tag{11}$$

$$w_j^{(\mathrm{DS})}(f,\theta) = \frac{1}{J}\exp\left(-i2\pi(f/N)f_{\mathrm{s}}d_j\sin\theta/c\right) \tag{12}$$

ここで $s_{\mathrm{DS}}(f,\tau)$ は最終出力音声信号，$\boldsymbol{w}_{\mathrm{DS}}(f,\theta)$ は DS のフィルタ係数ベクトル，θ は DS の目的音声方位を表し，ロボットのカメラの画像情報から得られるユーザ方位である．ここで f_{s} はサンプリング周波数，d_j $(j = 1,\ldots,J)$ はマイクロホン位置，N は DFT 長，c は音速を表す．

3 提案法 1: ガンマ分布に基づくパーミュテーション解決

パーミュテーション問題の解決法として様々な提案がなされている [6] [7] [8]．これらは点音源である音声と音声の分離問題においては有効な解決法であるが，音声と拡散性の環境雑音の分離問題においては，うまく機能しない．そこで本研究では，ICA におけるパーミュテーション問題を，ガンマ分布に信号をモデリングすることで解決する方法を提案する．本手法は分離信号の確率統計量を用いるので音声と拡散性雑音の分離問題にも有効であると考えられる．ガンマ分布は，一般に，パワースペクトル領域の音声信号や実環境の雑音信号を表現可能であると言われている [9]．また，ガンマ関数に基づく分布であるので，数学的に有用な性質が多く，高次統計量を表現する目的にも利用しやすい特徴を持つ．ガンマ分布の確率密度関数 (PDF) は，

$$P(x) = \frac{1}{\Gamma(\alpha)\theta^{\alpha}}x^{\alpha-1}e^{-\frac{x}{\theta}} \tag{13}$$

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1}e^{-x}dx \tag{14}$$

と表せる．ここで，$x \geq 0$ は信号のパワースペクトル系列であり，$\alpha > 0$ かつ $\theta > 0$ である．また，α は形状母

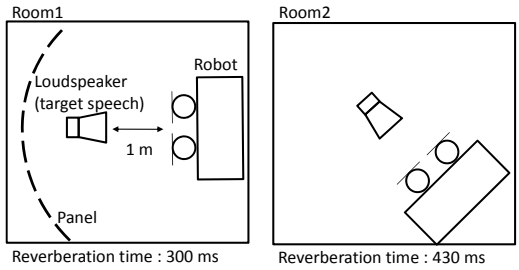


Figure 1: Layout of two reverberant rooms used in our simulation.

数，θ は尺度母数，$\Gamma(\alpha)$ はガンマ関数である．$\alpha = 1$ の場合，式 (13) は指数分布と一致することが知られており，これはガウス性信号のパワースペクトルに対応する．また，$0 < \alpha < 1$ の場合は，優ガウス性信号であることを示す．ガンマ分布の平均値は，以下の式で表現できる．

$$\mathrm{E}[P(x)] = \alpha\theta \tag{15}$$

ここで $\mathrm{E}[\cdot]$ は期待値演算子である．ガンマ分布によるモデリングは，観測される生のデータサンプルから，形状母数 α と尺度母数 θ を推定することで行われる．これらの母数は，以下のように最尤推定法に基づき推定される．

$$\hat{\alpha} = \frac{3 - \gamma + \sqrt{(\gamma - 3)^2 + 24\gamma}}{12\gamma} \tag{16}$$

$$\hat{\theta} = \frac{\mathrm{E}[x]}{\hat{\alpha}} \tag{17}$$

ここで $\gamma = \log(\mathrm{E}[x]) - \mathrm{E}[\log x]$ である．推定された $\hat{\alpha}$ の値が小さい程，優ガウス性が高い分布形状となり，$\hat{\alpha} = 1$ のとき，ガウス性の分布形状となる．一般に音声の PDF は優ガウス性の分布形状であり，拡散性雑音の PDF はガウス性の分布形状であると言われている．よって，それぞれの分離信号を用いて $\hat{\alpha}$ の値を求め，その大小を比較することによってパーミュテーション問題を解決する．

以上の手法の有効性を確認するために，予備実験として ICA による音源分離実験を行った．Fig.1 の Room2 で収録したインパルス応答をクリーン音声データベースに畳み込み，入力 SNR が 10 dB となるように拡散性雑音を付加した．拡散性雑音は，実収録による人ごみ雑音を用いた．マイクロホンアレーには 2 素子を用い，音声とマイクロホンアレーの距離は 1.0 m とした．比較手法として以下の 4 つを評価した．

- パーミュテーション問題未解決 (unprocessed).
- 方位特性に基づく解決法 (DOA-based) [6].
- ガンマ分布に基づく解決法 (proposed).
- 真の分離信号を用いる理想的な解決 (ideal).

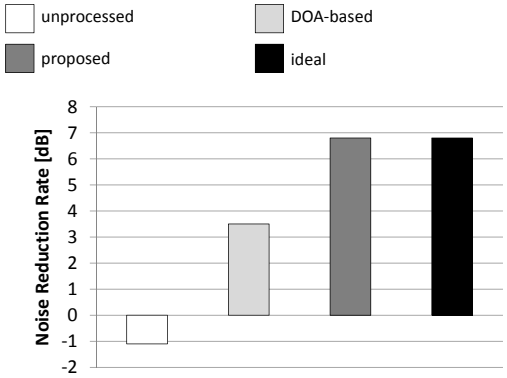


Figure 2: result of preliminary experiment 1.

また，分離性能の評価には雑音抑圧量 (noise reduction rate: NRR) を用いた．NRR は値が高いほど良い性能を示す．実験結果を Fig 2 に示す．この結果より，本稿のような音声と拡散性雑音の分離問題におけるパーミュテーション問題の解決法としては，提案法が有効であることが確認できる．

4 提案法 2: 画像情報に基づく第一発話処理

4.1 概要

本研究では，音声対話ロボットにはカメラが搭載されており，カメラから得られる画像情報からユーザ方位を瞬時に推定できると仮定する．予め過去に学習した ICA フィルタをユーザ方位にタグ付けをして保存することでフィルタバンクを作成し，そのフィルタバンクに存在する話者方位の ICA フィルタを初期値として使うことで雑音を推定する．さらに推定した雑音を用いて，マイクロホンアレーで観測した信号に WF を適用することで目的音抽出を行い，最後に DS によって目的音声強調をする手法を提案する．提案法における ICA のパーミュテーション解決法としては，ガンマ分布に基づく解決法を用いる．これらの処理によって得られた目的音声を Julius [11]によって機械音声認識することで，提案法の有効性を示す．アルゴリズムの詳細を以下で述べる．

4.2 予備実験

ここで，ある部屋で予め ϕ 方位のユーザに対して学習を行い $\boldsymbol{W}_{\mathrm{ICA}}(f, \phi)$ を保存し，それとは別の部屋で ϕ 方位のユーザに対して保存した $\boldsymbol{W}_{\mathrm{ICA}}(f, \phi)$ を用いて目的音声抽出を行う場合を考える．ICA は信号間の独立性のみを用いて分離学習を行うが，結果的には部屋の残響特性を含めたユーザ方位の音を抑圧するフィルタを学習している．よって，部屋の残響特性が変化すると過去に学習した ICA フィルタ $\boldsymbol{W}_{\mathrm{ICA}}(f, \phi)$ を用いても精度良く雑音推定できるとは限らない．ロボット音声対話を想定した場合，ユーザ方位から到来する音波は，順に直接波，ロボット本体によ

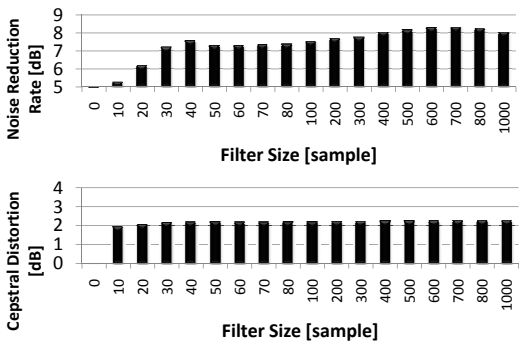


Figure 3: result of preliminary experiment 2.

る回折波，初期反射波，部屋の伝達特性による残響波であるが，部屋が変わることによって伝達特性に変化があるのは，初期反射波以降の部屋の伝達特性による波である．よって，この成分を抑圧する ICA フィルタの一部分は再度 ICA で学習させなければならないが，保存したフィルタをそのまま使って学習させるほうが良いか，初期反射以降の部屋の伝達特性を抑圧する部分を無くすために ICA フィルタを短くして学習させるほうが良いか，またその場合フィルタの長さはどれくらいが良いかを確認しなければならない．そこで最適な ICA フィルタの長さを確認するために予備実験を行った．まず，0 °方位の話者に対して ICA で学習を行い，そのフィルタを長さをそれぞれ操作したものを保存する．そしてそれとは別の残響特性のインパルス応答を用いて観測信号にそれぞれの長さの ICA フィルタを学習させずに適用させ，WF によって目的音声抽出を行う．ICA フィルタ長は最大で 1024 サンプルである．実験条件としては前節の予備実験と同じで，評価値には NRR 及びケプストラム歪み (cepstral distortion: CD) を用いた．CD は値が小さいほど良い性能を示す．この予備実験の結果を Fig. 3 に示す．一定のフィルタ長までは性能が上がっていくが，それ以降は変化がないことが分かる．この結果より，部屋の残響特性の変化による雑音抑圧性能は ICA フィルタの長さによらないということが言える．よって以降，本稿では ICA フィルタを保存する際，フィルタ長を操作しない方法を用いる．

4.3 リアルタイム処理のフロー

本手法では，以下のステップでリアルタイム音声強調を行う．処理フローを Fig. 4 に示す．

Step 1 事前フィルタバンクの構成

本手法では，画像情報から得られたユーザ方位を，正面方位を 0 度とし，-90 度から 90 度まで 15 度間隔で区切った 13 方位のうち，最も近い方位を θ とし，処理に使用する．13 方位全てにおいて，過去に十分学習を行ったユーザ方位 θ に関する ICA の分離行列フィルタ $\boldsymbol{W}_{\mathrm{ICA}}(f,\theta)$ を，3 節で述べたパーミュテーション解決法でパーミュテーション解決を行い，フィルタバンクに保存する．

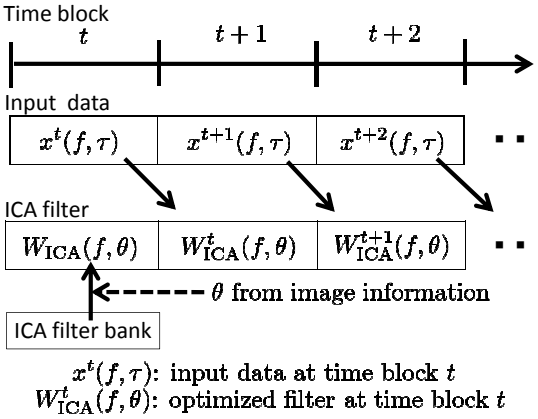


Figure 4: Signal flow of updating ICA filter in real-time simulation.

Step 2 画像による方位推定

ユーザがカメラに映ったときに，ユーザ方位推定を行い θ を取得する．

Step 3 第一発話目に対する処理

Step 2 で取得した θ に基づき，フィルタバンクから $\boldsymbol{W}_{\mathrm{ICA}}(f,\theta)$ を読み込む．それを用いて，現在の入力データブロックに対して雑音推定を行い，推定雑音を用いて入力データに対して各チャネル毎に WF を適用させて雑音抑圧処理を行い，強調音声を出力する．最後に DS をすることで目的音声強調を行い，最終出力音声信号を得る．

Step 4 第二発話目以降に対する処理

第二発話目以降については，Fig. 4 のように入力信号を時間ブロックに分け，各ブロックで ICA のフィルタを学習させ，更新していく．このときのパーミュテーション問題についても 3 節で述べた手法でパーミュテーション解決を行う．雑音抑圧処理は Step 3 と同様に行う．

Step 5 ユーザ発話終了後の処理

ユーザの発話が終了したときは，$\boldsymbol{W}_{\mathrm{ICA}}(f,\theta)$ を未来のユーザに対する第一発話目用のフィルタとして，フィルタバンクに上書きをする．ここでも 3 節で述べた手法でパーミュテーション解決を行う．その後，Step 1 へ戻る．

5 音声認識実験

5.1 実験条件

提案法の有効性を確認するため，Julius を用いて機械音声認識を行った．実験は以下に示す 3 つの手法を用いて行った．

- 正面方位の NBF を初期値として雑音推定を行った従来法 (Conventional).
- 画像により取得されたユーザ方位の NBF を初期値として雑音推定を行った手法 (Supervised).
- 提案法 (Proposed).

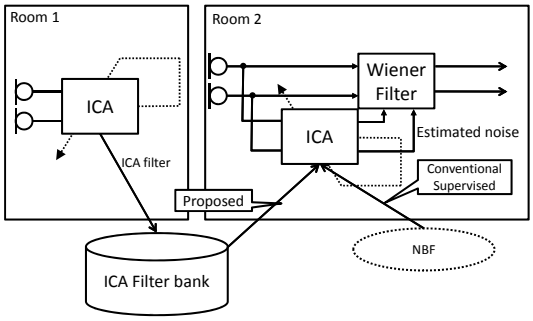


Figure 5: Block diagram of speech recognition experiment.

Table 1: 音声認識実験の条件

テストデータ	JNAS テストセット (男女話者による 200 文)
音声認識タスク	新聞記事読み上げ (語彙数: 20 k)
音響モデル	音素内タイドミクスチャーモデル (phonetic-tied mixture model: PTM) に基づく 25 dB オフィス雑音重畳モデル
音響モデルの学習データ	JNAS 260 話者 (1 話者あたり 150 文)
認識デコーダ	Julius ver. 3.4.2

提案法の処理ブロック図を Fig. 5 に示す．提案法における ICA フィルタのデータベース作成は Fig. 1 の Room1 で行い，実験は Room2 で行う．Room1 と Room2 では，パネルを用いて残響時間を変え，ロボットの位置も変えた．Room1，Room2 のいずれにおいても，各音響環境で収録したインパルス応答を JNAS のクリーン音声に畳み込んだ信号を目的音声信号とした．この信号に対して SNR が 10 dB となるように，実収録の駅環境雑音を付加した．マイクロホンアレーの素子数は 2 個で，SHURE 製の指向性マイクロホン MX-184 を使用した．実験における WF の処理強度パラメータは，いずれの手法においても音声認識精度を基にして最適な値を選んだ．ICA フィルタ学習時の時間ブロック長は 3 s とし，学習回数は 100 回とした．音声認識に実験の条件を表 1 に示す．なお，Conventional 及び Supervised におけるパーミュテーション解決法としては，方位特性に基づく解決法を用いた [6].

5.2 実験結果

Fig. 6 に音声認識結果を示す．この結果より，いずれの話者方位でも過去に学習した ICA フィルタを初期値にし，パーミュテーション解決法としてガンマ分布に基づく解決法を用いた提案法のほうが音声認識率が改善されていることがわかる．Conventional と Supervised の認識率がそれほど変わらない原因は，マイクロホンアレーに指向性マイクロホンを使用したこと及び目的信号のロボット本体による回折波成分が大きいため，NBF の死角が正しく形成されていないことが原因であると考えられる．従って，提案法はロボット音声対話におけるユーザの第一発話目の認識率向上に有効であると言える．

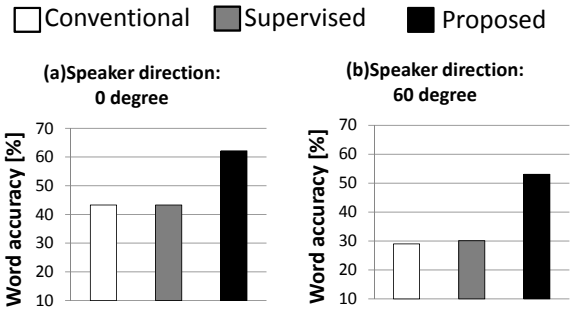


Figure 6: Word accuracy of different speaker directions; (a) 0 degrees and (b) 60 degrees.

6 まとめ

本稿では，音声対話ロボットにおいてユーザ方位が瞬時に推定できたときの，リアルタイムを想定した雑音抑圧におけるパーミュテーション解決法を提案した．音声認識実験により提案法の有効性を確認した．今後はさらなる音声認識率の向上を目指す．

謝辞 本研究の一部は総務省・戦略的情報通信研究開発推進制度 (SCOPE) の支援を受けた．

参考文献

[1] P. Comon, "Independent component analysis, a new concept", *Signal processing,* vol. 36, pp. 287–314, 1994.

[2] Y. Takahashi, et al., "Blind source extraction for hands-free speech recognition based on wiener filtering and ICA-based noise estimation," *Proc. HSCMA*, 2008.

[3] Y. Takahashi, et al., "Blind spatial subtraction array for noisy environment," *IEEE Trans. Audio, Speech, and Language Processing*, vol.17, no.4, pp.650–664, 2009.

[4] S. F. Boll, *IEEE Trans. Acoustics, Speech, Signal Proc,* vol. ASSP-27, no. 2, pp. 113–120, 1979.

[5] 高橋祐，他 "独立成分分析を導入した空間的サブトラクションアレーによるハンズフリー音声認識システムの開発," 電子情報通信学会論文誌 D, vol.J93-D, no.3, pp.312–325, 2010.

[6] H. Saruwatari, S. Kiumura, K Takeda, F. Itakura, and T. Nishikawa. "Blind source separation combining independent component analysis and beamform-

ing." *EURASIP Journal on Applied Signal Processing,* vol. 2003, pp. 1135–1146, 2003.

[7] N. Murata, S. Ikeda, and A. Ziehe. "An approach to blind source separation based on temporal structure of speech signal." *Neurocomputing,* vol. 41, no. 1-4, pp. 1–24, 2001.

[8] H. Sawada, R. Mukai, S. Araki, and S. Mikano. "A robust and precise method for solving the permutation problem of frequency-domain blind source separation." *IEEE Transactions on Speech and Audio Processing,* vol.12, no. 5, pp. 530–538, 2004.

[9] T. Inoue, Y. Takahashi, H. Saruwatari, K. Shikano and K. Kondo, "Theoretical analysis of musical noise in generalized spectral subtraction: why should not use power/amplitude subtraction?," *Proc. EUSIPCO European Signal Processing Conference*, pp. 994–998, 2010.

[10] P. Loizou, *Speech Enhancement: Theory and Practice,* CRC Press, 2007.

[11] A. Lee, et al., "Julius An open source realtime large vocabulary recognition engine," *Proc. Eur. Conf. Speech Commun. Technol.*, pp.1691–1694, 2001.

社団法人　人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B002-9

能動人工耳介

Active artificial pinnae

公文誠, 野田佳孝, 魚住守治
Makoto KUMON, Yoshitaka NODA and Shuji UOZUMI
熊本大学
Kumamoto University
kumon@gpo.kumamoto-u.ac.jp

Abstract

This paper proposes an active artificial pinna that is able to change the form imitating animals do in order to localize the sound source. The shape of the proposed pinna provides directivity to the sound source that locates in front of the pinna, and it has an ability to steer the direction by the active motion. A kinematic model to control the pinna is also proposed in this paper. In order to clarify the characteristics of the proposed pinna with respect to auditory functions, its directivity and the effect on the frequency response by the deformation were studied with the developed device.

1　はじめに

音源の位置や方向を認識する音源定位と呼ばれる機能は, 音情報を利用して環境を認識する上で基本的な情報を与える重要な聴覚機能の一つである. 人間や動物の音源定位では, 特徴量の一部に到来音に対する身体の伝達特性を利用していると言われるが, このような特徴量は特に耳に近い上体の寄与が大きい. 例えば定位感のある音再生を目指すなどの応用が考えられ, これらの影響を頭部伝達関数としてモデル化する様々な研究が精力的に行われている. 耳介は頭部伝達関数に寄与する身体器官の一つであり, 前方方向への指向性を高めるとともに, 音源方向に対する方向依存性フィルタとしての働きがある. 人間の場合, 耳介の影響は周波数領域においてノッチ周波数が音源の上下方向の関数となっていることが知られている[Shaw, 1968]. 耳が二つに限られる場合, 両耳間時間差のように両耳間の音信号の差に基づく特徴量では, 両耳から等距離にある正中面内の音源を区別することが出来ないことが多く, 耳介ノッチの周波数は音源定位にとって重要な特徴量の一つであると考えらえる.

この原理を利用し, 2 つのマイクロホンのみで構成されるバイノーラル聴覚において音源定位能をロボットで実現する試みが報告されている. 下田ら[Shimoda, 2006]は, 十分な周波数成分を含む音信号に対し, 複数の周波数帯域での耳介ノッチ周波数モデルを用いて音源方向を与える逆モデルを求め, フィードバック系を構成することで音源にロボット頭部を制御するサーボ系を提案した. Hörnstein[Hörnstein, 2006]は両耳の耳介周波数特性の差に基づいて, 音源の上下の情報を得て頭部を制御している. Finger[Finger, 2010]は, 両耳間周波数特性差が音源水平方向の関数となることを示し, 同一の耳介周波数特性を持つ左右の耳介を用いて上下方向の音源定位が可能であることを示している. このような特性は耳介形状に依存するため, 著者らは耳介形状について, 耳介ノッチを所望のものに近づける方法について検討している[Kumon, 2009].

ところで, 人間や猫などは頭部を動かすことで音源定位能が向上することが知られている. 例えば, 猫頭部を固定すると水平方向の音源定位における分解能が劣化することが報告されて[Populin, 1998]おり, 身体動作を伴う能動的な音源定位が重要であることが示唆される. ロボットによる音源定位においても, 同様に能動的な作用を考えることは有用であると考えられ, 実際, 佐々木ら[Sasaki, 2009]は移動ロボットによってマイクロホンアレイとロボットの移動による三角測量を基礎とした音源定位手法を提案している. 戸嶋ら[Toshima, 2006]は能動的なダミーヘッドによって, テレオペレーションにおける操作者の定位能向上に成功している.

しかしながら, このような身体動作は, エゴノイズと呼ばれるロボット自身の駆動に伴う騒音や, 身体動作によるマイクの配位変化に伴う集音環境の大きな変化を生じるなど, 特有の難しさを生む可能性もある. そこで, 本研究では頭部の一部だけが動作することで, エゴノイズの影響

を抑制し, 音響特性の変動を限定的にすることを考える. 具体的には猫や犬の耳介のように前方に指向性のある形状を考え, 頭部は動かずにこの耳介の形状のみが変化する機構を考察した. 指向性によって, 対象音信号を選択的に得ることが出来る一方, 耳介を対象音方向に向けることで, 様々な方向の到来音の受聴も可能になると期待されるので, 動きのない耳介に比べて優位性があると考えられる. しかし, 形状変化に伴う周波数特性の変化は複雑になる可能性が予想される. 原理上は数値的にこの特性を求めることが可能であるが, 耳介形状の正確なデータを得ることや耳介表面での反射を厳密にモデル化することは難しいため, 現実的には実際の装置に基づく検証が必要である. このような観点から, 本稿では, 能動的に形状を変化可能な耳介を実際に制作し, 計測によってその特性を明かにすることを目的とする.

本稿の構成は以下の通りである. 第 2 節で耳介ノッチによる音源定位について説明し, その後, 実際に制作した能動人工耳介の構造および機構を考察する (第 3 節). また, 聴覚における耳介の特性として周波数特性と指向性を調べたので, これを第 4 節で説明し, 提案耳介の性質を明らかにし, 最後にまとめを述べる (第 5 節).

2 耳介

2.1 耳介ノッチ

耳介は古くから「 集音器」として前方からの到来音に対して指向性を与える器官として認識されてきた[Batteau, 1967]. これに加え, 耳介表面が複雑な凹凸形状を持つため, 耳への入射音が反射や回折することで, 耳介は複雑な周波数特性を持つフィルタとして作用することが知られており, 音源位置推定に効果があると指摘されている. 特に耳介の周波数特性には音源の方向に応じてゲインが鋭く低減する周波数帯が存在し, この特徴は耳介ノッチ (Pinna Notch) と呼ばれている. 人間の耳介の場合, およそ 4kHz より高い周波数帯域で耳介ノッチ認められている[Butler, 1984]. 実際に, Shaw[Shaw, 1968]らは人間の耳介における周波数特性を測定し, 耳介ノッチの周波数が音源方向の関数になっていること, この周波数を計測することが出来れば, 音源方向を求めることが出来る可能性があることを示した. 耳介ノッチの生じるメカニズムについて, Lopez-Poveda[Lopez-Poveda, 1996] は回折と反射を考慮して, 簡単な一次反射の音波の加え合わせによる物理モデルを提案し, 簡略化された形状の耳介モデルにおいて, 耳介ノッチの周波数がモデルによるものと良く一致することを示している. ロボットで耳介ノッチの利用を目的とした研究に[Hörnstein, 2006]や[Shimoda, 2006]がある. 一例として Fig.1 に Shimoda ら [Shimoda, 2006]の用いた耳介での周波数応答の例を示す.

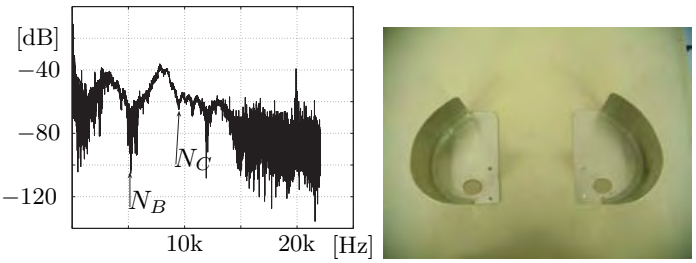


Figure 1: 耳介ノッチの例 (Shimoda らによる [Shimoda, 2006])
周波数応答 (左図) およびロボット用耳介 (右図)

2.2 動物の耳介

人間の耳介は耳介筋の発達が限定的なので, 通常はほとんど動かせないが, 他の多くの動物は耳介の向きを随意的に動かすことが出来, 選択的な集音効果を得ていると言われている. これらの動物の耳介そのものは皮膚と軟骨等から成り, 耳介の運動は頭部の複数の耳介筋が複合的に作用することで生じる. 猫の場合, 頭部および首からの筋肉によって[Ellsworth, 1902] ほぼ 180 度にわたって耳介の向きを変えることが可能で, その形状から開口部方向に指向性がある [Rice, 1992]. 動物が音源定位に際し, 耳介の向きを変えていることが Populin[Populin, 1998] によって報告されている.

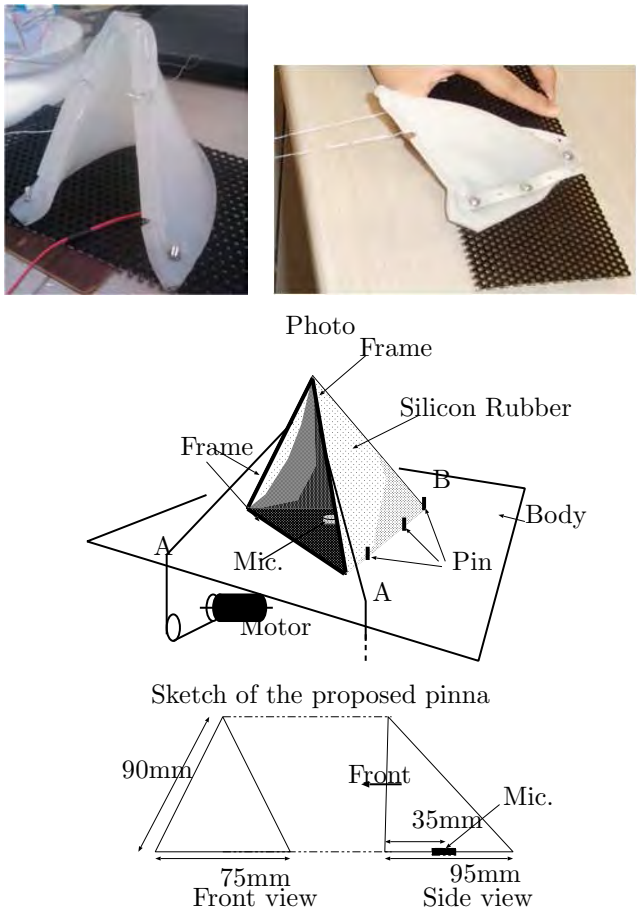
そこで, このような耳介をロボットで実現することを考え, 以下ではその設計と動作, 特性について考える.

3 能動人工耳介

本稿では, 動物にならって方向と形状を変化可能な耳介を提案する. ここで, 耳介がマイクロホンを覆うように設置されることから, 耳介そのものの駆動機構は静穏性の高い必要があるため, 耳介そのものは受動的な機構とし, ワイヤで耳介を牽引する方式とする. これにより, 騒音源となるモータとマイクロホンを十分に離れた配置とすることが出来る.

3.1 構造

実際に制作した能動耳介の外観とその概略を Fig.2 に示す. 皮膚に相当する部分はシリコンゴムで製作し, 開口部にアクリルの骨材が三角形の形状を成すよう埋め込まれており, 耳介形状を支持している. 上部の骨材の左右それぞれには駆動用ワイヤが結び付けられ, これらのワイヤは耳介の前方の取り付け穴 (図中 A 点) を経て, プーリを介して耳介下部に設置されたモータで駆動される. シリコンゴムに弾性があるため, モータはワイヤを牽引する方向にのみ駆動力を発揮すれば十分である. 予備実験においてワイヤを牽引した際, 最大の変位を得る時に約 1kgf の力で前方に引く必要があった (Fig.2 最上段右図). シリコンゴ



Size of the pinna (front and side view)

Figure 2: Active artificial pinna: photo and sketch

ムは耳介を固定する板にピン止めされており, 耳介後端の固定点 (図中 B 点) を中心に耳介を固定する板が水平面内で受動回転する構造になっている.

ワイヤは適当に減速されたステッピングモータ (オリエンタルモータ社) からのトルクで駆動され, モータの駆動信号はドライバから指令パルス信号として与える. モータの回転角は二相パルス信号を制御用プロセッサ (SH7145) でカウントすることで求める. マイクロホンからの音信号と制御用プロセッサからの信号はホストとする計算機 (PC) で処理され, 制御用プロセッサへ指令パルス値へとフィードバックされる構成とする (Fig.3).

3.2 運動学モデル

提案する耳介はワイヤ長を制御することで, 水平回転および前後方向の 2 自由度の運動が可能である. ここでは耳介の参照点を耳介頂点に取り, ワイヤ長と参照点間の運動学的関係を導く.

今, 原点が Fig.2 の B 点に対応し, X 軸を前方, Z 軸を上方に取る右手系として Fig.4(a) に示す座標系を考える. 耳介の下端固定部を Q, Q', 耳介とワイヤの取り付け点を R, R' とし, ワイヤは S, S' を通ってモータに接続してい

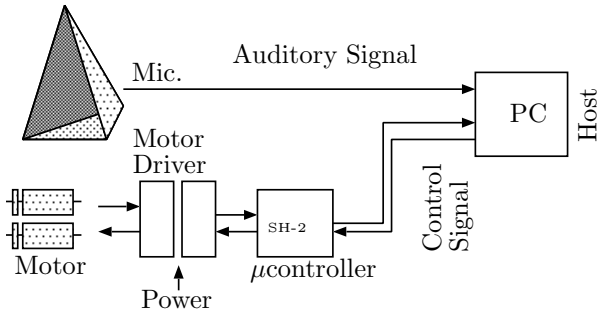
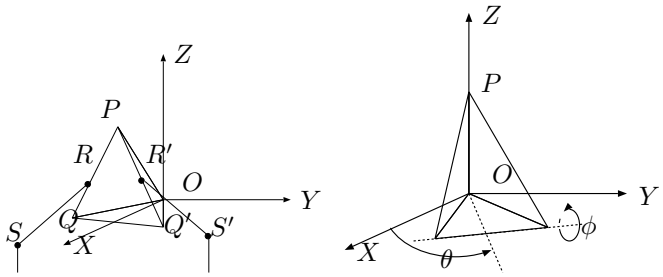


Figure 3: Control system



(a)Origin $(\theta, \phi) = (0, 0)$ (b) Definition of (θ, ϕ)

Figure 4: Kinematic model

るものとする. 耳介の姿勢を Z 軸まわりの回転角 θ と Y 軸に平行で耳介開口部を通る $X-Y$ 平面内の軸まわりの角度 ϕ で表わすこととする (Fif.4(b)). 耳介頂点の参照点を P と表し, $\theta = \phi = 0$ の時の P の座標を $(r, 0, a)^T$ と記述する. 姿勢 (θ, ϕ) の時 P の座標は

$$
\begin{aligned}
P &= R_z(\theta)\left[R_Y(\phi)\left\{\begin{pmatrix} r \\ 0 \\ a \end{pmatrix} - \begin{pmatrix} r \\ 0 \\ 0 \end{pmatrix}\right\} + \begin{pmatrix} r \\ 0 \\ 0 \end{pmatrix}\right] \\
&= \begin{pmatrix} -a\cos\theta\sin\phi + r\cos\theta \\ a\sin\phi\sin\theta - r\sin\theta \\ a\cos\phi \end{pmatrix} \qquad (1)
\end{aligned}
$$

である. ここで R_Z, R_Y はそれぞれ Z 軸と Y 軸まわりの回転変換を表す.

さて, $(\theta, \phi) = (0, 0)$ における点 Q の座標を $(r, d, 0)^T$ とすると, Z 軸まわりの回転によって $(r\cos\theta + d\sin\theta, -r\sin\theta + d\cos\theta, 0)^T$ へと移される. PQ の長さは $\sqrt{a^2+d^2}$ で, QR の長さを b とすると, R の座標は

$$
\begin{aligned}
R &= OQ + \frac{b}{\sqrt{a^2+d^2}}QP \\
&= \begin{pmatrix} \cos\theta & \sin\theta & -\cos\theta\sin\phi \\ -\sin\theta & \cos\theta & \sin\theta\sin\phi \\ 0 & 0 & \cos\phi \end{pmatrix} \begin{pmatrix} r \\ d - \frac{bd}{\sqrt{a^2+d^2}} \\ -\frac{ab}{\sqrt{a^2+d^2}} \end{pmatrix}
\end{aligned}
$$

となる.

S から R へのベクトルを l とすると

$$
l^2 = (\sin\phi, \cos\theta, \sin\theta, \sin\phi\sin\theta, \sin\phi\cos\theta)c + c_0 \qquad (2)
$$

と書ける. ここで c, c_0 は定数ベクトルと定数を表す. R' についても同様に (2) に相当する関係を得ることが出来るので, l, l' の長さが与えられた時, このモデルから得られる関係を連立して ϕ, θ を求める. 実際には非線形方程式になるため数値的に求めることになるが, $|\theta| \ll 1$, $|\phi| \ll 1$ が成立すれば (2) は θ, ϕ について一次になるので, l, l' の二つを観測することで姿勢の近似値を得ることが出来る.

また l の長さ ($|l|$ と表す), θ, ϕ が与えられた時

$$\frac{d}{dt}|l| = \frac{1}{2|l|} J^T(\theta, \phi) \frac{d}{dt} \begin{pmatrix} \phi \\ \theta \end{pmatrix}, \qquad (3)$$

の関係が得られる. これは ϕ, θ の運動を与えた時に, ワイヤ長をどのように変化させれば良いかを与える逆運動学モデルとなっている. ここで

$$J(\theta, \phi) = \begin{pmatrix} \cos\phi & 0 & 0 & \cos\phi\sin\theta & \cos\phi\cos\theta \\ 0 & -\sin\theta & \cos\theta & \sin\phi\cos\theta & -\sin\phi\sin\theta \end{pmatrix} c$$

である.

以上 (2) および (3) によって耳介とワイヤ長さの間の運動学モデルが得られた.

4 特性

前節までで説明した耳介について, 聴覚上の特性を調べる. ここでは, 耳介の大きな効用である耳介ノッチと集音効果に相当する性質を調べることとした. なお, 耳介ノッチは音源の上下方向に対する関数として周波数特性を考えることが多いが, 本稿では対象とする耳介の形状変化に注目しているので, 前節の ϕ の変化に伴う影響を考える. 他方の自由度 θ については姿勢を一定に保ったまま水平面内での指向性について調べ, 耳介の集音効果について着目した.

4.1 測定方法

4.1.1 周波数特性

周波数特性の測定は以下の手順で行った. TSP[Suzuki, 1992]信号を駆動信号とし, Fig.5 に示すように耳介のマイクロホンと耳介近傍のマイクロホンでこれらの信号を受聴した. なお, TSP 信号を受聴した信号は十分な周波数成分を含むと考えられるので, 耳介での信号と耳介近傍での信号の間の特性をクロススペクトル法で求めたものを耳介の周波数特性と考えた. これは収録環境の特性を除去することを期待したもので, 具体的な計算は MATLAB の tfesimate により 4096 点毎のオーバーラップを伴う 8192 点の FFT で行った.

実験ではスピーカと耳介間は 1m, 耳介近傍のマイクは耳介開口部 20cm 前方に設置した. 耳介は 頂点 P が開口部下端から後方 3cm, 真上, 前方 2cm, 同 5cm の 4 通り (ϕ の 21.5 度, 0 度, −14.1 度, −37.7 度 に対応) を測定し, それぞれの姿勢において TSP 応答を 9 回計測している. 信号は 44100Hz でサンプリングし増幅したものを収録しているため, 特性はアンプ等の特性を含んだものになっている.

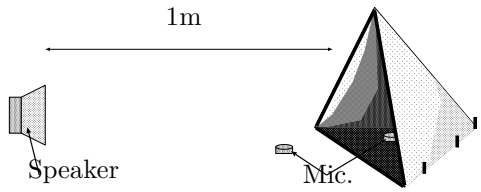


Figure 5: Impulse response measurement

4.1.2 指向性

次に水平方向の特性として指向性について調べた. ここではワイヤを自然な状態にし (頂点 P が開口部下端から後方 3cm に位置, $\phi = 21.5$ 度), 先の実験と同様, 音源を耳介前方 1m に設置した. 耳介を回転台に載せ, マイクロホンを中心として耳介そのものを回転させることで, 音源との相対的な水平方向を変化させた. 計測にあたっては 11.25 度刻みで 360 度全方向からの特性を測定した. なお, 音源方向毎のゲイン特性によって指向性を考えることとした. この実験でも, 前述の実験と同様, TSP 信号を駆動信号とし, 耳介前方に設置したマイクロホンと耳介で収録した信号の間の特性によって耳介の特性とした.

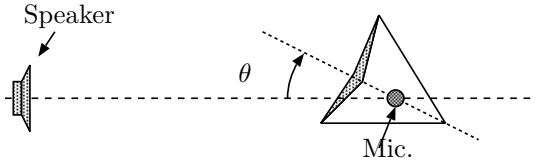


Figure 6: Directivity measurement

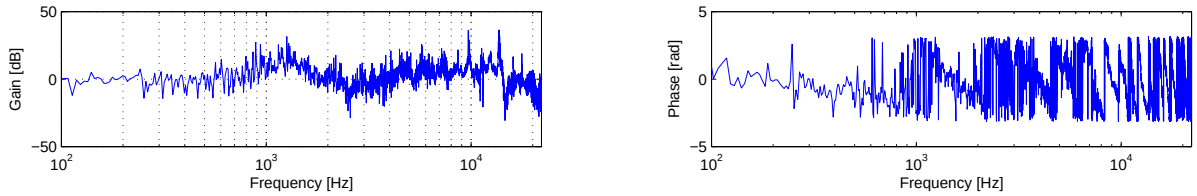
4.2 測定結果

4.2.1 周波数応答

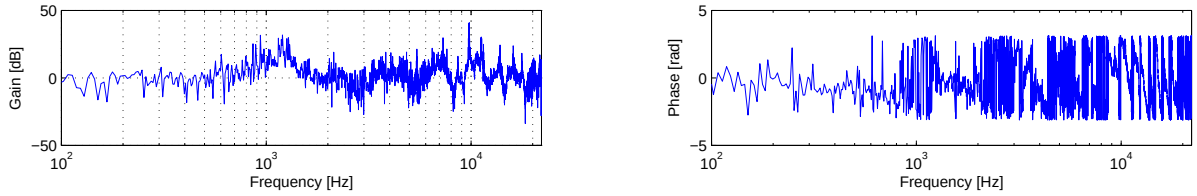
Fig.7 に測定された耳介の周波数応答を示した. ゲイン特性を見ると, ϕ に依らず 1kHz 付近のなだらかなピークと 2.5kHz から 3kHz にかけてのノッチが安定して見られる. 一方 5kHz より高い周波数帯域では, 耳介変形の影響を受けてゲイン特性が変化していることが分かる. 位相については, 高周波数帯域 (数 kHz 以上) ではっきりとした構造が見られ, ϕ の変化の影響と見られる変化がある.

4.2.2 指向性

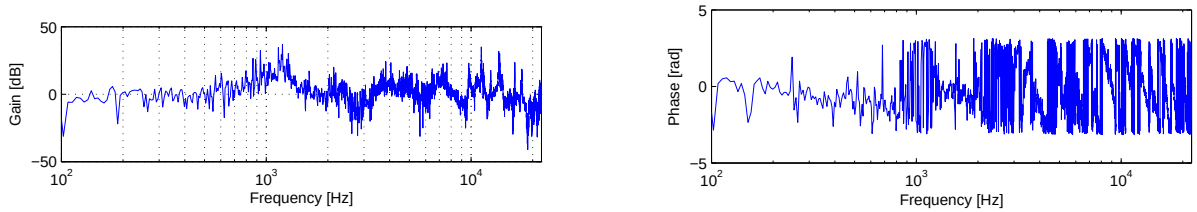
Fig.8 に耳介が回転した際のゲイン特性を示す. 図は横軸に耳介の方向, 縦軸に周波数を示したもので, 0 度が耳



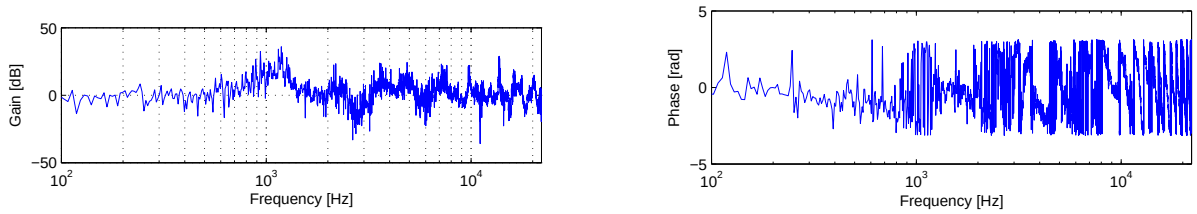
(a) $\phi = 21.5$deg : Gain(left) and phase(right)



(b) $\phi = 0$deg : Gain(left) and phase(right)



(c) $\phi = -14.1$deg : Gain(left) and phase(right)



(d) $\phi = -37.7$deg : Gain(left) and phase(right)

Figure 7: Frequency response of the proposed pinna

介正面方向にスピーカが位置する場合に対応し, 反時計まわりを正としている. 前方に対してある程度の指向性を示しており, 8kHz を越える高周波数帯域ではっきりとした特性が確認される. しかし, 8kHz 以下では指向性は認められず, 1kHz 付近のピークでは音源の方向に依らず一様なゲイン特性となっていた.

4.3 考察

これらの測定結果から, 指向性や形状変化への依存性といった特性は主に 数 kHz より高い周波数帯域で見られた. 周波数特性では ϕ の正負での特性変化がある程度見られたと言えるが, これは $\phi = 21.5$ 度の時, P 点がマイクロホンよりも後方にあり, マイクロホンが剥き出しに近い状態になっていたのに対し, それよりも小さな ϕ ではマイクロホンが耳介壁面に覆われる状態になっていたためと考えられる. また, 1kHz 付近に増幅特性があり, 方向や耳介形状に関係せず存在することが判った. これは耳介の効果ではあるが, 音源定位の観点からはあまり望ましくない.

所望の特性が高周波数帯域に制限されるのは, 波長の関係である程度仕方がないが, 指向性に関する結果から低周波域で耳介が機能していないことが想定される. これは, 変形を優先するために耳介の薄くした結果, 対応する周波数帯域の音波が透過している可能性がある.

5 おわりに

本稿では, 能動的音信号の受聴を目指し, ワイヤ駆動により形状を変える, 指向性が可変の耳介を提案し, その基本的な特性を実験によって調査し, 耳介の変形を導く簡単な運動学モデルを導いた. 得られた特性は, 前方に対してある程度の指向性を有するものであったが, 耳介の変形に伴う周波数特性の変化は高周波数帯域に限定的であった. この理由として、耳介が薄く、低周波帯域では十分な反射が得られていない可能性が考えられる。今後、耳介の厚さを変える、あるいは異った素材を用いるなどが必要となる。

また, 受聴した音信号に対する応答, 例えば音圧情報を手掛りに音源方向を探査する [Bernard, 2010]など, 耳介を運動学モデルに基づいた耳介の駆動制御を行う. 本稿で調べていないが, 音源の上下に対する耳介の影響も検討する必要がある。

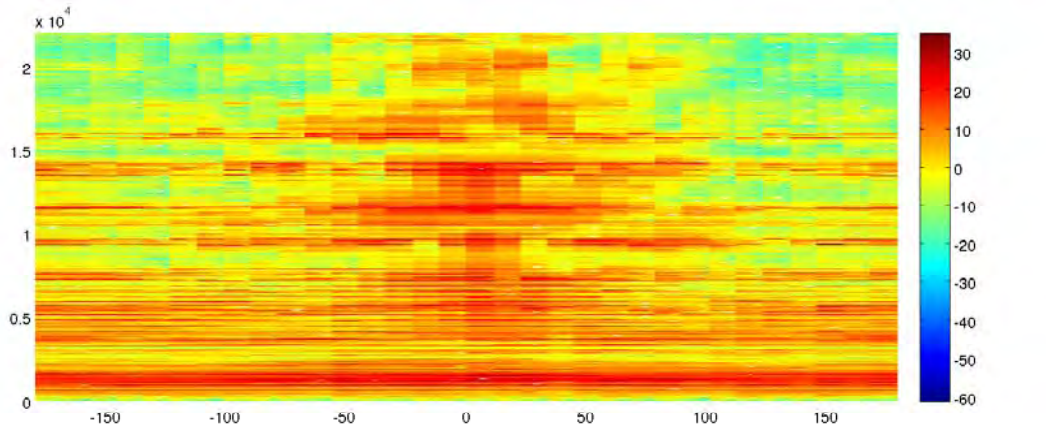


Figure 8: Directivity

参考文献

[Shaw, 1968] Shaw, E.A.G. and Teranishi, R.: Sound pressure generated in an external-ear replica and real human ears by a nearby point source, J. of Acoust. Soc. Am. 44 (1), 240–249, (1968).

[Batteau, 1967] Batteau, D.W.: The role of the pinna in human localization, Proc. of Royal Soc. of London, B 158, 158–180,(1967)

[Butler, 1984] Musicant,A.O. and Butler, R. A.: The influence of pinnae-based spectral cues on sound localization, J. of Accoust. Soc. Am. 75(4), 1195–1200, (1984).

[Lopez-Poveda, 1996] E.A. Lopez-Poveda and R. Meddis: A physical model of sound diffraction and reflections in the human concha, J. of Acoust. Soc. Am. 100 (5), 3248–3259, (1996).

[Shimoda, 2006] Shimoda, T., Nakashima, T., Kumon, M., Kohzawa, R., Mizumoto I., and Iwai, Z.: Spectral cues for robust sound localization with pinnae, Proc. of 2006 IEEE/RSJ Int'l Conf. Intell. Robot. and Sys., 386–391,(2006).

[Hörnstein, 2006] Hörnstein, J., Lopes, M., Santos-Victor, J. and Lacerda, F.: Sound localization for humanoid robots - building audio-motor maps based on the HRTF, Proc. of IROS 2006, 1170–1176, (2006).

[Finger, 2010] Finger, H., Ruvolo, P., Liu, S.C., Movellan, J.: Approaches and Databases for Online Calibration of Binaural Sound Localization for Robotic Heads, Proc. of IROS 2010, 4340–4345, (2010).

[Kumon, 2009] 公文誠, 石飛光章: 境界要素法を用いた音響解析による耳介形状の検討, 人工知能学会 SIG チャレンジ研究会, 14–19,(2009).

[Rice, 1992] J.J. Rice, B.J. May, G.A. Spirou and E.D. Young: Pinna-based spectral cues for sound localization in cat, Hearing Research, 58, 132–152 (1992).

[Populin, 1998] L.C. Populin and T. C. T. Yin: Pinna movements of the cat during sound localization, J. of Neuroscience,18(11), 4233–4243, (1998).

[Sasaki, 2009] Sasaki, Y., Kagami, S., Mizoguchi, H.: Online Short-Term Multiple Sound Source Mapping for a Mobile Robot by Robust Motion Triangulation, Advanced Robotics, 23, 1-2, 145-164, (2009).

[Toshima, 2006] 戸嶋巌樹, 青木茂明, 平原達也: 頭部運動を再現する改良型ダミーヘッドシステム：テレヘッド II, 日本音響学会誌, 62, 3, 244–254, (2006).

[Ellsworth, 1902] Ellsworth R.J. and Jennings,H. S. : Anatomy of the cat,Henry Holt and Co., (1902).

[Suzuki, 1992] 鈴木陽一: 時間引き伸ばしパルス設計法に関する考察, 信学技法, EA92–86, (1992).

[Bernard, 2010] Bernard, M., N'Guyen, P., Pirim, P., Gas, B. and Meyer J-A.: Phonotaxis Behavior in the Artificial Rat Psikharpax, Proceedings of ISIR 2010, (2010).

〒162 東京都新宿区津久戸町 4-7 OS ビル 402 号室 03-5261-3401 Fax: 03-5261-3402

(本研究会についてのお問い合わせは下記にお願いします.)

SIG-AI-Challenges home page (WWW): `http://winnie.kuis.kyoto-u.ac.jp/SIG-Challenge/`