人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B303

ＡＩチャレンジ研究会 (第３９回)

Proceedings of the 39th Meeting of Special Interest Group on AI Challenges

CONTENTS

日　時　2014 年 3 月 18 日　　場　所　京都大学 吉田キャンパス 総合研究 8 号館 講義室 1
Kyoto University, Tokyo, Mar. 18, 2014

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-B303-01

スマートポスターボード : ポスター会話のマルチモーダルなセンシングと解析

Smart Posterboard: Multi-modal Sensing and Analysis of Poster Conversations

河原達也
Tatsuya KAWAHARA
京都大学
Kyoto University

Abstract

学会等で一般的に行われているポスター発表における会話（ポスター会話）は、マルチモーダルな多人数インタラクションに関する様々な興味深い研究テーマを提供してくれる。本稿では、著者らが進めているポスター会話のマルチモーダルなセンシングと解析に関するプロジェクト、及び構築しているスマートポスターボードシステムの概要を紹介する。本システムは、ポスターボードに設置した複数のセンサを用いて会話を記録し、誰がポスターに来て、どのような質問やコメントを行ったかを容易に検索できるようにすることを目指している。そのためにまず、顔向き（視線）検出と話者区間検出を統合したマルチモーダルな信号処理を実現する。さらに、視線配布や相槌などの聴衆のマルチモーダルな振る舞いにに着目することで、興味・理解度の推定を試みる。

1 はじめに

マルチモーダルな信号・情報処理に関する研究は従来、人間型ロボットを含むヒューマンマシンインターフェースの高度化を主な目標として行われてきた。一方、画像処理や音声処理が高度になり、上記のようなインターフェースを意識しない人間の自然なふるまいも扱えるようになり、いわゆるアンビエントなシステムを目指した研究開発も可能になっている。実際に、ミーティング [1] や自由会話 [2] などの人間どうしの会話を対象とした研究も行われている。

我々はポスターセッションにおける会話（=ポスター会話）に焦点をおいたプロジェクトを進めている [3, 4, 5, 6]。ポスターセッションは、学会やオープンラボなどで一般的になっているが、未だに情報通信技術 (ICT) の導入がほとんどなされておらず、ICT 分野の会議でも紙のポスターを用いることが多い。一部液晶ディスプレイや携帯プロジェクタを用いる場合もあるが、センサを備えた環境は世界的にも前例がないと思われる。講演や講義の映像・音声が収録・配信されることが一般的になっているのに対して、ポスターセッションを収録して分析した研究も皆無に近い。

ポスター会話は、講演と会議の中間的な形態と捉えることができる。すなわち、発表者が自身の研究内容について少人数の聴衆に説明する一方、聴衆の側も相槌や頷きなどでリアルタイムにフィードバックし、時折質問やコメントも行う。また会議と違って、参加者は立っており、動くこともできるので、マルチモーダルなインタラクションを行うことが多い。さらに、ポスター会話を扱う利点としては、話題や他の参加者に対する親近性を制御しながら、（研究者を集めてくれば）自然でリアルなデータを収集することが非常に容易であることが挙げられる。

本プロジェクトの目標は、人間どうしのインタラクションの信号レベルのセンシングとより高いレベルの解析である。人間型ロボットを含むヒューマンマシンインターフェースと比較して、長時間にわたる自然な振る舞いを扱う点が最大の違いである。認識のタスクとしては、人物・視線・話者・発話区間などの検出がある。これらは会話アーカイブに対する新たなインデキシングの枠組みを提供する。例えば、自身あるいは同僚のポスターセッションが終わった後で、どのくらいの聴衆がやって来て、どのような質疑・コメントが行われたかを振り返ることができるようになる。

さらに、どの部分に興味を持ってもらえたか、どこがわかりにくいところであったか、といった解析も研究する。動画投稿サイトなどの事例からもわかるように、我々は他の人が興味を持ったものを視聴したくなるのが普通であるので、このようなアノテーションは有用であると考えられるが、アノテーションの基準や評価を含めて、これらを

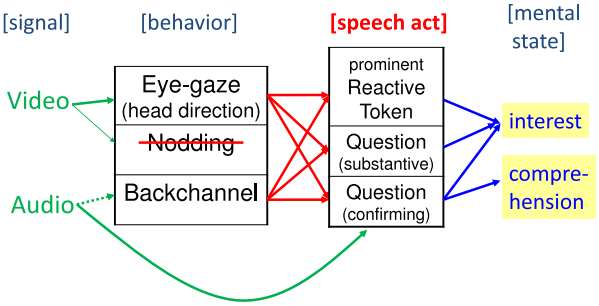


図 1: マルチモーダルなセンシングと解析の枠組み

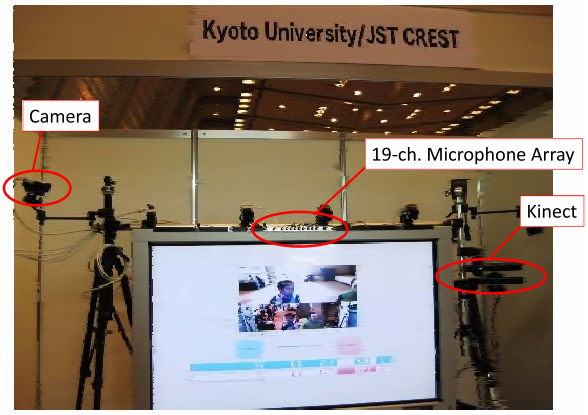


図 2: スマートポスターボードの概観

明確に定義するのは非常に困難である。そこで、これらの心的状態に関係すると考えられ、客観的に観測できる発話行為 (speech act) に着目する。具体的には、聴衆による質問と特定のパターンの相槌 (reactive token) に着目する。さらに、質問を確認質問 (confirming) と踏み込み質問 (substantive) に分類する。マルチモーダルな振る舞いからこれらの発話行為を予測することで、興味・理解度の推定を近似できると期待している。提案する枠組みを図 1 に示す。

本稿では、まず 2 章でセンシング環境と収集したコーパスの説明を行う。3 章では、図 1 の左側の部分、すなわち信号レベルから振る舞いの検出を行う処理について述べる。4 章では、図 1 の右側の部分、すなわち発話行為と心的状態を関係づけて、興味・理解度を定義する。5 章では、図 1 の中央の部分、すなわち、聴衆のマルチモーダルな振る舞いからの発話行為の予測について述べる。

2 ポスター会話のマルチモーダルコーパス

2.1 収録環境：スマートポスターボード

我々は、ポスター会話における音声・映像と振る舞いなどのマルチモーダルな情報を収録するための環境の構築を進めてきた [7, 8]。また、ポスターボードを大型液晶ディスプレイで構成し、これに多様なセンサを設置することで、ポスター会話をセンシングするシステム（=スマートポスターボード）を構築している。スマートポスターボードの概観を図 2 に示す。

音声に関しては、ポスターボードの上に設置するマイクロフォンアレイを設計した。映像に関しては、参加者全員とポスターをカバーできるように、6〜8 個のカメラをポスターボードに設置した。また、Kinect センサも設置した。[1]

ただしコーパス構築の上では、正確な情報 (ground truth) を取得する必要がある。そのために、各参加者にワイヤレスのヘッドセットマイクを装着してもらうとともに、様々なセンサを着用してもらった。当初はモーションキャプチャシステムや視線計測装置を使用したが、直近は磁気センサを使用している。

2.2 コーパスとアノテーション

上記の環境を用いて、これまで 5 ヶ年度にわたって合計 43 セッションのポスター会話を収集してきた。ただし、いくつかのセンサデータが欠損したものも含まれる。

各セッションにおいては、1 名の発表者（A と表記）が自身の研究に関する発表を、2 名の聴衆（B,C と表記）に対して行う。聴衆は、発表者についても研究内容についても初めて接する設定となっている。セッションの長さは制御しているわけではないが、おおむね 20〜30 分程度である。

音声データは、ヘッドセットマイクで収録されたものをポーズで区切られた発話単位 (IPU) に分割し、時間と話者ラベルを付与した上で、『日本語話し言葉コーパス』(CSJ) と同じ基準で書き起こしを行った。ただし、フィラー以外に相槌と笑いに対してもアノテーションを行った。

視線情報は、視線計測装置とモーションキャプチャシステム、または磁気センサのデータを用いて、視線ベクトルと他の参加者やポスターの位置との交差判定に基づいてアノテーションを行った。

以降の章の実験では、2012 年度に収集・アノテーションを行なった 4 つのセッションを主に用いている。これらのセッションにおける発表者と聴衆の組合せはすべて異なっている。

3 マルチモーダルな振る舞い（視線・発話）の検出

音声や映像の信号から、各会話参加者の振る舞いを検出する処理については、順次研究開発を進めている。ここでの

[1] 簡易版は Kinect センサのみを用いる。

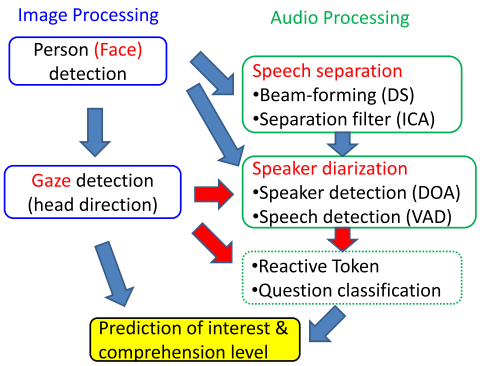


図 3: マルチモーダル統合による振る舞い検出の処理の流れ

目標は、参加者にマイクなどの装置を一切装着してもらうことなく、ポスターボード等に設置したセンサだけで処理を行うことである。具体的には、マイクロフォンアレイで収録される音声信号、及び複数台のカメラ（Kinect センサ）で得られる映像信号・距離情報を用いる。

各処理の高度化と適用を行なうとともに、これらのマルチモーダルな情報を統合する方法を検討している。全体の処理の流れを図 3 に示す。まず聴衆の人物（顔）を検出し、各人の顔向き（視線）を検出・追跡する。人物の位置情報は、音響処理、具体的には音声強調と話者同定において利用される。また、視線情報は発話区間検出や質問種類の同定に用いる。

3.1 Kinect センサによる視線（頭部方向）検出

本研究では Kinect センサを用いて、ポスター会話に適した実用的な視線推定を実現した。なお、画像の解像度及び眼鏡等の影響により眼球そのものを安定に撮影することは困難であるため、視線方向を頭部方向で代用する。視線と頭部方向のずれは平均 10 度程度で、ポスターを注視する状況ではさらに小さくなる傾向がある [9]。処理は以下の手順で行っている。

1. 顔検出

   Kinect センサで撮影したカラー画像及び距離画像から、Haar-like 特徴を利用した物体認識法を用いて正面顔探索を行う。同時に複数人を処理することが可能である。

2. 頭部モデル獲得

   顔検出結果に従って、距離画像から頭部の 3 次元形状を、カラー画像からその色情報を計算する。計算結果はポリゴンとテクスチャ情報に変換し、頭部モデルとする。

3. 頭部追跡

   頭部方向の推定を、画像への頭部モデルのフィッティングとして行う。具体的には、頭部を剛体とみなし、頭部の 3 次元位置と姿勢を表す 6 変数をパーティクルフィルタによる追跡処理で逐次計算する。

4. 注視対象特定

   頭部追跡処理で得られた 6 変数から、3 次元空間で視線に対応する半直線を求める。この半直線と、ポスターボードや他の参加者との交差判定を行うことで、注視対象を決定する。

上記の処理は、GPU を使うことで、オンライン・リアルタイムに実行可能である。

3.2 マイクロフォンアレイによる音声の分離・強調

音声の分離と強調は、ブラインド空間的サブトラクションアレイ (BSSA)[10] によって実現する。これは、マイクロフォンアレイで得られる信号に対して、遅延加算 (Delay-and-Sum) 型ビームフォーミングを行うとともに、独立成分分析 (ICA) に基づいて各会話参加者の音声と背景雑音を分離し、目的信号以外の抑圧を行うものである。ポスター会話の設定では、発表者・聴衆・背景雑音の 3 つの成分への分離を行う（聴衆間の分離は行われない）。その際に、画像処理によって得られる各参加者の位置情報を用いることで、ICA のフィルタ計算の高速化を実現している。この処理を逐次的に行うことで、参加者が移動しても追随できるようにしている。

19 チャネルのマイクロフォンアレイを用いる場合は、高い品質の音声強調ができるが、リアルタイムには処理できない。Kinect センサ内蔵の複数のマイクロフォンを用いる場合は、音質は低下するが、リアルタイム処理が可能である。

3.3 音響情報と画像情報を統合した話者区間検出

話者区間検出 (speaker diarization) は、“いつ誰が発話したか” を検出する処理で、話者同定 (speaker localization) と発話区間検出 (voice activity detection) の 2 つの要素からなる。そのために、マイクロフォンアレイで得られる音響情報（音のパワーと位相情報）に加えて、画像から得られる各参加者の位置情報を利用する。マルチモーダルな発話区間検出として、口唇の動きを用いることも考えられるが、解像度の高い正面画像が必要なため、ポスター会話においては現実的でない。

マイクロフォンアレイを用いた音源の到来方向推定 (DOA estimation) の代表的な手法である MUSIC 法 [11] を用いる。MUSIC 法は、観測信号の部分空間の直交性に基づいて、同時に複数の音源をリアルタイムに推定することができ、各時刻 t・各方向 θ に関して、そこに音源が存在する尤度 $P_{MU}(t,\theta)$ を求めることができる。

表 1: 話者区間検出結果

	F 値（発表者）	F 値（聴衆）	DER
音声分離+パワー	0.880	0.515	38.5%
MUSIC 法	0.920	0.581	47.5%
ルールベース統合	0.921	0.591	42.3%
確率的統合	0.887	0.686	29.9%

ベースライン手法では、この尤度 $P_{MU}(t,\theta)$ の極大値を探索し、しきい値以上となるものを音源とみなす。このときに、画像処理による顔検出で得られる参加者の位置情報を利用する。すなわち、$P_{MU}(t,\theta)$ がしきい値以上であり、かつ θ が参加者の推定位置からしきい値以内である場合に、発話がなされたと判定する。

さらに、確率的な統合 [12, 13] も検討する。すなわち、画像情報により得られる参加者の位置情報に関して、推定位置を平均、その信頼度を分散とする正規分布に基づいて尤度を算出し、$P_{MU}(t,\theta)$ と統合する。

4 セッションに対する話者区間検出の結果を表 1 に示す。発表者はマイクに近く、発話が多いため、90%に近い検出精度が得られるのに対して、聴衆の発話区間検出は困難である。前節の音声分離結果に対してパワーに基づいて単純に発話区間検出する方法では F 値が 50%程度になっている。MUSIC 法ではそれより高い精度が得られるが、単純に画像による位置情報をヒューリスティックに用いてもほとんど効果が見られない。それに対して、確率的な統合により、70%程度まで改善している。ただし、雑音を重畳すると性能が低下するので、その対策も検討する必要がある。

4 興味・理解度の定義

興味・理解度のアノテーションを行う最も自然な方法は、ポスターセッション終了後に聴衆の各人に、各々のスライド話題単位に対する興味と理解の評定を行ってもらうことである。しかしながら、このようなアンケート調査を大規模に行うことはあまり現実的でないし、既に収録済みのセッションに行うことは不可能である。またこのような評定は主観的で、その信頼性を評価することも難しい。

そこで本研究では、興味・理解度に関係が深いと考えられ、客観的に観測可能な発話行為に着目する。これまでに我々は、「へー」「あー」「ふーん」といった非語彙的・引き延ばし型で韻律的にも顕著な特徴を持つ相槌（=顕著な相槌）が聴衆の興味と関係があることを明らかにした [14, 15]。Ward ら [16] は英語の相槌に関して、そのパターンと役割の分析を行っている。

また経験的に、聴衆の質問の生起は興味と関係があると考えられる。すなわち、聴衆は発表に引きつけられるほど、より多くの質問をするものである。また、質問の種類を調べることで、理解度を推測することもできる。例えば、既に説明されたことを質問しているなら、理解が困難であったことを示唆している。

4.1 質問の種類のアノテーション

本研究では、質問を確認質問と踏み込み質問に分類した。確認質問は、現在の説明の理解が正しいか確認するために行うもので、「はい／いいえ」のいずれかで答えることができる。[2] これに対して踏み込み質問は、発表者の説明に含まれていなかったことに関して尋ねるもので、「はい／いいえ」のみで答えられるものでなく、何らかの補足説明が必要になる。踏み込み質問は、表層的には質問の形式をとっているが、実質的にコメントに近い場合もある。

4.2 質問の種類と興味・理解度との関係

2012 年度に収録した 4 つのセッションについては、終了後に聴衆の各人に各スライド話題単位に対する興味と理解の度合いを評定してもらった。そこで、このデータを用いて、評定と質問との関係を調べた。

図 4 に、2 種類の質問（confirming: 確認質問; substantive: 踏み込み質問）の生起毎、及び全話題セグメント（entire）の興味・理解度の分布を示す。興味度については、1（低い）から 5（高い）の 5 段階で評定してもらい、理解度については、1（低い）から 4（高い）の 4 段階で評定してもらっている。左のグラフから、質問の種類に関わらず、質問が生起している場合には全般に興味が高い (4 か 5) ことがわかる。また右のグラフから、確認質問の大多数 (86%) が理解度が低い (1 か 2) ことと相関があることがわかる。

この分析結果と顕著な相槌に関する先行研究 [15] を踏まえて、分析対象の全話題セグメントに対して、以下のアノテーションの枠組みを採用した。

- 興味が高い ← （種類に関わらず）質問もしくは顕著な相槌が生起している
- 理解度が低い ← 確認質問が生起している

これらのアノテーションは各話題セグメントに対して、聴衆 2 名の各人について行った。これらの心的状態の検出は、ポスター会話を後で振り返る際に有用であると考えられる。

5 マルチモーダルな振る舞いに基づく発話行為の予測

聴衆の興味・理解度の推定を、関係する発話行為の予測を行う問題として定式化した。すなわち、興味の推定は質問

[2] ただし、実際に発表者が「はい／いいえ」のみで答えたとは限らない。

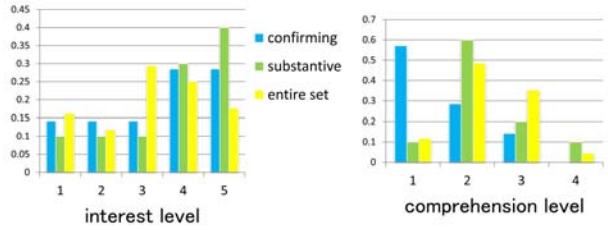


図 4: 質問の種類毎の理解・興味度の分布

と顕著な相槌の生起の予測に、理解度の推定は質問タイプの分類に帰着させた。この予測を、当該の発話行為が（話題セグメントの終わり頃に）実際に生起する前に、聴衆のマルチモーダルな振る舞いを元に行う。これにより、そのような発話行為が実際に生起しなくても、聴衆の心的状態の推定が可能になると期待できる。

まず、各聴衆の振る舞いを特徴量にする必要がある。マルチモーダルな振る舞いとして、相槌と視線配布に着目した。相槌については、発表者の発話で正規化した平均頻度を求めた。発表者に対する視線配布については、発表者の発話で正規化した出現頻度と継続時間割合を求めた。

次に、識別のための機械学習の方法については、ナイーブベイズ分類器を用いた。これは、学習データが少なく、各特徴量の重みなどのパラメータを推定することが困難であるためである。特徴量ベクトル $F = \{f_1, \ldots, f_d\}$ に対するナイーブベイズ分類は、以下の事後確率に基づいて行われる。

$$p(c|F) = p(c) * \prod_i p(f_i|c)$$

ここで、c は分類カテゴリであり、ここでは「興味を持ったか否か」である。また、$p(f_i|c)$ を計算するのには、ヒストグラム量子化を用いた。これは、特徴量の値を量子化ビンに割り当てるもので、確率密度関数を仮定しないためモデルパラメータの推定を必要としない。特徴量の分布ヒストグラムを単純に 3 ないし 4 に分割して量子化ビンを設定する。その上で、各ビンの相対的な出現頻度を確率値に変換する。

本実験では 2009 年度と 2011 年度に収録したものを含めて合計 10 セッションを用いた。この 10 セッションには計 58 個のスライド話題単位があった。各セッションに 2 名の聴衆がいるので、興味・理解度を推定すべきスロット（= 話題セグメント）が合計 116 個あることになる。評価実験は、セッション単位の leave-one-out クロスバリデーションにより行った。

5.1 質問・顕著な相槌の生起の予測：興味度の推定

まず、各話題セグメントにおける聴衆の興味度を推定する実験を行った。これは聴衆が質問ないし顕著な相槌を生成するかを予測する問題に帰着される。すなわち、当該の発話行為を行った聴衆は、その話題セグメントに「興味を持った」とみなす。

表 2: 質問・顕著な相槌を含む話題セグメントの予測結果（興味の推定）

	F 値	正解率
ベースライン	0.49	49.1%
(1) 相槌	0.59	55.2%
(2) 視線 頻度	0.63	61.2%
(3) 視線 時間	0.65	57.8%
(1)-(3) の組合せ	0.70	70.7%

表 3: 確認質問／踏み込み質問の同定結果（理解度の推定）

	正解率
ベースライン	51.3%
(1) 相槌	56.8%
(2) 視線 頻度	75.7%
(3) 視線 時間	67.6%
(1)-(3) の組合せ	75.7%

種々の特徴量に対する結果を表 2 に示す。正解率は計 116 の話題セグメントで正しい判定が得られたものの割合である。なお、すべての話題セグメントに「興味を持った」とした場合の (chance rate) ベースラインは、49.1%である。

相槌と視線の特徴を用いることで、有意に高い正解率が得られ、両者を組み合わせることで 70%を上回る結果となった。ただし、視線に関する 2 つの特徴量（頻度と時間）については一方を外しても結果は変わらなかった。以上、相槌と視線のマルチモーダルな統合効果を確認した。

5.2 質問の種類の同定：理解度の推定

次に、各話題セグメントにおける聴衆の理解度を推定する実験を行った。これは聴衆が質問を行った際に、質問の種類を同定する問題に帰着させる。すなわち、確認質問を行った聴衆は、その話題セグメントの「理解が困難であった」とみなす。

確認質問／踏み込み質問の分類結果を表 3 に示す。なお、このタスクでは各質問の出現頻度 $p(c)$ に基づく (chance rate) ベースラインは、51.3%である。

すべての特徴量が正解率の向上に一定の効果があったが、視線の出現頻度のみで最良の正解率が得られ、他の特徴量と組み合わせても相乗効果は見られなかった。質問が生起する直前の 2 発話における視線特徴量の時間的な変化を用いることも検討したが、改善は得られなかった。

6 おわりに

本稿では、ポスター会話をセンシングするスマートポスターボードの紹介を行なった。ポスター会話では、複数人が動きながら遠隔で発話を行なう点でヒューマノイドロボットの場合と同様であるが、長時間の自然な会話であり、しかも発表者に比べて聴衆の発話が圧倒的に少ないので、より挑戦的なタスクと考えられる。現時点では、発表者に対して約 90%、聴衆に対して約 70%の話者区間検出精度となっている。

さらに、視線情報などをインタラクションの高度な解析に用いることを研究している。これにより、会話の流れを捉えるとともに、興味・理解度の推定も可能になると考えている。興味・理解度ともに 70%程度の精度で推定可能なことを示した。

これらの情報を可視化するブラウザも構築しており、長時間のポスター会話を振り返る上で有用であるか検証を進めていく予定である。

謝辞

本研究は、JST CREST「人間調和型情報環境」領域ならびに科学研究費補助金の支援を受けて実施されたものである。特に、スマートポスターボード（2 章・3 章）の研究開発は、CREST プロジェクトに参画して頂いている京都大学の吉本廣雅研究員、Tony Tung 特定助教、若林佑幸研究員、井上昂治君と奈良先端科学技術大学院大学の猿渡洋准教授をはじめとする多くの方々の貢献によるものである。また、4 章・5 章の研究は、京都大学の高梨克也研究員と林宗一郎君におうものである。

参考文献

[1] S.Renals, T.Hain, and H.Bourlard. Recognition and understanding of meetings: The AMI and AMIDA projects. In *Proc. IEEE Workshop Automatic Speech Recognition & Understanding*, 2007.

[2] K.Ohtsuka. Conversation scene analysis. *Signal Processing Magazine*, Vol. 28, No. 4, pp. 127–131, 2011.

[3] T.Kawahara. Multi-modal sensing and analysis of poster conversations toward smart posterboard. In *Proc. SIGdial Meeting Discourse & Dialogue*, pp. 1–9 (keynote speech), 2012.

[4] T.Kawahara. Smart posterboard: Multi-modal sensing and analysis of poster conversations. In *Proc. APSIPA ASC*, p. (plenary overview talk), 2013.

[5] 河原達也. ［招待講演］ スマートポスターボード: ポスター会話のマルチモーダルなセンシングと認識. 電子情報通信学会技術研究報告, SP2012-51, 2012.

[6] 河原達也. ［特別講演］ スマートポスターボード: ポスター発表における場のマルチモーダルなセンシングと認識. 電子情報通信学会技術研究報告, PRMU2012-167, 2013.

[7] 瀬戸口久雄, 高梨克也, 河原達也. 多数のセンサを用いたポスター会話の収録とその分析. 情報処理学会研究報告, SLP-67-6, 2007.

[8] T.Kawahara, H.Setoguchi, K.Takanashi, K.Ishizuka, and S.Araki. Multi-modal recording, analysis and indexing of poster sessions. In *Proc. INTERSPEECH*, pp. 1622–1625, 2008.

[9] 矢野正治, 中田篤志, 福間良平, 角康之, 西田豊明. 非言語マルチモーダルデータを用いた会話構造の分析のための環境構築. 情処学研報, 2009-UBI-22-12, 2009.

[10] Y.Takahashi, T.Takatani, K.Osako, H.Saruwatari, and K.Shikano. Blind spatial subtraction array for speech enhancement in noisy environment. *IEEE Trans. Audio, Speech & Language Process.*, Vol. 17, No. 4, pp. 650–664, 2009.

[11] R. Schmidt. Multiple emitter location and signal parameter estimation. *IEEE Trans. Antennas & Propagation*, Vol. 34, No. 3, pp. 276–280, 1986.

[12] 中村圭佑, 中臺一博, 浅野太, 中島弘史, Ince Gokhan. マルチモーダル情報統合によるインテリジェント人追跡システム. 計測自動制御学会論文集, Vol. 48, No. 6, pp. 349–358, 2011.

[13] K.Nakamura, K.Nakadai, F.Asano, and G.Ince. Intelligent sound source localization and its application to multimodal human tracking. In *Proc. IROS*, 2011.

[14] 常志強, 高梨克也, 河原達也. ポスター会話におけるあいづちの韻律的特徴に関する印象評定. 人工知能学会研究会資料, SLUD-A901-06, 2009.

[15] T.Kawahara, K.Sumi, Z.Q.Chang, and K.Takanashi. Detection of hot spots in poster conversations based on reactive tokens of audience. In *Proc. INTERSPEECH*, pp. 3042–3045, 2010.

[16] N.Ward. Pragmatic functions of prosodic features in non-lexical utterances. In *Speech Prosody*, pp. 325–328, 2004.

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B303-02

発話者の音声に対応する動作生成と遠隔操作ロボットへの動作の付加効果

Online speech-driven head motion generation system and evaluation on a tele-operated robot

○境　くりま [*1,2], 石井　カルロス寿憲 [*2], 港　隆史 [*2], 石黒　浩 [*1,2]
Kurima SAKAI[*1,2], Carlos Toshinori ISHI[*2], Takashi MINATO[*2], Hiroshi ISHIGURO[*1,2]
ATR[*1], 大阪大学大学院 基礎工学研究科 [*2]
sakai.kurima@irl.sys.es.osaka-u.ac.jp, carlos@atr.jp, minato@atr.jp, ishiguro@sys.es.osaka-u.ac.jp

Abstract

本論文では，遠隔操作対話ロボットの頭部動作を操作者の音声情報のみから自動生成するシステムを提案する．遠隔対話では発話音声と一致した頭部動作の表現が必要となるため，発話の意味（相槌や発話の保持などの談話機能）を言語情報と韻律情報を用いてをリアルタイムで推定し，推定した談話機能に基づき頭部動作を生成する．提案システムには推定誤りが含まれ，対話に適さない動作が生成される場合がある．そのため，提案システムを用いた対話時の動作の印象を被験者実験により評価した．主観評価から，提案システムによる動作を付加することで，ロボットの動作がより対話に適したものになることが示された．

1　はじめに

電話やインターネットなどの通信技術の発達により，遠隔地にいる人といつでもどこでも対話することが容易になってきたが，そのようなコミュニケーションメディアを介した対話では，対話相手と対面しているように感じられず，円滑な対話が阻害される．円滑な遠隔対話の実現は，円滑な人間関係の構築につながる重要な課題である．我々は遠隔操作ロボットを用いて場の共有感や身体動作といった非言語情報を伝達することで，円滑な遠隔対話の実現を目指している．遠隔操作ロボットによる対話システムでは，操作者が遠隔地にいる人型ロボットを操作することで，ロボットが操作者の音声と身体動作を表現する．対話相手はそのロボットと対話することで，操作者と対面しているように感じながら対話を行うことができる（図 1）．

ロボットの身体動作の生成手法は，操作者の動きを計測しロボットの関節角にマッピングする手法が一般的である．しかし，この手法では「対話コンテキストに一致しない動作」と「ロボットの身体的制限による不完全な動作」の 2 つの動作が問題となる．例えば，前者の問題は，操作者が対話中に操作者の部屋の時計を見ると，ロボットも同じ動作をするが，遠隔地では操作者の部屋と同じ位置に時計があるわけではないため，対話者は動作の意味が分からなくなるという問題である．また，後者の問題は，ロボットの腕が可動範囲の制約で頭まで上がらないにもかかわらず，操作者が頭をかく動作を行うと，ロボットは不完全に動作を再現するため，対話者は意味が理解できなくなるという問題である．これら問題は，遠隔対話を阻害する要因となる．

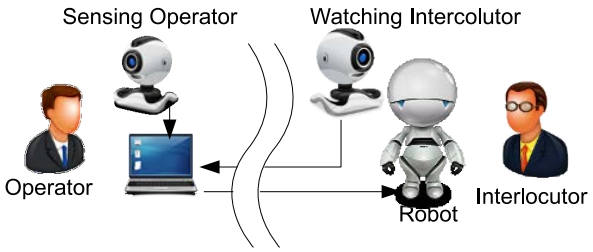


図 1:　Overview of the tele-communication by tele-operated robot.

解決策として，対話者が意味を理解できるようにロボットの動作を変換すればよいが，音声はそのまま伝達されるため，動作の変換が不適切で動作の意味が音声の意味と一致しないと，逆に対話者の理解を妨げることになる．音声との一致という制約を考慮すると，音声の意味を推定し，ロボットが表現可能な動作で表現すればよい．この方針の下，遠隔操作ロボットを用いた対話を行う際にシステムが満たすべき条件は，発話音声から発話の意味と一致する動作をリアルタイムで生成することである．本論文では，音声と身体動作の関係性の知見を基に，リアルタイムで発話の意味を推定し動作を生成するシステムを構築する．

音声情報から動作を生成するシステムはいくつか提案

されている．オフライン処理システムでは，ユーザの音声の韻律情報から，コンピュータグラフィックス上のアバタの動作を生成する試みがある．Sargin et al.[1] や Busso et al.[2] は，ユーザの音声の韻律情報から，アバタの頭部動作を生成するシステムを提案している．韻律情報に加え，Foster et al.[3] は対話コーパスを用いる手法を提案している．ロボットの動作生成では，Ishi et al.[4, 5, 6] は日常会話の頭部動作と談話機能（発話の意味）の関係性に基づく頷きと首傾げ動作生成モデルを提案している．

しかし，リアルタイムで音声から動作を生成する研究は少ない．Le et al.[7] は音声の韻律情報（ピッチとパワー）と頭部動作（首の角度，速度，加速度）を機械学習を用いて事前に学習し，リアルタイムで頭部動作を生成するシステムを提案している．Watanabe et al.[8, 9] は，音声の ON-OFF 情報を入力とした頷き動作の予測モデルを構築し，コンピュータグラフィックスのエージェントやヒューマノイドロボットの頷き動作をリアルタイムで生成するシステムを構築した．

これらの従来研究は，発話の意味に基づく動作生成とリアルタイム性を両立するものはない．Watanabe et al.[8] の手法では，頷き動作しか生成できず，発話者の否定の意味や困惑などを表現することができない．また，日本語（本論文の対象）では韻律特徴と頭部動作の相関が低いことが報告されている [10]．そのため，Le et al.[7] の手法のように，韻律情報を用いる手法は日本語音声から動作を生成する場合には適さないと考えられる．Ishi et al.[6] の手法を用いれば，発話の意味に合う動作が生成できると考えられる．彼らは，音声に手動で談話機能をラベリングしモデルを評価したが，リアルタイムで動作を生成するシステムは構築していない．そのため，本論文では発話者の音声からリアルタイムで談話機能を推定し，ロボットの頭部動作を生成するシステムを提案する．

2　頭部動作システムの構築

提案システムは，操作者の発話音声から談話機能を推定し，推定した談話機能に基づき頭部動作を生成する．本節では，まず Ishi et al.[4, 5, 6] の談話機能に基づく頭部動作生成モデルを説明する．次に，発話音声から談話機能を推定するための発話音声と談話機能の関係性，言語情報の抽出手法と韻律情報の抽出手法を説明する．最後に，言語情報と韻律情報を組み合わせ動作を生成するシステムを説明する．

2.1　談話機能と頭部動作の関係性

Ishi et al. はマルチモーダル対話音声データベースを用いて，頭部動作と談話機能の関係性を解析した [4, 5]．データベースには以下の Ishi et al. が提案した談話機能タグ [11, 12] が付与されている．

- k(keep): 発話権の保持（ポーズないしはっきりしたピッチのリセットが伴う強い句境界）
- k2(keep): 発話権の保持 2（発話文の中にある弱い句境界）
- k3(keep): 発話権の保持 3（話者が発話末の音節を伸ばし，考えていることや発話の途中であることを表現する場合）
- f(filler): 「えっとー」「あのー」などの感嘆詞を伴った，考え中であることの表現（フィラー）
- f2(conjunctions):「じゃ」などの感嘆詞を伴った，考え中であることの表現（短いフィラー）
- g(give): 発話権の譲渡（当話者の発話が終了し，発話権を対話相手への譲渡する場合）
- q(question): 発話権の譲渡 2（対話相手確認するなど応答を求める場合）
- bc(backchannels): 「うん」「はい」などの感嘆詞を伴った相槌の表現
- su(admiration/surprise/unexpectedness): 「へー」「うそ！」「ああ！」などの感嘆詞を伴った，驚きや感心の表現
- dn(denial, negation): 「いいえ」「ううん」などの感嘆詞を伴った，否定の表現

談話機能と頭部動作の関係性の分析によると，頷き動作が対話の中で最も多く生起し，特に頷き動作は相槌 (bc) や強い句境界 (k,g,q) で多く見られることが報告されている [4]．また，発話者が考えていたり，次の発話の準備をしているなどの場合では，語尾を延ばすことが多い．それら弱い句境界 (f,k3) では，首かしげ動作が最も多く出現したことも報告されている．さらに，驚きや感心を表す感嘆詞 (su) においても，顔上げ動作や首かしげ動作が頻繁に見られる．

2.2　談話機能と発話の関係性

日本語対話における談話機能と言語情報・韻律情報の関係性は分析されている．相槌 (bc) は「うん」「ええ」「ああ」「はい」のような感嘆詞を下降調トーンで発話することが多く，驚きや感心 (su) は「ええ」「へえ」「うん」を上昇調トーンで発話することが多く，フィラー (f) は「ええ」「へえ」「ううん」などを長く平坦調に発話することが多いことが報告されている [11, 12]．発話権を保持する強い句境界 (k) では,「で」「から」「けど」などの接続助詞を伴い，句末の音節のトーンが大きく下降する傾向にあることが報告されている [4]．これらの知見に基づき本論文では，“bc”，“su”，“f”，“k” の談話機能を推定し，頭部動作を生成するシステムを構築する．

2.3 言語情報の抽出

2.2 節で説明したように，談話機能の推定には言語情報と韻律情報が必要となる．本論文では，オープンソースである大語彙連続音声認識エンジン Julius[13] を用いて操作者の音声から言語情報を抽出する．Julius に付属する音響モデルは読み上げ音声を用いて作成されている．しかし，自然会話の「ああ」「ううん」「ええ」などの感嘆詞は，はっきり発音されないことが多いため，付属の音響モデルでは正しく認識することが困難である．感嘆詞の認識率を向上させるために，我々は自然対話データベースの音声データから感嘆詞を抽出し音響モデルを作成した．音響モデルの学習には，4406 フレーズ（男性:1903，女性:2503）の音声を用いた（「ああ」「ええ」「はい」「はあ」「へえ」「ほお」「うわあ」「わあ」「ううん」「いや」「いいえ」）．感嘆詞のモノフォン HMM（隠れマルコフモデル）モデルの作成には HTK(`http://htk.eng.cam.ac.uk/`) を用いた．韻律特徴は 12 MFCC（メル周波数ケプストラム），12 delta-MFCC，1 delta-power を用い，HMM の状態数は各感嘆詞の音素数にあわせ 8~16 とした．

遠隔対話における自由会話の言語モデルを作成することは困難であるため，本論文では言語モデルは検出したい相槌 (bc) や感心 (su) やフィラー (f) で見られる感嘆詞をキーワードとし，それ以外の音節を組み合わせる記述文法を用いた．また，漸次認識結果に音素アライメントを出力させるよう Julius のソースコードの改良も行った．

2.4 韻律情報の抽出

2.2 節で説明したように，談話機能を推定するには韻律情報（音調）も必要となる．そのため，基本周波数 (F0) の抽出を用いて，音調の識別を行った．

まず，F0 の値の抽出には，32 ms のフレーム幅で 10 ms 毎に LCP(Lear Predictive Coding) 逆フィルタによる残差波形の自己相関関数の最大ピークに基づいた処理を行う．さらに，人間のイントネーションの知覚特性と一致するよう，F0 の値を対数スケールに変換した．

$$F0[\text{semitone}] = 12 \times \log_2(F0[\text{Hz}]) \tag{1}$$

次に，音節内で F0 の変化量を表す $F0move$(人間の音調の知覚に基づくパラメータ [14]) を抽出した．$F0move$ は音節の後半の F0 の近似直線上の音節末の F0($F0tgt2b$) と前半部の F0 平均値 ($F0avg2a$) との差分を用いて計算する（式 2）．そして，音節の音調は式 3 に応じて，上昇調，下降調，平坦調に分類した．

$$F0move = F0tgt2b - F0avg2a \tag{2}$$

$$tone = \begin{cases} rising\ (Rs) & (F0move > 1 \text{ semitone}) \\ falling\ (Fa) & (F0move < -2 \text{ semitones}) \\ flat\ (Ft) & (\text{otherwise}) \end{cases} \tag{3}$$

2.5 リアルタイム音声駆動頭部動作生成システム

図 2 に実装したシステムの概要を示す．adintool（Julius に付属）はマイクロフォンから操作者の音声信号を取得し，音声のパワーと零交差に基づき音声区間のセグメンテーションを行う．音声情報は言語情報を取得するために Julius に送られ，また韻律情報を取得するために F0 抽出部へ送られる．リアルタイムで処理するために，Julius は 100 ms 毎に漸次認識結果を出力する．F0 値は 10 ms 毎に取得できる．動作生成部では，Julius からの音節アライメント情報に基づき，F0 情報を用いてキーワード区間の音調を識別する．抽出した言語情報と韻律情報に基づきロボットの頭部動作を生成し，ロボットにモータコマンドを送信する．すべてのモジュール間のデータ通信は TCP/IP を用いた．

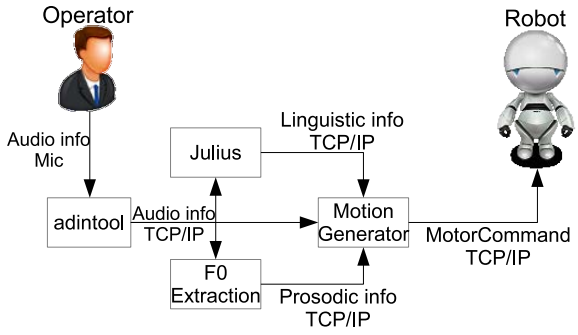


図 2: Overview of our online speech-driven head motion generation system.

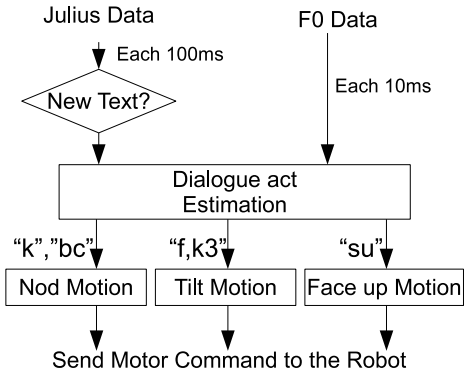


図 3: The system flow of the Motion Generator module.

図 3 に動作生成部のシステムフローを示す．Julius からの漸次認識結果は 100 ms 毎に取得できるため，新しく認

識された言語のみ処理する必要がある．そのため，漸次認識結果の最後の音節がすでに処理されたものかを確認するルールを設け，“bc”，“f”，“k3”，“su” は 100 ms 毎に処理される．一方で，“k” は Julius が発話終わりまたはショートポーズを認識した際に処理される．

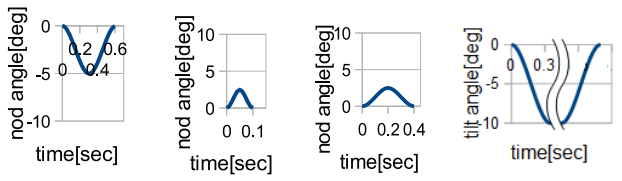
以下に各談話機能の推定と動作生成のルールを説明する．生成する動作の大きさと時間は従来研究の分析に基づいたものを用いた [4, 5]．

“bc” の推定は，「ああ」「あー」「ええ」「はい」「はあ」「へえ」「ほお」「ううん」「うん」といった感嘆詞が認識され，その発話区間の音調が下降調である場合とする．その際，図 4(a) に示す頷き動作を生成する．従来研究では，頷き動作には微かに頭部を上げてから下げる動きが観測されている [4]．しかし，首上げ動作は小さく，また動作生成の遅延をできる限り小さくするために，本論文では首上げ動作はないものとして首下ろし動作のみ実装した．

“k” の推定には，本来は「で」「から」「けど」などの接続助詞の抽出が必要になるが，現状の音声認識では接続助詞の抽出は現状困難である．そのため，句末やショートポーズの前の音節が下降調に発話された場合を “k” とした．

“su” の推定は，「ええ」「へえ」「ほお」といった感嘆詞が認識され，その発話区間の音調が上昇調である場合とする．その際に 2 種類の動作を生成する．1 つは図 4(b) に示す驚きを表す動作であり，他方が図 4(c) に示す感心を表す動作である．驚き動作は発話区間が短い場合に生成し，感心動作は発話区間が長い場合に生成する．

“f” と “k3” の推定では，一般的な会話での母音区間の長さは 200~300 ms であるため，1 音節の閾値を 350 ms とし，1 音節が閾値以上長く平坦調である場合を “f” と “k3” した．この際生成される首傾げ動作を図 4(d) に示す．従来研究 [6] 同様に，首傾げ状態は発話が終わるまで維持される．



(a) (b) (c) (d)

図 4: Head rotation angle shapes used in our head motion generation system: (a)Nod Motion; (b)Face up Motion (surprise); (c)Face up Motion (admire); (d)Tilt Motion.

3 遠隔操作ロボットを用いた提案システムの評価実験

3.1 実験目的

本節では提案システムの実用性を評価する．従来研究において，談話機能を手動でラベリングし，談話機能の認識に誤りがない場合，談話機能に基づき生成された頭部動作が自然と評価された [6]．しかし，提案システムは言語情報の制限や音声認識の誤りによる談話機能の推定に誤りがあるため，対話に不適切な動作が生成される可能性がある．それらは対話相手に対話コンテキストに適さない不自然な動作という悪い印象を与えてしまう．そのため，本実験では提案システムを用いた対話条件と用いない対話条件を比較し，提案システムにより生成される動作の対話への適切さを評価した．

3.2 実験設定

本実験では遠隔操作ロボットテレノイドを用いた（図 5）．テレノイドは首に 3 自由度（ピッチ軸，ロール軸，ヨー軸）と口に 1 自由度（上下開閉）のアクチュエータを持つ．また，テレノイドの外見は操作者に対する印象形成を乱さないような外見になっているため [15]，テレノイドを用いることで操作者の音声とロボットの外見との不適合による印象への影響を軽減することができる．

提案システムは操作者が発話しないとロボットが全く動かないため，正規雑音モデルに従って生成した指令値を首関節に常に入力し，人間の微小な不随意運動を表現した．500 ± 150N（N は標準偏差 0.1 の正規雑音）msec かけて 0.5 ± 0.5N（N は標準偏差 0.1 の正規雑音）度，首の 3 回転軸（ピッチ軸，ヨー軸，ロール軸）を移動させ，同じ時間をかけて元の位置に戻るというものである．不随意動作を頷き・首傾げ・首上げ動作に解釈されることを避けるため，約 1 秒のゆっくりした動きを用いた．また，テレノイドの口の動きは発話者音声に基づいて口唇動作を制御する手法 [16] を用いて，操作者の音声のみから口の開閉動作を生成する．

本実験では，操作者（実験協力者）が操作するテレノイドと対話する人が被験者となり，テレノイドの動作の印象を評価した．操作相手に対する印象を統制するために，操作者は 1 人（女性）とした．

被験者はテレノイドを通して操作者と対話し動作の適切さを評価した．比較条件は，テレノイドが口の開閉動作に加え上記の不随意運動を行う条件 (Noise) と，さらに提案手法による頭部動作を加えた条件 (Voice+Noise) である．被験者の話題に対する興味を統制するため，事前に被験者に「好きなスポーツ」「好きな映画」「好きな本」の 3 つから対話のテーマを 2 つ選ばせた．また，被験者自身が対話テーマを選ぶため，被験者に話し始めるよう教示した．この教示により被験者が話を率先することになり，被

験者はテレノイドから伝わる非言語情報から自分の話題に対する操作者の反応を把握しようとすると考えられる．このようにして，被験者にテレノイドから伝わる非言語情報に自然に注意を向けるように仕向けた．被験者には 2 条件とも体験してもらい被験者内比較を行った．また，操作者には自分がどちらの条件でテレノイドを操作しているのかわからないようにした．

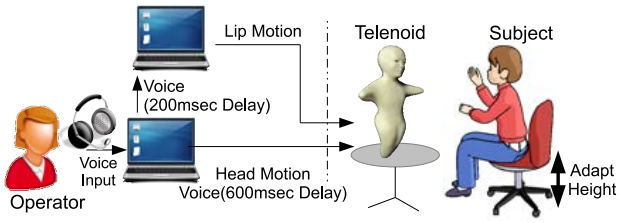


図 5: Experiment setup for evaluation of the proposed system using a tele-operated robot.

3.3 実験システム

図 5 に実験システムを示す．音声と動作を同期させるために 600 ms の音声遅延を設けテレノイド内のスピーカーから出力する．操作者の反応が対話者に 600 ms 送れて伝わるため，対話者の反応も 600 ms 程度遅れて操作者に伝わることになるが，予備的試行では，操作者はこの程度の遅れがあっても自然に振る舞うことができていた．また，Ishi et al.[16] の手法による口動作の生成の遅延は 400 ms の音声遅延が必要となる．そのため，頭部動作に合わせ口動作を生成するために，口動作を生成する PC には 200 ms の遅延をかけて操作者の音声を入力する．ロボットの見え方を統制するために，被験者は椅子の高さを調節しテレノイドと視線を合わせた（図 5）．

3.4 実験手順と評価指標

実験手順を以下に示す．

1. 対話ロボットの説明と遠隔操作の実演
2. 女性実験者が操作するロボットと自由に対話（3 分）
3. 1 回目テーマトーク（3~5 分）
4. 2 回目テーマトーク（3~5 分）
5. 1 回目と 2 回目のロボットの動作についての比較アンケート記入

1 回目のテーマトークと 2 回目のテーマトークでの動作条件はカウンタバランスを取った．被験者は 1 回目と 2 回目の動作のうちどちらが対話に適しているかを 7 段階（1: 1 回目の対話の方，4: どちらとも言えない，7: 2 回目の対話の方）で評価した．

3.5 主観評価実験結果

実験の被験者は 15 人（男 : 7 人，女 : 8 人，平均年齢 : 21.7，標準偏差 : 2.1）であった．動作音がうるさかったなどと，動作の評価に影響しそうなことに言及した被験者（2 人），実験中ロボットが故障した際の被験者（1 人）は解析から除いた．実験後のインタビューとアンケートの結果を比較し，Noise 条件の方が VoiseNoise 条件よりも動きが多かったと答えた被験者（2 人）も解析から除いた（被験者が順序を混同した可能性があるため）．解析に用いた被験者は 10 人（男 : 5 人，女 : 5 人，平均年齢 : 21.9，標準偏差 : 1.7）であった．

図 6 に主観評価の結果のヒストグラムを示す．アンケート結果の平均値が 4 より大きければ大きいほど，Voice+Noise 条件の動作がより適していることを表し，4 より小さければ小さいほど，Noise 条件の動作がより適していることを表す．そこで，平均値が 4 よりも大きいかどうかの検定を行った．解析データには Shapiro-Wilk 検定より正規性が認められたため，t 検定を行い，有意差が認められた ($t(9) = 3.10, p < 0.01$)．平均値は 5.4 であり，提案システム (Voice+Noise) の動作がより対話に適切であることが示された．

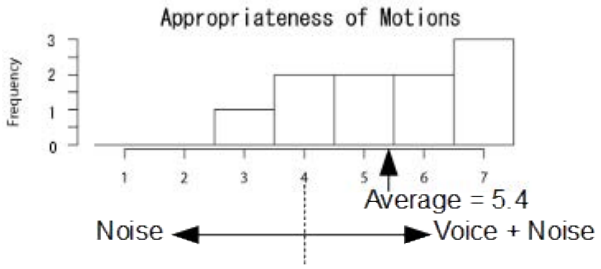


図 6: Subjective preference scores between the baseline motion (“Noise”) and the motion generated by the proposed system (“Voice+Noise”).

3.6 談話機能の誤認識分析

実験中の操作者の音声を録音し，Ishi et al.[5] の手法に従って別の実験協力者（1 人）が手動で談話機能タグを付けたものを正解データとした．この実験協力者はどの音声データがどちらの条件のものかは知らされていない．“bc” と “k” は同じ頷き動作を生成するため，“bc” と “k” の混同は対話者にとっての印象形成には影響しない．そのため，対話相手の印象形成への影響を考慮し，談話機能の推定の誤りではなく，生成された動作の誤りについて分析した．

表 1 に正しい談話機能を基にした動作生成タイミング (expected motions) において，システムが生成した動作 (generated motions) の個数を示す．頷き動作は 55%，首上げ動作は 50%で提案システムは正しく動作を生成でき

ていた．これは，頷き (bc) と驚き・感心 (su) についての感嘆詞とイントネーションを正しく検出できたためであると考えられる．しかし，首傾げ動作については 10%以下の正解率であった．これは，フィラー (f) や言いよどみの推定には音節の長さと音調のみ使用したため，音節の長さが短い場合の言いよどみが検出できなかったと考えられる．一方で，多くの頷きタイミングで首傾げ動作を生成していたり，どの動作も期待されないフレーズで頷き動作の生成も見られた．

表 1: Distributions of the generated motions for each of the expected motions (when the dialogue act tags are given).

	expected motion			
generated motion	nod	tilt	face up	insertions
nod	354	0	35	341
tilt	43	3	18	13
face up	113	0	83	36
no motion (deletions)	135	36	28	

4 考察

本実験により，提案システムによりリアルタイムで音声から動作が生成できることが確認できた．さらに，本システムにより動作を付加することで，誤り動作が含まれるにもかかわらず，より適切な印象を与えることが明らかになった．

提案手法は談話機能と頭部動作を一対一でマッピングしている．しかし，人間同士の対話では相槌の際に頷かない場合があるなど，提案システムのように談話機能と頭部動作が一対一に対応しているわけではない [11]．そのため，削除誤り（動作が生成されない場合）があっても評価が下がらなかったと考えられる．一方で，置換誤り（例：首傾げタイミングで頷いてしまう）や挿入誤り（例：フィラー以外で首をかしげる）は対話の意味と異なる意味を伝達するため，遠隔対話を行う上でより問題となる．しかし，頷き動作の意味は相槌のみならず，発話の強調や相手の頷きに対する反応や相手の応答を促すなど多岐にわたるため [17, 18]，頷き動作の挿入誤りや置換誤りが不自然と評価されなかったと考えられる．一方で，首傾げ動作は自然会話で生起するタイミングは頷き程多くはなく，その意味も頷き動作ほど多くはない．そのため，実験インタビューにおいても「頷き動作は変ではなくリアルで自然であった．首の動かし方は人それぞれなので不気味とかは感じなかったが,「首をこのタイミングで動かすか」と思うことはあった.」と答える被験者がいた．ただし，首傾げ動作とかわいらしい声という関係性から「声がかわいらしかったのでそれにマッチするかわいらしいしぐさ」と首傾げ動作を解釈する被験者もいた．前者の被験者は，誤認識動作を不適切と思っているが，後者の被験者はそれを操作者の癖と思っていると考えられる．そのため，誤認識があっても適切と判断される理由として，後者も一要因として考えられる．さらにこの結果から，ロボットの動作生成においては，音声との適合だけでなく，操作者のイメージとの適合も重要な要素であることが判った．

多少の誤り動作も許容されることが分かったが，置換誤り・挿入誤りは対話相手にとっては解釈できないものとなり得るため，将来研究としてそれら誤りを減らすよう制約を加えることが重要となる．また，ロボットの見た目によって求められる人間らしい動作の程度が変わると考えられるため（例：人らしいほど人間らしく動いてほしい），本論文で用いたテレノイド以外のロボットを用いた際の印象評価も行う．本実験では，操作者を 1 人に固定しているため，複数の操作者を用いた場合における提案システムの実用性の評価も重要である．さらに，頭部動作のみならず表情や視線といった他のモダリティの自動生成も行う．

5 まとめ

本論文では，円滑な対話を実現するために，遠隔操作ロボットの新たな動作自動生成手法を提案し，その有効性を心理実験により検証した．操作者音声の談話機能に適合した動作（頷きや首傾げなどの頭部動作）を自動生成したところ，操作者本人動作と同じ動作でないにも関わらず，対話において適切な動きと評価されることが判った．これはボディジェスチャや視線や表情といったモダリティを音声から生成できる一例となる．将来研究として，音声からそれら他のモダリティの生成が課題となる．

提案手法は操作者の音声情報のみ用いるため，操作者は携帯電話などの小型通信機器での遠隔操作も可能である．また，四肢が不自由なため身体動作による意思疎通が困難な人たちも，音声のみでロボットアバタに動作を表現させることで，円滑なコミュニケーションが可能である．このように，提案手法は新たな遠隔コミュニケーションサービスに役立つはずである．

参考文献

[1] Mehmet Emre Sargin, Oya Aran, Alexey Karpov, Ferda Ofli, Yelena Yasinnik, Stephen Wilson, Engin Erzin, Yücel Yemez, and Ahmet Murat Tekalp. Combined Gesture-Speech Analysis and Speech Driven Gesture Synthesis. *International Conference on Multimedia and Expo*, 2006.

[2] Carlos Busso, Zhigang Deng, Michael Grimm, Ulrich Neumann, and Shrikanth Narayanan. Rigid

Head Motion in Expressive Speech Animation: Analysis and Synthesis. *Transactions on Audio, Speech, and Language Processing*, Vol. 15, No. 3, 2007.

[3] Mary Ellen Foster and Jon Oberlander. Corpus-based generation of head and eyebrow motion for an embodied conversational agent. *Language Resources and Evaluation*, Vol. 41, No. 3-4, pp. 305–323, 2007.

[4] Carlos Toshinori Ishi, Judith Haas, Freerk P. Wilbers, Hiroshi Ishiguro, and Norihiro Hagita. Analysis of head motions and speech, and head motion control in an android. In *IROS2007*, pp. 548–553, 2007.

[5] Carlos Toshinori Ishi, ChaoRan Liu ChaoRan Liu, H Ishiguro, and N Hagita. Head motion during dialogue speech and nod timing control in humanoid robots. In *HumanRobot Interaction*, pp. 293–300, 2010.

[6] Chaoran Liu, Carlos Toshinori Ishi, H Ishiguro, and N Hagita. Generation of nodding, head tilting and eye gazing for human-robot dialogue interaction. In *HumanRobot Interaction*, pp. 285–292, 2012.

[7] Binh Huy Le, Xiaohan Ma, and Zhigang Deng. Live Speech Driven Head-and-Eye Motion Generators. *Transactions on Visualization and Computer Graphics*, Vol. 18, No. 11, pp. 1902–1914, 2012.

[8] Tomio Watanabe, Masashi Okubo, Mutsuhiro Nakashige, and Ryusei Danbara. InterActor: Speech-Driven Embodied Interactive Actor. *International Journal of Human-Computer Interaction*, Vol. 17, No. 1, pp. 43–60, 2004.

[9] Hiroki Ogawa and Tomio Watanabe. InterRobot: a speech driven embodied interaction robot. *RO-MAN2000*, pp. 322–327, 2000.

[10] Kevin G. Munhall, Jeffery A. Jones, Daniel E. Callan, Takaaki Kuratate, and Eric Vatikiotis-Bateson. Visual prosody and speech intelligibility: head movement improves auditory speech perception. *Psychological science*, Vol. 15, No. 2, pp. 133–137, 2004.

[11] Carlos Toshinori Ishi, Hiroshi Ishiguro, and Norihiro Hagita. Analysis of prosodic and linguistic cues of phrase finals for turn-taking and dialog acts. *INTERSPEECH*, 2006.

[12] Carlos Toshinori Ishi, Hiroshi Ishiguro, and Norihiro Hagita. Automatic extraction of paralinguistic information using prosodic features related to F0, duration and voice quality. *Speech Communication*, Vol. 50, No. 6, pp. 531–543, 2008.

[13] Akinobu Lee, Tatsuya Kawahara, and Kiyohiro Shikano. Julius — an Open Source Real-Time Large Vocabulary Recognition Engine. In *EUROSPEECH*, pp. 1691–1694. ISCA, 2001.

[14] Carlos Toshinori Ishi. Perceptually-Related F0 Parameters for Automatic Classification of Phrase Final Tones. *IEICE transactions on information and systems*, Vol. 88, No. 3, pp. 481–488, March 2005.

[15] Kaiko Kuwamura, Takashi Minato, Shuichi Nishio, and Hiroshi Ishiguro. Personality distortion in communication through teleoperated robots. In *RO-MAN2012*, pp. 49–54, 2012.

[16] Carlos Toshinori Ishi, Chaoran Liu, Hiroshi Ishiguro, Norihiro Hagita, Intelligent Robotics, and Communication Labs. Evaluation of formant-based lip motion generation in tele-operated humanoid robots. *IROS2012*, pp. 2377 – 2382, 2012.

[17] Joseph D. Matarazzo, Arthur N. Wiens, George Saslow, Bernadene V. Allen, and Morris Weitman. Interviewer mm-hmm and interviewee speech durations. *Psychotherapy: Theory, Research & Practice*, Vol. 1, No. 3, pp. 109–114, 1964.

[18] 庵原彩子, 堀内靖雄, 西田正史, 市川熹. 自然対話におけるうなずきの機能に関する考察 (分析、生成と評価)(音声とコミュニケーション及び一般). 電子情報通信学会技術研究報告. HCS, ヒューマンコミュニケーション基礎, Vol. 104, No. 445, pp. 13–18, 2004.

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B303-03

耳珠のある能動耳介システムとその動作について

Active pinnae with tragus and their motion

公文誠，尾堂航，木元大輔
Makoto KUMON, Wataru ODO, Daisuke KIMOTO
熊本大学大学院自然科学研究科
Graduate School of Science and Technology, Kumamoto University
kumon@gpo.kumamoto-u.ac.jp

Abstract

本研究では二つのマイクロホンを用いて音源定位を行うロボットシステムとして，マイクロホンの周囲に可動式の反射板である能動耳介を利用するものを考える．耳介の姿勢に伴って伝達特性が変化することを活用することで，音源定位性能の向上を目指すが，このためには伝達特性の変化が顕著であることが重要である．動物では耳珠が耳介の効果を強調することが知られているため，本研究では音源定位に向けた能動耳介の効果改善のため耳珠に相当する部位を導入することを提案する．実際に耳珠を備えた能動耳介システムを開発し，音源定位のための動作に反映するための基礎性能を調べたのでこれを報告する．

1 はじめに

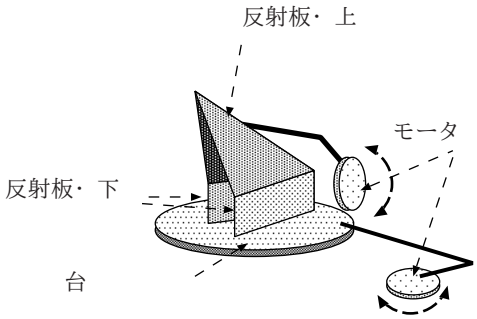
ロボットが自律的に動作するには，判断のためにロボット周囲の環境を適切に認識することが不可欠である．環境認識の代表的な方法はカメラやレーザ距離計などを用いた光学的な方法が数多く提案されているが，人間の住環境のように壁面など障害物によるオクルージョンの問題や，カメラの視野外は認識できないなど課題もある．一方，音信号を用いることが出来れば，対象が障害物の隠れるような場合でも，これを見つけることが可能であると同時に，自動車のクラクションや電話の呼び出し音のような音記号の理解を通じた環境理解も可能となる．音を通じた環境理解の重要な基礎能力に，音源の位置や方向を推定する音源定位があり，ロボットに向けてはマイクロホンアレイと音響信号処理による方法が種々提案されている．

さて，生物の場合の音源定位を考えると，多くは実用的な範囲を二つの耳で実現している．二つに限られた受聴点の信号から音源定位を実現する生物の能力は興味深く，このような機能を実現する機序の一つに，頭部など身体を動かすことが指摘されている (Blauert[Blauert 96] など)．さらに，ネコなど一部の動物では，耳介そのものを随意筋によって能動的に動作させることが可能であるが，この耳介動作によっても音源定位能が改善されることが知られている (Populin[Populin 98],Heffner[Heffner 82] など)．

このような生物に倣って，二つあるいは少数のマイクロホンを用いたロボットシステムで高性能の音源定位を実現しようとする試みがある．例えば Kim ら [Kim 13] は頭部動作を利用することで音源方向の推定性能が向上することを示している．また，マイクロホンが二つに限られると正中面内の方向が認識できないことが指摘されているが，著者ら [Kumon 05] や Hörnstein[Hörnstein 06] などは耳介を利用することで音源の上下を区別する方法を提案している．あるいは，動く耳介を活用する先行研究として，著者ら [Kumon 11] や 金ら [金 12] の研究が挙げられる．

これらの動作する耳介を用いたロボットシステムでは，耳介の動作に伴って音源からマイクロホンまでの音信号の伝達特性が変化するが，この伝達特性の変化が既知であれば，例えば音信号の到来方向を認識するなどが可能となる．また，著者らの音源の上下を区別する手法 ([Kumon 05]) では，耳介による伝達要素の周波数特性を利用するものだったので，より望ましい特性を呈する耳介の姿勢を求められれば，音源定位能が改善されることが期待される．一方，このことが機能するためには，耳介の姿勢変化によって十分に伝達特性の変化が明瞭であることが必要である．これまでの著者らの結果 [Kumon 11] では，主に高周波数帯域のパワーが姿勢変化に伴ってなだらかに変化することが示されていたが，この変化を顕著にすることが定位性能に貢献することが考えれる．

ところで，動物の外耳には耳介以外にも様々な部位があり，耳珠と呼ばれるものが存在する．耳珠とは動物などの耳に見られる耳孔前方にある小突起で，耳介の効果を強調するとの指摘がある [本田 85]．そこで本研究では能



(a) 機構

(b) 能動耳介全景

(c) 片側の拡大図

図 1: 耳珠を持つ能動耳介システム

動耳介に合わせて耳珠を耳介開口部に設け，姿勢による周波数特性の変化を強調することを考える．まず，これまでに提案した能動耳介のシステムに耳珠に相当する構造を取り付けた耳介システムを開発し，この特性を調べることとした．また，得られた特性の評価として，いくつかの耳介の姿勢で音収録を行い音源の方向を推定することで行った．

2 能動耳介と耳珠

2.1 実験装置

まず本研究で用いる能動耳介装置について説明する．図 1 は装置についてまとめたものである．

図 1(a) に示すように耳介は基礎を成す台の上に取り付けられており，アルミ板の反射板を組み合わせたテント型の構造になっている．反射板は上下に分かれており，上部はモータによって姿勢を前後に動かすことが出来，下部は台に固定されている．本研究では用いないが，耳介を取り付けている台もモータによって動かすことが可能で，耳介の左右方向の自由度に対応する．下部の台の中央にはマイクロホンが上向き（耳介の上部方向）に取り付けられている．このような能動耳介を二つ組み合わせ，マネキン頭部に搭載した状態で実験に供した (図 1(b))．

本研究の中心となる耳珠 [本田 85]. は，図 1(c) に示される写真の白い部分であり，上述する耳介の開口部の下半分ほどを覆う正方形の板で模擬した．各耳珠は 4mm 厚のプラスチック板で，以下では 30mm 四方のものと 40mm 四方のものの二種類を検討している．

2.2 耳珠の効果：周波数特性

耳介の姿勢の変化で周波数特性がどのように変化するかを確認するため耳珠を取り付けた場合と取り付けていない場合での伝達特性を検証した．音源は頭部の正面方向 1.5m の距離に設置するものとし，頭部の正面高さを基準位置として仰伏角方向に $-20°$ ～ $20°$ まで $10°$ 刻みで計 5 点から白色雑音を試験信号として用いた．また，それぞれの音源位置ごとに耳介の姿勢を右耳 6 パターン，左耳 6 パターンの計 36 パターンで収録した．なお，音信号はサンプリング周波数 44.1kHz で録音しており，再生装置の制限で呈示される白色信号は 16kHz までの帯域に制限されている．

音源の方位角については, IPD を用いて比較的良好に推定ができると想定できるので，ここでは特に二つのマイクロホンで難しいと考えられる音源の仰角方向について検証することとし，ILD に相当するパワースペクトルに着目する．

図 2 に収録した白色信号のスペクトログラムを示す．図の列は，左から順に耳珠を取り付けなかった場合，小型 (30mm 四方) の耳珠を取り付けた場合，大型 (40mm 四方) の耳珠を取り付けた場合の結果を示しており，図の行は音源の仰角に対応する．各図の横軸は時間，縦軸が周波数を表わし，図中の色によってパワーを示している．

いずれの場合も縞状のパターンが見られ，安定したスペクトル構造が得られている. また，耳珠を取り付けることでパターンが変化していることも確認される．特に耳介の効果は高周波数帯域 (10kHz 以上) で顕著であり，耳珠も同じ帯域で影響が際立っている（黒枠）．具体的な変化を挙げると，音源方向が $-20°$ ～ $-10°$ のとき耳珠のない場合では周波数特性が酷似している一方，耳珠を取り付けた場合，13kHz 付近のパワーが小さくなっており，これらの別が明瞭であること，また，音源方向が $0°$ ～ $20°$ のときには 14 ～ 16 kHz にノッチがはっきりするなどがある．音源方向の推定法では，この周波数特性の違いを手がかりに定位を行うので，このように特性の変化が顕著になることは重要である．この観点からすると，逆に耳珠の大きなものでは，$-10°$ ～ $0°$ の方向での周波数特性の変化が乏しく，耳珠の小さなものに比べ適当とは言い難い．このことから，単に耳珠を取り付ければ良いというわけではなく，適当な大きさのものを選択する必要があることが示唆される．

先の条件に替えて，耳介を前方に傾斜させた姿勢 ($\boldsymbol{\rho}_2$) でも同様の試験を行った．この結果を図 3 に示すが，先と同様に耳珠の影響が確認される．例えば音源方向が $-20°$ および $-10°$ ～ $0°$ のとき，それぞれ 10 ～ 16kHz と 1.5 ～ 1.6kHz にノッチが見られることが挙げられる．特に重要なのは，同じ音源方向であっても耳介の姿勢が $\boldsymbol{\rho}_1$ から $\boldsymbol{\rho}_2$ に変化することで周波数特性がはっきり変化する

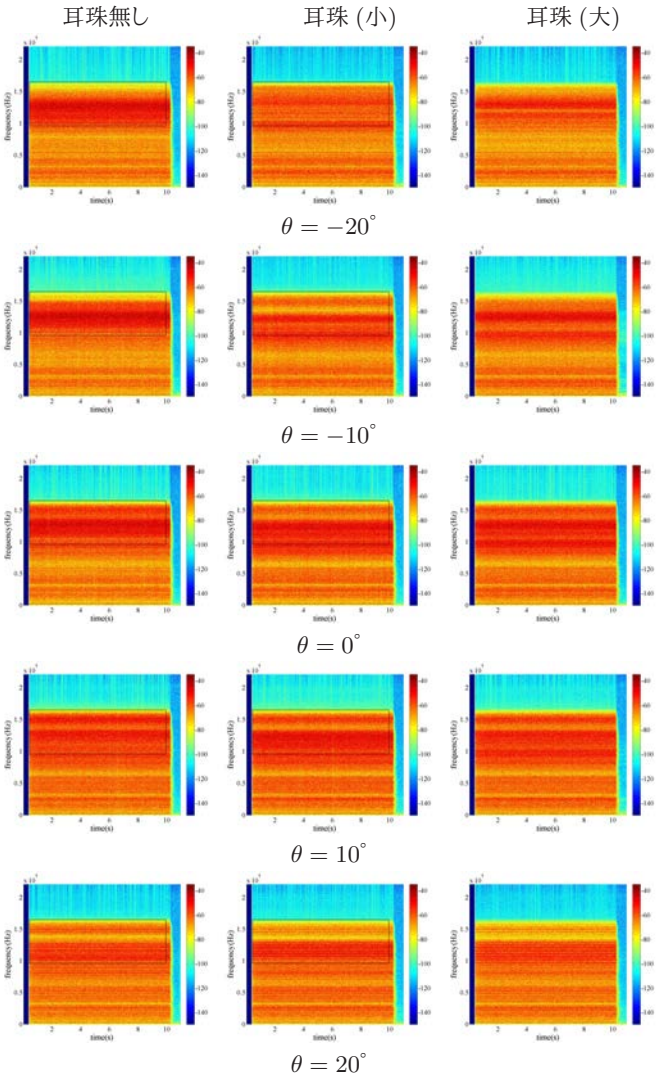


図 2: スペクトログラム(耳珠の影響: 姿勢 $\boldsymbol{\rho}_1$)

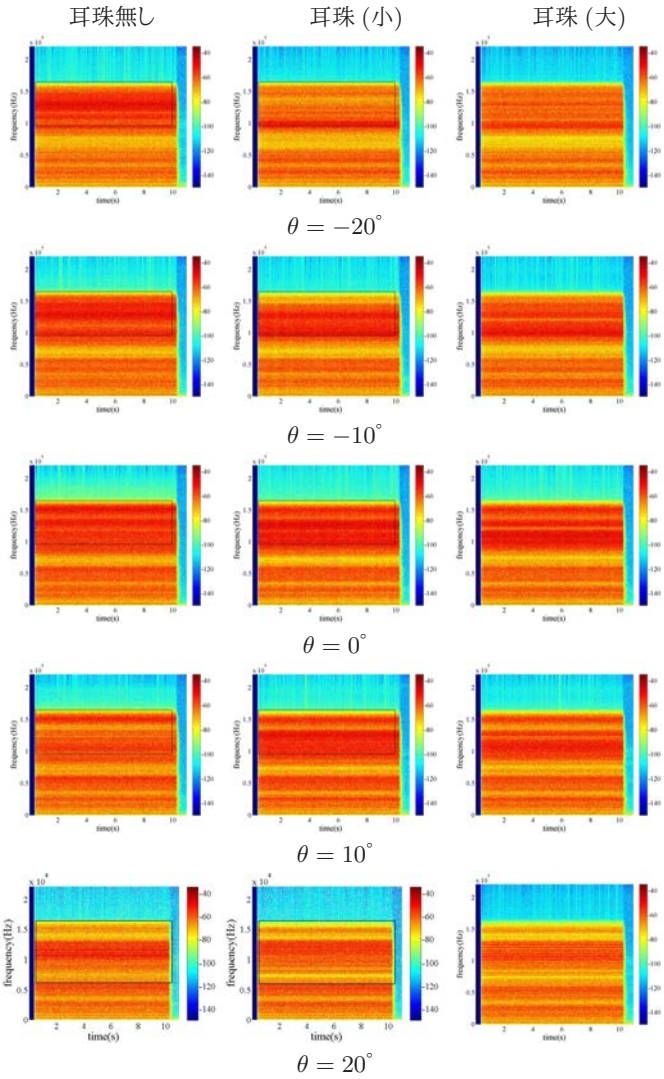


図 3: スペクトログラム(耳珠の影響: 姿勢 $\boldsymbol{\rho}_2$)

ことである．例えば $\theta = -20°$ の場合，図 2，3 を比較すれば，耳珠を取り付けない場合では周波数特性はほとんど変化しないが，耳珠 (小) を取り付けた場合，耳介を傾ける ($\boldsymbol{\rho}_2$, 図 3) と 10kHz 付近にピークが出現するといった明瞭な変化がある．音源方向を識別する上で耳介の方向を能動的に操作する効果が生じることになるので，能動耳介と耳珠の組み合わせが有用であると言える．

3 音源方向の推定

前節に説明した耳珠を有する能動耳介システムを用いて音源方向の推定を行う．本研究では頭部正面を基準とする座標系から見た音源の方向を仰角，方位角で考え，それぞれを θ, ϕ と表すことにする．また，左右の耳介の姿勢を $\boldsymbol{\rho} := (\rho_l, \rho_r)$ で表すものとする．

以下ではまず定位の方法を説明し，その後，実際の定位結果を示す．

3.1 推定方法

二つのマイクロホンによる音源方向の推定に用いる特徴量として両耳間レベル差 (Interaural Level Difference; ILD) と位相差 (Interaural Phase Difference; IPD) が良く知られている [Blauert 96]．ここでは，ILD, IPD を周波数領域上の N 個の周波数点上で表現した特徴量ベクトルとして扱うこととし，これらの量を $\boldsymbol{z}_{\mathrm{ILD}}, \boldsymbol{z}_{\mathrm{IPD}}$ と表す．

定位を行うために，音源方向と対応づけて，これらの音響特徴量を事前に測定したデータベースを作成しておく．これらの特徴量のことを以下では規範の特徴量と呼び，(θ, ϕ) 方向の規範の ILD, IPD を $\boldsymbol{z}^d_{\mathrm{ILD}}(\theta, \phi, \boldsymbol{\rho})$, $\boldsymbol{z}^d_{\mathrm{IPD}}(\theta, \phi, \boldsymbol{\rho})$ のように添字 ^{d} を付して表すものとする．

耳介の姿勢 $\boldsymbol{\rho}$ にあって測定された ILD, IPD の特徴量 $\boldsymbol{z}_{\mathrm{ILD}}, \boldsymbol{z}_{\mathrm{IPD}}$ から音源方向を推定するため，ILD, IPD について次のような尤度を考える．

$$
\begin{aligned}
& l_{\mathrm{ILD}}(\theta, \phi | \boldsymbol{\rho}, \boldsymbol{z}_{\mathrm{ILD}}) \qquad (1)\\
&= \sum_{i=1}^{N} \exp\left\{ -\left(z_{\mathrm{ILD},i} - z^d_{\mathrm{ILD},i}(\theta, \phi, \boldsymbol{\rho}) \right)^2 \right\},
\end{aligned}
$$

$$l_{\mathrm{IPD}}(\theta,\phi|\boldsymbol{\rho},\boldsymbol{z}_{\mathrm{IPD}}) \quad (2)$$
$$= \sum_{i=1}^{N} \exp\left\{\cos\left(z_{\mathrm{IPD},i} - z^{d}_{\mathrm{IPD},i}\left(\theta,\phi,\boldsymbol{\rho}\right)\right) - 1\right\}.$$

ここで i は各特徴量の i 番目の要素を示す．次に，これらの尤度をまとめて結合尤度を

$$l(\theta,\phi|\boldsymbol{\rho},\boldsymbol{z}_{\mathrm{ILD}},\boldsymbol{z}_{\mathrm{IPD}}) \quad (3)$$
$$= l_{\mathrm{ILD}}(\theta,\phi|\boldsymbol{\rho},\boldsymbol{z}_{\mathrm{ILD}}) l_{\mathrm{IPD}}(\theta,\phi|\boldsymbol{\rho},\boldsymbol{z}_{\mathrm{IPD}}).$$

とする．

今，k 回の観測の後，音源方向について事前分布 $p_k(\theta,\phi)$ が与えられたとし，$k+1$ 回目の観測によって (3) の結合尤度が得られたとすると，繰り返しベイズ推定による音源定位法 [Kumon 13] に従って事後分布を

$$p_{k+1}(\theta,\phi) = \frac{l(\theta,\phi|\boldsymbol{\rho},\boldsymbol{z}_{\mathrm{ILD}},\boldsymbol{z}_{\mathrm{IPD}})p_k(\theta,\phi)}{\int l(\theta,\phi|\boldsymbol{\rho},\boldsymbol{z}_{\mathrm{ILD}},\boldsymbol{z}_{\mathrm{IPD}})p_k(\theta,\phi)d\theta d\phi} \quad (4)$$

のように更新する．

3.2 定位実験

ここでは，音源を仰角方向に 5 通り，方位角方向に 5 通りの計 25 点を対象に音源方向の推定（音源定位）を行うこととし，データベースには先と同じ白色雑音による規範データを作成し，試験信号には図 4 に示す音楽信号を用いることとした．

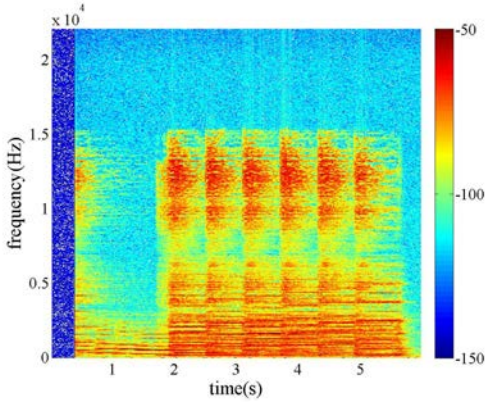
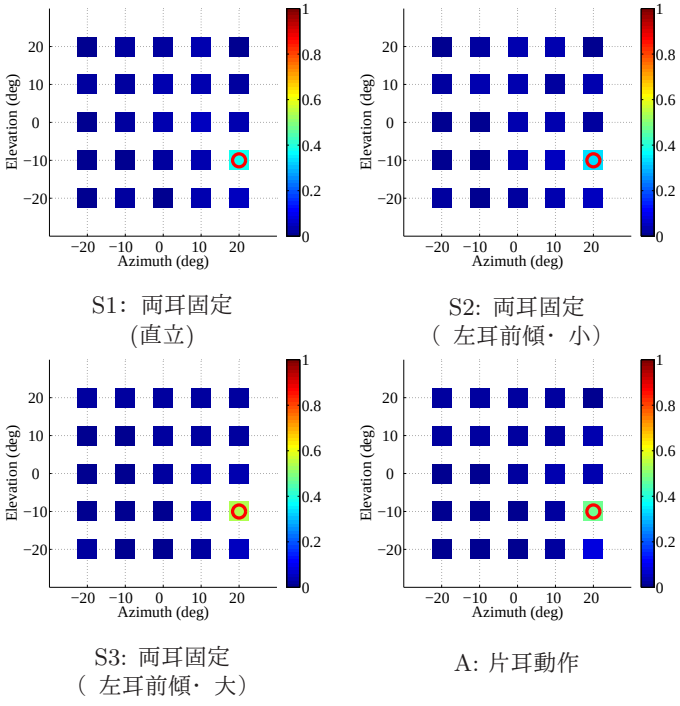


図 4: 定位実験に用いた信号

まず，音源を $(\theta,\phi) = (-10°, 20°)$ に置いた時，耳介を S1: 両方の耳を直立させた場合 (姿勢 $\boldsymbol{\rho}_1$ とする), S2:右耳を直立，左耳を前方に小さく傾斜させた場合（同 $\boldsymbol{\rho}_2$), S3:右耳を直立，左耳を前方に大きく傾斜させた場合（同 $\boldsymbol{\rho}_3$)，A: 右耳を直立させた状態，左耳を観測毎に直立させた状態から前方に6 段階に傾斜させた場合の 4 つの条件で推定を行った．観測の回数は S1, S2, S3, A のいずれの場合も 6 回であり，初期の事前分布には一様分布を用いた．ここで A は片耳が取る全ての場合で一度ずつ音信号を取得するもので，基本的な耳介動作の一つである．



S1: 両耳固定（直立）　S2: 両耳固定（左耳前傾・小）

S3: 両耳固定（左耳前傾・大）　A: 片耳動作

図 5: 音源方向の推定結果

適当な時間区間の信号に対する定位結果として，各条件での6 回目の事後分布 p_6 を図 5 に示す．各図は 25 個のセルからなっており，各セルが仰角・方位角に対応し，そのセルの色が音源の事後分布の値を示している．図中の赤丸が音源の正解方向を示しており，S1,S2,S3,A のいずれの場合も音源の方向に対応するセルの値だけが大きくなっている．この例では耳介の姿勢や動作に依らず，音源方向が正しく定位されているとも見えるが，$\boldsymbol{\rho}_3$ では推定された値がやや小さいなどの違いがある．紙面の都合で割愛したが，他の音源方向では推定結果にばらつきがあるなど，図 5 のような事後分布だけでは性能を測ることが出来ないため，定位性能を評価する系統的な指標を導入することとした．

収録された音信号に独立な白色雑音を混合し，騒音下での音源定位を模擬した場合を考える．なお，各試行毎に S/N 比 (SNR) が一定となるよう収録音の各時間区間のパワーに合わせて白色信号のパワーを調整している．また，音源定位性能については以下で定義される F 値 [Chinchor 92] を用いることとした．

$$F = \frac{2RP}{R+P}, \quad (5)$$

ここで

$$P = \frac{\#\text{ of estimated sound sources}}{\#\text{ of estimated sound source candidates}}$$
$$R = \frac{\#\text{ of estimated sound sources}}{\text{Total }\#\text{ of sound sources}}.$$

で，F 値は $[0,1]$ の範囲を取り，大きいほど定位性能が良いことを示す指標である．

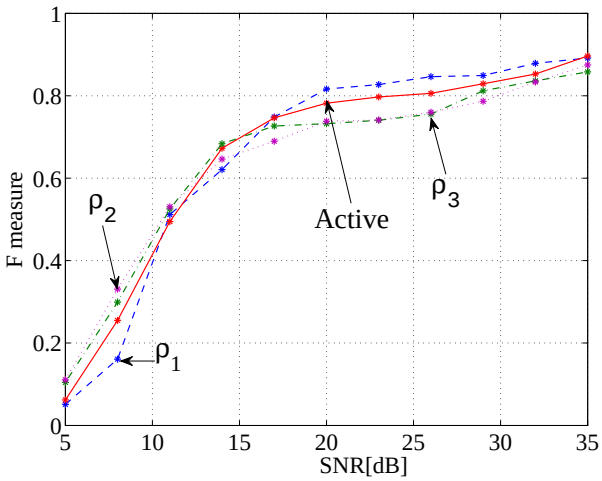


図 6: 音源定位のロバスト性

SNR の変化の下での定位性能を図 6 に示す．図の横軸が SNR，縦軸に F 値を示しており，先の S1 と S2，S3, A の各条件での性能をまとめて表示している．

収録環境が良好な場合 (SNR が大きい場合)，どの場合でも同じ程度の定位性能が得られている．20dB 以上の領域では，両耳を直立させた $\boldsymbol{\rho}_1$ が最良の結果を示しており，次いで動作させた場合 (A), 片方の耳介を前に傾斜させた場合 ($\boldsymbol{\rho}_2$, $\boldsymbol{\rho}_3$) の順になっている．これは，音信号が正確に得られる場合は，耳介を立て開口部を広く保つ $\boldsymbol{\rho}_1$ が有効であるので，自然な結果とも言える．一方，SNR が 15dB 付近の悪条件では状況が異なり，$\boldsymbol{\rho}_1$ は SNR の低下とともに急激に定位性能が劣化し，逆に耳介を伏せた $\boldsymbol{\rho}_2$ や $\boldsymbol{\rho}_3$ の方が F 値の点ではやや良い結果となった．ここで，耳介を観測毎に前方へ傾斜させた場合 (A の場合)，$\boldsymbol{\rho}_1$ と $\boldsymbol{\rho}_2$ や $\boldsymbol{\rho}_3$ の中間的な性質を示しており，SNR が高いところでは $\boldsymbol{\rho}_1$ に近い性能を呈するとともに，SNR の低下時は $\boldsymbol{\rho}_2$ や $\boldsymbol{\rho}_3$ に近い性能であった．このことは，耳介の姿勢を変化させることで，音源の方向や雑音の程度に応じて伝達特性を変化させることで，それぞれに適した観測を活用し，ロバストな音源定位を実現できる可能性を示唆している．なお，SNR が 12dB 以下になると F 値は 0.5 以下と定位性能は悪く，ここでの F 値の大小を論じることは意味のあることとは言えない．

4 おわりに

本研究では動物などをヒントに耳介を能動的に動作させることで，音源定位性能を改善することを目的とし，耳珠の導入を提案するとともに，実際の音源定位においてその効果を調べた．その結果，耳珠を取り付けることで，単に伝達特性が変化するだけでなく，特に仰角方向の音源方向の推定に重要な耳介ノッチが明瞭になること，また耳介動作による伝達特性の変化が明らかになることを実際の装置によって確認した．また，耳介を観測の度に少しずつ前方に傾斜させる動作を考え，この動作の下で音源定位を行った場合の定位性能を調べた所，雑音下でロバスト性を改善出来ることが示され，耳介動作が音源定位性能の向上に繋がる例を挙げることが出来た．

今回は最も基本的な動作として前方に傾斜し続ける動作を考えたが，実際には定位情報にあたる事後分布を用いて次に取るべき姿勢を計算することが効果的だと考えられる．この音源定位の観点からどのような姿勢を良いとするかは今後の検討が必要である．

また，SNR が著しく低下した場合では定位性能が悪くなるが，耳介姿勢を適切に選ぶことで改善が可能か，可能であればどのようにすれば良いかについても今後の課題である．

参考文献

[Blauert 96] Blauert, J.: *Spatial Hearing - Revised Edition: The Psychophysics of Human Sound Localization*, The MIT Press, rev sub edition (1996)

[Chinchor 92] Chinchor, N.: MUC-4 EVALUATION METRICS, in *Proceedings of Fourth Message Understanding Conference (MUC-4)*, pp. 22–29 (1992)

[Heffner 82] Heffner, R., Heffner, H., and Stichman, N.: Role of the elephant pinna in sound localization, *Animal Behaviour*, Vol. 30, No. 2, pp. 628–629 (1982)

[Hörnstein 06] Hörnstein, J., Lopes, M., Santos-Victor, J., and Lacerda, F.: Sound localization for humanoid robots - building audio-motor maps based on the HRTF, in IEEE, ed., *Proceedings of 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1170–1176 (2006)

[Kim 13] Kim, U.-H., Nakadai, K., and Okuno, H. G.: Improved Sound Source Localization and Front-Back Disambiguation for Humanoid Robots with Two Ears, in *IEA/AIE*, pp. 282–291 (2013)

[Kumon 05] Kumon, M., Shimoda, T., Kohzawa, R., Mizumoto, I., and Iwai, Z.: Audio Servo for Robotic Systems with Pinnae, in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 885–890 (2005)

[Kumon 11] Kumon, M. and Noda, Y.: Active Soft Pinnae for Robots, in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 112–117 (2011)

[Kumon 13] Kumon, M., Kimoto, D., Takami, K., and Furukawa, T.: Bayesian Non-Field-of-View Target Estimation Incorporating an Acoustic Sensor, in *Proceedings of 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3425–3432 (2013)

[Populin 98] Populin, L. and Yin, T.: Pinna movements of the cat during sound localization, *Journal of Neuroscience*, Vol. 18, No. 11, pp. 4233–4243 (1998)

[金 12] 金 天海, 中臺 一博, 辻野 広司: ウェアラブル人工可動耳介-音追従動作による音源定位能力の向上-, 第 30 回日本ロボット学会学術講演会, 日本ロボット学会 (2012)

[本田 85] 本田 学: 耳珠のはたらき, 耳鼻咽喉科臨床, Vol. 78, (1985)

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-B303-04

Impact of Reverberation to the Energy Transfer of Connected Words

Randy Gomez, Keisuke Nakamura, Takeshi Mizumoto and Kazuhiro Nakadai

***Abstract*— In this paper, we present a method in suppressing speech degradation affecting human-robot communication in real reverberant environment condition. The novelty of the proposed method is the mechanism to indirectly incorporate language information in the enhancement process. This is achieved by considering the effects of the acoustic energy transfer of two consecutive words. The proposed method is categorized into two namely, a one-time computation of the smearing coefficients (offline mode) and the actual enhancement (online mode). In the offline mode, word pair smearing coefficients reflective of the inter-word energy transfer within a word pair are calculated and stored, creating a pool of smearing priors database. Then, the online robust enhancement process integrates this information to the conventional framewise enhancement method. In theory, the proposed method outperforms the conventional framewise-only enhancement since it is able to dynamically update the enhancement parameters based on the actual acoustic energy transfer between successive words during testing. Experiments using a humanoid robot inside a reverberant room confirms the effectiveness of the proposed method. Our method renders human-robot speech communication robust to the effect of reverberation as opposed to the conventional method.**

I. Introduction

In recent years, interest towards humanoid robots has gained a dramatic impact in the field of robotics. A humanoid robot is specifically built to resemble the shape of a human body for functional designs replicating human actions such as bipedal movements, arm manipulation, interaction among others. In short, humanoid robots may perform human tasks in manufacturing, assembly line operation or at reception desks to entertain guests. Thus, the notion of a humanoid robot companion assisting humans in day-to-day tasks is not far-fetched. Because of these endless humanoid practical applications, the fascination towards the development of this type of robot is gaining momentum.

Harnessing the potential of a humanoid robot opens a variety of challenging research topics in human-robot interaction. In this paper, we focus on the speech communication interaction with emphasis on robustness in real reverberant environment condition. The very idea that humanoid robot follows the form of a human being makes it more endearing to us, in which the need for it to deliver a more gratifying interaction experience is inevitable. And there is no better interaction experience other than speech communication [1]. The humanoid robot featured in movies that can talk, understand, and execute speech commands is becoming more of a reality than science fiction these days. Recent semiconductor design developments resulting to a fast and more power-efficient processors have significantly improved the robot's ability to process mathematical tasks necessary for speech recognition and understanding. The advancement in computing technology continually benefits the improvement of the speech communication interaction capability of the humanoid robot.

Honda Research Institute Japan Ltd. Co., Honcho Wako-shi, Japan

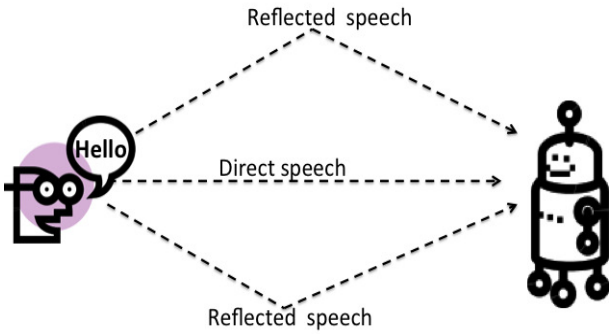


Fig. 1. Reflections of the speech signal inside an enclosed room.

Although system performance using close-talking microphone reaches 90-94 % in recognition accuracy [2], this is only applicable in ideal condition where everything is controlled. In real world however, the pursuit towards a seamless human-machine interaction through speech communication is often plagued with robustness problems [3]. In the real world, humans prefer to communicate with the robot hands-free (not close-talking), which gives the user the freedom to be at some distance away from the robot without being constrained by the microphone [4] as shown in Fig. 1. Hands-free speech communication often use multi-microphone sensors (e.g. microphone array). And as the distance between the user and the microphone array increases, the observed power at the microphones also decreases which makes it more vulnerable to contamination. Although this mode of communication offers more degree of freedom to the user as far as location is concerned, hands-free speech communication is sensitive to the effects of reverberation caused by the reflection of the speech signal in an enclosed environment. In a typical room, the acoustic speech may be reflected on the walls, ceiling, floors, etc. As a result, the speech reflections arrive at different time delays as observed by the microphones mounted on the robot, creating a smearing effect (Fig. 2) to the speech known as reverberation. This phenomenon drastically degrades the performance of the speech recognizer, affecting human-robot interaction experience.

To mitigate the effects of reverberation, speech enhancement is employed. Conventional speech enhancements op-

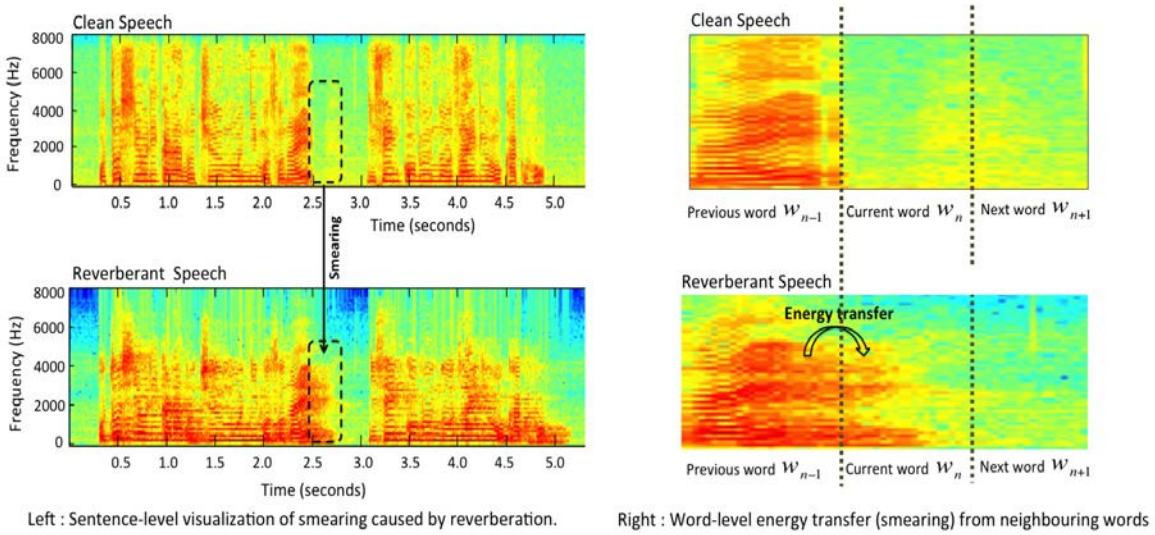


Fig. 2. The smearing effects of reverberation.

erate in framewise manner (e.g. 25 msec. window frame). This design does not distinguish a frame to a word. The immediate concern of speech enhancement is to make the processed speech intelligible to the human ear by focusing on the acoustic quality primarily. Besides, the human brain can infer language from sequence of sounds automatically. In short, the effects of reverberation are considered purely in the acoustics domain. This concept was adopted in speech recognition research. However, speech recognition systems (i.e., model-based systems) are knowledge-based and requires both the acoustic (acoustic waveform) and language (word sequence) information. Thus, simple adoption of the conventional enhancement approach is not sufficient because it only addresses acoustic waveform requirements. In this paper, we propose to indirectly incorporate language information in the speech enhancement process to match it with the speech recognition system. Using the training database, the inter-word relationships among word pairs are analyzed. This enables us to gather prior information of the actual impact of smearing between two successive words. The smearing phenomenon infers the actual energy transfer of two consecutive words due to reverberation. Then, smearing coefficients are calculated for all of the word pairs and a database of smearing coefficients is created. All of these processes are done during offline mode. In the actual testing prior to input to the speech recognizer (online mode), the system dynamically selects the appropriate smearing coefficients and integrate to the framewise enhancement technique for improved recognition performance in real world using a humanoid robot. In our method, the speech enhancement technique can be treated as a black box since the concept is applicable to any conventional speech enhancement platforms employing framewise processing. For simplicity, we focus on Spectral subtraction [5] to explain the proposed concept.

This paper is organized as follows; in Section II, the proposed method is discussed involving both the offline and online procedures. Followed by the experimental set-up in Section III. In Section IV, the results from real-world experiment are presented. Finally, we conclude the paper in Section V.

II. METHODS

A. Offline Smearing Coefficient Training

In Fig. 3, the process of obtaining the smearing coefficients of two neighbouring words due to reverberation are explained as follows,

1) Training Database: The clean speech database s is composed of speech recording (waveform) from different speakers using a close-talking microphone. This is a standard database used in speech recognition applications. Consequently, this is transcribed into word-level text transcripts t. Thus, each waveform speech utterance has a corresponding word level text transcription w_n. The clean speech database is re-played using a loudspeaker inside a reverberant room and a microphone embedded on the robot's head, located at a distance away from the loudspeaker is used to capture both the direct and reflected speech. This set up is used to create a realistic reverberant speech database r, which is needed to analyze the actual smearing effect between neighbouring words.

2) Word pair Extraction: Using the training transcripts originally transcribed in words $w_1, w_2, ..., w_{n-1}, w_n$, for $n = 1 : N$ words, word pair tokens are extracted as $w_1w_2, ..., w_{n-1}w_n$. The segmented transcript of a word pair token j is defined as

$$t_j \triangleq w_j w_{j+1} \quad for \quad j = 1 : N-1. \tag{1}$$

Word-pair extraction is applied to all of the transcripts in the database resulting to intermediate word pairs. Using the information from the word and word pair transcripts, the clean waveform database is segmented into word token

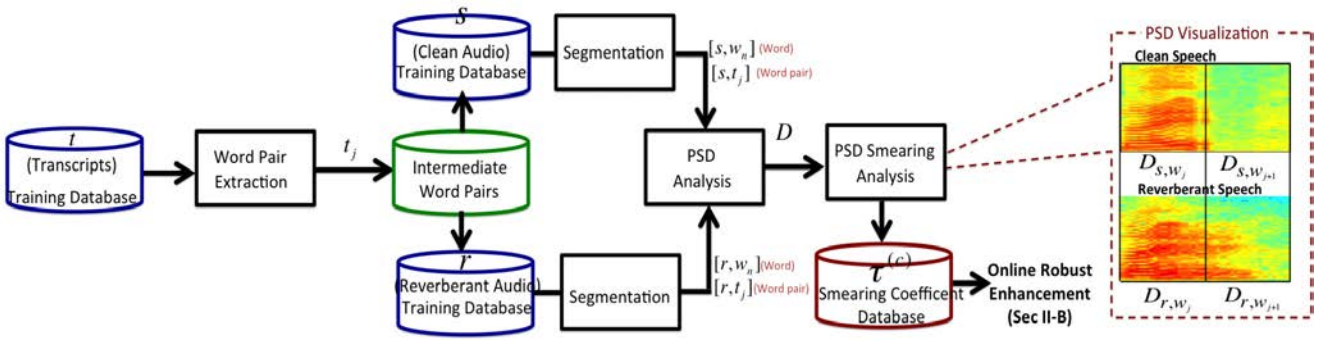


Fig. 3. The training component of the system.

$[s,w_n]$ and word pair token $[s,t_j]$. The segmented reverberant database as $[r,w_n]$ and $[r,t_j]$ for single word and word pair tokens, respectively.

3) Power Spectral Density (PSD) Analysis: The clean and reverberant speech are aligned together, then power spectral density (psd) using the Welch's method is applied to both and calculate the energy distribution in the frequency domain. The Welch's method is preferred since different words vary in duration and the Welch's method is designed to operate by dividing the time domain into successive blocks of periodograms and average across time. The averaging minimizes the impact of the variable word duration. Thus if a word w_n has a total duration of M_{w_n} frames, the PSD of the speech signal s is defined as

$$D_{s,w_n}(\omega) \triangleq \frac{1}{M_{w_n}} \sum_{m=0}^{M_{w_n}-1} P_s(\omega,m) \tag{2}$$

where P_s is the periodogram of the *mth* frame of the word level segmented speech signal. The word level periodogram of the reverberant signal $D_{r,w_n}(\omega)$ is computed in the same manner as Eq. (2). In addition, the word pair psd is also calculated. From the word level psd in Eq. (2), the psd for word pairs are computed by simply expanding the limit M_{w_n} to M_{t_j} to accommodate both two neighbouring words. Specifically,

$$D_{s,t_j}(\omega) \triangleq \frac{1}{M_{t_j}} \sum_{m=0}^{M_{t_j}-1} P_s(\omega,m)$$

and

$$D_{r,t_j}(\omega) \triangleq \frac{1}{M_{t_j}} \sum_{m=0}^{M_{t_j}-1} P_r(\omega,m)$$

are the word pair psd of the clean speech and reverberant speech, respectively. Word pair psd is used for classification while word level psd is used to calculate for the smearing coefficient.

The intermediate word pairs are divided into two categories, namely the frequently occurring pairs and the infrequently occurring pair (including single pairs) duplicates. The former is referred to as base class $c = 1:C$ while the latter as infrequent pairs class. Suppose that there are $c = 1:C$ base classes and $l = 1:L$ infrequent pairs classes, word pair acoustic similarity is used to re-assign the latter into the base classes. Similarity measure is given as

$$sim^{(c,l)}(\omega) \triangleq \tilde{D}^{(c)*}_{s,t_j}(\omega)\tilde{D}^{(l)}_{s,t_j}(\omega), \tag{3}$$

where $\tilde{D}^{(c)*}_{s,t_j}(\omega)$ and $\tilde{D}^{(l)}_{r,t_j}(\omega)$ are the psd representatives from classes c (base classes) and l (infrequent occurring class), respectively. The objective is to distribute the word-pairs in (l) to the base classes in (c) through the similarity measure in Eq. (3). The word pair in l with corresponding c which results to a maximum value in the similarity measure in Eq. (3) for all base classes $1:C$ will be assigned to the corresponding base class c. This process is repeated until all entries in l are exhausted and assigned to the base class accordingly.

4) PSD Smearing Analysis: It is important to analyze the effects of smearing in the utterance empirically which is achieved by analyzing the transferred power using the actual sequence of words leading to the computation of the word pair smearing coefficient. This method establishes a direct link between acoustic power transfer impacted by the language itself. In the conventional methods [6][7], only the acoustic contribution is addressed without consideration of the word sequences. Smearing coefficients characterizes the actual energy contribution of preceding word w_{n-1} to the current word w_n. Since we are not dealing with extremely huge rooms that are very echoic, this assumption is valid. The smearing coefficient for two neighbouring words is calculated as follows,

$$\tau^{(c)}_j \triangleq \frac{D_{r,w_{j+1}}(\omega) - D_{s,w_{j+1}}(\omega)}{D_{s,w_j}(\omega)} \quad for \quad j = 1:N-1, \tag{4}$$

where $D_{r,w_{j+1}}$, $D_{s,w_{j+1}}$ and D_{s,w_j} are the psds of the reverberant and clean speech of the current word in consideration while D_{s,w_j} is the psd of the previous word, respectively.

B. Online Enhancement

Although the proposed robustness method works on different enhancement platforms, the spectral subtraction (SS) [5] enhancement platform is used for simplicity. The SS method is one of the most simple denoising technique for decades

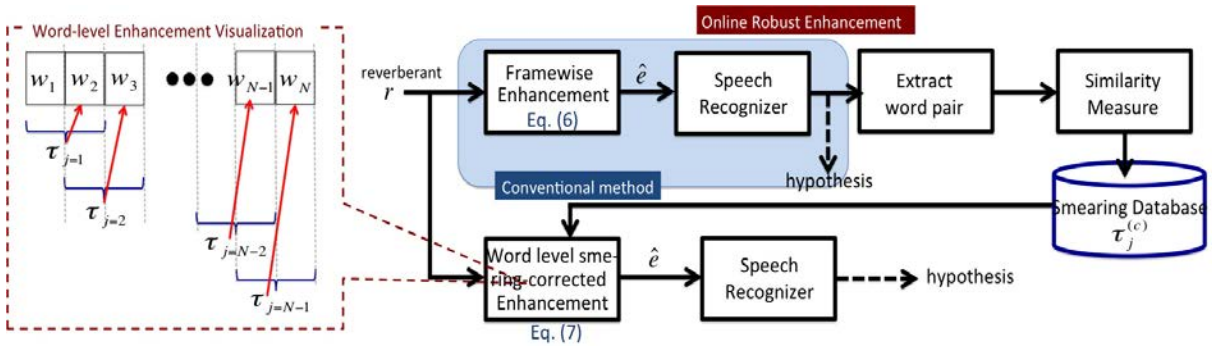


Fig. 4. Robust enhancement in real environment condition.

and has been expanded for dereverberation application [6]. The reverberant speech is modelled as a superimposition of the early and late reflections [8], in which the former is composed mostly of the direct speech signal and the latter is treated as noise. The reverberant speech model in framewise manner m is given as

$$r(\omega,m) = e(\omega,m) + l(\omega,m) \tag{5}$$

where $e(\omega,m)$ and $l(\omega,m)$ are the early and late reflections, respectively. Speech enhancement for speech recognition is defined as suppressing the late reflection and recovering the early reflection [6]. Figure 4 is the block diagram of the the online robust enhancement method.

1) Conventional Method: In our previous work [6], the enhanced speech signal is given as

$$|\hat{e}(\omega,m)|^2 = \begin{cases} |r(\omega,m)|^2 - \delta_b |l(\omega,m)|^2 \\ \quad \text{if } |r(\omega,m)|^2 - \delta_b |l(\omega,m)|^2 > 0 \\ \beta |r(\omega,m)|^2 \quad \text{otherwise.} \end{cases} \tag{6}$$

where β is the flooring coefficient, δ_b is the dereverberation parameters in bands $b = 1,...,B$ and the late reflection power denoted as $|l(\omega,t)|^2$. The process of estimating these parameters is described in detail in [6][7]. Note that Eq. (6) is purely implemented in framewise manner without distinction to the actual word sequence in the actual speech utterance. Thus, the inter-word smearing effect is not accounted for.

2) Proposed Robust Method with Word-level Smearing Compensation: Right after the framewise speech enhancement, the processed reverberant speech is fed into the speech recognition system, the hypothesis from the recognizer is used as a preliminary information to extract the candidate sequence of words. The hypothesis may be imperfect due to acoustic ambiguity in which a framewise processing is unable to address. This is true because reverberation usually spills over several frames rendering framewise processing insufficient. Thus, the hypothesis is further processed to extract word pairs and then calculate similarity measure (see Eq. (3)) against the base class $c = 1:C$. Then, the corresponding smearing coefficient of the selected base class is used in conjunction with the framewise enhancement to include word level smearing effects. Thus Eq. (6) becomes

$$|\hat{e}(\omega,m,w_j)|^2 = \begin{cases} |r(\omega,m,w_j)|^2 - \delta_b \tau_j^{(c)} |r(\omega,m,w_j)|^2 \\ \quad \text{if } |r(\omega,m,w_j)|^2 - \\ \quad \delta_b \tau_j^{(c)} |r(\omega,m,w_j)|^2 > 0 \\ \beta |r(\omega,m,w_j)|^2 \quad \text{otherwise.} \end{cases} \tag{7}$$

It is obvious that the SS in Eq. (7) is capable of resolving reverberation effects both framewise and word level manner through the introduction of $\tau_j^{(c)}$ which is not possible in the conventional method given in Eq. (6).

III. Experimental Set-up

A. Realistic Environment Condition

In our experiment, the human speaker is positioned in front of the robot. The room set-up is shown in Fig. 5 (right). Two rooms were considered with approximately 240 msec (Room A) and 640 msec. (Room B) of reverberation time (RT). The distances between the robot and the speaker are 0.5 m, 1.0 m., 1.5 m. and 2.0 m., respectively. Occlusions due to refrigerator, board, chairs, etc. are considered during testing to recreate a realistic environment. In the experiment, we use the proprietary humanoid robot of Honda Research Institute named Hearbo as shown in the same figure (left). This experimental robot platform has 8 microphones embedded in a circular fashion along its spherical head. The microphone array technology uses HARK [9] for sound source separation. The separated speech is used as the reverberant speech signal to be enhanced by the proposed method.

B. Speech Recognition

The task is composed of 2000 word vocabulary using continuous speech recognition [2]. The topic of the human-robot interaction is about fish varieties used in preparing sushi and sashimi (Japanese traditional dish). The human-robot interaction is initiated by the speaker by asking the robot questions pertaining a fish and the robot answers back. Due to the effects of reverberation, speech recognition may fail which leads to the failure of the robot to give the correct answer by failing to recognize the fish being asked. Thus, the

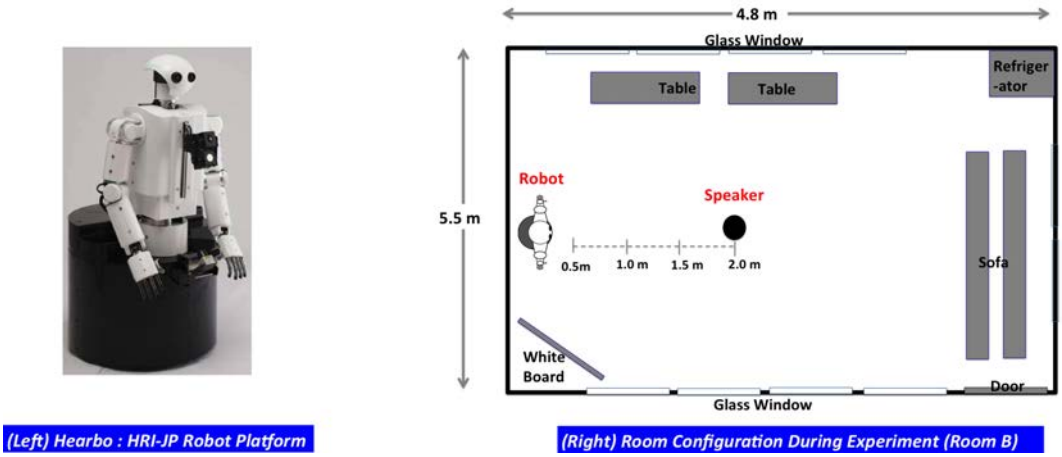


Fig. 5. Room configuration in our experiment

TABLE I

WORD CORRECT RATE IN ROOM A WITH REVERBERATION TIME RT = 240 MSEC..

Methods	*0.5 m*	*1.0 m*	*1.5 m*	*2.0 m*
(A) No Enhancement	90.2 %	84.3 %	74.4 %	69.7 %
(B) Blind Dereverberation [10]	90.6 %	86.4 %	77.3 %	72.6 %
(C) Conventional Wiener Filtering (Framewise only) [7]	90.6 %	86.7 %	78.9 %	75.3 %
(D) Proposed Robust Wiener Filtering (Framewise + Word-level)	**91.0 %**	**87.2 %**	**80.2%**	**78.9 %**
(E) Conventional Spectral Subtraction (Framewise only) [6]	90.6 %	86.9 %	79.0 %	76.2 %
(F) Proposed Spectral Subtraction (Framewise + Word-level)	**91.1 %**	**88.8 %**	**81.9 %**	**79.7 %**

TABLE II

WORD CORRECT RATE IN ROOM B WITH REVERBERATION TIME RT = 640 MSEC..

Methods	*0.5 m*	*1.0 m*	*1.5 m*	*2.0 m*
(A) No Enhancement	81.2 %	65.3 %	44.5 %	28.7 %
(B) Blind Dereverberation [10]	83.6 %	73.5 %	58.1 %	45.3 %
(C) Conventional Wiener Filtering (Framewise only) [7]	84.9 %	76.9 %	60.1 %	48.2 %
(D) Proposed Robust Wiener Filtering (Framewise + Word-level)	**86.7 %**	**83.9 %**	**70.1%**	**61.4 %**
(E) Conventional Spectral Subtraction (Framewise only) [6]	85.9 %	78.6 %	62.3 %	49.3 %
(F) Proposed Spectral Subtraction (Framewise + Word-level)	**87.5 %**	**84.5 %**	**72.4 %**	**63.7 %**

proposed method is used to enhance it. During interaction, the system puts the robot into listening mode while the user is speaking and then switches into speaking mode as soon as it is ready to talk. An example of the conversation is shown as follows,

- *Human>> Hearbo, my friend and I went to a sushi bar yesterday and ordered Sweetfish. Can you give us information of that fish ?*
- *Hearbo>>Sweetfish is common in South East Asia. An edible fish known to its distinctive sweet flavour with melon and cucumber aromas.*
- *Human>> We ate it with maki-sushi. Hearbo, can you give me more information about maki-sushi ?*
- *Hearbo>> Maki-sushi is a rolled rice with other ingredients using a sheet of nori. There are many varieties of maki-sushi like chu-maki, futo-maki, temaki, uramaki among others.*
- *Human>> Hearbo, what is Tororo Kombu ?*
- *Hearbo>> Tororo kombu is made from thinly sliced kombu with vinegar flavour and dried.*

The actual conversation may be longer than the ones above. A total of 20 speakers participated in the human-robot interaction experiment (not included in training). Each speaker asks 10 questions to Hearbo in a freestyle conversation like the ones shown in the conversation example. The only condition is that each question should contain a fish name. We used English triphone Hidden Markov Models (HMMs) acoustic model trained with the Wall Street Journal database.

IV. RESULTS AND DISCUSSION

The ASR results in terms of word correct are shown in Tables 1 and 2 for reverberation times RT=240 msec. and RT=640 msec., respectively. The result in method (A) is when the reverberant speech is not processed prior to input to the speech recognizer. Method (B) is the result using a blind dereverberation approach based on Linear Prediction residual [10]. This is a speech enhancement method that exploits the characteristics of the vocal chords to remove the effects of reverberation. Methods (C) and (E) are the results of two different enhancement platforms based on Wiener filtering [7] and Spectral Subtraction [6] discussed in Eq. (6). Both

of these are based on the conventional framewise processing. The methods in (D) and (F) are based on the same platform in methods (C) and (E) but expanded to include the word-level smearing effect using prior information from word pair database. In these results, it is obvious that the proposed robust enhancement method is outperforming the conventional methods. The rate of improvement due to the proposed method is more evident in Table 2 than in Table 1 since the latter is not so reverberant. Thus, the proposed method works well in very reverberant environment. Consequently, since shorter distances have less reverberation effect, the benefit of speech enhancement is not very obvious (i.e., 0.5 m.) as opposed to farther distances (i.e., 2.0 m.). The superior performance of the proposed method as shown in methods (D) and (F) can be attributed to the following,

- More realistic characterization of smearing (transfer of energy) in the word level which is not considered in the conventional method.
- Dynamic update of smearing parameters in the word level enables the system to adapt to the current environment changes as opposed to being constant in the conventional method
- HMM speech recognition system is primarily defined by both the acoustic and language models. The word level sequence treatment in the proposed method creates a synergistic effect to the language model. Note that the language model is derived from word sequences as well.

It is important to note that even though the hypothesis in which the word pair extraction is based may contain wrong pairs due to the erroneous speech recognition results, the system may still recover when applying the the smearing factor since it alters the acoustic characteristics of the word and impacts recognition performance. Bottom line, processing the acoustics of a known misrecognized word is always better than using the same unprocessed acoustics. After all, speech recognition is probabilistic in nature.

V. CONCLUSION

In this paper, we have indirectly considered the contribution of language (sequence of words) in enhancing the reverberant signal through realistic energy transfer of two consecutive words. By using the smearing prior, we can effectively design and improve any framewise speech enhancement platform for speech recognition application. This is true because reverberation is nothing more of a transfer of sound energy, and framewise processing is not sufficient since smearing affects more than a single frame and sound units impact other sound units differently due to its unique spectral energy characteristics. By knowing before hand the smearing dynamics of two neighbouring words, enhancement can be corrected to reflect the actual energy transfer. Currently we are only limited to two neighbouring words and in the future we will expand this to using more appropriate and well-defined sound unit. Also, in our future works we will further investigate the transference of energy in a more general context.

REFERENCES

[1] R. Gomez, T. Kawahara, Keisuke Nakamura and Kazuhiro Nakadai, "Multi-party Human Robot Interaction with Distant-talking Speech Recognition" *In Proceedings of IEEE Human Robot Interaction (HRI)*, 2012.

[2] Akinobu Lee, *Multipurpose Large VocabularyContinuous Speech Recognition Engine*, 2001.

[3] S. Vaseghi "Advanced Signal processing and Digital Noise reduction", Wiley and Teubner, 1996.

[4] H. Nakajima, K. Nakadai, Y. Hasegawa and H. Tsujino, "Adaptive Step-size Parameter Control for real World Blind Source Separation" *In Proceedings IEEE Int. Conf. Acoust., Speech, Signal Proc. ICASSP*, 2008.

[5] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction" *In Proceedings IEEE Int. Conf. Acoust., Speech, Signal Proc. ICASSP* ,1979.

[6] R. Gomez and T. Kawahara, "Robust Speech Recognition based on Dereverberation Parameter Optimization using Acoustic Model Likelihood" *In Proceedings IEEE Transactions Speech and Acoustics Processing*, 2010.

[7] R. Gomez and T. Kawahara , "Optimizing Spectral Subtraction and Wiener Filtering for Robust Speech Recognition in Reverberant and Noisy Conditions" *ICASSP*, 2010.

[8] E. Habets, "Single and Multi-microphone Speech Dereverberation Using Spectral Enhancement" *Ph.D. Thesis*, June 2007.

[9] http://winnie.kuis.kyoto-u.ac.jp/HARK/

[10] B. Yegnanarayana and P. Satyaranyarana, "Enhancement of Reverberant Speech Using LP Residual Signals", *In Proceedings of IEEE Trans. on Audio, Speech and Lang. Proc.*, 2000.

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-B303-05

実世界知識を扱う音声対話技術とクラウドロボティクスへの展開

Grounded Spoken Dialogues with Robots: Cloud Robotics Tools and Service Robot Applications

杉浦孔明

Komei Sugiura

情報通信研究機構

National Institute of Information and Communications Technology

komei.sugiura@nict.go.jp

Abstract

ロボットとの音声によるコミュニケーションは，ユーザにとって手軽であるというメリットがあるが，実現は簡単ではない．頑健な音声認識が必要とされるだけでなく，発話の解釈が実世界情報や履歴により影響を受けるためである．このような背景から，音声・画像・動作・コンテキスト情報を用いてユーザの発話を解釈するロボット対話技術 LCore を開発している．本稿では，音声対話を通じた実世界知識の学習と行動生成について述べたのち，サービスロボットの競技会であるロボカップ@ホームにおけるロボット対話技術の応用について紹介する．また，ロボットの音声対話機能の技術動向，特にクラウドロボティクスへの展開と課題について述べる．

1 はじめに

人口構造の変化や労働形態の多様化とともに，生活環境で人間と共存・支援するロボットへの期待が高まっている．実際に，掃除ロボットの国内市場規模は 2006 年に約 2 万台であったものが，2012 年には約 38 万台になっている．また，2014 年にはリモートプレゼンスロボットがノート PC 程度の価格帯で発売されたことから，普及が期待される．家庭用ロボットで使われる技術はますます高度化，多目的化すると考えられる．Google 会長のエリック・シュミットらは，当面は高度な機能を有する多機能ロボットは一般消費者には高くて手が届かないものの，将来的に一般家庭でも数台の多目的ロボットを持つようになると予想している [Cohen 13].

一方，どれほど豊富な機能を持つロボットを構築したとしても，人が手軽に機能を使えなければ普及しないであろう．つまり，多機能なロボットが日常生活に浸透するためには，人とのコミュニケーション機能が課題となる [西田 03]．優れた音声コミュニケーション機能を有するロボットが現れれば，日常生活が大きく変わる可能性がある．

ロボットとの音声によるコミュニケーションは，ユーザにとって手軽であるというメリットがあるが，実現は簡単ではない．もちろん，雑音抑圧，発話検出，話者分離などは音声処理における重要な課題である．しかし，本稿で強調したい本質的な課題は，発話の解釈が実世界情報や経験により影響を受けることである．例えば「コップ取って」という命令を実行するには，どのコップを取るのか，「取る」とはどのような動作軌道を表すのか，を推論しなければならない．一方，家族にコップを取ってもらう場合は省略した言い方でも通じることが多いうえ，わからなければ聞き返すだろうという期待がある．つまり，人間はそれまでの経験から，相手が自分の発言をどれくらい理解できるかに関して暗黙的な知識がある．しかし，ロボットと対話するユーザにとって，ロボットが状況をどれだけ理解しているかを推定するのは難しい．

現状のロボット対話処理機構では，動作コマンドの伝達を目的とすることが多いにも関わらず，動作情報と音声認識は別々に処理されることが多い [Hartanto 11]．それに対し，ロボットの対話機構にグラウンドしない知識を導入することが，シンボルグラウンディング問題を孕むことは古くから指摘されている [Harnad 90,Pfeifer 99]．したがって，ロボットに人間と自然にコミュニケーションさせるためには，ユーザや状況への適応手法が重要課題となる．

本稿では，実世界知識を扱う音声対話技術について概説する．まず，ロボットとの音声対話における研究を分類し課題を整理する．3 節では，著者らの研究グループがこれまでに行なってきたロボット対話技術について述べる．4 節では，それらの技術の実世界への応用例として，ロボカップ@ホームタスクへの適用例を紹介する．5 節では，著者らが開発したサービスロボット向けのクラウド型音声コミュニケーションツールキットについて説明する．

2 ロボットとの音声対話

実世界にグラウンドした対話を実現するためには，実世界のオブジェクトや動作を記号化・言語化することが極めて重要な課題である．ロボティクスの分野では，動作-記号の相互変換に関する試みが近年注目されてきている [Inamura 04, Ogata 07, 高野 09]．高野らは，運動の分節化を通じてヒューマノイドロボットが獲得した原始シンボルを用いて，運動認識・生成を行っている [高野 09]．Ogata らは，動作系列と記号列の間の多対多対応問題を扱うリカレントニューラルネットに基づく手法を提案している [Ogata 07]．Kollar らは，ロボットに与える移動指示に関して，ランドマークオブジェクトや動作にグラウンドした言語表現を学習する手法を提案した [Kollar 10]．学習されたモデルを用いることにより，指示が示す最も確からしい経路を推論する．

オブジェクト-記号の相互変換に関しては，人工知能分野で多くの研究が行われてきた [山肩 04, Roy 02, Dale 95]．詳細については，[長井 12] が詳しい．山肩らはコップ類の名称とイメージモデルの参照関係における曖昧性に一貫した個人差があることを示している [山肩 04]．Roy は，ディスプレイ上の複数の長方形のうちひとつを指示する言語表現をテキストベースで生成する手法を提案している [Roy 02]．[Roy 02] で提案された手法では，単語のカテゴリは教師なし学習の枠組みでクラスタリングされるため，設計者が属性を用意する必要がない．Yu らは，動画と文を入力として，文節が動画中のどの領域に対応するかを学習させている [Yu 13].

音声対話を行うロボットの先駆的事例としては，Jijo-2 が挙げられる [松井 00]．また，稲邑らは，移動ロボットの障害物回避において，ロボットが段階的に行動決定モデルを獲得する機構を提案した [稲邑 01]．センサ値と行動の関係をベイジアンネットを用いてモデル化し，推論結果の確信度を用いて応答決定を行う．

一方，対話システムの分野では，ホテル検索やバスの経路検索などが代表的なタスクである [Komatani 00, Bohus 06, Kawahara 98]．対話システムの評価タスクとしては，1990 年に始まった Loebner Prize[1]がある．Loebner Prize では，端末を通じてシステムと人間が対話を行い，チューリングテストに合格した場合は 10 万ドルが与えられる．これまでに合格したシステムはないものの，毎年最も人間に近い動作をしたと判定されたシステムには賞金が与えられる．音声対話システムの評価タスクとしては，2010 年に行われた “Spoken Dialog Challenge” [Black 11] がある．システムのタスクはバスの経路案内であり，実際に自動応答サービスとして実装された．システムの評価尺度として，単語誤り率やタスク達成率などが用いられている．Spoken Dialog Challenge の後継として，REAL Challenge[2]が企画されている．

スマートフォンを始めとする種々のデバイスに音声インタフェースが導入され，広く一般に認知されるようになってきた [河原 13, 松田 13]．検索や対話に代表されるサービスの多くは，クラウド型サービスとして実装されている．ロボティクスにおいて，クラウド型サービスの利活用を目指す分野はクラウドロボティクス [Kuffner 10] と呼ばれる．代表的な研究としては，Google Goggles を用いたマニピュレーション [Kehoe 13] や，クラウド型知識共有を行うプラットフォームおよびインタフェース言語 RoboEarth [Tenorth 12] などがある．また，著者らはクラウド型の音声コミュニケーションツールキット rospeex を公開している [杉浦 13b].

3 実世界にグラウンドした音声対話

ロボットが生活環境に浸透するに従い，ロボットのコミュニケーション能力が課題となる．現状の技術では,「コップ（テーブルに）置いて」などの曖昧な発話の解釈は非常に難しい．省略された語が表すオブジェクトを推定する必要があるうえ，環境中に存在する「コップ」の候補から正しいオブジェクトを選択しなければならない．

本節では，著者らが開発してきたコミュニケーション学習基盤 LCore [Iwahashi 09] を紹介する．LCore は，画像，動作，アフォーダンス，履歴などの実世界情報を学習し，発話・動作の生成が可能である．

3.1 コミュニケーション学習基盤 LCore

現状のロボットの対話処理機構では，動作コマンドの伝達を目的とすることが多いにも関わらず，動作情報と音声認識は別々に処理されていることが多い [Hartanto 11]．ユーザの発話の意味はグラウンドされない知識に基づいて解釈されるため，動作が状況にふさわしいかどうかは音声認識時には考慮されない．しかしながらこのような手法では，ユーザの発話の意味が状況に応じて適切に理解されない，という問題がある．

LCore では，マルチモーダル入力（音声・画像・予測動作など）から学習されたモデルを用いてユーザの発話を理解する．以下では，各モダリティに対応するモデルをモジュールと呼ぶこととする．

まず，音声発話 s から最適行動 $\hat{a}$ を出力する場合について考える．マルチモーダル発話理解スコアを表す関数 Ψ を，各モジュールの重み付き和として定義する．

$$\Psi(s,a_k,O,q^{\langle i_t \rangle}) = \max_z (\gamma_1 B_S + \gamma_2 B_I + \gamma_3 B_M + \gamma_4 B_R + \gamma_5 B_H) \quad (1)$$

[1] http://www.loebner.net/Prizef/loebner-prize.html

[2] https://dialrc.org/realchallenge/

ここに，z は各単語の概念構造への分割であり，$(s, a_k, O, q^{\langle i_t \rangle})$ はそれぞれ，発話，行動，状況，行動コンテキストを表す．また，$\gamma = (\gamma_1, ..., \gamma_5)$ は各モジュールに対する重みであり，MCE 学習 [Katagiri 98] を用いて学習される．各モジュールが出力するスコアは以下のように定義される．

- 音声スコア B_S
  単語の n-gram，および節の連接確率として学習される．B_S は，発話 s に対する概念構造 z の条件付き確率の対数として表す．
- 視覚スコア B_I
  ガウス分布により学習される．B_I は，オブジェクトの視覚特徴量が与えられたときの対数尤度である．
- 予測動作スコア B_M
  Reference-Point-Dependent HMM（RPD-HMM）[Sugiura 11a] により学習される．B_M は，可能な行動に対して軌道を仮想的に生成したうえで，その軌道の尤度として得られる．
- 動作-オブジェクト関係スコア B_R
  「平らなものはオブジェクトを載せられやすい」など，動作と視覚的特徴の関係を表す．B_R は，2 個のオブジェクトの視覚特徴量に対するガウス分布の対数尤度である．
- 行動コンテキストスコア B_H
  B_H は，あるコンテキスト（「把持されている」,「直前に操作された」など）のもとでの指示対象としてのオブジェクトの適切さ (スコア) を表す．

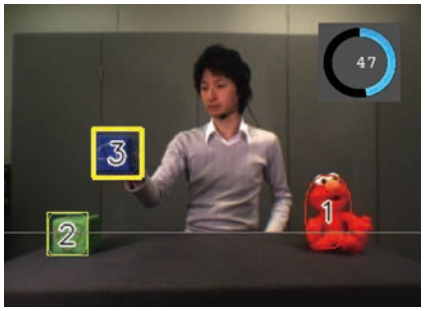
以上より，コンテキスト q，状況 O，発話 s が与えられたときの最適行動 $\hat{a}$ は以下で得られる．

$$\hat{a} = \underset{k}{\operatorname{argmax}} \Psi(s, a_k, O, q) \tag{2}$$

3.2 確信度に基づく動作と発話の生成

前節までの手法は，ユーザから発話が入力され，ロボットが動作を出力するという一方向的過程であった．本節では，入力された発話に対し，ロボットが応答（発話または行動）を出力する場合について考える．例えば，ユーザが「コップ（テーブルに）置いて」などの曖昧な発話を行ったとする．曖昧性を解消するためには,「赤いコップですか，青いコップですか」のような確認発話を毎回行なってもよいが，曖昧なときのみ確認発話を行う方が望ましい．

著者らは，発話理解確率（発話を正しく解釈できる確率）を推定し，効用最大化により応答を生成する手法を提案した [Sugiura 11b]．発話が曖昧であるとき，共有信念関数の第 1 候補と第 2 候補のスコアの差（マージン）が小さいことから，これを曖昧性の尺度として用いている．



【状況】オブジェクト 2 が直前に操作された
U: ハコ エルモ ちかづけて.
R: ミドリハコをちかづけて？
U: いいえ.
R: アオイハコをちかづけて？
U: はい.
R: (動作実行：オブジェクト 3 をオブジェクト 1 に近づける)

図 1: グラウンドした語彙による確認発話の生成

提案手法では，統合確信度を発話理解確率の推定値としてモデル化した．統合確信度は，ベイズロジスティック回帰により学習される．動作応答 b_1 と確認発話応答 b_2 は，対応する期待効用 $\mathbb{E}[R_i](i = 1, 2)$ の最大化により選択される．

$$\mathbb{E}[R_i] = r_{i1} f(d; w) + r_{i2}(1 - f(d; w)) \tag{3}$$

$$d = \Psi(s, \hat{a}, O, q) - \max_{j \neq k} \Psi(s, a_j, O, q) \tag{4}$$

ここに，$f(d; w)$ は，$w = (w_0, w_1)$ をパラメータとするロジスティックシグモイド関数である．また，r_{i1}, r_{i2} はそれぞれ，行動 $\hat{a}$ がそれぞれ正解，不正解のときの応答 b_i に対する効用である．

実験では，オブジェクトを操作するよう，ユーザはロボットに指示を与える．両者は音声対話により曖昧性を解消し，ロボットがユーザの意図した行動をとればタスク成功とした．ロボットが失敗行動（正解でない行動）を行ったときは，アームに取り付けたセンサを叩くことで教師信号を与えることができる．このようにして得たマージンと教師信号の組を用いて，ベイズロジスティック回帰により $f(d)$ を学習させた．

図 1 は，ユーザ (U) とロボット (R) の対話例を示したものである．図において，右上の数値は統合確信度 $f(d)$ を表す．図 1 では，最適行動の確信度は $f(d) = 0.478$ であり，確認発話「アオイ ハコをちかづけて」が最適応答であった．この言語表現は，オブジェクト 2 と 3 の視覚的特徴のなかで最も異なる属性[3]について述べており，ユーザにとって理解しやすい．ランドマークについては確認発話を行わなくても確信度に影響はないため，確認を省略していると考えられる．

[3] カラー画像ではオブジェクト 2 は緑，オブジェクト 3 は青色である．

図 2: 2012 年世界大会に参加したロボット

実験の結果，動作や視覚などの情報を用いず，音声のみで発話理解を行った場合の行動失敗率は 83.4%であった．ベースライン手法（発話を行わず動作のみで応答する）における行動失敗率は 12.0%である一方，提案手法による行動失敗率は 2.6%であった．このことから，提案手法はベースライン手法に比べて行動失敗率を大幅に低減できたといえる．

4 実世界への適用：ロボカップ@ホーム

サービスロボットの研究開発においては，独自の環境や評価尺度が用いられることが多く，一般的に手法同士の比較が難しい．一方，タスクを標準化することで比較評価のコストを低減すれば，コミュニティ全体の研究開発に貢献できるであろう．ロボカップ@ホームは，サッカーやレスキューと並ぶロボカップ [浅田 10] のリーグのひとつであり，生活支援ロボットの競技である [Iocchi 10, 杉浦 12]．各チームのロボットは，日用品の探索，棚からユーザに言われたものを取ってくる，人を追従する等，日常生活に役立つ機能を制限時間内にどれだけ達成できるかを競う．

4.1 ロボカップ@ホームとは

2012 年のロボカップ@ホーム世界大会に出場したロボットを図 2 に示す．ロボカップ@ホームでは，家庭・オフィス・スーパーマーケットなどにおけるロボットの応用を想定したタスクが設定されている．中心課題は，モバイルマニピュレーションとヒューマンロボットインタラクション (HRI) である．後述するように，未知環境における地図作成・移動，日用品の物体認識・把持，高騒音環境における音声認識などを含む．各タスクはベンチマークテストとして明文化されると同時に，複数の技術的課題を含んだストーリーになっており，観客を飽きさせないよう努力されている．

ロボカップ@ホームは 2006 年に始まり，我々は 2008 年より参加してきた．2009 年からは，世界大会のルール策定やジャパンオープン (日本大会) の運営にも参与している．2012 年メキシコで開催された世界大会には 9 カ国から 18 チームの参加があり，ジャパンオープンには 3 カ国から 10 チームの参加があった．各チームは 6〜10 人程度で構成されていることが多い．

世界大会では，2 日間のセットアップ期間にフィールドやオブジェクトが発表されるので，参加者は環境地図構築，オブジェクト登録を事前に行うことができる．マニピュレーションの対象であるオブジェクトは，ペットボトルや菓子などの日用品である．各オブジェクトには，名称（「コーンフレーク」など）とカテゴリ名（「食べ物」など）が定義されている．

ロボカップ@ホームに関する日本語による文献には [岡田 10, 杉浦 12] などがある．また，2011 年世界大会については [Stückler 12] が詳しい．これまでの世界大会における得点傾向が [Iocchi 10] にまとめられているほか，各チームの獲得スコア情報が大会ウェブサイト上で公開されている．また，インターネット上にアップロードされた過去の大会の動画は，イメージをつかむ手段として効果的である．最新版の公式ルールは公式サイト[4]からダウンロードできる．

4.2 タスク環境

タスク環境として 2LDK 程度のモデルルームが用意され，部屋構成や家具・食器等は毎年変更される．2012 年世界大会 (メキシコ) で用いられたタスク環境を図 3 に示す．図 3(a)(b) に示す環境は，9 種類のタスクのうち 7 種類を行うメイン環境（以下,「フィールド」と呼ぶ）である．2012 年世界大会のフィールドは，ロビー，リビングルーム，キッチン，ベッドルームの 4 部屋から構成されている．

実際の使用シーンを想定した環境で性能評価を行う意図から，一部のタスクは店舗などフィールド外で行われる．2012 年世界大会では，Restaurant タスクをレストラン (図 3(c)) で行なった．また，2010 年は玩具店，2011 年はスーパーマーケットにおいてタスクを行なっている．これらのタスクの主眼は，未知環境でのオンライン SLAM 機能とモバイルマニピュレーション機能の評価であるため，事前に環境地図を作成することは許可されていない．図 3(c) の環境にはガラスの仕切りや金属製のチェアなどが存在するため，測距センサのレーザが透過あるいは反射してしまい，環境地図構築が非常に難しい．Follow Me タスク (図 3(d)) では，100 人以上の観客が「ノイズ」となり，音声認識，顔認識，人追従を難しくしている．

4.3 日用品マニピュレーションの模倣学習

家庭内でタスクを行うロボットにとっては,「食器棚からコップを取り出す」などの物体の操作は必要不可欠な機能であり，これらの動作を言語で指示できることが望ましい．一方，各種の日用品や棚に対応する動作を事前にプログラムするコストは非常に大きいうえ，事前にプロ

[4] http://www.robocupathome.org/rules

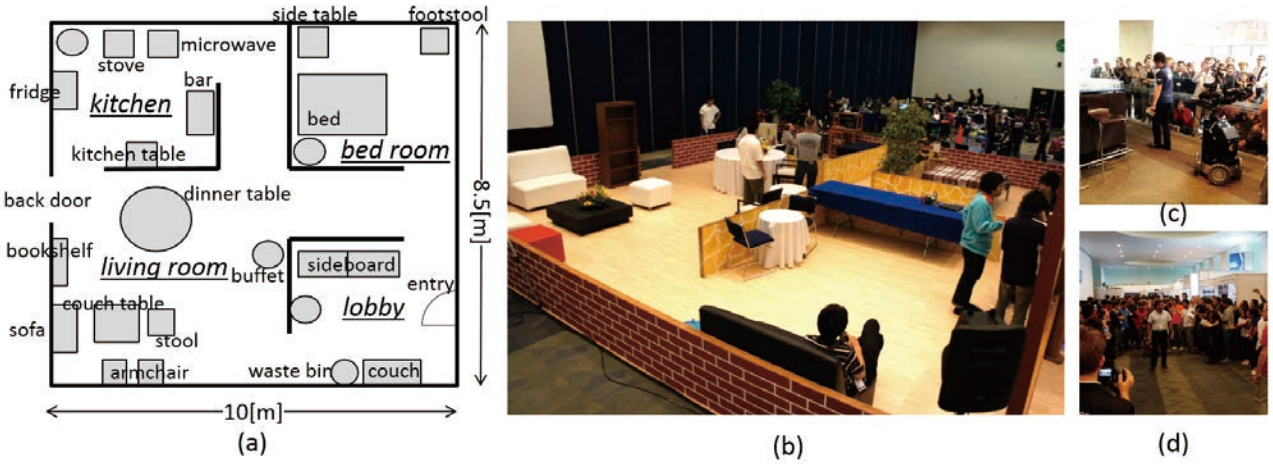


図 3: 2012 年世界大会のタスク環境．(a) 家具配置，(b) メイン環境（フィールド），(c) Restaurant タスクの環境，(d) Follow Me タスクの環境.

図 4: ロボカップ@ホーム環境における動作「捨てる」の学習

グラムされた動作がユーザにとってイメージしにくいものであった場合，安心して動作指示できないという問題がある．そこで我々は，模倣学習の枠組みにより物体操作を学習する手法の開発を行なってきた．このような学習手法を構築することで，プログラミングスキルが必要とされないユーザフレンドリな動作教示方法を実現できる．ロボカップ@ホーム環境における家事動作の学習への適用例を図 4 に示す.

「X を Y にのせる」や「Z を回す」など参照点に依存した動作の模倣では，世界座標系での動作軌道の模倣に意味はなく，適切な座標系を推定し軌道を汎化しなければならない．このために，参照点に依存した隠れマルコフモデル（RPD-HMM）を開発した [Sugiura 11a]．RPD-HMM は，物体位置の時系列を入力として，動作をモデル化するための最適な座標系を EM アルゴリズムにより推定し，軌道のモデルを学習している.

動作の生成時には，学習時とまったく同じように物体が配置されている訳ではないため，たとえ同じ「載せる」動作であっても，学習時の軌道をそのまま用いることは無意味である．学習した RPD-HMM は固有座標系において汎化されたものであるので，固有座標系 C から世界座標系 W へ変換する．RPD-HMM の位置・速度・加速度の平均ベクトルおよび共分散行列は，同時変換行列により C から W 上のモデルに変換される．HMM から滑らかな軌道を生成するために，音声合成の分野で用いられる HMM 軌道生成 [Tokuda 00] を用いる．実験の結果，7 回程度の教示で生成軌道の誤差が収束することが確認できた.

5 クラウド型音声コミュニケーションツールキット “rospeex”

近年，音声対話システムの分野では，開発者が容易に利用できるツールキットが公開されている（例えば [大浦 13]）が，人とロボットのインタラクションでは，高性能な音声認識・合成を容易に利用できる状況ではない．ロボットとの高度な音声インタラクションを可能とするためには，音声処理とロボティクスの深い知識を要求されるのが現状である．このような背景のもと，著者らは音声対話機能の開発コストを下げるべく，クラウド型音声対話ツールキット “rospeex” を開発・公開している．ロボカップ@ホームなどのサービスロボット開発では，開発コストを低減できることから，RT ミドルウェアや ROS（Robot Operating System）などのミドルウェアの利用が一般的になってきている．rospeex は ROS 上で利用可能なクラウド型音声コミュニケーションツールキットであり，学術研究用途に限り無料かつ非登録で利用可能である.

5.1 rospeex の機能

rospeex が提供する機能と想定する標準的な構成を図 5 に示す．発話理解（言語理解），対話制御，応答生成については，ユーザが記述するものとする.

rospeex では，雑音抑圧と発話区間検出はネットワーク

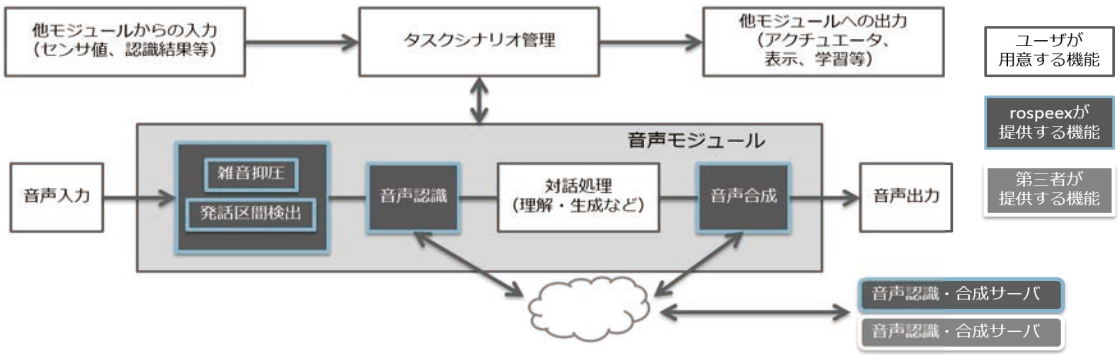


図 5: rospeex の構成の概略

上サーバで行わない設計としている．これらをサーバで処理するとネットワーク由来の遅延によりリアルタイム性の確保が難しくなるためである．また，一般的に発話区間検出の精度はそれほど高くないため，後段の処理でロボット名を含む発話のみ受け付けるなどの工夫が必要である．

rospeex は複数のクラウド型音声サービスに接続可能であり，それらを切り替えて使用できる．本節では，NICT が提供する音声認識・合成サービスについて説明する．これらは，ROS を経由せずに単体としても利用可能である．現時点では，学術研究目的に限り無償・登録不要で公開している．本サービスは，JSON ファイルをインタフェースとする．ユーザが用いるプログラミング言語には依存しないため，C++や Python など各種のプログラミング言語を利用可能である[5]．10 行程度で簡単な対話（時刻の問い合わせなど）を行う関数を記述することができる．

対話管理にマークアップ言語（VoiceXML など）を利用するソフトウェアと異なり，rospeex は対話管理の簡単なインタフェースを用意していない．これは，想定ユーザとして，複雑な対話管理を必要としないロボット開発者を念頭に置いたためである．一方，ROS 上で Python や C++ で開発したソフトウェア資産があれば，rospeex と簡単に組み合わせることが可能であるという利点がある．また，現状では，音源定位などの音響処理は統合されていない．しかしながら，HARK [Nakadai 10] など音響処理を扱うモジュールが提供されているので，rospeex の前段に容易に組み込むことが可能であると考えられる．

5.2 非モノローグ音声合成

ロボットのコミュニケーション機能の開発においては自然な音声合成が求められているが，一般的な音声合成器は人-ロボット対話に最適化されている訳ではない．rospeex では，ロボットとの対話に特化して開発されたボイスフォントを利用可能である．本ボイスフォントは，非モノローグ HMM 音声合成 [杉浦 13a] により生成される．以下では，非モノローグ HMM 音声合成について概説する．

[5]サンプルコードを `http://komeisugiura.jp/software/nm_tts.html` から入手可能である．

表 1: 学習セットの比較

システム	収録スタイル	学習セットサイズ
(0) AS	分析合成音	-
(1) Mono-176 (ベースライン)	モノローグ	176 分 (2359 文)
(2) NonM-176 (提案手法)	非モノローグ	176 分 (4485 文)
(3) NonM-325 (提案手法)	非モノローグ	325 分 (8861 文)
(4) NonM-433 (提案手法)	非モノローグ	433 分 (14179 文)

HMM の学習セットとして，声優による掛け合い対話コーパスを作成した．表 1 に示すように，声優による掛け合い対話コーパスとしては，最大級のものを用いている．サービスロボットへの応用を想定した被験者実験を行い，ベースライン手法に比べて品質が優れるという結果を得た．図 6 に提案手法とベースラインの MOS 値を示す．エラーバーは 95%信頼区間を示す．2 つの信頼区間が重なっていなければ，統計的有意差があるといえる．図より，提案手法（NonM-176，NonM-325，NonM-433）の MOS 値はベースラインと比べて高く，分析合成音（理論上の上限）に近い値であることがわかる．実験の詳細は，[杉浦 13a] を参照されたい．

非モノローグ音声合成は，ブラウザベースのサービスとして 2013 年 9 月 5 日に公開された．約半年間における音声合成サービス利用データを表 2 に示す．平均すると 1 日あたり 400 件程度の音声合成リクエストを処理している．

表 2: 実証実験の概要

実験期間	2013/9/5-2014/3/4
音声合成ユニーク IP 数	2862
音声合成リクエスト数	33320

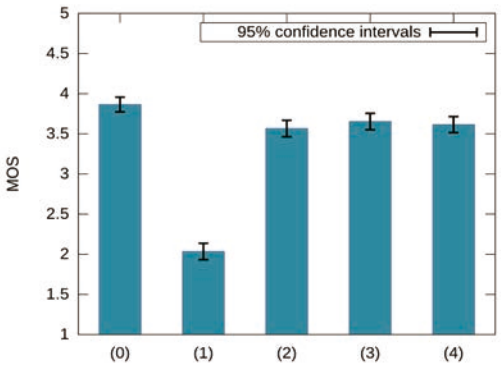


図 6: (0) 分析合成音 (理論上の上限)，(1) Mono-176 (ベースライン)，(2) NonM-176，(3) NonM-325，(4) NonM-433 に対する MOS 値.

6 おわりに

本稿では，実世界にグラウンドした音声対話に関する著者らの取り組み，日常環境における生活支援ロボットのベンチマークテストであるロボカップ@ホーム，ならびに rospeex について紹介した．高齢化社会における QOL(Quality of Life) 向上が社会的に急務であることを考えると，ロボットの日常環境への適用，ロボットによる自立支援の促進などは，ロボット開発において今後も重要な課題であろう．

将来的に家庭用多機能ロボットの普及を目標とすると，全ての機能をスタンドアロン機能として実装することはコスト面から現実的でない．一方，ネットワークへの接続を前提とすれば，音声認識や画像認識などに関する高度な技術を安価で導入することが可能である．実際に，音声での検索サービスや対話サービスの多くは，クラウド型サービスとして実装されている．現状ではクラウドロボティクスの事例は多くないものの，今後主要な分野になると考えられる．

また，ロボットに使用される技術が高度化，多機能化していくに従い，そのような複雑な機能を手軽に使用できることがますます重要になる．結果として，ユーザフレンドリなインタフェースを持つことがロボットの普及に大きな意味を持つと予想される．今日，スマートフォン上のサービスに代表されるように，適切な入出力インタフェースを選択することがユーザ体験の向上につながることは広く認識されている．ロボティクスにおいても音声と種々の入出力インタフェースをうまく統合させることが求められるようになるであろう．

参考文献

[Black 11] Black, A. W., Burger, S., Conkie, A., Hastie, H., Keizer, S., Lemon, O., Merigaud, N., Parent, G., Schubiner, G., Thomson, B., et al.: Spoken dialog challenge 2010: Comparison of live and control test results, in *Proceedings of the SIGDIAL 2011 Conference*, pp. 2–7 (2011)

[Bohus 06] Bohus, D., Langner, B., Raux, A., Black, A., Eskenazi, M., and Rudnicky, A.: Online supervised learning of non-understanding recovery policies, in *Proceedings of the IEEE/ACL Workshop on Spoken Language Technology*, pp. 170–173 (2006)

[Cohen 13] Cohen, J. and Schmidt, E.: *The New Digital Age: Reshaping the Future of People, Nations and Business*, John Murray (2013)

[Dale 95] Dale, R. and Reiter, E.: Computational interpretations of the Gricean maxims in the generation of referring expressions, *Cognitive Science*, Vol. 19, No. 2, pp. 233–263 (1995)

[Harnad 90] Harnad, S.: The Symbol Grounding Problem, *Physica D*, Vol. 42, pp. 335–346 (1990)

[Hartanto 11] Hartanto, R.: *A Hybrid Deliberative Layer for Robotic Agents*, Springer (2011)

[Inamura 04] Inamura, T., Toshima, I., Tanie, H., and Nakamura, Y.: Embodied symbol emergence based on mimesis theory, *International Journal of Robotics Research*, Vol. 23, No. 4, pp. 363–377 (2004)

[Iocchi 10] Iocchi, L. and Zant, van der T.: RoboCup@Home: Adaptive Benchmarking of Robot Bodies and Minds, in *Proceedings of the International Conference on Simulation, Modeling and Programming for Autonomous Robots*, pp. 171–182 (2010)

[Iwahashi 09] Iwahashi, N., Taguchi, R., Sugiura, K., Funakoshi, K., and Nakano, M.: Robots that Learn to Converse: Developmental Approach to Situated Language Processing, in *Proceedings of International Symposium on Speech and Language Processing*, pp. 532–537 (2009)

[Katagiri 98] Katagiri, S., Juang, B., and Lee, C.: Pattern Recognition Using a Family of Design Algorithms based upon the Generalized Probabilistic Descent Method, *Proceedings of the IEEE*, Vol. 86, No. 11, pp. 2345–2373 (1998)

[Kawahara 98] Kawahara, T., Lee, C., and Juang, B.: Flexible speech understanding based on combined key-phrase detection and verification, *IEEE Transactions on Speech and Audio Processing*, Vol. 6, No. 6, pp. 558–568 (1998)

[Kehoe 13] Kehoe, B., Matsukawa, A., Candido, S., Kuffner, J., and Goldberg, K.: Cloud-Based Robot Grasping with the Google Object Recognition Engine, Proc. ICRA (2013)

[Kollar 10] Kollar, T., Tellex, S., Roy, D., and Roy, N.: Toward understanding natural language directions, in *Proceeding of the 5th ACM/IEEE international conference on Human-robot interaction*, pp. 259–266 (2010)

[Komatani 00] Komatani, K. and Kawahara, T.: Flexible mixed-initiative dialogue management using concept-level confidence measures of speech recognizer output, in *Proceedings of the 18th conference on Computational Linguistics*, pp. 467–473 (2000)

[Kuffner 10] Kuffner, J.: Cloud-Enabled Robots, in *Proc. Humanoids* (2010)

[Nakadai 10] Nakadai, K., Takahashi, T., Okuno, H. G., Nakajima, H., Hasegawa, Y., and Tsujino, H.: Design and Implementation of Robot Audition System'HARK'—Open Source Software for Listening to Three Simultaneous Speakers, *Advanced Robotics*, Vol. 24, No. 5-6, pp. 739–761 (2010)

[Ogata 07] Ogata, T., Murase, M., Tani, J., Komatani, K., and Okuno, H. G.: Two-way translation of compound sentences and arm motions by recurrent neural networks, in *Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and System*, pp. 1858–1863 (2007)

[Pfeifer 99] Pfeifer, R. and Scheier, C.: *Understanding Intelligence*, MIT Press, Cambridge, MA. (1999)

[Roy 02] Roy, D.: Learning visually grounded words and syntax for a scene description task, *Computer Speech and Language*, Vol. 16, No. 3, pp. 353–385 (2002)

[Stückler 12] Stückler, J., Holz, D., and Behnke, S.: RoboCup@Home: Demonstrating Everyday Manipulation Skills in RoboCup@Home, *Robotics & Automation Magazine, IEEE*, Vol. 19, No. 2, pp. 34–42 (2012)

[Sugiura 11a] Sugiura, K., Iwahashi, N., Kashioka, H., and Nakamura, S.: Learning, Generation, and Recognition of Motions by Reference-Point-Dependent Probabilistic Models, *Advanced Robotics*, Vol. 25, No. 6-7, pp. 825–848 (2011)

[Sugiura 11b] Sugiura, K., Iwahashi, N., Kawai, H., and Nakamura, S.: Situated Spoken Dialogue with Robots Using Active Learning, *Advanced Robotics*, Vol. 25, No. 17, pp. 2207–2232 (2011)

[Tenorth 12] Tenorth, M., Perzylo, A. C., Lafrenz, R., and Beetz, M.: The RoboEarth Language: Representing and Exchanging Knowledge about Actions, Objects, and Environments, in *Proc. ICRA*, pp. 1284–1289 (2012)

[Tokuda 00] Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., and Kitamura, T.: Speech Parameter Generation Algorithms for HMM-Based Speech Synthesis, in *Proceedings of ICASSP*, pp. 1315–1318 (2000)

[Yu 13] Yu, H. and Siskind, J. M.: Grounded language learning from video described with sentences, in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Vol. 1, pp. 53–63 (2013)

[稲邑 01] 稲邑 哲也, 稲葉 雅幸, 井上 博允：ユーザとの対話に基づく段階的な行動決定モデルの獲得, 日本ロボット学会誌, Vol. 19, No. 8, pp. 983–990 (2001)

[高野 09] 高野 渉, 中村 仁彦：統計的相関に基づく動作パターンのリアルタイム教師なし分節化と原始シンボルの自律的獲得, 日本ロボット学会誌, Vol. 27, No. 9, pp. 1046–1057 (2009)

[長井 12] 長井 隆行, 中村 友昭：マルチモーダルカテゴリゼーション, 人工知能学会誌, Vol. 27, pp. 555–562 (2012)

[西田 03] 西田 豊明：人とロボットの意思疎通, 情報処理, Vol. 44, No. 12 (2003)

[松井 00] 松井 俊浩, 麻生 英樹, John, F., 浅野 太, 本村 陽一, 原 功, 栗田 多喜夫, 速水 悟, 山崎 信行：オフィス移動ロボット Jijo-2 の音声対話システム, 日本ロボット学会誌, Vol. 18, No. 2, pp. 142–149 (2000)

[山肩 04] 山肩 洋子, 河原 達也, 奥乃 博, 美濃 導彦：音声対話システムにおける物体指示のための信念ネットワークを用いた曖昧性の解消, 人工知能学会論文誌, Vol. 19, No. 1, pp. 47–56 (2004)

[岡田 10] 岡田 浩之, 大森 隆司：ロボカップ@ホーム：人とロボットの共存を目指して, 人工知能学会誌, Vol. 25, No. 2, pp. 229–236 (2010)

[河原 13] 河原達也：音声対話システムの進化と淘汰–歴史と最近の技術動向–, 人工知能学会誌, Vol. 28, No. 1, pp. 45–51 (2013)

[松田 13] 松田繁樹, 林輝昭, 葦苅豊, 志賀芳則, 柏岡秀紀, 安田圭志, 大熊英男, 内山将夫, 隅田英一郎, 河井恒, 中村哲：多言語音声翻訳システム“ VoiceTra ”の構築と実運用による大規模実証実験, 電子情報通信学会論文誌, Vol. J96-D, No. 10, p. in print (2013)

[杉浦 12] 杉浦孔明：ロボカップ@ホームリーグ, 情報処理, Vol. 53, No. 3, pp. 250–261 (2012)

[杉浦 13a] 杉浦孔明, 志賀芳則, 河井恒, 翠輝久, 堀智織：サービスロボットのための非モノローグ HMM による音声合成, 第 31 回ロボット学会学術講演会資料, pp. 2C1–02 (2013)

[杉浦 13b] 杉浦孔明, 堀智織, 是津耕司：rospeex: クラウド型音声コミュニケーションを実現する ROS 向けツールキット, 信学技報（CNR2013-10）, 第 113 巻, pp. 7–10 (2013)

[浅田 10] 浅田稔, 松原仁：ロボカップ創世記, 情報処理, Vol. 51, No. 9, pp. 1195–1200 (2010)

[大浦 13] 大浦圭一郎, 山本大介, 内匠逸, 李晃伸, 徳田恵一：キャンパスの公共空間におけるユーザ参加型双方向音声案内デジタルサイネージシステム, 人工知能学会誌, Vol. 28, No. 1, pp. 60–67 (2013)

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B303-06

騒音下における声の張り上げ現象の計算機による実現に向けて

Towards Computational Implementation of Phenomenon of Raising Voice in Noisy Environment

北原 鉄朗 † **小暮 計貴** ‡ **吉永 眞宏** † **鈴木 光** †

Tetsuro Kitahara† **Kazuki Kogure**‡ **Masahiro Yoshinaga**† **Hikaru Suzuki**†

† 日本大学文理学部 ‡ 日本大学大学院総合基礎科学研究科

†**College of Humanities and Sciences, Nihon University**

‡**Graduate School of Integrated Basic Sciences, Nihon University**

{kitahara, kogure, yoshinaga, hikaru}@kthrlab.jp

Abstract

雑音の大きい環境では，人間は自然と声を張り上げてしまうことがある．このことは，人間による音声発話には聴覚系からのフィードバックが有することを示唆しており，雑音の大きい環境で対話相手に確実に聴こえる発話をするのに役立っていると思われる．本研究では，雑音の大きい環境で有効に動作する音声対話システムを実現する上で，この現象を計算機上で再現することが鍵になると考え，そのための課題と予備的検討について述べる．

1 はじめに

ヘッドフォンをして音楽を聴いている状態で話しかけられると，妙に大きな声で返事をしてしまう場合がある．これは，音声発話に聴覚系のフィードバックが強く働いていることを示している．ヘッドフォンに限らず，周囲の雑音が大きい環境にいると，自然と声を張り上げてしまうことはよく知られている．これにより，静かな場所で発声された音声に比べてインテンシティが大きくなるだけでなく，基本周波数やフォルマント周波数が高くなるなど，様々な音響的特徴が変化する．このことは**ロンバード効果** [Lane 71] と言われている．

一方，音声対話システムにおける音声発話部には，このような特徴はもちろんない．周囲の雑音状況とは無関係に，あらかじめ決められた声色で音声を合成し，あらかじめ決められた音量でそれを再生する．そのため，雑音の状況が動的に変わるような環境では，周囲が静かなときには声が大きすぎ，うるさいときには逆に聞こえないという事態になりかねない．携帯電話のようなユーザが個人的に所有・使用するような場合はユーザが自ら音量調整することもできるが，駅での運行案内など，公共の場で用いられることを想定したシステムでは，ユーザが音量を調整するのは容易ではない．

音声対話システムが広く社会で用いられるようになる上で，雑音耐性が重要であることは言うまでもない．これまで雑音下音声認識については非常に多くの研究がなされてきたが，雑音の状況が動的に変化する環境で，システムの発話を確実にユーザに聴こえるようにする工夫については，あまり研究されてこなかった．音声強調や音声明瞭化などの研究は様々なものが存在する（e.g., [Arai 02, 荒井 07, 竹山 06]）が，雑音が動的に変化する環境で，音量やその他の音響的特徴を自動的に調整して，ユーザが確実に発話内容を聞き取れるようにする試みではなかった．

我々は，このようなことの実現を目指す上で，上述のロンバード効果が参考になると考えている．つまり，ロンバード効果を計算機上で再現することが，動的に変化する雑音状況に適切に対処する音声発話への近道だと考えている．本稿では，ロンバード効果について簡単にまとめた後，それを計算機上で実現する上での課題について述べる．その後，できるだけ単純化して実現した場合の予備的な検討結果について述べる．最後に，その検討結果によって分かった問題点を挙げ，その解決案について議論する．

2 ロンバード効果について

ロンバード効果については様々な研究結果があるが，ここではその一例として程島らによる研究結果 [程島 09] を紹介する．

程島らは，静かな環境（Q），雑音のある環境（N），2種類の残響のある環境（R1，R2）で，東京方言話者 4 名（男女 2 名ずつ，22~37 歳）に様々な単語や音素バランス文を発声してもらった．雑音は白色雑音を使用し，発話者の耳元で平均 80dBA になるように騒音計を用いて音量を調整した．

その結果，基本周波数（F0）と第 1 フォルマント（F1）については，Q 条件に比べて N 条件，R1 条件，R2 条件

いずれも有意に上昇した．一方，第 2 フォルマント（F2）については，N 条件，R1 条件，R2 条件いずれも Q 条件に比べて有意差はなかった．子音と母音のインテンシティ比（CVR）は，N 条件，R1 条件，R2 条件いずれも Q 条件に比べて減少した．音圧レベルは，Q 条件と比較して N 条件，R1 条件，R2 条件いずれも増加したが，Q 条件に対する増加量は R1 条件，R2 条件の方が N 条件よりも少なかった．

3 ロンバード効果を計算機上で実現する上での課題

ここでは，このロンバード効果を計算機上で実現する上で解決すべき課題について述べる．課題は，大きく次の 2 つに分けることができる．

課題 1 雑音測定

課題 2 発話パラメータ設定

課題 1 は，その名の通り，どのように雑音の音量を測定するかである．課題 2 は，次の 3 つに細分化される．

課題 2-1 どの音響的特徴を変化させるか．

課題 2-2 どのタイミングで音響的特徴を変化させるか．

課題 2-3 どの程度の値に音響的特徴を変化させるか．

課題 1 に対して最も単純な方法が

案 A 音声対話（ユーザ音声の入力）用以外に雑音測定用のマイクロフォン（あるいはマイクロフォンアレイ）を用意し，それで計測された音響信号の振幅を求める

という方法である．この方法は単純で実装も容易であるが，自己発話（システム自身が発話した音声をこう呼ぶこととする）やユーザ発話も雑音とみなしてしまう場合がある．そのため，システムの判断によって自己発話の音量を大きくすると，それによって雑音が大きくなったと判断されるので，より一層自己発話の音量を大きくしようとし，発散してしまうという問題がある．この問題を解決しようとするのが次の案である．

案 B 自己発話やユーザ発話を抑制してから振幅を求める

自己発話については，雑音に重畳される自己発話の音響信号は既知なので，その分を減算して抑制することで，雑音のみの音量をより正確に推定できると考えられる．また，ユーザ発話については音声対話（ユーザ音声入力）用のマイクロフォンから得られた音響信号を参照信号として同様の処理を行う方法が考えられる．

課題 2-1 については，

案 A 音量のみ制御する

案 B 基本周波数，フォルマントも制御する

の 2 案が考えられる．案 A の場合は，音声合成エンジンによって生成された合成音声の再生系を制御すればよいので，音声合成エンジンと実装を切り離すことができ，比較的容易に実装できるというメリットがある．案 B の場合は，音声合成エンジンに対して基本周波数やフォルマントを制御する必要があるので，そのような制御が可能な音声合成エンジンを使用する必要がある．

課題 2-2 については，

案 A 音声発話開始時にのみパラメータ設定を変更する

案 B 音声発話中も時々刻々と動的にパラメータ設定を変更する

の 2 案が考えられる．案 A は実装が単純化されるだけでなく，上で述べたような，システムによる発話の音量が上がることによって雑音の音量が上がったと判定されてシステム発話の音量を上げてしまい，これが繰り返すことによって音量の設定が発散する事態を防ぐことができる．しかし，発話開始直後に大きな雑音が突発に発生しても対処できないという問題がある．そのため，長い発話には特に不向きである．

課題 2-3 は，たとえば雑音が 80dBA だと分かったときに，システム発話の音量やその他の音響的特徴をどれだけ上げ下げしたら，容易に聞き取れて大きすぎない音声になるか，という課題である．最も単純な方法は，

案 A 雑音の音圧レベルをいろいろ変えてみて，各音圧レベルに対してちょうどいい音量設定（音量以外も変えるならそのパラメータ）を実験的に調査する

という方法であろう．この案の最大の問題点は，環境依存になってしまうことである．音の聴こえ方は雑音の種類，対話音声用のスピーカーの設置角度など，様々な要因によって変化してしまうため，運用環境ごとに調査が必要となってしまい，汎用性に問題が生じる．また，調整すべきパラメータが増えたときに調査は大変困難になる．そこで，次のような案が考えられる．

案 B 音声の明瞭度を何らかの基準で定義し，その基準を満たすようにパラメータを最適化する

システム発話の音響信号は既知であるので，その音響信号が他の音源に比べて十分に優勢であるかを何らかの方法で測定できれば，その優勢度を音声の明瞭度として用い，この値が一定値を超えるように音量やその他のパラメータを自動的に最適化することができるであろう．たとえば，システム発話の音響信号のスペクトルピークとそれ以外（雑音）のスペクトルピークを比較し，SN 比を算出してこれを明瞭度とみなすなどの方法が考えられよう．

案 C 音声認識させてみて一定以上の精度が出るようにパラメータを最適化する

ユーザが音声を聞き取れるか（内容を認識できるか）が重要であると考えるのであれば，雑音入りの音声をシステムで認識させてみて，その認識が成功するように音量などを調整するという方法も考えられる．しかし，システムによる音声認識の精度は人間によるそれに比べて（特に雑音環境下では）低く，人間がきちんと聞き取れれば十分という観点では，システムによる音声認識精度を基準とするのは，過剰要求であるとの考え方もあるであろう．

このように，ロンバード効果を計算機上で実現するには，様々な課題がある．我々は現在，これらの課題を解決すべく，検討を進めている．まずは最も単純な方法（各々の案 A）を試し，その後，より複雑な方法（案 B，案 C など）を試すという方針で進めている．

4 予備的検討

本章では，3. の議論に基づいて行った予備的検討 [鈴木 14] について述べる．この予備的検討では，次の方針を採用した．

課題 1 案 A を採用.

課題 2-1 案 A を採用.

課題 2-2 案 B を採用.

課題 2-3 案 A を採用.

課題 2-2 のみ案 B を採用したのは，この予備的検討に先だって行った実験で案 A を採用したところ，発話開始直後に大きな雑音が発生してシステム発話が聞き取れない事態が頻出したためである．

4.1 システム構成

実験用システムの構成を図 1 に示す．この実験用システムは，利用者が発話用マイクロフォンの手前に位置して音声対話を行うことを想定している．発話用マイクロフォンの近くにシステムによる音声発話用のスピーカーが設置され，発話用マイクロフォンとは別に，雑音測定用にマイクロフォンアレイが設置されている．音声対話の内容は東京都内の乗り換え案内とし，利用者は「○○駅から▲▲駅まで行きたい」のように発話を行うと，システムは「駅すぱあと Web API」を用いて最短経路を取得し，音声合成による発話を行う．ただし，今回の実験用システムでは，システムによる発話が聞き取れるかどうかのみに目的を限定し，発話用マイクロフォンは使用しないものとする．また，後述のように，実際の経路を探索して案内するのではなく，あらかじめ用意した音声を聞かせるものとする．

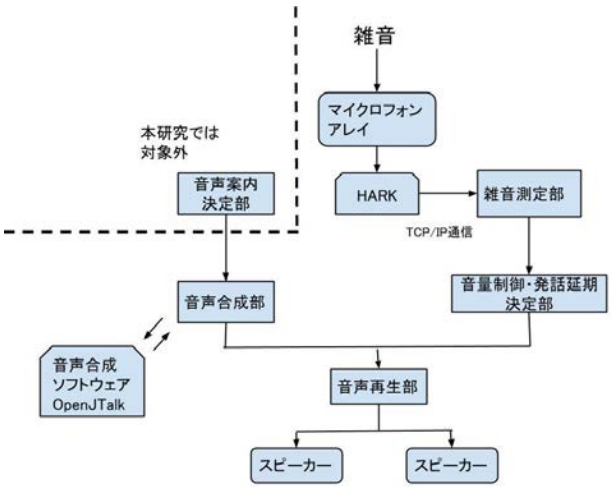


図 1: システム構成図

4.1.1 雑音レベルの計測

雑音レベルは，マイクロフォンアレイから得られる音響信号に基づいて推定する．現在の実装では，7ch のマイクロフォンアレイ「Microcone」からロボット聴覚オープンソフトウェア「HARK」[奥乃 10] を利用して約 1 秒間毎に音響信号を取得する．それに対して RMS を計算し，あらかじめ騒音計を用いて作成した RMS と騒音レベル (dB) の変換式に代入し，騒音レベル (dB) を算出する．

4.1.2 再生音量の変更

音量の変更は計算した雑音レベルを元に行う．システムの発話音量より周囲の雑音の方が大きい場合，雑音と同じ値まで発話音量を増幅する．また，周囲の雑音がシステムの発話音量より小さい場合は発話音量の縮小も行う．これにより環境に最適な発話音量の自動調整を実現する．

4.1.3 発話の延期

音量調整による雑音対策の他に発話の延期による対策も施す．これは，電車の警笛など最大音量を超える突発的な雑音に対応するためである．現在の実装では，64dB を超える雑音を感知した場合は発話を中断し，1 秒毎に雑音の計測を行い，64dB を下回ったときに発話を行うようになっている．

4.2 実験

提案手法によって利用者がシステムの発話を聞き取りやすくなったかどうかを実験する．

4.2.1 実験方法

実験は外からの騒音が入りにくい部屋で行った．被験者は 21 歳から 24 歳の正常な聴力を有する男性 3 人，女性 3 人の計 6 人である．被験者の位置を中心に 60 度おきに 6 箇所にスピーカーを設置した（図 2）．以下の流れで実験を行った．

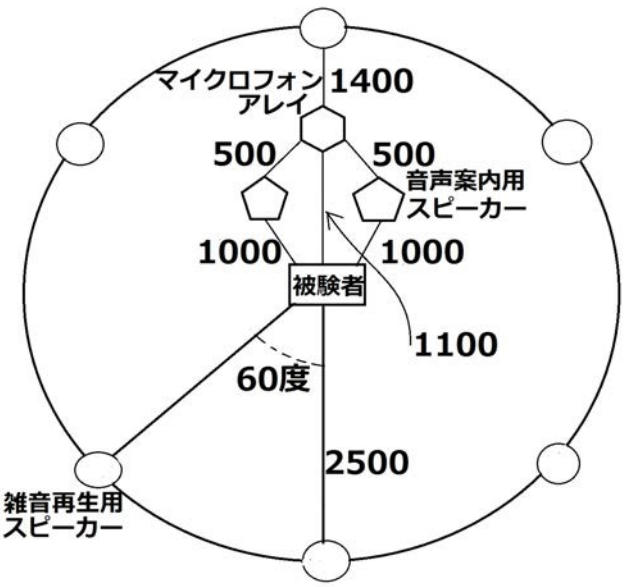


図 2: 実験時の機器配置図 [mm]

1. スピーカーから雑音を再生する.
2. 雑音に慣れてもらう (30 秒間).
3. 音量調整を行わない音声発話を再生.
4. 提案手法の音声発話を再生.

雑音は Microcone を使って東京都内の駅のホームで録音したものを再生し，発話には OpenJTalk で音声合成した女声を用いた．発話頻度は 10 秒に 1 回である．被験者は実験中に音声発話が聞こえたかどうかを 6 段階で評価する．それぞれの評価は以下のようにした.

1 まったく聞こえなかった

2 声は聞こえるが何を言っているのかわからない

3 聞き取れるが，大部分が聞き取りにくい

4 聞き取れるが，一部聞き取りにくい

5 聞き取れる

6 聞き取れるが，音量が大きすぎる

発話内容は,「○○から△△までの料金は□□円です.」とし，次のような 4 つの選択肢から聞き取ったものを選んでもらった.

- 「青砥から青井までの料金は 230 円です.」
- 「青砥から青井までの料金は 230 円です.」
- 「青井から青砥までの料金は 530 円です.」

表 1: 通常の音声発話による結果．各セルの左は各回答番号を選んだ回数，右はその中で選択問題に正答した割合を表す.

被験者	回答番号 6		5		4		3		2		1	
A	1	100%	10	90%	5	80%	3	67%	1	—	0	—
B	0	—	17	94%	2	100%	2	50%	0	—	0	—
C	0	—	12	100%	5	100%	3	100%	0	—	0	—
D	0	—	14	93%	1	100%	2	50%	4	—	0	—
E	0	—	16	94%	1	0%	0	—	0	—	3	—
F	0	—	15	93%	2	0%	2	0%	2	—	0	—
平均	0.2	100%	14	94%	2.7	75%	2	58%	1.2	—	0.5	—

表 2: 提案手法による音声発話の結果．各セルの左は各回答番号を選んだ回数，右はその中で選択問題に正答した割合を表す.

被験者	回答番号 6		5		4		3		2		1	
A	1	100%	15	93%	5	0%	0	—	0	—	0	—
B	0	—	16	94%	5	100%	0	—	0	—	0	—
C	0	—	19	100%	2	100%	0	—	0	—	0	—
D	0	—	16	100%	4	100%	0	—	1	—	0	—
E	0	—	21	100%	0	—	0	—	0	—	0	—
F	0	—	20	95%	0	—	0	—	0	—	0	—
平均	0.2	100%	17.8	97%	2.7	69%	0	—	0.2	—	0	—

- 「青井から青砥までの料金は 530 円です.」

音声発話を聞き取れるかどうかを確かめるのが目的なのでこの実験では「駅すぱあと Web API」を用いず，また，元々料金を知っていることの効果を防ぐため，でたらめな料金を案内することとした．他に「音量は適切だったか」,「発話の遅延は適切だったか」など設問回答型のアンケートも実施した.

4.2.2 実験結果と考察

通常の音声発話を再生した結果を表 1，提案手法による音声発話の結果を表 2 に示す.

表は横軸が聞き取りやすさの違いによる 6 段階の評価，縦軸が各被験者を表している．結果の左側がそれぞれの評価が記録された回数，右側が類似文章による選択問題の正答率となっている．それぞれの結果で，音声発話が正常に行われなかったデータは削除している．また，6 段階評価の [1], [2] については聞き取れなかった評価のため，4 択問題による聞き取り判断は行わないものとする.

表 1 の結果から，音量調整を行わない場合は 6 人中 4 人が [1] または [2] の評価をしているため，聞き取りに困難を感じていると判断できる．平均に着目すると，すべての人が 1 回以上は聞き取りに困難を感じている結果となった．また，評価 5 を選択した回数は提案手法を用いた場合，通常再生より平均で 3.8 回増加し，6 人中 5 人で評価 5 を選択した回数が増加している．類似文章による選

表 3: 再発話

被験者	発生発話数	評価 5	評価 4
A	6	5	1
B	6	5	1
C	6	6	0
D	7	6	1
E	7	7	0
F	6	5	1
平均	6.3	5.7	0.7

択問題の正答率も上昇していることから，提案手法によって聞き取りやすくなっていると言える．

６人の被験者のうち，被験者 B だけが提案手法を利用して適切な音量で聞き取れたと回答した回数が低下した．ここで被験者 B の類似文章による選択問題の正答率に注目すると聞き取れ方が不安定な時の場合，通常の音声発話の正解率が 50 %なのに対して提案手法の正答率は 100 %となった．これは聞き取りやすく感じた回数は低下したが，実際に正しく聞き取れた回数は増加したことを意味している．

また，聞き取りが不安定な時の正答率の平均が低下した原因については以下のようなことが考えられる． 提案手法が音量調整を行う際に周囲の環境音が静かなときに必要以上に再生音量を小さくしてしまう． 実際に，アンケートでは提案手法で周りが静かになったときに音量が音声発話の音量が小さくなりすぎていたとの回答が複数得られた．

提案手法による延期が発生した発話の全てが評価５または評価４であった．その結果を表３に示す．左の列から「被験者」,「延期が発生した発話数」,「延期が発生した発話で評価５が記録された回数」,「延期が発生した発話で評価４が記録された回数」となっている．通常の音声発話と提案手法を比較すると，評価５の出現回数が平均 3.8 回程度増加し，評価４以下の出現回数が平均 3.5 回減少していることから，発話の延期が行われることによって聞き取りづらく感じていた発話が聞き取りやすくなったと考えられる．

5　今後の改良に向けて—まとめに代えて

4. で述べた予備的検討では，自己発話が雑音と判断されて自己発話の音量上昇によって発話の音量を上げ続けてしまう現象を防ぐ根本的な解決は行わなかったため，雑音が 64dB を超える場合は音声発話を行わず，雑音が収まるまで発話を延期させる方策をとった．これにより，電車の通過のような突発的な雑音が発生した場合でも発話を聞き取れるようになったが，一方，待たされる時間が長いという意見があった．このことから，3. での議論の通り，課題 1 に対して案 A では不十分で，案 B を検討することが重要であることが明らかになった．課題 2-3 に対しては，今回 1ヶ所でしか実験を行っていないので，設置環境に対する汎用性については未検証であるが，今回取った方法（案 A：音圧レベルごとにちょうどいい音量などの設定値を実験的に調査する）は事前調査に要する手間が大きく，案 B あるいは案 C を検討する必要があることが分かった．課題 2-2 に対しては，音声発話開始後に突発的な雑音が発生したときに，すぐにそれに合わせて音量が上昇する点はよかったが，雑音の音量変化に合わせてシステム発話の音量を細かく上下させたことにより，発話が不自然になることがあった．また，発話途中に雑音が収まると，それに合わせて音量が低下するために，聞き取りにくくなる場合があって．そのため，音量を上昇させる場合は素早く，低下させる場合はゆっくりと行うなどの工夫が必要であることが分かった．

このように，3. で述べた課題の解決が重要であることが 4. の予備的検討で明らかになった．雑音環境下で有効に働く音声対話システムの実現のため，3. での議論に従って研究を進めていきたい．

謝辞

本研究は，SCAT 研究助成による助成を受けて実施されたものである．また,「駅すぱあと Web API」をご提供くださった（株）ヴァル研究所に感謝する．

参考文献

[Arai 02] Arai, T., Kinoshita, K., Hodoshima, N., Kusumoto, A., and Kitamura, T.: Effects of Suppressing Steady-state Portions of Speech Intellibility in Reverbant Environments, *Acoust. Sci & Tech.*, Vol. 23, No. 4 (2002)

[Lane 71] Lane, H. and Tranel, B.: The Lombard Sign and the Role of Hearing in Speech, *J. Speech Hear. Res.*, Vol. 14, pp. 677–709 (1971)

[奥乃 10] 奥乃 博：ロボット聴覚の現状と展望, 日本ロボット学会誌, Vol. 28, No. 1, pp. 2–5 (2010)

[荒井 07] 荒井 隆行：音声に関するバリアフリー, 音響研資, H-2007-66, pp. 377–382 (2007)

[竹山 06] 竹山 佳成：騒音環境下における車室内発話音声の分析とその合成に関する研究, Master's thesis, 北陸先端科学技術大学院大学 (2006)

[程島 09] 程島 奈緒, 荒井 隆行, 栗栖 清浩：雑音・残響下における発話の音響的特徴の話者変動, 信学技報, SP2009-69, pp. 43–48 (2009)

[鈴木 14] 鈴木 光, 吉永 眞宏, 小暮 計貴, 北原 鉄朗：雑音環境下のための音声案内システム：周囲の雑音レベルに合わせた音量の自動調整, 情処全大, 6S-1 (2014)

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-B303-07

周波数比の素数指数表現に基づく調性理解モデルとその応用可能性の検討

A Tonality Understanding Model Based on Prime Factor Representation of Frequency Ratio and Its Application Potentiality

白松 俊, 大囿 忠親, 新谷 虎松

Shun SHIRAMATSU, Tadachika OZONO, Toramatsu SHINTANI

名古屋工業大学 大学院工学研究科 情報工学専攻

Graduate School of Engineering, Nagoya Institute of Technology

{siramatu, ozono, tora}@nitech.ac.jp

Abstract

我々が提案する調性理解モデル PFG Tonnetz は，音名や階名といったシンボルからではなく，協和音程の知覚に関わる物理量 (音の周波数比) から直接導かれる．本稿ではその工学応用として，音楽知識を必要としない身体動作により入力された旋律線 (pitch contour) から，調性感を損なわない音高を出力する手法を示す．さらに，即興合奏のような多人数音楽インタラクションへの参加支援という応用を検討する．

1 はじめに

調性感は，音楽を聴いたり演奏する上で重要な認知的要素の 1 つであり，主に協和音程／不協和音程の知覚や，更にその下位の物理現象 (複数音間の周波数比や調波構造の重なり) によって説明される現象である．文献 [波多野 87] によると，旋律歌唱能力の発達過程や旋律聴取の認知的処理は，(1) リズム (rhythm) 構造の処理，(2) 旋律線 (pitch contour) 構造の処理，(3) 調性 (tonality) 構造の処理，という 3 つの処理側面から捉えられる (図 1)．本稿では特に (3) の調性構造に着目し，(音名や階名といったシンボルからではなく) 調の主音に対する周波数比から導かれる調性理解モデルについて述べる．

その前にまず，調性理解モデルが満たすべき要件を考える前提として，以下の 2 つの工学的応用を想定する．

1.1 多人数音楽パフォーマンスへの参加支援

近年，地域活性化のために住民参加型の音楽イベントが多く企画されている [堺都市 11]．地域活性化の効果を向上させるには，イベント参加者の裾野を広く，間口を広くする仕掛けが望まれる．例えば「音楽は好きだが，音楽知識／音楽経験は豊富でなく，そのイベントで演奏される楽曲にも詳しくない」という層が，情動の赴くまま自由かつ

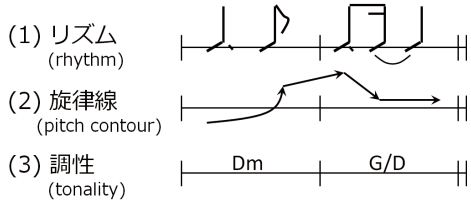


図 1: 旋律歌唱や旋律聴取における 3 つの処理側面

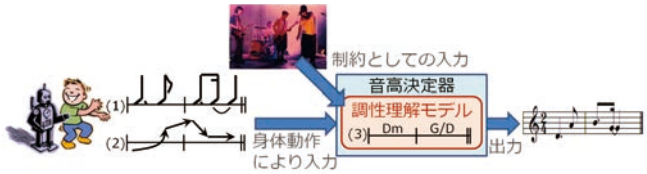


図 2: 想定する応用: 身体運動で入力された旋律線から，調性感を損なわない音高を出力

能動的に演奏に参加できる仕掛けがあれば，地域活性化という趣旨に沿った新たな参加型音楽パフォーマンスや多人数音楽インタラクションの場をデザインできる可能性がある．そのためには，そのような人々の音楽知識や音楽経験の不足を補う技術が必要となる．

通常，そのような人々が演奏時に参加する手段は，手拍子や掛け声，あるいはリズムに乗って体を動かす程度である．これらは，上記 3 つの側面のうち (1) のリズム構造だけを使った参加形態であり，比較的自由度が高い．一方，持続的な調波音を用いた参加形態としては，例えばステージ上の演奏者が同じ旋律パターンを繰り返し提示した上で，聴衆にもその旋律の繰り返しを要求し，それに応えた聴衆がユニゾンで歌唱するといった形態が一般的であるが，これは自由度が低い．持続的な調波音を自由に発し，かつ演奏音との齟齬を生じさせないためには，(3) の調性構造の理解が不可欠である．しかし，これは音楽経験が豊富でない人々にとって敷居が高い．この解決のため，本稿では図 2 のように，(1) リズムと (2) 旋律線をユーザが自由に入力すると，(3) 調性の整合性が保証された旋律に変

換されるような機構を考える．

ここで (2) の旋律線 (pitch contour) とは，前ページ図 1 に示したように，音高の上昇，同音，下降の組み合わせによる運動のパターンである．なお本稿では，2 度進行，3 度進行のような旋法に関する要素は (3) の調性にも関係してくるため，(2) 旋律線とは区別して考える．すなわち，運動のパターンのおおまかな概形として旋律線を捉える．この前提に立つと，リズムと旋律線の処理は，調性の処理よりも音楽知識や音楽経験の影響が少ない．リズムと旋律線は，音楽知識が乏しくても身体動作を介して生成しやすい構造であろう．特に旋律線と身体動作の親和性については文献 [菅 08] でも指摘されている．そこで，自由な身体動作でリズムと旋律線を入力できるよう，加速度センサーやセンシング入力デバイスを用いて身体動作の上昇・下降を検出する．ここで入力された旋律線から音高への変換を担うのが調性理解モデルであるが，周囲の演奏音との調性感を損なわないために重要なのは調推定やコード名推定の精度向上ではなく，周囲の演奏音との協和／不協和を制御する性能であろう．つまり，音響信号から抽出された基本周波数からシンボル処理を介さずダイレクトに調性を処理するモデルが適していると考える．

1.2 即興合奏に参加可能な音楽ロボット

調性を損なわずに即興合奏に参加できるロボットを実現する上でも，上記 3 種の構造に対応する処理モジュールをどう構成すべきか考える必要がある．3 つの構造のうちリズムと旋律線については，身体動作や情動の動きとの関係が指摘されており [Brown 00, Dogantan-Dack 13]，そのような認知機構を考慮に入れた生成機構が必要となろう．そこで生成されたリズムと旋律線を入力とし，調性感を損なわない音高を出力する調性理解モデルを考えると，やはり 1.1 節で示したような要件を満たす必要があると考えられる．

2 PFG Tonnetz: 素数指数表現に基づく調性理解モデル

ある調の主音の周波数 f_{tonic} と，協和音程の周波数 f の関係は，$f = \frac{3}{2} f_{\text{tonic}}$ (完全 5 度) や $f = \frac{4}{3} f_{\text{tonic}}$ (完全 4 度) のような単純な比になることが知られている．この周波数比は，

$$f = \Big(\prod_{p \text{ は } n \text{ 以下の素数}} p^{(z_p)} \Big) \cdot f_{\text{tonic}} \quad (z_p \text{は整数}) \qquad (1)$$

のような素数の積で表すことができる．例えば素数の上限を $n = 5$ とおくと，完全 4 度の場合は

$$f = (2^2 \cdot 3^{-1} \cdot 5^0) \cdot f_{\text{tonic}} \qquad (2)$$

となる．

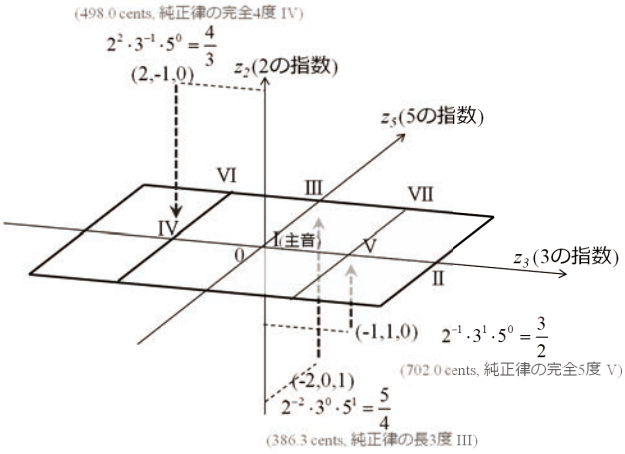


図 3: オクターブ般化のための $z_3 z_5$ 平面への投影

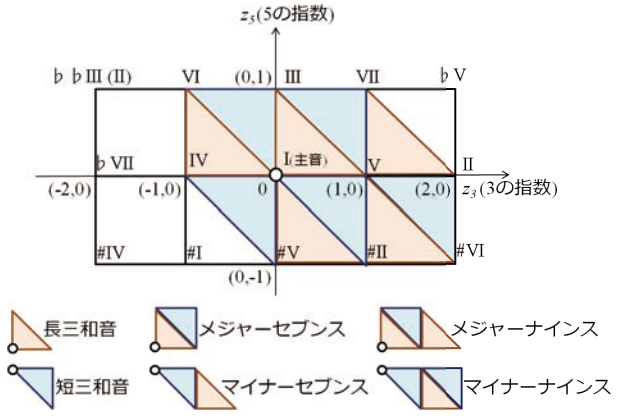


図 4: 調性理解モデル PFG Tonnetz (5-limit)

このような主音 f_{tonic} と f の関係を，各素数 p の指数 z_p から成るベクトル $(p_2, p_3, \cdots, p_n)$ で表すことを考える．このベクトルは，整数論で用いられる素数指数表現 [Graham 93] を拡張し，指数 z_p が負の整数になるのを許容することで協和音程の単純な比を表現可能にしたものである．例えば上記の完全 4 度の場合，主音との関係を表す $n = 5$ の素数指数表現は $(2, -1, 0)$ となる．このベクトルを $z_2 z_3 z_5$ 空間にプロットすると，原点が主音 f_{tonic} を表し，原点近傍の整数格子点が主音との協和音程を表すことになる（図 3）．

ここで，オクターブ隔たった音（$f_1 = 2^z f_2$ の関係にある f_1 と f_2）は音楽的に等価であることを考慮し，オクターブ関係にある音をまとめるオクターブ般化 (octave generalization)[Burns 87] を行う．具体的には，図 3 のように 2 の指数 z_2 を 0 にして $z_3 z_5$ 平面へ投影する．上記の完全 4 度の場合は，$z_3 z_5$ 平面上の $(-1, 0)$ へ投影される．

このように，主音から 1 オクターブ内におさまる原点近傍の整数格子点[1] (z_2, z_3, z_5) をオクターブ般化した (z_3, z_5) は，図 4 のように $z_3 z_5$ 平面に配置される．このとき，図中の赤い三角形，つまり $\big[(a, b), (a, b+1), (a+1, b)\big]$ の 3

[1] つまり $1 \le 2^{z_2} \cdot 3^{z_3} \cdot 5^{z_5} < 2$ となる原点近傍の整数格子点

点を辿った三角形は明るい響きの長三和音の表現形式となり，青い三角形，つまり $\big[(a,b),\,(a+1,b-1),\,(a+1,b)\big]$ の 3 点を辿った三角形は暗い響きの短三和音の表現形式となる．さらに，赤い三角形と青い三角形を交互に右方向（z_3 軸の正の方向）へ積み重ねていくことで，セブンスコードやナインスコード，さらにその上のテンションコードの表現形式が得られる．これを定式化するために，根音 (a,b) に対して以下のような整数格子点列 $\mathrm{code}(a,b,\delta,m)$ を考える．

$$\mathrm{code}(a,b,\delta,m)=\Big[(a,b)+\sum_{i=0}^{k}\delta(i)\Big]_{k=0,1,\cdots,m} \tag{3}$$

$$\delta_{\mathrm{maj}}(i)=\begin{cases}(0,1) & (i\text{ が奇数のとき})\\ (1,-1) & (i\text{ が偶数のとき})\end{cases} \tag{4}$$

$$\delta_{\mathrm{min}}(i)=\begin{cases}(1,-1) & (i\text{ が奇数のとき})\\ (0,1) & (i\text{ が偶数のとき})\end{cases} \tag{5}$$

このとき格子点列 $\mathrm{code}(a,b,\delta_{\mathrm{maj}},m)$ は，$m=2$ のとき長三和音，$m=3$ のときメジャーセブンスコード，$m=4$ のときメジャーナインスコードの表現形式となる．これらメジャーコードの構成音は，根音 (a,b) の上側 $b\le z_5\le b+1$ に分布する．一方，格子点列 $\mathrm{code}(a,b,\delta_{\mathrm{min}},m)$ は，$m=2$ のとき短三和音，$m=3$ のときマイナーセブンスコード，$m=4$ のときマイナーナインスコードの表現形式となる．これらマイナーコードの構成音は，根音 (a,b) の下側 $b-1\le z_5\le b$ に分布する．根音 (a,b) を調の主音 $(0,0)$ で置き換えて考えると，明るい長音階の構成音（I, II, III, IV, V, VI, VII）が原点の上側 $0\le z_5\le 1$ に分布し，暗い短音階の構成音（I, II, #II, IV, V, #V, #VI）が原点の下側 $-1\le z_5\le 0$ に分布するのが見て取れる．

ここまで，調性の表現形式を導く過程で用いた前提は，以下のような周波数比に基づく 2 つの認知的原理のみである．

1. 調性音程の周波数比は単純な整数比で表せる．よって，周波数比を素数指数表現 (z_2,z_3,z_5) で表せる（ただし $z_2,z_3,z_5<0$ も整数なら許容）．
2. 周波数比がちょうど 2^z（z は整数）である音程はオクターブ等価である．よって，素数指数表現 (z_2,z_3,z_5) の 2 の指数 z_2 を捨象し z_3z_5 平面上に投影することで，オクターブ般化できる．

つまり，音名，階名，コード名のようなシンボルを前提としては用いていない．まとめると，導出の過程で現れた以下の表現形式は，上記の認知的原理および導出されたモデルの観察から自然に導かれたものである．

- 整数格子点 (z_3,z_5): 音階の構成音，階名
- 三角形 $\big[(a,b),\,(a,b+1),\,(a+1,b)\big]$:
  整数格子点 (a,b) を根音とする長三和音
- 三角形 $\big[(a,b),\,(a+1,b-1),\,(a+1,b)\big]$:
  (a,b) を根音とする短三和音
- 整数格子点列 $\mathrm{code}(a,b,\delta_{\mathrm{maj}},m)$:
  (a,b) を根音とする長和音
- 整数格子点列 $\mathrm{code}(a,b,\delta_{\mathrm{min}},m)$:
  (a,b) を根音とする短和音
- $a\le z_5\le b+1$: (a,b) を根音とする長和音の分布領域
- $a-1\le z_5\le b$: (a,b) を根音とする短和音の分布領域
- $0\le z_5\le 1$: 原点を主音とする長音階の分布領域
- $-1\le z_5\le 0$: 原点を主音とする短音階の分布領域

この調性理解モデルは，著名な数学者 Leonhard Euler が 1739 年に考案し（図 5），音楽学者 Hugo Riemann が 1880 年に図 6 のように発展させた Tonnetz [Behringer 10] というモデルに似ている．Tonnetz は 1980 年代以降，数学的に定式化された新リーマン理論 (Neo-Riemannian theory) へと発展し，様々な拡張が行われた [Hewlett 07, Tymoczko 12] が，基本的には図 5，6 のように音名同士を繋いだモデルであり，調の主音は表現できていない．また，物理現象や認知機構に根ざした説明になっておらず，ヒューリスティックな前提が多い．

一方，提案する図 4 のモデルは，まず原点に調の主音という役割があり，そこからの相対的な位置関係によって調性が決まるという性質があるため，主音を変えても同じ座標系で表現でき，式 (3), (4), (5) のような定式化も容易となる．また，これまでは素数の上限 $n=5$ としてきたが，$n=7$ にして $z_3z_5z_7$ 空間へ拡張すると 7-limit 純正律 [Partch 74] を表現でき，$n=11$ にすると 11-limit 純正

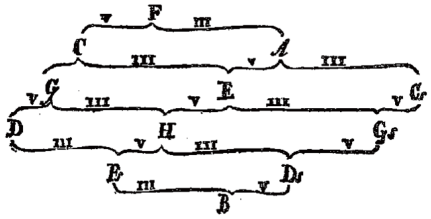


図 5: オイラーの Tonnetz

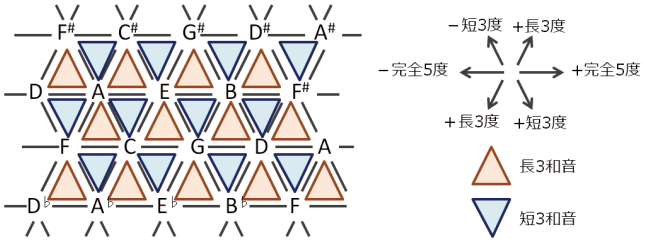


図 6: リーマンの Tonnetz

律を表現できる．さらに，周波数比という物理量と協和音程に関するシンプルな原理から導かれたモデルであり，従来の Tonnetz よりも前提が少ない．

これらのことから，提案するモデルは従来の Tonnetz の一般化になっていると考え，我々は過去の研究 [白松 12] で PFG Tonnetz (Prime-Facter-based Generalized Tonnetz) と名付けた．また，n-limit 純正律のアナロジーにより，$z_2 z_3 \cdots z_n$ 空間のモデルを n-limit PFG Tonnetz と呼ぶ．特に $n = 5$ の 5-limit PFG Tonnetz（図 4）は，通常の純正律（5-limit 純正律）に対応しており，$z_3 z_5$ の 2 次元平面で可視化でき，従来の Tonnetz とのトポロジー的類似性が理解しやすい等の利点があり，扱いやすい．よって，以後 n に関して指定のない PFG Tonnetz は，全て 5-limit PFG Tonnetz のことを指すものとする．また，以後，$z_3 z_5$ 平面の整数格子点とその集合を $g = (z_3, z_5) \in G$ で表す．

3 PFG Tonnetz を用いた旋律線からの音高決定

1 節で示したように，身体操作により入力されたリズムと旋律線から，周囲の演奏音の調性を損なわない音高を出力するために，調性理解モデル PFG Tonnetz を用いる．そのためには，周囲の演奏音の音響信号から調性の制約を決定する手法と，その制約と旋律線に合わせて音高を決定する手法が必要となる．

3.1 音響信号からの調性制約の決定

まず前処理として，観測された周囲の多重奏音響信号に含まれる複数の基本周波数を推定する．この前処理を行うツールとして，Durrieu らによるオープンソースの Vamp プラグインである IMMF0salience[2] [Durrieu 11, Salamon 14] を用いる．IMMF0salience は，IMM (Instantaneous Mixture Model) と呼ばれる手法により，多重奏音楽音響信号中の複数の基本周波数 F0 を推定し，中間的な解析結果を可視化・出力する．具体的には，周波数ビン毎に基本周波数 F0 の顕著性スコア (F0 salience representation; F0 顕著性表現)，言い換えると F0 存在可能性の推定結果を出力する．これは北原らの Instrogram[Kitahara 07] による可視化と似た中間的表現であるが，IMMF0salience の出力は各楽器音に分離する前の F0 推定結果である．通常，IMMF0salience の実行は Sonic Visualizer の GUI 上で行われるが，ここでは Sonic Annotator[3] を用いて以下のように CUI 上で実行し，前処理をバッチ処理化できる．

```
$ sonic-annotator -d vamp:f0salience:f0salience
WAV_FILE -w default > OUTPUT_FILE
```

デフォルト設定では 100Hz から 800Hz までの帯域で F0 推定が行われ，その結果が図 7 のような形式で出力される．feature タグがフレーム毎の推定結果を表し，timestamp タグが当該フレームの時刻を，values タグが各周波数ビンの F0 顕著性スコアを表す．

```
⋮
<feature>
<name>Salience of F0s.</name>
<timestamp> 12.445895691R</timestamp>
<values>100 Hz:2629.86 100 Hz:192.319 101 Hz:40.9141
102 Hz:7.2439 102 Hz:2.62986 103 Hz:2.62986 104
Hz:2.62986 105 Hz:2.62986 ··· 777 Hz:137.789 782
Hz:64.2944 788 Hz:164.495 0 Hz:25.5596</values>
</feature>
<feature>
<name>Salience of F0s.</name>
<timestamp> 12.469115646R</timestamp>
<values>100 Hz:26568.5 100 Hz:4487.88 101 Hz:116.993
102 Hz:116.993 102 Hz:116.993 103 Hz:116.993 104
Hz:116.993 ··· 777 Hz:116.993 782 Hz:116.993 788
Hz:4033.32 0 Hz:116.993</values>
</feature>
⋮
```

図 7: IMMF0salience が出力する F0 推定結果

このように得られた，フレーム t における周波数ビン f の顕著性スコアを $s(t, f)$ とする．この $s(t, f)$ から，以下のようなフレーム t における調性制約 tnl_t を推定する．

$$tnl_t = \left\langle f_{\text{tonic}}, h_t(f_{\text{tonic}}) \right\rangle \quad (6)$$

$$f_{\text{tonic}} = \arg\min_{f_{\text{cand}}} \sum_{g \in G, t=t_0,\cdots,t} w(g) \cdot h_t(f_{\text{cand}})(g) \quad (7)$$

$$h_t(f_{\text{cand}})(g) = \sum_{f \in F(f_{\text{cand}}, g)} s(t, f) \quad (8)$$

$$F(f_{\text{cand}}, g) = \{(2^{z_2} \cdot 3^{z_3} \cdot 5^{z_5}) \cdot f_{\text{cand}}\}_{z_2 \in \mathbb{Z}} \quad (9)$$

$$w(g) = \min_{k_1,k_2,k_3 \in \mathbb{Z}} (|k_1| + |k_2| + |k_3|) \quad (10)$$

$$g = (z_3, z_5) = k_1(1, 0) + k_2(0, 1) + k_3(1, -1) \quad (11)$$

ただし，f_{tonic} は調の主音の周波数，$h_t(f_{\text{tonic}})$ はフレーム t の F0 顕著性スコアの G 上の分布，f_{cand} は主音の候補の周波数, $w(g)$ は格子点 g の重み，$F(g)$ は格子点 g にオクターブ般化される周波数の集合，t_0 は推定に用いるフレーム列の開始番号，$\mathbb{Z}$ は整数の集合である．格子点 g の重み $w(g)$ は，完全五度に対応するベクトル $(1, 0)$，長三度に対応するベクトル $(1, 0)$，短三度に対応するベクトル $(1, -1)$ をネットワークのリンクと見なした時に，原点から g を辿れる最短ステップ数となる．

[2] https://github.com/wslihgt/IMMF0salience, http://www.durrieu.ch/research/jstsp2010.html
[3] http://www.vamp-plugins.org/sonic-annotator/

3.2 調性を損なわない音高決定

身体動作により入力された時刻 t の旋律線の高さを $x(t)$ とする．これを，調性制約 tnl_t を考慮した周波数 $f(t)$ に変換する．

まず，$tnl_t = \langle f_{\text{tonic}}, h_t(f_{\text{tonic}}) \rangle$ のうち主音の周波数 f_{tonic} のみを考慮した $f'(t)$ への変換を行う．

$$f'(t) = f_{\text{topic}} \exp(\alpha(x(t) - x_{\text{tonic}})) \tag{12}$$

ただし x_{tonic} は主音 f_{tonic} に対応付けた基準位置，α は位置変化と周波数変化の比率を調整する係数である．この $f'(t)$ を，$h_t(f_{\text{tonic}})$ を考慮して $f(t)$ に音高補正するには，例えば以下のような手法が考えられる．

$$f(t) = \underset{g \in Scale(t)}{\arg\min} \Big(|f(g) - f'(t)| + \beta \sum_{g' \in G} w(g - g') h_t(f_{\text{tonic}})(g') \Big) \tag{13}$$

$$Scale(t) = \{ g | \sum_{t=t_0,\cdots,t} h_t(f_{\text{tonic}})(g) \geq \theta \} \tag{14}$$

$$f(g) = (2^{z_2} \cdot 3^{z_3} \cdot 5^{z_5}) \cdot f_{\text{tonic}}, g = (z_3, z_5), z_2 \in \mathbb{Z} \tag{15}$$

ここで，$Scale(t)$ は取りうる格子点の集合，θ は時刻 t までの F0 顕現性スコア総和の閾値，β は調性をどれだけ重視するかという重みを表すパラメータである．ここでも式 (10) の重みを用いているが，不協和度 [藤澤 06] に基づく重みに置き換えることも考えられる．

なお本手法は未検証であり，今後，実験による性能評価が必要である．

4 本研究の位置付けと関連研究

4.1 調性の表現手法

素数指数表現に基づく音の表現方法は，従来研究でも提案されている [Monzo 99, Keislar 87]．しかし，これらの従来研究は Tonnetz との関連について触れておらず，素数指数表現に基づいて Tonnetz を一般化したのは本研究の独自の着眼点である．

調性の表現形式に関しては，五度圏 (Cycle of Fifth) など古くから多くのモデルや理論が存在し，五度圏を用いた調推定手法 [Inoshita 09] も提案されている．また Tonnetz に関しては，新リーマン理論への発展に伴いトーラス面上の Tonnetz 表現形式 [Hewlett 07] 等へ拡張され，さらに isomorphic keyboard [Milne 07] 等の楽器インタフェースにも応用されている．しかし，本稿で提案したような，観測音響信号から推定される周波数比を直接用いて調性制約を得るという手法へは応用されていない．

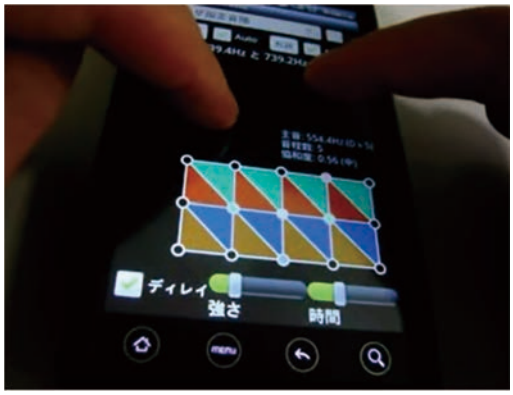


図 8: PFG Tonnetz に基づくスマートフォンアプリ TonalityTouch

4.2 スマートフォンアプリへの応用

図 8 は，我々が過去の研究で開発した TonalityTouch というスマートフォンアプリである [白松 12][4]．これは，PFG Tonnetz を用い，スマートフォン上のタッチ動作やスワイプ動作を調性感ある音高へ補正して発音するという機能を持つ．マルチタッチによる協和音生成も可能である．これにより，音楽知識や音楽経験に乏しいユーザでも調性感ある演奏が可能であるほか，PFG Tonnetz を用いて新たな音階を試行的に生成し，ユーザが気に入った新音階を用いて音高補正することができる．しかし，周囲の演奏音を考慮に入れていないため，1 節で述べたような即興合奏への参加支援のためには不充分である．

なお，スマートフォンを用いて参加型音楽パフォーマンスを支援するシステムの研究事例としては，massMobile[Weitzner 13] や SWARMED[Hindle 13] がある．これらは，本研究で扱っているような無作為な身体動作によって調性感を伴う即興合奏参加が可能になる機構は有していない．この点が本研究独自の着眼点となっている．

4.3 地域振興音楽イベントへの応用

我々はこれまで，地域住民向けのワークショップからの意見集約や，公共圏での協働を支援する技術を研究してきた [Shiramatsu 13]．地域コミュニティでの協働を円滑にするには，課題解決のための議論だけが重要なわけではない．緊張関係をほぐして場の空気を和ませるために，アイスブレイクと呼ばれる仕掛けが用意される場合がある．音楽は，身体動作の共有を通じてポジティブな情動を喚起する [寺澤 13] という社会的機能を持つので，1.1 節で触れたような音楽イベントが地域コミュニティの円滑化に寄与し，広い意味でのアイスブレイクの役割を果たす可能性は大いにある．調性理解モデル PFG Tonnetz を用いる

[4] https://play.google.com/store/apps/details?id=org.toralab.music.beta

ことで自由な身体動作が許容されるが，この機能がコミュニティ内でのアイスブレイク効果を向上させるかどうか，今後，改めて実証していく必要がある．

5 おわりに

本稿では，協和音程が単純な周波数比で表せるというシンプルな認知的原理から導かれる調性理解モデル PFG Tonnetz について述べた．これは，従来の Tonnetz の一般化になっており，また多重奏音楽音響信号からの基本周波数推定結果との親和性も高いモデルであると考えられる．さらにこの提案モデルを即興合奏への参加支援へ応用するために，身体動作によって入力された旋律線から，周囲の演奏音との調性感を損なわない音高を決定する手法を検討した．今後は提案手法の性能評価実験を行い，加速度センサーやセンシング入力デバイスを用いた旋律線入力機構と統合したシステムを開発する予定である．

参考文献

[Behringer 10] Behringer, R. and Elliot, J.: *Linking Physical Space with the Riemann Tonnetz for Exploration of Western Tonality*, chapter 6, pp. 131–143, Nova Science Publishers (2010)

[Brown 00] Brown, S.: The “musilanguage” model of music evolution, in *The origins of music*, pp. 271–300, MIT Press (2000)

[Burns 87] Burns, E. M. and Ward, W. D.: 音程，音階，調律, 第 8 章, pp. 301–334, 西村書店 (1987)

[Dogantan-Dack 13] Dogantan-Dack, M.: Tonality: the shape of affect, *Empirical Musicology Review*, Vol. 8, No. 3-4, pp. 208–218 (2013)

[Durrieu 11] Durrieu, J., David, B., and Richard, G.: A musically motivated mid-level representation for pitch estimation and musical audio source separation, *IEEE Journal of Selected Topics in Signal Processing*, Vol. 5, No. 6, pp. 1180–1191 (2011)

[Graham 93] Graham, R. L., Knuth, D. E., and Patashnik, O.: コンピュータの数学 (1993)

[Hewlett 07] Hewlett, W., Selfridge-Field, E., and Correia, E.: *Tonal Theory for the Digital Age*, Vol. 15 of *Computing in Musicology*, Center for Computer Assisted Research in the Humanities, Stanford University (2007)

[Hindle 13] Hindle, A.: SWARMED: Captive Portals, Mobile Devices, and Audience Participation in Multi-User Music Performance, in *Proceedings of the 13th International Conference on New Interfaces for Musical Expression*, pp. 174–179 (2013)

[Inoshita 09] Inoshita, T. and Katto, J.: Key Estimation Using Circle of Fifths, in *Advances in Multimedia Modeling*, pp. 287–297, Springer (2009)

[Keislar 87] Keislar, D.: History and principles of microtonal keyboards, *Computer Music Journal*, pp. 18–28 (1987)

[Kitahara 07] Kitahara, T., Goto, M., Komatani, K., Ogata, T., and Okuno, H. G.: Instrogram: probabilistic representation of instrument existence for polyphonic music, *IPSJ Journal*, Vol. 48, No. 1, pp. 214–226 (2007)

[Milne 07] Milne, A., Sethares, W., and Plamondon, J.: Isomorphic controllers and dynamic tuning: Invariant fingering over a tuning continuum, *Computer Music Journal*, Vol. 31, No. 4, pp. 15–32 (2007)

[Monzo 99] Monzo, J.: *JustMusic: A New Harmony Representing Pitch as Prime Series*, J. Monzo, 4th edition (1999)

[Partch 74] Partch, H.: *Genesis of a music: an account of a creative work, its roots and its fulfilments*, Da Capo Press (1974)

[Salamon 14] Salamon, J., Gómez, E., Ellis, D. P., and Rechard, G.: Melody Extraction from Polyphonic Music Signals: Approaches, applications, and challenges, *IEEE Signal Processing Magazine*, Vol. 31, No. 2, pp. 118–134 (2014)

[Shiramatsu 13] Shiramatsu, S., Ozono, T., and Shintani, T.: Approaches to Assessing Public Concerns: Building Linked Data for Public Goals and Criteria Extracted from Textual Content, in *Electronic Participation. Proceedings of the 5th IFIP WG 8.5 International Conference, ePart 2013*, Vol. 8075 of *Lecture Notes in Computer Science*, pp. 109–121, Springer (2013)

[Tymoczko 12] Tymoczko, D.: The Generalized Tonnetz, *Journal of Music Theory*, Vol. 56, No. 1, pp. 1–52 (2012)

[Weitzner 13] Weitzner, N., Freeman, J., Chen, Y.-L., and Garrett, S.: massMobile: towards a flexible framework for large-scale participatory collaborations in live performances, *Organised Sound*, Vol. 18, No. 01, pp. 30–42 (2013)

[堺都市 11] 堺都市 政策研究所 : 市民主体の地域活性化（音楽イベントを通じたまちおこし）に関する調査研究業務報告書, http://www.sakaiupi.or.jp/30.products/31.resarch/H22/H22_music.pdf (2011)

[寺澤 13] 寺澤 洋子, 星-柴 玲子, 柴山 拓郎, 大村 英史, 古川 聖, 牧野 昭二, 岡ノ谷 一夫 : 身体機能の統合による音楽情動コミュニケーションモデル, 認知科学, Vol. 20, No. 1, pp. 112–129 (2013)

[菅 08] 菅 道子 : 身体表現を取り入れた参加型音楽コンサートの可能性: カノンの理解を目指した 「追いかけっこしよう」の事例から, 和歌山大学教育学部教育実践総合センター紀要, Vol. 18, pp. 121–129 (2008)

[藤澤 06] 藤澤隆史, ノーマン D. クック : 和音性の計算法と曲線の描き方: 不協和度・緊張度・モダリティ, 情報研究: 関西大学総合情報学部紀要, Vol. 25, pp. 35–51 (2006)

[波多野 87] 波多野 誼余夫 : 音楽と認知, 認知科学選書, 第 12 巻, 東京大学出版会 (1987)

[白松 12] 白松 俊, 大囿 忠親, 新谷 虎松 : TonalityTouch: 素数指数表現に基づく一般化 Tonnetz を用いたスマートフォン楽器, 情報処理学会研究報告 音楽情報科学, Vol. 2012-MUS-96, No. 20 (2012)

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B303-08

音声可視化デバイス「カエルホタル」による ニホンアマガエル合唱の時空間構造解析

Spatio-temporal Analysis of Frog Choruses using Sound-to-Light Conveting Device *Firefly*

水本 武志 [1]　**合原　一究** [2]　**奥乃 博** [3]

Takeshi MIZUMOTO Ikkyu AIHARA Hiroshi G. OKUNO

[1] 株式会社ホンダ・リサーチ・インスティチュート・ジャパン [2] 理化学研究所 [3] 京都大学

[1] **Honda Research Institute Japan, Co., Ltd.** [2] **RIKEN** [3] **Kyoto University**

t.mizumoto@jp.honda-ri.com　ikkyu@brain.riken.jp　okuno@i.kyoto-u.ac.jp

Abstract

本稿では生物の合唱可視化システムと, それを用いたニホンアマガエルの合唱の時空間構造解析について報告する. ニホンアマガエルのような小型の夜行性生物の生息地における合唱の時空間構造は, 各発声の位置と時刻の計測が困難なため明らかではなかった. そこで我々は, 音を光に変換するデバイス「カエルホタル」によって野外の合唱を可視化し, 同期状態を解析するシステムを開発した. フィールド実験の結果, 6 匹の合唱を観測した. その解析の結果, 隣り合う個体同士が逆相同期しており, 距離の近いペアほど同期が安定していた.

1　はじめに

日本では, 梅雨の時期にニホンアマガエル (*Hyla japonica*) の合唱が広くみられる (図 1). 彼らの合唱をよく聴いていると, 互いにタイミングを合わせているように聞こえることがある. 実際, 室内で 3 匹のニホンアマガエルを 1 列に配置すると, 隣り合う個体同士が交互に鳴いたり (1:2 逆相同期), 各個体がワルツのように順番に鳴く (3 相同期) 現象を観察できる [1]. 合唱する種は, カエルに限らず, コオロギ, バッタ, ゴリラ, コウモリなど多く知られている. 合唱の目的は, 繁殖, 縄張りの維持, 餌の定位など多岐にわたり, 合唱中の個体位置 (空間構造) も, その時間間隔 (時間構造) も様々である.

合唱の時空間構造, すなわち各個体の鳴く位置と時刻, の計測は彼らの音声コミュニケーションを明らかにする上で重要である [2]. 典型的には観測者の耳に頼った計測方法が用いられていたが, 得られる情報の精度と質には限界があった. 特にカエルをはじめとする小型の夜行性生物は, 目視が困難な夜間に鳴くうえ, 人が近づきすぎると鳴き止むので距離を取らなければならず, 野外での詳細な音声の計測は困難であった. そのため, 音声コミュニケーションの実験は主に屋内で行われていた [2].

図 1: ニホンアマガエル (*Hyla japonica*)

本稿で対象とするニホンアマガエルとその合唱の特徴は以下のとおりである. 繁殖期は春から夏であり, その時期に水辺, 主に水田に集まって交配する. このときの集団での発声行動が**合唱**と呼ばれる. 鳴くのはオスのみで, 鳴き声の種類には, *advertisement call* と呼ばれるメスへアピールする鳴き声や, *aggressive call* と呼ばれる縄張り維持のための鳴き声などが知られている [2]. 各個体は 1 秒間に 3-4 回という高頻度で鳴くので, 合唱中は常に複数個体が同時に発声している. 体長は通常 3-4cm と小さく, 水田では 1-2m の間隔を空けて疎に分布している [3].

以上の議論から, 小型夜行性生物の合唱における時空間構造を計測するシステムは, 次の 4 つの要求条件を満足する必要がある.

1. 対象種の生息地で使用可能
2. 複数個体の同時発声を検出可能
3. 動物行動への影響の最小化
4. 多様な空間配置の動物へ適用可能

本稿では, 安価で容易に使用できる合唱可視化システム (*Sound Imaging System*) と, それを用いたニホンアマガエルの合唱構造の解析について述べる. 本システムは, 数十の音声可視化デバイス **カエルホタル** (図 2) と, 市販のビデオカメラから構成される. まずカエルホタルを対象生物の合唱するフィールドに配置する. 各カエルホタルは

周囲の音を検出して LED 発光に変換するので, ビデオカメラでその発光パターンを録画する. 次に動画から LED の輝度時系列を抽出し, それを表示することで合唱の時空間構造を可視化する. さらに各発声時刻の系列から合唱中の同期状態を解析する. 本システムは, 次のとおり要求条件を満たしている. (1) 生物の生息地に設置して使用可能. (2) 各カエルホタルのマイクは近い位置の音声のみを検出するので, 離れた個体が発声すればその個体の近くの LED が光る. したがって, 光のパターンから同時発声を検出可能. (3) 発光パターン計測中はフィールドから離れてよい. (4) カエルホタルの配置数や間隔を変えれば異なる空間配置へ応用可能.

本稿の構成は以下のとおりである. まず 第 2 節 で関連研究をまとめる. 次に 第 3 節 で合唱可視化システムについて述べる. 第 4 節でフィールド実験とその結果について議論し, 第 5 節で結論を述べる. なお, 本論文は文献 [4,5] の概説である. 詳細は文献を参照されたい.

2 関連研究

生物学の観点からの音声コミュニケーションについては, 古くから様々な種について調べられてきた. たとえば, カエルの合唱や [6,7], コオロギの合唱 [8], コウモリのエコーロケーション [9,10] などの報告がある.

音声コミュニケーションの計測手法はさまざまな手法が提案されているが, いずれもニホンアマガエルの合唱の計測への適用は困難である. まず, 人の耳で聞いて記録する方法は, カエルが多数の個体が同時かつ頻繁に鳴く特徴を持つので, 現実的ではない. 近年, バイオロギングと呼ばれる動物にデータロガーを装着する方法が広く使われており, ウミガメの回遊ルート [11] やイルカの音声コミュニケーション [12] などが計測されている. この手法の問題はデータロガーのサイズと個体数である. ニホンアマガエルは小型動物なので装着すると行動に影響を与えてしまい, そのうえ数十の全個体に装着しなければ各鳴き声の時空間情報が得らないため, 現実的ではない.

マイクアレイ処理による音源定位や音源分離は有望なアプローチである. 到達時間差による音源定位が最も広く用いられており, ウシガエルや [13] ヒキガエル [14] などの発声タイミングの推定と行動解析の報告がある. また, 複数マイクアレイとその全地球測位システム (GPS) を組み合わせた鳴き声の定位手法も提案されている [15]. これらの手法の問題点は, 発声の時間的スパースネスを仮定していることである. ウシガエルやヒキガエルなどの発声間隔の長い種では同時に鳴くことは少ないので問題ないが, ニホンアマガエルではほぼ常に同時発声が生じているので仮定が成立しない. より高度な信号処理手法, たとえば独立成分分析を用いたブラインド音源分離 [16] やロボット聴覚ソフトウェア HARK [17] の使用も考えられるが,

図 2: カエルホタルの写真

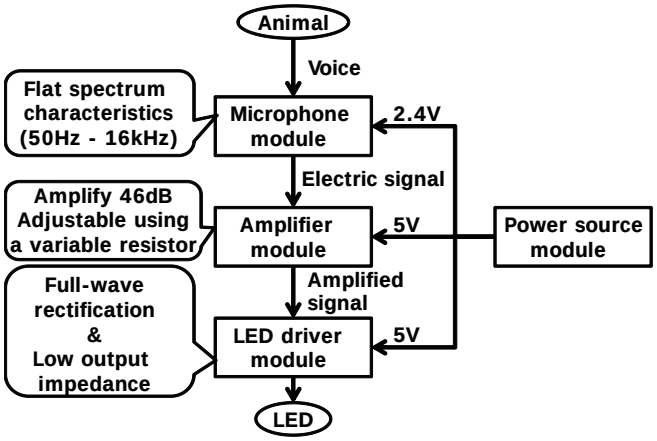


図 3: カエルホタルのブロック図

少なくとも筆者らの経験では合唱中の各音声の信号対雑音比が低すぎて十分な性能は得られなかった.

3 音声可視化システム

本節では, 音声可視化システムについて概説する. 詳細は [4,5] を参照されたい.

3.1 システム概要

本システムは市販のビデオカメラと多数のカエルホタルから構成される. ビデオカメラは, ソニー社のハンディカム (HDR-XR520; 29.97 fps) を用いた. 実験では HD モード (解像度 1480×1080) に設定し, 三脚で地面に固定して使用した.

カエルホタルは, マイクと調整用可変抵抗, LED, 単 4 充電池 2 本がプラスチックケースに収納されている. 図 2 左側にカエルホタルの内部回路を, 右側に外観を示す. 内部処理は 図 3 に示す通り, 音声収録, 増幅, 全波整流と LED 駆動の各回路で構成されている. 部品の誤差などによる個体差は, 可変抵抗の手動調整で合わせる. LED には赤色を使用する. これは, 筆者らの経験からニホンアマガエルは光に近寄る特性 (photopositive) を持つことが分かっており, photopositive な生物は赤色への感度が低いからである [18]. したがって, 赤色の使用で行動への影響の最小化が期待できる. また, ビデオカメラはカエルホタルの側面から撮影するため, 光の側面からの視認性を高め

る必要がある. そこで, LED にシリコン製キャップを装着して光を拡散している.

3.2 音声可視化システムによる計測

音声可視化システムによる合唱の解析は, (1) データ収録, (2) 時空間構造可視化, (3) 同期状態解析から構成される.

3.2.1 データ収録

まずカエルホタルをフィールドに配置し, 発光パターンをビデオカメラで収録する. カエルホタルは小雨程度なら動作するのでデータ収録は可能であるが, その場合はビデオカメラを防水ケースで覆うことで故障を防ぐ必要がある.

データ収録で重要な点は次の 3 点である. 2, 3 については対象種に対する知見に基づいているので, 他の種の合唱を計測する場合はそれに合わせる必要がある.

1. ビデオカメラはすべてのカエルホタルが見渡せる位置に三脚で固定する.
2. ニホンアマガエルは水田の縁 (あぜ道) で鳴くことが知られているので, カエルホタルはあぜ道に沿って等間隔に配置する.
3. ニホンアマガエルは日没後に鳴くので, 実験はすべて日没後に行う.

3.2.2 時空間情報可視化

時空間情報の可視化には動画像からの LED の位置検出と, 輝度時系列の抽出が必要となる. このとき問題となるのは, (1) LED の光が弱いので単独フレームのみでは検出が困難, (2) カエルホタルの輝度特性の個体差の存在の 2 点である. それぞれに対して, (1) 30 秒程度の平均フレームを用いてコントラストを向上することで検出を可能とし, (2) 各輝度時系列の平均値の正規化によって個体差を吸収することで解決する.

時空間情報可視化は 5 ステップから構成される.

(1) フレーム分割 まず, 動画をフレームごとに分割する. `ffmpeg` をはじめとして様々なツールがあるが, 我々は Pegasys 社の TMPGEnc を使用した.

(2) 平均フレーム計算 各フレームから NTSC 規格 [19] に基づいて輝度を計算し, グレイスケールに変換する.

$$\begin{aligned} I_g(x,y,t) &= 0.2989 I_R(x,y,t) + 0.5870 I_G(x,y,t) \\ &\quad + 0.1140 I_B(x,y,t) \end{aligned} \tag{1}$$

ただし, $I_g(x,y,t)$, $I_R(x,y,t)$, $I_G(x,y,t)$, $I_B(x,y,t) \in [0,255]$ はそれぞれ t フレーム目の 座標 (x,y) のピクセルの輝度, 赤成分, 緑成分, 青成分を表す.

こうして求めたグレイスケールのフレームを用いて平均フレーム $\bar{I}(x,y)$ を計算する.

$$\bar{I}(x,y) = \frac{1}{N} \sum_{t=1}^{N} I_g(x,y,t) \tag{2}$$

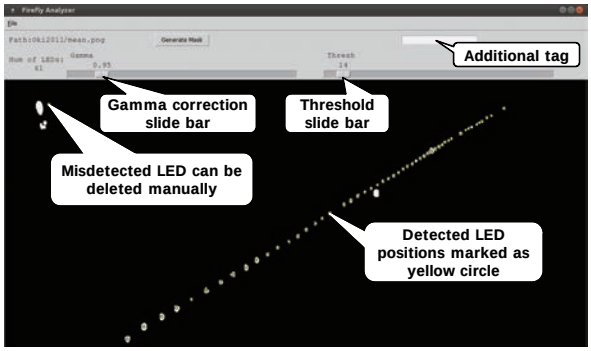


図 4: マスク修正 GUI ツールのスクリーンショット. マウス操作で, 誤検出されたマスクの削除や検出に漏れたマスクの追加が可能

ただし, N は 30 秒分のフレーム数をあらわす.

(3) LED 位置検出とマスク生成 LED 部分を抽出するため平均フレームを 2 値化する. フレームの輝度ヒストグラムは, ほぼすべての領域が 0 にすべき小さい値を持ち, 狭い領域のみが大きな値を持つという特性がある. そこで, 次のアルゴリズムで閾値を求める.

まず, 閾値 θ を輝度ヒストグラムの最大値に設定する. 次に, 頻度が単調減少する間 θ を増加させる. 単調減少が停止したときの値 θ を 2 値化の閾値とする. この方法で, 0 にすべき小さい値の山を 0 にし, 1 にすべき大きい値の山の手前に閾値を設定できる. 擬似コードは次のとおりである.

$\theta \leftarrow \underset{i}{\text{argmax}}\ h(i)$
while $h(\theta) > h(\theta+1)$ and $\theta < 255$ **do**
  $\theta \leftarrow \theta + 1$
end while

ただし, $i \in [0,255]$ を輝度, $h(i)$ を頻度とする.

こうして得られたバイナリ画像を 4 連結で結合し, 各連結成分にラベル (通し番号) を与える. そして誤検出されたラベルを GUI ツール (図 4) によって手作業で取り除くと, 各 LED に対応するマスクが得られる. LED 周辺に漏れた光も抽出するため, 各ラベルには膨張処理を施す. ビデオカメラは固定しているので, こうして得られたマスクは同一動画内では同じものが使える.

(4) 輝度時系列抽出 マスクのラベルを m で表すと, 第 m マスクのフレーム t での輝度 $l(m,t)$ は, 次式で求まる.

$$l(m,t) = \sum_{x=0}^{W} \sum_{y=0}^{H} \Big(I_{mask}(x,y,m) I_g(x,y,t) \Big) \tag{3}$$

ただし, W, H はフレームの幅と高さとし, $I_{mask}(x,y,m) \in [0,1]$ は座標 (x,y) が第 m マスク内のとき 1, マスク外のとき 0 の値を取るとする.

このままでは, カエルホタルごとの輝度の違いがあるために発声位置・時刻が特定できない. そこで, 各マスクご

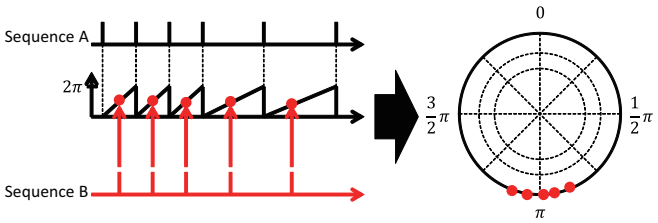


図 5: ストロボスコープ法による同期解析の概要

とに輝度の平均値を求め, それを減算することで個体差を正規化する. そして, 輝度の値域を $[0, 1]$ とするため, 輝度の総和で除算する.

(5) 発声位置・時刻検出　最後に, 発声時刻と位置を $l(m, t)$ のローカルピークから求める. 1 辺 L の正方形の窓を, l 上にスライドさせ, 窓の中央の座標 (m, t) の輝度 $l(m, t)$ が窓内で最大の場合, (m, t) をローカルピークとする. L は実験的に 3 とした.

3.2.3　同期状態解析

発声位置・時刻からカエルの同期状態を統計的に解析する. まず, 近接する 2 匹のカエルの発声系列の同期状態をストロボスコープ法 [20] で求める. 図 5 にストロボスコープ法の概要を示す. 2 つの発声系列を A, B とし, A の位相を, 発声時刻で 0, 次の発声時刻で 2π となるように線形な単調増加の組み合わせで定義する (図左側). 次に, B の各発声時刻における A の位相をプロットすることで, A と B の位相差を計算できる (図右側). 図 5 の場合, 位相差が π に偏っているので, A と B は逆相同期している.

次に, この位相差データから同期状態を統計的に検定する. 直感的には, 位相差が一様分布であれば発声系列は同期しておらず, いずれかの値に偏っていればその位相差で同期しているといえる. 方向統計で用いられる Rayleigh 検定 [21] は, 帰無仮説が「位相データは一様分布に従う」なので, これを用いれば位相差が統計的に有意に偏っているかを検定できる. もし有意に偏っているのであれば, そのカエルの組は同期しているといえる.

4　フィールド実験

合唱可視化システムを用いた野外でのニホンアマガエルの合唱の解析を行い, 本手法の有効性を示す.

4.1　実験条件

フィールド実験は 2011 年 6 月 9 日から 16 日にかけて島根県隠岐島の水田で行った. 実際に使用した水田の外形を図 6 に示す. なお, カエルホタルを配置したあぜ道は湿った土と高さ 10cm ほどの草に覆われていた. カエルホタルは, 図 6 中の赤丸の位置から, 同図の矢印で示すあぜ道に沿って 33.8m に渡って 40cm 間隔で配置した. ビデオカ

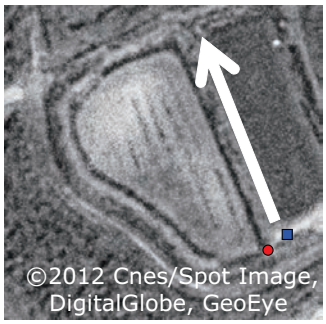


図 6: 実験を行った水田の概形

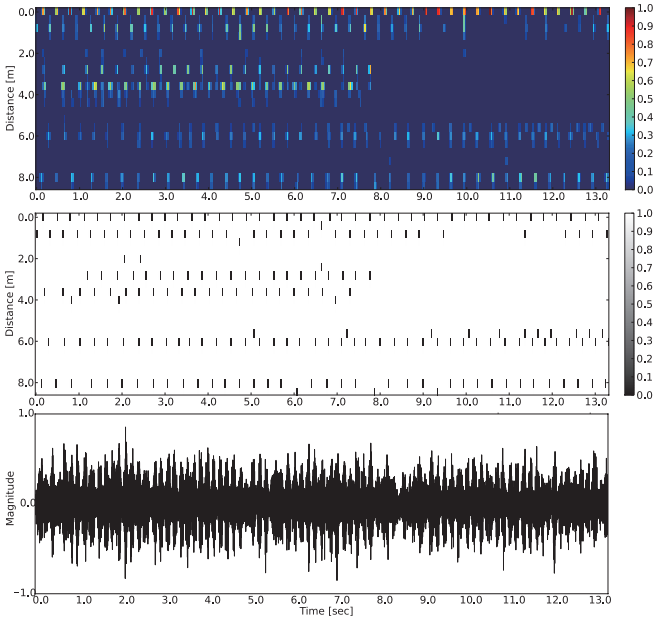


図 7: 6 匹の合唱の可視化　(2011 年 6 月 12 日)

メラは水田より 2m 程度高い図中の青い四角の位置に設置した.

4.2　実験結果

複数のカエルの合唱の可視化例を図 7 に示す. 図上段は輝度時系列, 図中段は輝度時系列のピーク, 図下段は同時に録音した音声波形である. 上段と中段の縦軸はカエルホタルの位置を, 下段は振幅をそれぞれ表す. 横軸はすべて時間を表す. まず, 下段をみると多数のピークが見られる. 各ピークがカエルの発声時刻に対応しているので時間情報はわかるが, 空間情報が欠落している.

図 7 中段には, 安定した発声が 0, 0.5, 2.5, 3.5, 6, 8[m] の位置に 6 系列見られる. 各系列をそれぞれ a, b, c, d, e, f と名付ける. ニホンアマガエルは鳴きながら動かないことから, 1 つの系列が 1 匹のカエルに対応する. したがって, 図には 6 匹のカエルがいたと推測できる. カエル間距離が近いので (a, b), (c, d), (e, f) がそれぞれペアを構成していると仮定すると, 各ペアが逆相同期していることが定性的に観察できる. このことは, 2 匹のカエルが逆相同

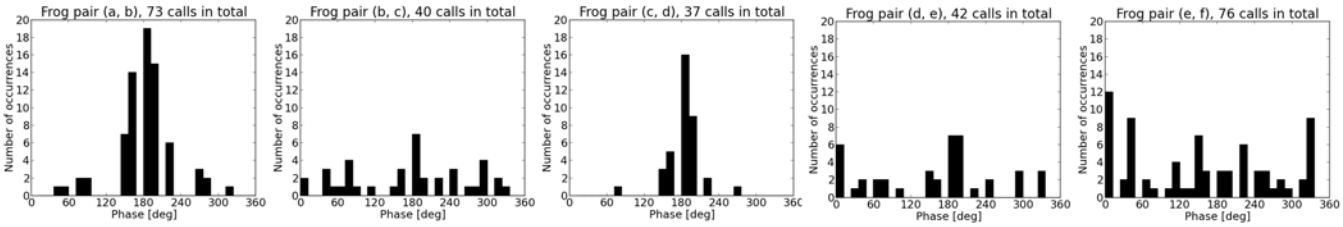


図 8: 隣り合うペアの位相差ヒストグラム

表 1: 同期解析結果　($p < 0.01$)

系列 A	系列 B	距離 [m]	平均位相 [deg]	p 値
a	b	0.8	179.4	< 0.01*
b	c	2.0	179.5	0.13
c	d	0.8	181.2	< 0.01*
d	e	2.4	185.1	0.32
e	f	2.0	4.5	0.34

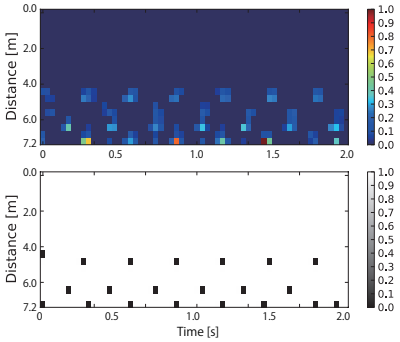


図 9: 1:2 逆相同期の観測例

期しやすいという室内実験の結果に対応している [1].

次に, 定量的に議論するために同期解析を行う. 図 8 に隣り合う全ペアの位相差ヒストグラムを, 表 1 にペアの距離と平均位相, p 値を示す. 統計的に有意 ($p < 0.01$) に位相差に偏りがある行には $*$ を付している. 表より, a から e まではどのペアも平均的に逆相同期しており, カエル間距離が 0.8m と短い 2 つのペア (a, b) と (c, d) が有意に逆相同期している. したがって, 合唱中でも隣り合う個体同士は逆相同期する傾向があり, それは個体間距離が近いほど安定すると考えられる.

他の観測例について議論する. まず, 図 9 に 1:2 逆相同期の観測例を示す. これは屋内で観測した 1:2 逆相同期が野外でも再現していることを表している. 次に図 10 に同相同期の観測例を示す. ニホンアマガエルは逆相同期が安定状態にもかかわらず同相同期が生じたのは, 合唱の開始時点では同期が不安定であることが理由だと考えられる. 実際, 図下側のカエルは 1.5 秒付近から鳴き始め, 同相同期を経て 4.0 秒付近では逆相同期になっている. この観測は, 逆相同期には数回の鳴き交わしが必要であることを示唆している.

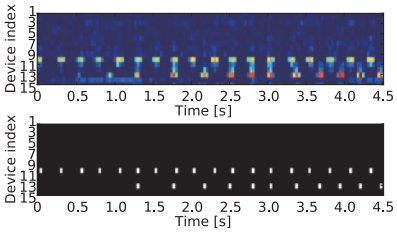


図 10: 同相同期の観測例

4.3　考察

まず, 合唱の解析結果について議論する.

実験結果は　野外における大規模な合唱においてもニホンアマガエルの数理モデルによる予測や屋内実験の結果 [22] と同様に, カエルたちは隣り合う個体と相互作用し逆相同期していることを示唆している.

繁殖行動の観点ではこの同期現象は理にかなっている. なぜなら, 隣り合うオス同士の発声タイミングをずらせば発声の時間的な重なりが減るので, メスに自分の位置を知らせやすいからである. 一方で, 離れたオス同士の安定した同期状態は見られない. この現象は, ツンガラガエルで報告されている, ある音量より大きな音にしか反応して鳴かない *selective attention* [23] に対応していると考えられる. なぜなら, 個体同士の距離が離れるほど伝わる音量は下がるからである.

次に, 本システムの利点と限界について議論する.

本システムの利点のひとつは, マイクアレイ処理では必須のセンサ間同期が不要となることである. この理由は, 音声の存在を LED の光で伝えることにある. 光の速度は音声に比べて非常に速いので, 水田の幅 20-30m 程度であれば, 光の到達時間差はほぼ無視してよい. したがって, 離れた場所の LED であっても, 同一フレーム内で光っていれば実際に同時に光っていたと言える. この利点のおかげで, 各カエルホタルは相互に接続することなく自由な位置に配置できるので, データ収録はマイクアレイの収録に比べて容易である.

本手法の限界は, (1) 使用できるフィールドに条件がある点と, (2) 生物の空間配置に条件がある点である. (1) については, まずに, フィールドは LED が全て見える平坦な場所にビデオカメラを設置できる必要がある. ただし, 複数カメラを同期収録できるなら, それらを組み合わせるこ

とで凹凸のあるフィールドでの収録も可能である. 次に, フィールドが背の高い草地で覆われている場合は LED が隠れるので, 台に乗せるなど高さを上げる工夫が必要となる. 天候に関しても, 強い雨の場合は回路が破損する恐れがあるので使用できず, 強風の場合はマイクへの風音が原因で鳴き声を検出できないという制約がある. (2) については, まず, 対象生物の空間配置はおおよそ既知である必要がある. なぜなら, 安定した定位のためにはカエルホタルは生物の空間配置の 5-6 倍の密度で設置する必要があるからである (詳細な解析は [5]). ニホンアマガエルの場合は, およそ 1-2m の間隔で交互に鳴く, すなわち実質的な個体間距離が 2-4m なので, 40cm 程度の距離で設置した. 次に, カエルホタルは地面に配置するので平面上に分布する種でなければ計測できない. したがって, 鳥やコウモリの鳴き声の時空間構造の可視化は困難である.

5 まとめ

本稿では, 音声可視化デバイス「カエルホタル」を用いた合唱可視化システムと, それを用いたニホンアマガエルの合唱のフィールド実験について述べた. 実験の結果, 6 匹の合唱で隣り合う個体たちと逆相同期していることが明らかになった. 多個体の合唱の数理解析は [24] を参照されたい. 今後は, より大規模なカエル合唱の解析や異なる種のカエル同士のコミュニケーションの解析, またカエル以外の生物の合唱の解析を行う予定である.

本研究は JSPS 基盤研究 S(No. 24220006), 特別研究員奨励費 (No. 08J00608), 理化学研究所・基礎科学特別研究員制度の支援をうけた.

参考文献

[1] I. Aihara, R. Takeda, T. Mizumoto, T. Otsuka, T. Takahashi, H. G. Okuno, and K. Aihara. Complex and transitive synchronization in a frustrated system of calling frogs. *Phys. Rev. E*, 83(3):031913 (5 pages), 2011.

[2] K.D. Wells. *The ecology and behavoir of amphibians*. The University of Chicago Press, Chicago, 2007.

[3] M. Matsui. *Natural history of the amphibia*, pages 150–152. University of Tokyo Press, 1996.

[4] T. Mizumoto, I. Aihara, T. Otsuka, R. Takeda, K. Aihara, and H. G. Okuno. Sound imaging of nocturnal animal calls in their natural habitat. *J Comp Physiol A*, 197(9):915–921, 2011.

[5] T. Mizumoto. *Temporal Synchronization among Interacting Individuals in Human-Robot Ensembles and Frog Choruses*. PhD thesis, Kyoto University, 2013.

[6] A.S. Feng, P. M. Narins, C-H Xu, W-Y Lin, Z-L Yu, Q. Qiu, Z-M, and J-X Shen. Ultrasonic communication in frogs. *Nature*, 440:2333–2336, 2006.

[7] P. M. Narins, A. S. Feng, R. R. Fay, and A. N. Popper, editors. *Hearing and Sound Communication in Amphibians*. Springer, 2007.

[8] B. Hedwig and JFA Poulet. Complex auditory behavior emerges from simple reactive steering. *Nature*, 430(7001):781–785, 2004.

[9] J. A. Simmons, M. B. Fenton, and M. J. O'Farrell. Echolocation and persuit of prey by bats. *Science*, 203(4375):16–21, 1979.

[10] H. Riquimaroux, S. J. Gaioni, and N. Suga. Cortical computational maps control the auditory perception. *Science*, 251(4993):565–568, 1991.

[11] J. Okuyama, K. Kataoka, K. M Kobayashi, O. Abe, K. Yoseda, and N. Arai. The regularity of dive performance in sea turtles: a new perspective from precise activity data. *Animal Behaviour*, 84(2):349–359, 2012.

[12] P. Tyack. An optical telemetry device to identify which dolphin produces a sound. , 78(5):1892–1895, 1985.

[13] A. M. Simmons, J. A. Simmons, and M. E. Bates. Analyzing acoustic interactions in natural bullfrog (*Rana catesbeiana*) choruses. *J Comp Physiol*, 122(3):274–282, 2008.

[14] D. L. Jones and R. Ratnam. Blind location and separation of callers in a natural chorus using a microphone array. , 126(2):895–910, 2009.

[15] D. T. Blumstein, D. J. Mennill, P. Clemins, L. Girod, K. Yao, G. Patricelli, J. L. Deppe, A. H. Krakauer, C. Clark, K. A. Cortopassi, S. F. Hanser, B. McCowan, A. M. Ali, and A. N. G. Kirschel. Acoustic monitoring in terrestrial environments using microphone arrays: applications, technological considerations and prospectus. *Journal of Applied Ecology*, 48(3):758–767, 2011.

[16] A. Hyvarinen, J. Karhunen, and E. Oja. *Independent component analysis*. Wiley-Interscience, New York, 2001.

[17] K. Nakadai, H. G. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino. Design and implementation of robot audition system "HARK". *Advanced Robotics*, 24:739–761, 2009. doi:10.1163/016918610X493561.

[18] J. P. Hailman and R. G. Jaeger. Phototactic responses to spectrally dominant stimuli and use of color vision by adult anuran amphibians: a comparative survey. *Anim Behav*, 22:757–795, 1974.

[19] ITU-R. *Recommendation ITU-R BT.606-6: Studio encoding parameters of digital television for standard 4:3 and wide screen 16:9 aspect ratios*. International Telecommunication Union Radiocommunication Sector, 2007.

[20] A. Pikovsky, M. Rosenblum, and J. Kurths. *synchronization: a universal concept in nonlinear sciences*. Cambridge University Press, 2001.

[21] K.V. Mardia and P. E. Jupp. *Directional Statistics*. Wiley, 2000.

[22] I. Aihara. Modeling synchronized calling behavior of Japanese tree frogs. *Phys. Rev. E*, 8:011918, 2009.

[23] M. D. Greenfield and A. S. Rand. Frogs have rules: Selective attention algorithms regulate chorusing in *Physalaemus* pustulosus (loptodactylidae). *Ethology*, 106:331–347, 2000.

[24] I. Aihara, T. Mizumoto, T. Otsuka, H. Awano, K. Nagira, H. G. Okuno, and K Aihara. Spatio-temporal dynamics in collective frog choruses examined by mathematical modeling and field observations. *Scientific Reports*, 4(3891), 2014. doi:10.1038/srep03891.

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B303-09

振動子モデルと音声可視化システムを用いたアマガエルの合唱法則の解析

Theoretical and Experimental Studies on Frog Choruses

Based on the Phase Oscillator Model and Sound-Imaging Method

合原　一究[1]、粟野　皓光[2]、 水本　武志[3]、坂東　宜昭[2]、
大塚　琢馬[2]、柳楽　浩平[2]、奥乃　博[2]

Ikkyu AIHARA, Hiromitsu AWANO, Takeshi MIZUMOTO, Yoshiaki BANDO,
Takuma Otsuka, Kohei NAGIRA, Hiroshi G. OKUNO

1 理化学研究所、2 京都大学、3 （株）ホンダ・リサーチ・インスティチュート　ジャパン

1 RIKEN, 2 Kyoto University, 3 Honda Research Institute Japan Co., Ltd.

ikkyu@brain.riken.jp

Abstract

ニホンアマガエルは鹿児島県から北海道までの日本の広範囲に生息しており、春から夏にかけてオスのニホンアマガエルが集団で鳴く様子を観察できる。オスのニホンアマガエルは単独では周期的に鳴く一方で、鼓膜を持っており周囲の音を認識できる。そのため、野外での集団発声行動は、単独では周期的に振る舞う素子が互いに影響を及ぼし合う結合振動子系として数理的に理解できるだろう。我々はニホンアマガエルの集団発声行動を記述する振動子モデルを考案し、数値計算を行うことで、全体が２つの集団に分かれて鳴き交わす同期現象が観測される可能性を示した。一方で、周囲の音声を検知して LED が発光する音声可視化装置「カエルホタル」を用いた野外調査を行った。その結果、近くの個体が交互に鳴き、全体では２つの集団に分かれる同期状態を観測した。

1　はじめに

たくさんのカエルによる集団発声行動、すなわちカエルの合唱は、水田や川、池など、様々な場所で観察できる。カエルの場合、一般に鳴くのはオスのみであり、メスは鳴かない。野外で最も頻繁に聞かれる鳴き声は広告音と呼ばれる種類であり、「メスを呼び寄せる役割」と「他のオスに縄張りを主張する役割」があると考えられている, [松井 1996, Gerhardt and Huber, 2002, Wells, 2007]。このような広告音を用いたカエルの合唱は、主に生物学者に興味を持たれ、様々な種について研究されてきた。例えば、２種類の広告音を使い分ける Coqui ガエル [Narins and Capranica, 1976]、20kHz 以上の超音波の成分を含む広告音を使うカエル [Feng et al., 2006] などが知られている。このようにカエルの合唱は世界各地で観測できる。しかし、野外での個体毎の発声タイミングと空間配置は詳しく調べられてこなかった。これは、野外では鳴いている個体数が多く、さらにその空間的な分布も広範囲に渡るためである。

ニホンアマガエル（以下、アマガエル）は日本で最もよく見られる種のカエルであり、南は鹿児島県から北は北海道まで生息している [前田、松井、1989]。野外での合唱を観察できる期間も春から夏にかけてと長く、主に水田で集団で鳴いている様子を観察できる。オスのアマガエルは単独では周期的に鳴く一方で、鼓膜を持っており周囲の音を認識できる。そのため、野外での集団発声行動は、単独では周期的に振る舞う素子（振動子）が互いに影響を及ぼし合う結合振動子系として数理的に理解できるだろう。このような結合振動子系の理論 [Kuramoto, 1984] は多くの研究者に興味を持たれており、共通ノイズ下での振動子集団の振る舞い [Teramae et al., 2009] や、空間的に運動する振動子集団の振る舞い [Tanaka, 2007, Fujiwara et al., 2011]、結合ネットワークを動的に変化させる振動子モデル [Aoki and Aoyagi, 2009] などが研究されてきた。一方で、結合振動子系の理論は、ホタルの集団発光やコオロギの鳴き交わしなど動物行動との関連も示唆されている。しかし、特に野外における集団行動を、結合振動子系の理論を応用して解析する数理研究はこれまでほとんど行われてこなかった。アマガエルの集団発声行動の詳細を音声可視化システムによって明らかにし、そのコミュニケーション機構を結合振動子系の考え方をもとに解析するのが本研究の目的である。これによりアマガエルの未知の行動を観測し、さらにはそのメカニズムの数理的な理解が進むことが期待できる。

本稿では、アマガエルの集団発声行動に関する数理研究と音声可視化装置「カエルホタル」を用いた野外調査結果を概説する [Aihara et al., 2014, Mizumoto et al., 2011]。

Figure 1: ニホンアマガエル。鹿児島県から北海道までの日本の広範囲に生息する。春から夏にかけて、主に水田で多数のオスのニホンアマガエルが鳴いている様子を観察できる。

2　振動子モデルを拡張した数理研究

本章では、アマガエルの合唱に関する数理研究 [Aihara et al., 2014] を概説する。まずアマガエル２匹の発声行動を定性的に説明する振動子モデルを紹介し (第 2.1 節)、その上で、２匹では交互に鳴く行動学的な理由を考察する (第 2.2 節)。さらに、第 2.2 節の考察に基づいたアマガエルの合唱の振動子モデルを紹介し、アマガエルの合唱において２種の同期現象が観測される可能性を示す (第 2.3 節)。

2.1　２匹の発声行動の数理モデル

室内で２匹のアマガエルを鳴かせる行動実験および独立成分分析法による音声分離解析から、アマガエルは２匹では交互に鳴くことがわかった [Aihara et al., 2011]。このような交互に振動する現象を、同じ位相で振動する同相同期現象と比較して、逆の位相で振動するという意味で逆相同期現象と呼ぶこととする。本節では、このアマガエル２匹の逆相同期現象を振動子モデルにおける安定平衡点として説明した研究成果を紹介する [Aihara et al., 2008, Aihara and Tsumoto, 2008, 合原、2013]。

まず、アマガエル２匹の発声タイミングを位相 θ_n ($n = 1, 2$) を用いてモデル化する。位相 θ_n は 0 から 2π までの値をとる変数であり、$\theta_n = 2\pi$ で個々のカエルが鳴くと仮定する。この θ_n を用いて、アマガエル２匹の発声行動を次式で与えられる振動子モデル [Kuramoto, 1984] で記述する：

$$\frac{d\theta_n}{dt} = \omega_n + \sum_{m=1,m\neq n}^{N} \Gamma_{nm}(\theta_m - \theta_n) \qquad (1)$$

ここで ω_n ($n = 1, 2$) は個々のカエル単独での固有周期を表すパラメータで、正の定数とする。アマガエルは単独では 1 秒間に 4 回程度鳴くことから [Aihara, 2009]、$\omega_n = 8\pi$ と仮定する。また、$\Gamma_{nm}(\theta_m - \theta_n)$ はカエル同士の相互作用を表す関数で、2π の周期関数で定義される [Kuramoto, 1984]。

この関数 $\Gamma_{nm}(\theta_m - \theta_n)$ を一次の sin 関数と仮定することで、アマガエル２匹の逆相同期現象を説明できる [Aihara et al., 2008, Aihara and Tsumoto, 2008, 合原、2013]。まず２匹の鳴くタイミングの差を表す変数として、位相差 $\phi \equiv \theta_1 - \theta_2$ を定義する。さらに、相互作用関数を一次の sin 関数を用いて、$\Gamma_{12}(-\phi) = -\Gamma_{21}(\phi) = K\sin\phi$ と定義する。ここで K はカエル同士の相互作用の強さを表すパラメータで、正の定数とする。このとき、位相差 ϕ のダイナミクスは、式 (1)、$\omega_n = 8\pi$、$\Gamma_{12}(-\phi) = -\Gamma_{21}(\phi) = K\sin\phi$ から、次式で与えられる：

$$\frac{d\phi}{dt} = 2K\sin\phi. \qquad (2)$$

この式において、$\phi = \pi$ は $\frac{d\phi}{dt} = 0$ を満たすので、$\phi = \pi$ は平衡点である。さらに、$\phi = \pi$ での勾配 $\frac{\partial}{\partial\phi}\frac{d\phi}{dt}$ は負なので、$\phi = \pi$ は安定な平衡点となる。このモデルでは $\theta_n = 2\pi$ となる度に個々のカエルが鳴くと仮定しているので、$\phi = \pi$ の逆相同期状態はアマガエル２匹が交互に鳴く状況を定性的に説明することになる。

2.2　２匹の逆相同期現象に関する考察

第 2.1 節で述べたように、アマガエルは２匹では交互に逆位相で同期して鳴く傾向がある。では、なぜ２匹では交互に鳴くのだろうか？ ここで扱っているアマガエルの鳴き声は広告音と呼ばれ、「メスを呼び寄せる役割」と「他のオスに縄張りを主張する役割」があると考えられている [松井 1996, Gerhardt and Huber, 2002, Wells, 2007]。そのため、オスのアマガエルによる逆相同期現象は、それぞれの縄張りを主張する役割に関係するものと考えられる。ここで、２匹のアマガエルが同時に同じ位相で同期して鳴く場合を考えよう。そのような場合、自分自身の鳴き声も大きいために、相手の鳴き声が聞き取りずらくなるであろう。これに対して、２匹が交互に逆位相で同期して鳴く場合は、それぞれの鳴き声のオーバーラップはほとんどなく、そのために相手の鳴き声も聞き取りやすくなるであろう。このように、アマガエル２匹の逆相同期現象は、お互いの鳴き声を聞き取りやすくし、それによって自身の縄張りを強く主張するのに有効であると予想している [Aihara, 2009, Aihara et al., 2011, 合原、2013]。

2.3　合唱の数理モデル

第 2.1 節でモデル化したアマガエル２匹の発声行動に関しては、それぞれのアマガエルをケースに入れていたので、その位置関係は実験中にはほとんど変化しなかった [Aihara et al., 2011]。しかし、野外ではアマガエルは自由

に移動できる。そのため、合唱のモデル化には、アマガエルの空間配置のダイナミクスも考慮する必要があると考えた。そこで、個々のアマガエルの発声タイミングを位相 $\theta_n (n = 1, 2, ..., N)$ で、そして個々のアマガエルの空間配置を２次元のベクトル $\boldsymbol{r}_n (n = 1, 2, ..., N)$ で記述し、それぞれの時間発展を次式でモデル化した [Aihara et al., 2014]:

$$\frac{d\theta_n}{dt} = \omega_n + \sum_{m=1,m\neq n}^{N} \Gamma_{nm}(\theta_m - \theta_n, \boldsymbol{r}_m - \boldsymbol{r}_n), \quad (3)$$

$$\frac{d\boldsymbol{r}_n}{dt} = \sum_{m=1,m\neq n}^{N} \boldsymbol{F}_{nm}(\theta_m - \theta_n, \boldsymbol{r}_m - \boldsymbol{r}_n) + \boldsymbol{G}_n(\boldsymbol{r}_n). \quad (4)$$

ここで ω_n は、第 2.1 節で紹介したモデルと同様、n 番目のアマガエルの固有周期を表す正のパラメータであり、$\omega_n = 8\pi$ に固定した。

式 (3) および (4) の $\Gamma_{nm}(\theta_m - \theta_n, \boldsymbol{r}_m - \boldsymbol{r}_n)$ と $\boldsymbol{F}_{nm}(\theta_m - \theta_n, \boldsymbol{r}_m - \boldsymbol{r}_n)$ は、n 番目と m 番目のアマガエルの相互作用を表す関数である。まず、$\Gamma_{nm}(\theta_m - \theta_n, \boldsymbol{r}_m - \boldsymbol{r}_n)$ に関しては、アマガエルは２匹では交互に逆位相で同期して鳴くこと、そして音はアマガエル間の距離 $|\boldsymbol{r}_m - \boldsymbol{r}_n|$ の二乗で減衰することから、次式でモデル化した：

$$\Gamma_{nm}(\theta_m - \theta_n, \boldsymbol{r}_m - \boldsymbol{r}_n) = -\frac{K_{nm}}{|\boldsymbol{r}_m - \boldsymbol{r}_n|^2} \sin(\theta_m - \theta_n). \quad (5)$$

ここで一次の sin 関数を使ったのは、第 2.1 節で述べたように、２匹が交互に鳴く現象を安定な逆相同期状態として説明できるからである。次に、$\boldsymbol{F}_{nm}(\theta_m - \theta_n, \boldsymbol{r}_m - \boldsymbol{r}_n)$ は次式でモデル化した：

$$\boldsymbol{F}_{nm}(\theta_m - \theta_n, \boldsymbol{r}_m - \boldsymbol{r}_n) = \frac{K_{nm}}{|\boldsymbol{r}_m - \boldsymbol{r}_n|^2}(1 - \cos(\theta_m - \theta_n))\boldsymbol{e}_{nm}. \quad (6)$$

この関数に関しても、アマガエルの音声コミュニケーションを表しているので、アマガエル間の距離 $|\boldsymbol{r}_m - \boldsymbol{r}_n|$ の二乗で相互作用が弱まるとした。さらに、cos 関数を使うことで、$\theta_m - \theta_n = \pi$ の逆相同期状態だと、アマガエル同士が最も強く反発するとした。これは、第 2.2 節にある「２匹のアマガエルは交互に鳴くことで、それぞれの縄張りを強く主張している」という考察のモデル化に対応する。$\boldsymbol{e}_{nm}$ は n 番目と m 番目のアマガエル間の単位ベクトルで、$\boldsymbol{e}_{nm} = -\frac{\boldsymbol{r}_m - \boldsymbol{r}_n}{|\boldsymbol{r}_m - \boldsymbol{r}_n|}$ と定義する。

一方、我々は野外調査の際に、稲が成長しきっていない状態だと水田内部には物理的に捕まるものがなく、アマガエルの空間配置は水田の周囲のあぜ道に集中することを観測した [Mizumoto et al., 2011]。式 (4) の $\boldsymbol{G}_n(\boldsymbol{r}_n)$ は、このような水田のあぜ道に沿った空間分布を説明するための関数であり、次式でモデル化した：

$$\boldsymbol{G}_n(\boldsymbol{r}_n) = (L - |\boldsymbol{r}_n|)\boldsymbol{e}_n. \quad (7)$$

ここでは、まず簡単のために水田の形状を円形と仮定し、その半径を正のパラメータ L で記述した。$\boldsymbol{e}_n$ は水田の中心を表す原点 $\boldsymbol{0}$ と n 番目のアマガエル間の単位ベクトルであり、$\boldsymbol{e}_n = \frac{\boldsymbol{r}_n}{|\boldsymbol{r}_n|}$ と定義する。この関数において、$L - |\boldsymbol{r}_n|$ の項は $|\boldsymbol{r}_n| = L$ を境に符号が変化する。すなわち、アマガエルが水田の内側に居る場合には中心から反発してあぜ道に寄って行き、アマガエルが水田の外側に居る場合には中心に引きつけられることで同様にあぜ道に寄って行くことになる。

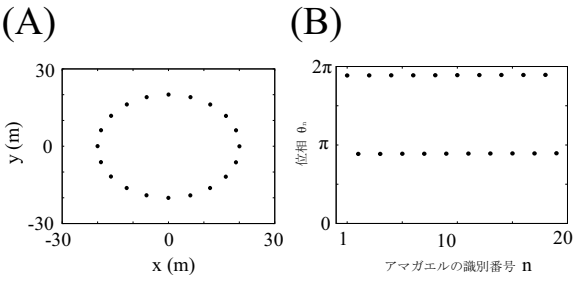


Figure 2: 数理モデルにおける２クラスタ逆相同期状態。左図が数値計算によって得られたアマガエルの空間配置を、右図がアマガエルの位相を表す。空間配置に関しては、水田の淵に沿って等間隔に並んでいる様子がわかる。同期状態に関しては、最近接の個体同士が逆位相で同期し、そのために全体では２つの集団が交互に鳴く状態が本数理モデルにおいて安定に測定された。

次に、本数理モデルを用いた数値計算を行い、実際に起こりえる同期状態を予測した。調査地である島根県・隠岐の島の水田は周囲が 100m 以上で、そのあぜ道に沿って 10 匹から 20 匹程度のアマガエルが鳴いていた [Mizumoto et al., 2011]。そこで、式 (5)–(7) のパラメータを $L = 20$、$N = 20$ に固定した。一方で、相互作用の大きさを表すパラメータ K_{nm} は、既知の観測データからの推定が困難であったので、簡単のために $K_{nm} = 1$ と固定した。

以上のパラメータを用いて数値計算を行った結果、２種の同期状態が本数理モデルにおいて安定に計測された。まず１つ目の同期状態を Figure 2 に示す。アマガエルは円形の水田のあぜ道に沿って分布している様子がわかる。同期状態に関しては、最近接の個体同士が逆位相で同期して、結果的に２つの集団にわかれている。このような同期状態を、２クラスタ逆相同期状態と呼ぶこととする。次に、２つ目の同期状態を Figure 3 に示す。空間配置に関しては、２クラスタ逆相同期状態と同様に、水田のあぜ道に沿って全ての個体が並んでいる。同期状態に関しては、最近接の個体同士がほぼ逆位相で同期しているものの、同一集団内で少しずつ位相がずれている様子がわかる。このような同期状態を、位相波状態と呼ぶこととする。

これらの数値計算により、本数理モデルにおいては２クラスタ逆相同期状態と位相波状態の２種の同期状態が

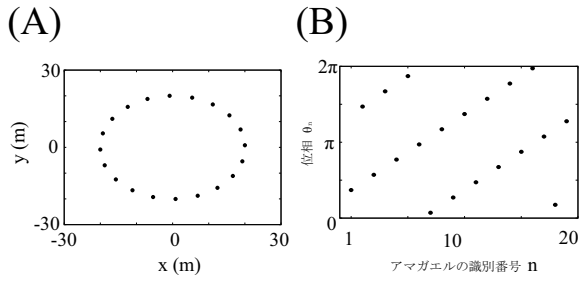


Figure 3: 数理モデルにおける位相波状態。左図が数値計算によって得られたアマガエルの空間配置を、右図がアマガエルの位相を表す。空間配置に関しては、Figure 2 と同様に、水田の淵に沿って等間隔に並んでいる様子がわかる。同期状態に関しては、最近接の個体同士がほぼ逆位相で同期し、それぞれの集団内で位相波状態が生じている。

初期条件に応じて安定に計測されることがわかった。しかし、実際の水田ではどちらがより安定に観測されるのだろうか？ 次に、水田の形状を円形から実際の水田の形状に近い長方形に変えた上で、同様の数値計算を行った。具体的には、長方形の水田の形状を長辺の長さ L_x と短辺の長さ L_y で記述し、$L_x + L_y = 60\mathrm{m}$ という拘束条件のもとでそれぞれの値を変化させた。さらに、個々の (L_x, L_y) に対して、初期条件をランダムに変化させる数値計算を繰り返し行った結果、より多くの形状で２クラスタ逆相同期状態を頻繁に検出した [Aihara et al., 2014]。

3 音声可視化装置「カエルホタル」を用いた野外調査

本章では、音声可視化装置「カエルホタル」を用いたアマガエルの合唱の野外調査を概説する [Mizumoto et al., 2011, Aihara et al., 2014]。まず、周囲の音を検知して LED が発光する音声可視化装置「カエルホタル」を紹介し（第 3.1 節）、島根県• 隠岐の島で行った野外調査の方法を説明する（第 3.2 節）。その上で、動画解析によって明らかになったアマガエルの合唱における同期現象を説明する（第 3.3 節）。

3.1 音声可視化装置「カエルホタル」

カエルは一般に夜行性であり、野外では数多くの個体が広範囲で鳴き交わす。そのため、野外での合唱における個体識別はこれまで困難であった。我々は、そのように複雑な音環境において、個々のカエルの発声タイミングおよび空間配置を計測するために、音声可視化装置「カエルホタル」（以下、カエルホタル）を開発した [Mizumoto et al., 2011]。カエルホタルはマイクロフォン、LED などから構成される電子回路であり、周囲で音が鳴るとその音量に応じた輝度で LED が明滅する。電源は再充電可能な単四電池

Figure 4: 音声可視化装置「カエルホタル」を用いた野外調査（京都大学構内の水田にて撮影）。カエルホタル数十台をあぜ道に沿って並べ、その様子をビデオカメラで撮影した。

２本であり、一度に多くの装置を野外調査に用いることができる。また、カエルホタルには可変抵抗が付いており、入力音声を光に変換する際の増幅率を、この可変抵抗をマイナスドライバーで回すことにより調整している [Mizumoto et al., 2011]。現時点ではこの可変抵抗の調整は手作業で行っているため、カエルホタル毎の音への応答には多少違いがある。

一方、稲が成長しきっていない状態だと、多くのニホンアマガエルは主に水田のあぜ道に並んで鳴いている。そこで、水田のあぜ道に沿ってカエルホタルを数十台並べて、その明滅をビデオカメラで撮影する実験手法を考案した [Mizumoto et al., 2011]。これによって並べたカエルホタルの内、どれがいつ光っているかを実験後の動画解析によって推定することで、合唱中の個体識別が可能になると考えた。

3.2 野外調査方法

カエルホタルを用いた野外調査を、2011 年の６月に島根県• 隠岐の島の水田で行った [Aihara et al., 2014]。調査地には複数の水田が存在したが、その中でより多くのアマガエルが鳴いている水田を選び、かつより多くのアマガエルが鳴いている一辺に沿ってカエルホタルを 85 台もしくは 86 台並べた。その上で、三脚の上部に固定したビデオカメラ（HDR- XR550V, SONY）によって、毎秒 29.97 フレームの時間解像度でカエルホタルの明滅を撮影した。この際、並べたカエルホタルを全て撮影できるよう、ビデオカメラを設置する位置および高さを調整した。撮影は１回１時間以上で、計６日間行った。

3.3 同期状態の解析

撮影した動画は研究室に持ち帰り、カエルホタルの輝度の時間変化を調べる解析を行った [Aihara et al., 2014]。ま

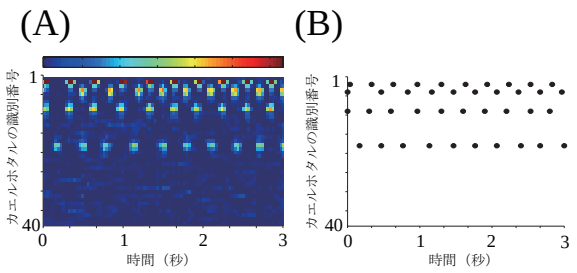


Figure 5: 音声可視化装置「カエルホタル」を用いた野外調査結果。左図がカエルホタルの輝度の時系列データを、右図が推定した発声タイミングおよび空間配置を表す。近接するアマガエル同士が交互に鳴いている様子がわかる。

ず、撮影した動画の内、最初の 15 分を画像に分割した。次に、分割した画像を足し合わせて、輝度の合計値の大きな箇所をカエルホタルの位置として推定した。その結果、少なくともビデオカメラに近い 40 台分のカエルホタルの明滅は安定して撮影できていたことがわかった。この 40 台よりも遠くにあるカエルホタルでは、LED の明かりが弱すぎてビデオカメラで安定して撮影できなかったものが存在した。そのため全ての動画に関して、ビデオカメラに近い 40 台分のカエルホタルを選んで輝度の時系列データを計算した。また、カエルホタルの可変抵抗の調整は手作業のため、無音状態での LED の輝度には若干の差が生じる。この差を補正するため、輝度の時系列データを計算するときに、15 秒毎に個々のカエルホタルの輝度の平均値を引いた。

Figure 5(A) がカエルホタル 40 台の輝度の時系列データを表す。複数台のカエルホタルが特に強い光を周期的に発している様子がわかる。次に、カエルホタル毎に輝度の合計値を計算し、輝度のピークに対応するカエルホタルを推定した [Aihara et al., 2014]。カエルホタルが発する光の強さは入力された音量に依存する。そのため、特に強く光っているカエルホタルの近くでアマガエルが鳴いているものと考えられる。最後に、鳴いているアマガエルに近いと推定されたカエルホタルそれぞれに対して、強く光ったタイミングを推定した。具体的には、最大輝度の 50% という閾値をそれぞれに設定し、この閾値を超える度に個々のアマガエルが鳴いたと判断した。

Figure 5(B) が推定した発声タイミングを表す。近接するアマガエル同士が交互に鳴いている様子がわかる。同様の方法で計 6 日間分の動画データを解析し、さらには種々のオーダーパラメータも計算することで、Figure 5(B) のような全体が 2 つの集団にわかれて交互に鳴く状態が繰り返し起こっていることを示した [Aihara et al., 2014]。

4　今後の課題

本章では以上の成果を踏まえて、今後の研究課題を議論する。

4.1　音声可視化システムの拡張

我々は音声可視化装置「カエルホタル」を用いて、アマガエルの発声タイミングおよび空間配置を計測した。このような音源定位の手法としては、複数のマイクロフォンを用いた録音実験も考えられる。カエルの発声行動に関しても、マイクロフォンアレイを用いた野外調査結果が最近報告された [Jones et al., 2014]。カエルホタルは音源位置や音が鳴ったタイミングは推定できるものの、音の高さは現時点では計測できない。カエルの鳴き声の高さを調べたい場合は、マイクロフォンアレイを併用した実験が必要になるであろう。カエルホタルで推定した位置情報に基づいてマイクロフォンアレイを用いたビームフォーミングを行う実験手法を確立することで、より詳細にカエルの発声行動を計測できる可能性がある。

4.2　カエルの行動研究の展望

第 3 章では、オスのニホンアマガエルの広告音に関する野外調査結果を紹介した。第 2.2 節で述べたように、広告音には「メスを呼び寄せる役割」と「他のオスに縄張りを主張する役割」があると考えられている。本研究では、オス同士のコミュニケーションを調べたが、それ以外の役割についてはまだわかっていない。例えば、野外で鳴いているオスの内、どの個体がメスに選ばれやすいのだろうか。また、カエルは鳴くことでタヌキやヘビなどの捕食者に自身の位置を知られてしまう可能性がある。このように、メスのカエルそして天敵が合唱中のオスガエルをどのように識別し特定の個体を選んでいるのか調べるのは、今後の研究課題である。

4.3　カエルホタルの改良

カエルに関しては、複数種が同一環境に生息する例が数多く報告されている。例えば、ニホンアマガエルの場合でも、ツチガエルやシュレーゲルアオガエルが同時期に同じ水田で観察できる。このような環境においては、それぞれの種が別種の鳴き声を認識し、種間でコミュニケーションをとっている可能性がある。

カエルは一般に、種が異なれば鳴き声の高さも異なる [前田、松井、1989]。カエルホタルにバンドパスフィルター機能を搭載し、特定の周波数を持つ音が入力された場合のみ LED が発光するように改良することで、野外でカエルの種を判別できる可能性がある [Mizumoto et al., 2012]。このような改良を行い、カエルの種間コミュニケーションを調べるのは今後の研究課題である。

4.4 数理モデルの改良

第 2 章では、アマガエルの合唱を結合振動子系と考えてモデル化することで、2 種の同期状態が起こりうる可能性を示した。さらに、第 3 章で紹介したように、数理モデルで示唆された同期状態の内、2 クラスタ逆相同期状態を実際に野外で観測した。このように本数理モデルはある程度の予測性を備えていると解釈できる。しかし、より正確にアマガエルの行動を記述するために改良すべき点が残されている。

まず本数理モデルでは、アマガエルの位相 θ_n と空間配置 $\boldsymbol{r}_n$ が同時に変化すると仮定した。しかし、アマガエルは鳴いている間はその場に留まっており、空間的に移動しない。そのため、鳴いている間に周囲の個体と相互作用することで自身の適切な空間配置を判断し、その後、鳴き止んだ上で空間配置を変化させるモデルに改良したほうが、実際の現象を正確に記述できる可能性がある。sin 関数と仮定した相互作用関数を直接実験データから推定する研究課題 [Aihara et al., 2011] も含めて、より正確にアマガエルの行動機構を記述する数理モデルに改良していくのは、今後の課題である。

4.5 音声可視化装置「カエルホタル」および数理モデルの応用

本稿で紹介した数理モデルおよび音声可視化装置「カエルホタル」は、他種のカエルや昆虫の行動解析に応用できる可能性がある。例えば、アマガエル以外にも周期的に信号を出し、かつ相互作用する動物が存在する。日本の清流に生息するゲンジボタルやヘイケボタルは単独では周期的に発光しながら集団で飛び回る。また、コオロギなどの昆虫にも周期的に鳴きながら、草むらなどを移動する種が存在する。本数理モデルにおいて、動物の生息地の形状や相互作用関数を変えることで、動物集団における同期状態の発現を予想できる可能性がある。

また、カエルホタルは音を光に変換するシンプルな装置なので、アマガエルに限らず夜行性で音を発する動物の行動研究全般への応用が期待できる。動画解析の方法を簡略化し、カエルホタルのチューニング方法も自動化することで、多くの研究者に扱いやすい実験手法に改良していくのは今後の課題である。

5 まとめ

本稿では、数理モデルおよび音声可視化装置「カエルホタル」を用いたニホンアマガエルの集団発声行動の研究成果 [Aihara et al., 2014, Mizumoto et al., 2011] を概説した。まずアマガエルの合唱の数理モデルを考案し、数値計算を行うことで、2 クラスタ逆相同期状態と位相波状態が観測される可能性を示唆した。次に、水田の形状を変化させる数値計算を行い、2 種の同期状態の内、2 クラスタ逆相同期状態がより安定に観測される可能性を示唆した。その上で、音声可視化装置「カエルホタル」を用いた野外調査を行い、2 クラスタ逆相同期状態を観測した。今後は、数理モデルとカエルホタルを用いた野外調査法の改良に加えて、これらの研究成果を他種の動物の行動研究に応用していく予定である。

本研究は、理化学研究所• 基礎科学特別研究員制度および科研費基盤 (S) No.24220006 の支援を受けた。

参考文献

[松井 1996] 松井正文著：両生類の進化、東京大学出版会 (1996).

[Gerhardt and Huber, 2002] Gerhardt, H.C., and Huber, F. *Acoustic Communication in Insects and Anurans*, (University of Chicago Press, Chicago, 2002).

[Wells, 2007] Wells, K.D. *The Ecology and Behavior of Amphibians*, (The University of Chicago Press, Chicago, 2007).

[Narins and Capranica, 1976] Narins, P.M., Capranica, R.R. Sexual differences in the auditory system of the tree frog Eleutherodactylus coqui, *Science* **192**, 378-380 (1976).

[Feng et al., 2006] Feng, A.S. et al. Ultrasonic communication in frogs, *Nature* **440**, 333-336 (2006).

[前田、松井、1989] 前田憲男、松井正文著：日本カエル図鑑、文一総合出版 (1989).

[Kuramoto, 1984] Kuramoto, Y. *Chemical Oscillations, Waves, and Turbulence*, (Springer-Verlag, Berlin, 1984).

[Teramae et al., 2009] Teramae, JN., Nakao, H., and Ermentrout, G.B. Stochastic Phase Reduction for a General Class of Noisy Limit Cycle Oscillators, *Phys. Rev. Lett.* **102**, 194102 (2009).

[Tanaka, 2007] Tanaka, D. General chemotactic model of oscillators, *Phys. Rev. Lett.* **99**, 134103 (2007).

[Fujiwara et al., 2011] Fujiwara, N., Kurths, J., and Guilera, A.D. Synchronization in networks of mobile oscillators, *Phys. Rev. E* **83**, 025101 (2011).

[Aoki and Aoyagi, 2009] Aoki, T., and Aoyagi, T. Coevolution of phases and connection strengths in a network of phase oscillators, *Phys. Rev. Lett.* **102**, 034101 (2009).

[Aihara et al., 2014] Aihara, I. et al. Spatio-Temporal Dynamics in Collective Frog Choruses Examined by Mathematical Modeling and Field Observations, *Scientific Reports*, **4**, 3891 (2014).

[Mizumoto et al., 2011] Mizumoto, T. et al. Sound imaging of nocturnal animal calls in their natural habitat, *J. Comp. Physiol. A* **197**, 915-921 (2011).

[Aihara et al., 2011] Aihara, I. et al. Complex and transitive synchronization in a frustrated system of calling frogs, *Phys. Rev. E* **83**, 031913 (2011).

[合原、2013] 合原一究、辻繁樹、香取勇一、合原一幸（三村昌泰編）：現象数理学入門、東京大学出版会 (2013).

[Aihara et al., 2008] Aihara, I. et al. Mathematical Modeling of Frogs' Calling Behavior and its Possible Application to Artificial Life and Robotics, *Artificial Life and Robotics*, Springer, Vol.12, No.1-2, pp29-32 (2008).

[Aihara and Tsumoto, 2008] Aihara, I., and Tsumoto, K. Nonlinear dynamics and bifurcations of a coupled oscillator model for calling behavior of Japanese tree frogs (*Hyla japonica*), *Math. Biosci.* **214**, 6-10 (2008).

[Aihara, 2009] Aihara, I. Modeling synchronized calling behavior of Japanese tree frogs, *Phys. Rev. E* **80**, 011918 (2009).

[Jones et al., 2014] Jones, D.L., Jones, R.L., Ratnam, R. Calling dynamics and call synchronization in a local group of unison bout callers, *J. Comp. Physiol. A* **200**, 93-107 (2014).

[Mizumoto et al., 2012] Mizumoto, T. et al. Sound imaging system for visualizing multiple sound sources from two species, Proc. of the 10th International Congress of Neuroethology, University of Maryland, College Park, MD, USA (2012).

〒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/`