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。

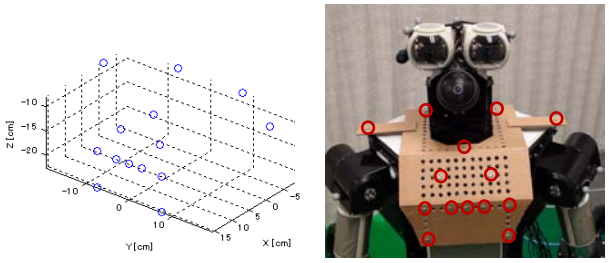

Fig. 1. (a) The geometry of the 14-element microphone array. (b) Robovie wearing the microphone array.

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。

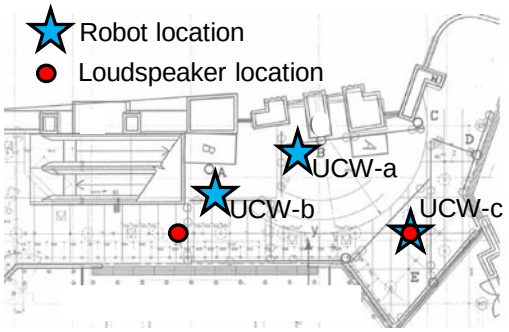

Fig. 2. A map of the UCW hallway, with locations of the robot and the ceiling loudspeakers.

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。

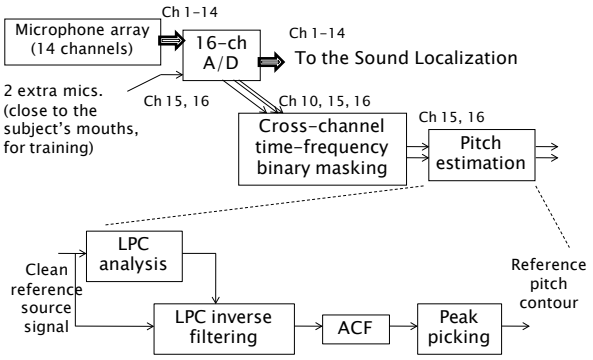

Fig. 3. Obtaining the reference pitch contours from the reference microphones.

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。

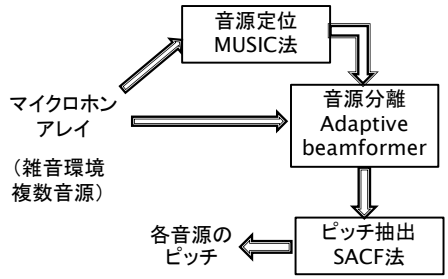

Fig. 4. Overall block diagram of the evaluated pitch extraction.

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。

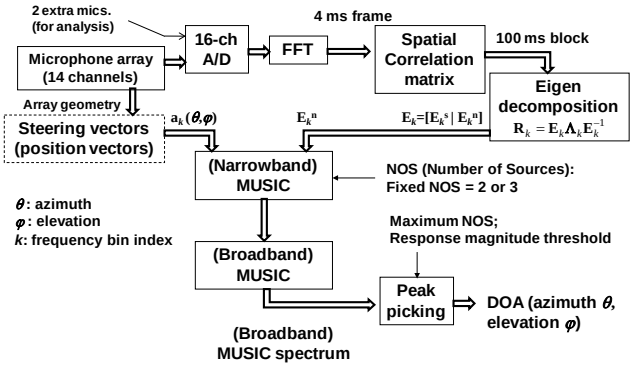

Fig. 5. The MUSIC-based sound localization algorithm, and related parameters.

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。

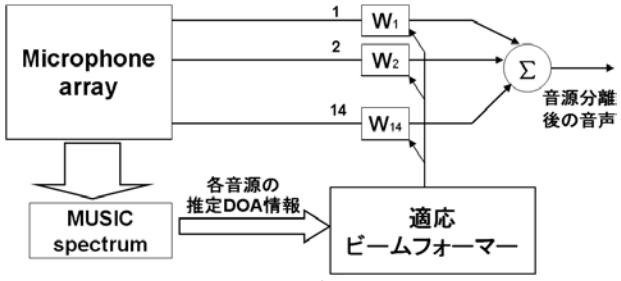

Fig. 6 Speech separation using adaptive beamformer

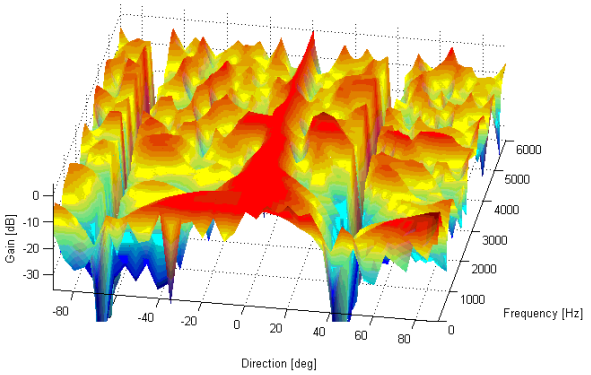

Fig. 7 Example of beamformer gain for a target source close to 0 degrees, and interference sources at 50 and -60 degrees.

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。

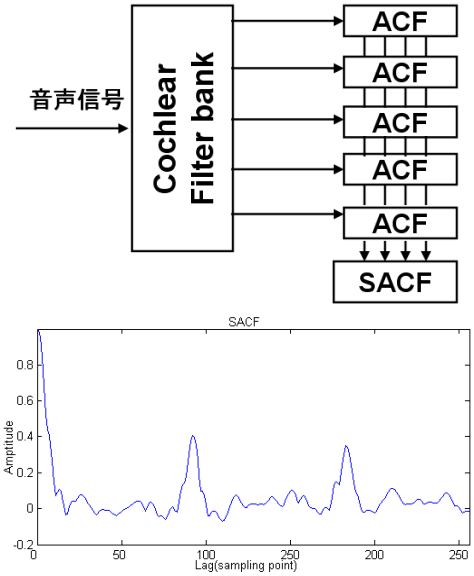

Fig. 8　Pitch extraction method based on SACF.

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。

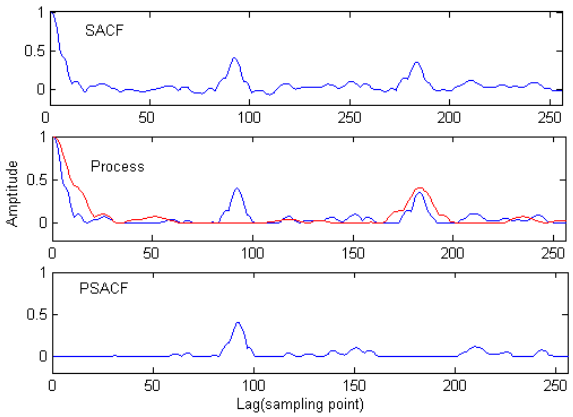

Fig. 9　Example of SACF peak pruning.

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。

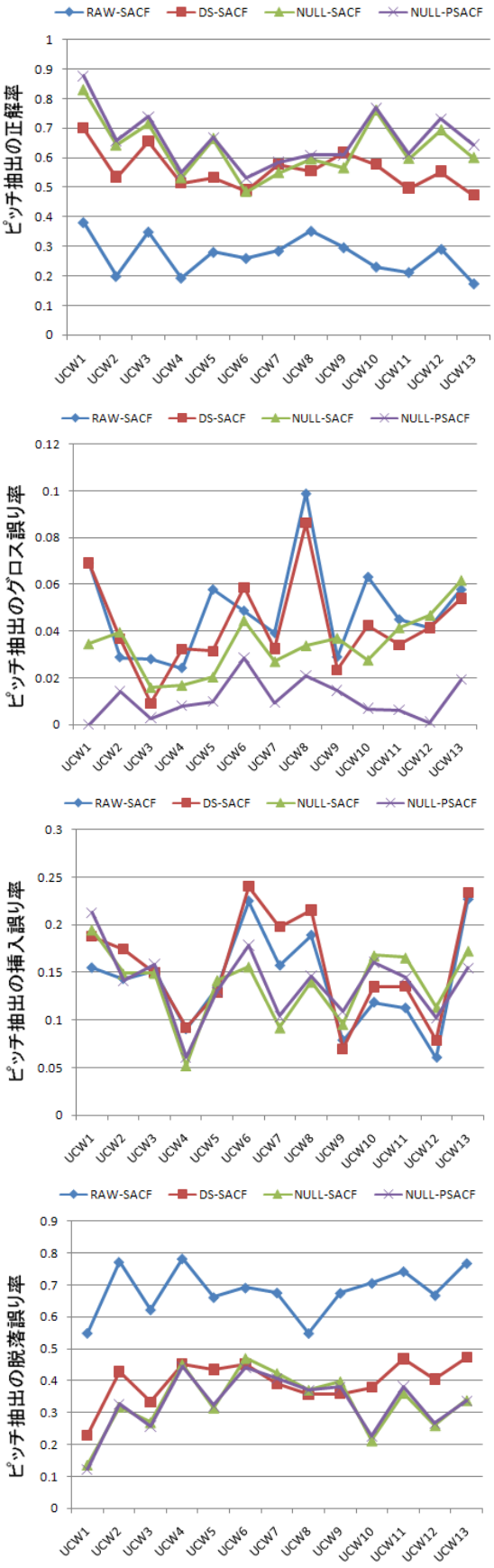

Fig. 10　Pitch extraction performance results for each trial in UCW.

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