韻律および声質を表現した音響特徴と対話音声におけるパラ言語情報の知覚との関連

石井 カルロス寿憲† 石黒 浩† 萩田紀博†

これまでのパラ言語情報の抽出に関する多くの研究は，韻律特徴を重視してきたが，自然対話音声を対象とした場合，気息性や非周期性などを含んだ non-modal な声質が現れやすく，基本周波数さえ測定できない場合も多い．ゆえに，韻律情報以外に声質情報を考慮することが重要となる．本稿では，発話スタイルを韻律と声質の特徴によって表現することを提案し，対話音声におけるパラ言語情報（発話意図，態度，感情など）との関連を考察する．具体的には，さまざまなパラ言語情報を意図して発声された音声データを対象に，知覚実験および音響分析を行い，韻律特徴と声質特徴のパラ言語情報への影響を調べた．その結果，韻律特徴は**肯定**，**聞き返し**，**フィラー**，**否定的な表現**のような機能的なパラ言語情報を識別するのに有効である可能性を示すことができた．また一方，強い気息性，強い非周期性，または喉頭を力んだ発声を含んだ声質特徴は**驚き**，**嫌悪**，**疑い**，**感心**など，比較的強い感情や態度を表すパラ言語情報に出現することが示せた．さらに追加実験として，自然対話音声に現れる non-modal な声質を含んだ発話を分析し，意図して発声された音声データと同様な傾向があることを示した．これらの報告とともに，音響特徴による声質の自動検出に関して，それぞれの声質に応じた各アルゴリズムの性能について報告する．

Acoustic Representation of Prosodic and Voice Quality Features and their Relationship with Perception of Paralinguistic Information in Dialog Speech

Carlos Toshinori Ishi,† Hiroshi Ishiguro† and Norihiro Hagita†

To date, most works dealing with paralinguistic information extraction have focused only on prosodic features like fundamental frequency (F0), power and duration. However, when analyzing natural conversational speech data, the presence of several voice qualities (caused by non-modal phonations) is often observed, mainly in expressive speech utterances. In the present work, the use of voice quality features in addition to classical prosodic features is proposed for automatic extraction of paralinguistic information (intentions, attitudes and emotional expressions) in dialog speech. Perceptual experiments and acoustic analyses are conducted for monosyllabic utterances spoken in several speaking styles (acted) in order to produce different paralinguistic information. Acoustic parameters related with prosodic and voice quality features potentially representing the variations in speaking styles are evaluated. Experimental results indicate that prosodic features are effective for identifying some groups of paralinguistic information carrying specific functions, while voice quality features are useful for identifying utterances with an emotional or attitudinal expressivity. Experiments are also conducted on natural conversational speech data with emphasis on utterances containing non-modal voice qualities. Results of these natural speech data showed the same trends as those of the acted ones. Evaluation results on the proposed algorithms for automatic detection of each voice quality are also reported.

1. はじめに

ロボットなどのような機械と人間の間で，音声対話を介して円滑なコミュニケーションを実現するには，言語情報の理解とともに，発話意図や話者の態度・感情などを表現するパラ言語情報の理解も重要となる．

パラ言語情報の識別に関しては，これまでにも多くの研究がある．意図・態度・感情を表現するさまざまな項目の中でも，特に，{ 怒り，悲しみ，喜び } などの感情識別に着眼した研究が多い[1)~4)]．感情以外のパラ言語情報の識別に関する研究においては，{ 肯定

† ATR 知能ロボティックス研究所
ATR Intelligent Robotics and Communication Laboratories

的，否定的 } 発話態度の認識を試みた研究[5] や，{ あいづち，理解，気づき，フィラー } など，発話機能の識別を試みているもの[6] があげられる．また文献 7) では，話者自身の感情状態を，快/不快，覚醒/睡眠のレベル，対人関係を，支配/服従，信頼/不信のレベル，態度を，関心/無関心，肯定的/否定的のレベルによって記述する方法を提案している．これらの研究に対して，本研究では機械と人間の間の円滑なコミュニケーションの実現を目標に，喜怒哀楽のような感情よりもむしろ文献 5)，6) が識別しようとする，発話の機能を表すパラ言語情報に焦点を当てる．

これまでのパラ言語情報の識別に関する研究には，基本周波数（F0）・パワー・持続時間などの韻律特徴（prosodic features）を利用したものが多く[4)~6),8)]，またケプストラムなどのスペクトル情報に基づいた分節的特徴を利用したものも存在する[2),3),8)]．しかし一方で，自然発話を分析した最近の研究では，声帯音源に関連する声質情報の重要性も指摘されている[9)~12)]．特に表現が豊かな発話音声では，気息性や非周期性などを含んだ non-modal な声質となりやすく，F0 さえも測定できない場合が多いため[13)]，韻律情報以外に，声質情報を考慮することは重要となる．

一般に，“声質”（“voice quality”）は，話者特有の声の特徴や，声道・鼻腔・声帯の音声器官全体の特徴を表した声の質を広く意味する[14)]．これに対して本稿では，狭義での声帯振動のモードによって特徴付けられる声の質（laryngeal voice quality）を指す．文献 14) では，声帯振動のモードと知覚的印象により，modal（地声），breathy および whispery（気息性のある声），vocal fry または creaky（基本周波数が通常発声よりも低く，パルス的な声），harsh および ventricular（雑音的で耳障りのある声），およびこれらの組合せとして，声質を分類表現することが提案されている．

近年のパラ言語情報の識別に関する研究では，声質に関連する音響的特徴を利用する研究も増えている．たとえば文献 1) では，韻律と声質に関連するさまざまな音響特徴を用いて，{ 恐怖，怒り，悲しみ，喜び，平常 } の感情識別を試みた結果，韻律よりも声質に関連するパラメータがより良い識別能力を示すことが報告されている．また，文献 15) では，気息性を表現する音響パラメータが快/不快の知覚と関連することを示している．

一方，著者の過去の研究[16)~19)] でも，韻律およびさまざまな声質に関連する音響パラメータを提案している．本研究ではそれらのパラメータを基に，**図 1** に示されるような，発話スタイルを韻律特徴と声質特徴で表現した構造を提案するとともに，さまざまなパラ言語情報の表現に必要な音響特徴を探るため，知覚実験および音響分析を行った．

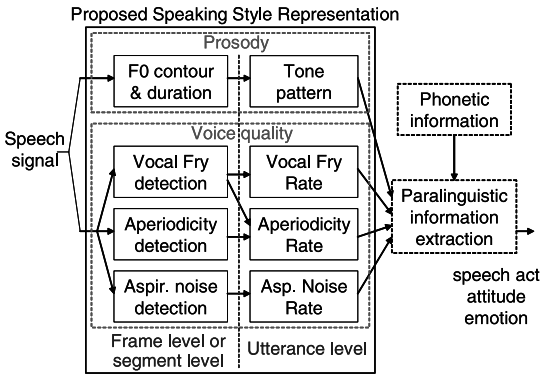


図 1 提案する韻律と声質情報を考慮した発話スタイルの表現とパラ言語情報抽出の構造

Fig. 1 Framework for paralinguistic information extraction including the proposed speaking style representation using prosodic and voice quality features.

本稿の構成は以下のとおりである．続く 2 章では音声データと知覚ラベルについて述べる．2.1 節ではパラ言語情報の種類を定義し，パラ言語情報の観点からバランスの良い音声データの作成について説明する．2.2 節と 2.3 節ではパラ言語情報と声質の知覚ラベルについて説明し，パラ言語情報の知覚における声質の役割を示す．3 章では韻律と声質に関連する音響パラメータを導入し，知覚されたパラ言語情報の識別性・関連性について報告する．4 章では自然対話音声に現れる non-modal な声質を対象に，パラ言語情報との関連および音響分析について報告する．最後に 5 章で結論と今後改善すべき点を述べる．

2. 音声データと知覚ラベル

2.1 パラ言語情報の種類と音声データ

人間同士の対話では，「えー」，「あー」，「うーん」などのような非語彙的な発話が頻繁に用いられる．これらの単語には特定の意味はないが，その発話スタイル（言い方の違い）によって何らかの意図，態度，感情などのパラ言語情報を伝達している．ある発話が伝達可能なパラ言語情報の種類は，図 1 でも考慮しているように，その発話の内容を表す音素情報にも依存することが考えられるが，本稿では，対話音声に頻繁に現れ，発話スタイルによって豊富な種類のパラ言語情報が伝達される，発話「え」に着目して検討した．

新明解国語辞典では，強い感動・驚き・疑問などを表す「え」，肯定・承諾を表す「ええ」（1 型），フィ

表 1 発話「え」におけるパラ言語情報のリスト
Table 1 List of paralinguistic information for the utterance “e”.

肯定・承諾 (affirm)	感心 (admiration)	驚き・びっくり (surprise)
同意・納得・理解 (agree)	羨望 (envy)	
相槌・頷き (backchannel)	同情・共感 (sympathy)	意外 (unexpected)
聞返し (ask for repetition)	戸惑い・躊躇・困惑 (embarrassment)	疑い (suspicion)
考え中・時間稼ぎ・フィラー (thinking)	不満 (dissatisfaction)	非難・拒絶 (blame)
		嫌悪・いや (disgust)

ラーの「ええ」(0 型) のように区別されている．しかし，「え」や「ええ」の表記以外にも，「え?」，「え!?」，「えーー」，「えっ!」など，文字やアクセント型だけでは表現しきれないさまざまな発話スタイルが存在する．ゆえに本稿において，「え」はこれらのさまざまな発話スタイルを含むものとする．

また，自然発話では国語辞典に載っていないパラ言語情報も多く存在する．本稿で対象とする発話「え」が伝達可能なパラ言語情報の種類については，CREST/ESP の自然対話音声データベース[20] に示されている相槌の発話行為ラベルセットを基にした．このラベルセットの作成においては，対話音声データに現れる「え」に関して，それぞれの発話が伝達するパラ言語情報について被験者 4 名が自由筆記で回答している．ラベル付与作業には文脈が考慮され，被験者には会話の流れを聞くことも許されている．また，付与されたすべての用語は，重複が少なくなるように被験者 4 名の話し合いによって整理されている．このような作業を経て得られた結果を本研究で用いた．**表 1** のリストにパラ言語情報の用語をまとめた．このリストは，「え」によって表現可能なパラ言語情報を必ずしもすべて含むものではないが，コミュニケーションにおける発話機能の表現に関して，十分豊かなものであると考えている．

表 1 のリストには，{ 肯定，聞返し } のような何らかの意図を示すものや，{ 疑い，非難 } のような態度的なもの，{ 驚き，嫌悪 } のように何らかの感情を表現するものも含まれている．これらの項目は喜怒哀楽のような感情よりも発話の機能的な役割を表すパラ言語情報を表現するものが多い．しかし，すべての項目を意図・態度・感情によって明確に分類するのは難しいため，本稿ではこれらの項目を総称して “パラ言語情報” と呼ぶ．

分析や評価用の音声データとしては，パラ言語情報の観点からバランスの良いデータを求めるために，表 1 に示すパラ言語情報を表現した発話音声を新たに収録した．そのために，指定のパラ言語情報を表現した発声を誘導するような台本を準備した．各パラ言語情報それぞれに対して例文は 2 つ準備した（付録の A 発話を参照）．

録音は次のように行った．まず，台本を基に特定の話者が発声したもの（誘導発話）を録音する．次に，録音された誘導発話を別途募った被験者にヘッドホンを通して聞かせ，指定のパラ言語情報を発話「え」によって表現するよう被験者に発声してもらった．より自然な発声が得られるように，「え」に続いて，指定のパラ言語情報をより強めるための短い後続発話も考案した（付録の B 発話を参照）．ただし，「え」と後続発話の間には短いポーズを入れるよう指示した．また，「え」で表現し難い場合は「へ」と発声することを許した．そのほか，追加発声として，自然発話では頻繁に現れるが，このような意図した発声では現れにくい喉頭を力んだ発声[21] を「え」と「へ」で発声してもらった．

話者 6 名（15 歳から 35 歳の男性 2 名，女性 4 名）に，以上の手順でさまざまなパラ言語情報を意図して発声してもらった．実際には 9 名の音声を収録したが，うち 3 名は棒読みのような不自然な発声となったので，分析データから外した．収録された音声データから「え」もしくは「へ」の部分を手動で切り出した総 207 発話を分析対象とした．

2.2 パラ言語情報の知覚ラベル

パラ言語情報の知覚ラベルを付与する理由は 2 つあげられる．1 つ目は，特定のパラ言語情報を意図して発声された発話「え」が，文脈なしでどの程度聞き手に伝わっているのかを調べることである．もう 1 つの理由は，文脈によって同じ発話スタイルでも異なったパラ言語情報が表現可能なので，その表現性の曖昧さを調べることである．ここでは 2.1 節で切り出された「え」または「へ」の部分のみの発話を聞いて，どのパラ言語情報が知覚されるのかを記録した．

切り出された 207 発話をランダムに並べ替え，訓練されていない被験者 4 名が各発話を聞いて，文脈なしでその発話のみから知覚されるパラ言語情報を表 1 に示したリストから選択した．ただし，文脈なしではパラ言語情報を唯一に特定することが難しい場合もあり，また，リスト中のパラ言語情報もすべて独立とは限らないので，複数の項目を選択可能として回答させた．そして，その結果 3 名以上が一致したものを，パラ言語情報の知覚ラベルとして扱うことにした．**表 2** に，発声時に意図したパラ言語情報（1 番目の列）と，知覚されたパラ言語情報との一致（2 番目の列）および不一致（3 番目の列）の結果をまとめる．省略のため，表 1 のリストで 1 つのパラ言語情報について複数の用語が存在する場合は，最初の用語のみを表 2 お

表 2 意図したパラ言語情報と知覚されたパラ言語情報との一致・不一致・曖昧さ
Table 2 Matches, mismatches and ambiguities between intended and perceived paralinguistic items.

Total number of intended SA	N. of matches	Number of mismatches or ambiguities
肯定(12)	(12)	同意(12) 相槌(12)
同意(9)	(9)	肯定(9) 相槌(9)
相槌(12)	(8)	肯定(6) 同意(7)
聞返し(12)	(11)	意外(1) 驚き(1)
感心(12)	(10)	羨望(3) 驚き(2) 意外(1)
驚き(12)	(10)	意外(6) 非難(1)
考え中(10)	(8)	戸惑(1) 不満(1) 嫌悪(1)
嫌悪(12)	(8)	非難(6) 不満(2) 疑い(1)
不満(12)	(7)	非難(5) 疑い(4) 嫌悪(2)
羨望(12)	(5)	不満(3) 意外(3) 驚き(2)
非難(12)	(5)	嫌悪(3) 疑い(2) 驚き(2) 意外(2)
疑い(12)	(4)	不満(5) 非難(4) 驚き(2)
意外(12)	(4)	驚き(4) 聞返(2) 疑い(2)
戸惑い(12)	(2)	考え中(5) 不満(6)
同情(12)	(2)	不満(4) 感心(3) 意外(2)
力んだ「え」(7)	-	嫌悪 (5) 考え中(2)
力んだ「へ」(5)	-	感心 (5)

よびこれ以降の表や図に表示する．

まず，意図して発声したパラ言語情報がどの程度聞き手に正しく伝わったかを示す 2 番目の列に注目すると，**肯定**，**同意**，**相槌**，**聞返し**，**感心**，**驚き**，**考え中**は文脈なしでも正しく伝わっており，**嫌悪**と**不満**はある程度伝わっているといえる．しかし，**戸惑い**，**同情**，**意外**，**非難**，**羨望**においては，発話の多数が他のパラ言語情報として知覚された．これらの項目の不一致および曖昧さを 3 番目の列で見ると，**戸惑い**の多くは**考え中**，または**不満**と知覚され，**意外**の多くは**驚き**と知覚された．意外だと感じた場合，驚いてしまうという状況は十分ありうるので，この 2 つの項目が同時に現れることは十分考えられる．また，戸惑いながら考える，不満を感じて戸惑うという状況もありうる．しかし，**同情**の場合は**不満**，**感心**，**意外**など，異なった意味を表した項目との不一致が多く，文脈なしで「え」の発話スタイルのみから認識することは難しいと考えられる．**羨望**の場合は，**不満**，**意外・驚き**と知覚され，これも後続の発話（つまり，文脈）によってパラ言語情報が明確になるものと考えられる．

喉頭を力んだ発声に関しては，自然発話ではよく見られるのであるが，意識して発声できない話者もいたのでサンプル数が少数となった．力んだ発話のうち，「え」は**嫌悪**に，「へ」は**感心**に知覚される傾向が見られた．

ここで注意していただきたいのは，本稿では文脈なしの発話「え」のみからどの程度パラ言語情報が認識できるのかという問題を重視している点である．したがって，本稿で議論する音響分析においては，意図されたパラ言語情報の分類ではなく，知覚されたパラ言語情報による分類を用いる．

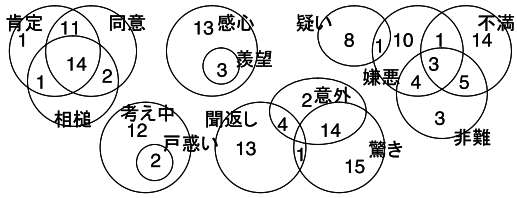


図 2 知覚ラベルによるパラ言語情報の項目の分類
Fig. 2 Grouping of the paralinguistic information items according to the perceptual data results.

表 3 音響分析に用いる知覚されたパラ言語情報の分類
Table 3 Groups of perceived paralinguistic information used for acoustic analysis.

Perceived speech act group	Number of utterances
肯定・同意・相槌	29
考え中・戸惑い	14
感心・羨望	16
聞返し	13（*聞返しのみ*）
驚き・意外	36
疑い	8 （*疑いのみ*）
非難・嫌悪・不満	27（*不満*のみを除く）
不満	14（*不満*のみ）

各パラ言語情報が知覚された発話数を**図 2** のように整理した．複数選択を許した結果がパラ言語情報間の重なりとして表されている．また，3 名以上の一致が得られなかった 50 発話は図から除外されている．

図 2 に示される重なりの部分の発話数が，各パラ言語情報の個別の発話数よりも多い場合は，パラ言語情報の項目をひとまとまりにした．その結果，**表 3** のような分類が得られた．これ以降の音響分析には，表 3 のように分類された 157 発話を評価対象とする．

2.3 声質の知覚ラベルとパラ言語情報との関係

声質特徴の知覚ラベルを付与する理由として，声質とパラ言語情報との関係を調べることと，声質に関連する音響パラメータを評価することがあげられる．

声質は知覚的に明確な分類が難しいので，ここでは声質の分類に経験のある被験者 1 名（著者本人）が音声を聴取し，波形やスペクトログラムを見ながら付与したラベルを用いることとした．音声サンプルは著者らが準備したホームページ[22] のリンクから聞くことができる．

声質ラベルとしては，modal（***m***，地声），whispery（***w***，気息性のある声），aspirated（***a***，発話末に現れる強い息漏れ），creaky（***c***，非常に低くパルス的な声），harsh（***h***，雑音的で耳障りのある声），pressed（***p***，喉頭を力んだ声）のカテゴリを準備し，これら単独または組合せ（***hw***，***pc*** など）によって表現されるものとした．

知覚によって表 3 のように分類されたパラ言語情報

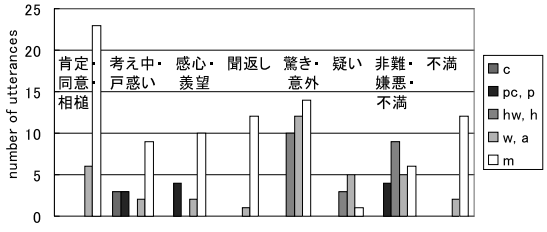


図 3 知覚されたパラ言語情報の項目における知覚された声質の分布
Fig. 3 Distribution of the perceived voice qualities, for each perceived paralinguistic information group.

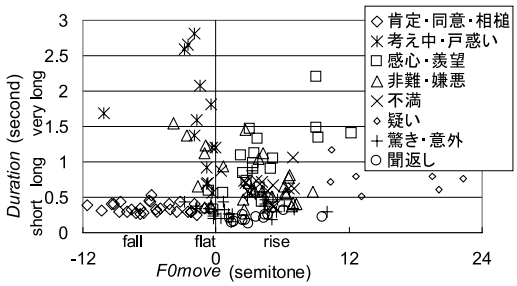


図 4 韻律パラメータによるパラ言語情報の分布
Fig. 4 Distributions of the prosodic parameters for each perceived paralinguistic information group.

のグループと，知覚された声質との関係を**図 3** に示す．***w*** と ***a*** は，知覚印象は異なるが，パラ言語情報の観点から著しい違いが見られなかったので，図 3 ではひとまとまりにした．

図 3 の結果から，比較的強い non-modal な声質（***h***，***hw***，***a***，***w***，***pc***）が知覚された発話は，比較的強い感情や態度を表現するパラ言語情報（**驚き・意外**，**疑い**，**嫌悪・非難**，**感心・羨望**）に現れることが推察できる．気息性（***w***）に関しては，**肯定・同意・相槌**でも多少知覚されたが，これは感情ではなく，丁寧さを表現するために生じたものと考えられる[23]．これらの結果はパラ言語情報の識別における声質情報の重要性を示している．

ただし，これらの強い感情や態度を表現するパラ言語情報において，図 3 の ***m*** カテゴリに示されるように modal 発声の発話も多数出現した．このことから，non-modal な発声は特定のパラ言語情報の表現において必要不可欠ではないが，non-modal な発声が起きた場合，これらの強い感情や態度を表したパラ言語情報が表現されている可能性が高いと理解できる．つまり，声質特徴はパラ言語情報の表現（生成）には必要不可欠ではないが，パラ言語情報の認識・理解においては重要な役割を果たしているといえる．

3. 音響パラメータとパラ言語情報との関連

前章ではパラ言語情報の項目と声質の関係を知覚の観点から調べた．本章では，さまざまな発話スタイルを表現するための韻律および声質に関連する音響パラメータを導入し，知覚されたパラ言語情報との関連について述べる．

3.1 韻律に関連する音響パラメータとパラ言語情報との関連

韻律特徴の基本パラメータとなる F0 の抽出には，LPC 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っている．ただし，特に non-modal な区間では誤った値が抽出されやすいので，これらの誤りの後続処理への悪影響を防ぐために，自己相関関数で F0 の sub-harmonic に対応するピークも，ある閾値を満たさなければならないという制約を追加した[17]．

韻律パラメータとして，先行研究[16] で提案した *F0move* と発話の持続時間を用いた．*F0move* は，ピッチ知覚を考慮し，音節内のピッチの動き（方向と度合い）を semitone 単位で表すパラメータである．具体的には音節を 2 等分し，各区間において代表的な F0 の値を抽出し，これらの差分をとったものである．先行研究[16] では，各区間の代表的な F0 としてさまざまな候補が評価されているが，ここではピッチ知覚に最も対応した前半区間の平均値（*F0avg2a*）と後半区間のターゲット値（*F0tgt2b*）を用い，*F0move*=*F0tgt2b*−*F0avg2a* として *F0move* を算出する．F0 抽出法や F0 のターゲット値の具体的な求め方については，文献 16) を参照のこと．

持続時間に関しては，発話「え」は単音節なので，人手によって区切られた情報をそのまま使うことも可能だが，発話前後に無音区間が多少入ってしまう場合がある．そこで母音区間のみを抽出するためにパワー情報を利用した．具体的には，発話前後のパワーが発話の最大パワーより 20 dB 以上になっている位置まで，境界を自動的に補正した．これによって得られた境界を用いて発話の持続時間（*duration*）を測定した．

図 4 に韻律パラメータ（*F0move* vs. *duration*）によるパラ言語情報の分布を示す．

図より，韻律特徴は，**肯定・同意・相槌**（短下降型），**聞返し**（短上昇調），**疑い**（動きの幅が広い上昇調），**考え中・戸惑い**などフィラー的な曖昧な表現（平坦，長下降調），それ以外の否定的または曖昧な表現（長上昇調，長平坦調）というように，主には機能的な項目を識別するのに有効である可能性を示している．しかし，長上昇調ではさまざまな項目（**非難・嫌悪**，**感心・羨望**，**不満**，**驚き・意外**）が混合しており，韻律

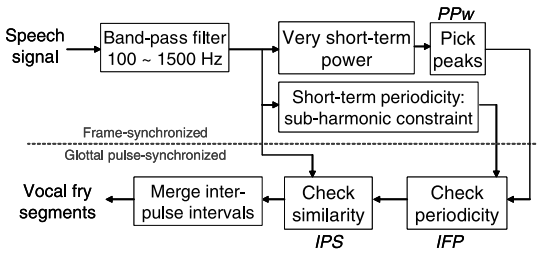


図 5 Vocal fry 区間の自動検出アルゴリズム
Fig. 5 Simplified block diagram of the vocal fry detection.

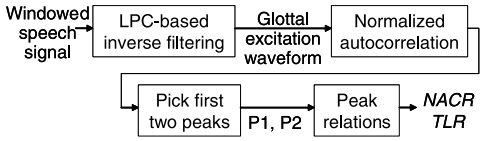


図 6 非周期性・ダブル周期性に関連する音響パラメータの推定法
Fig. 6 Simplified block diagram of the parameters for aperiodicity/double-periodicity detection.

特徴のみでの識別は難しい．また，短上昇調の中でも，**聞返し**と**驚き・意外**の違いは明確ではない．この結果は，韻律特徴のみでのパラ言語情報の識別には限界があることを示している．

また，F0 抽出には注意したが，主に harsh と creaky の区間で，F0 の抽出誤りが *F0move* に反映されてしまうような発話が**嫌悪・非難**で少数現れた．今後，これらの non-modal な声質を考慮して F0 抽出方法を改良する必要がある．

3.2 声質に関連する音響パラメータ

本節では，声質に関連する音響パラメータを 3.2.1〜3.2.3 項で導入し，韻律特徴のみでは表現できないパラ言語情報の項目を，声質特徴によってどの程度表現できるかを 3.2.4 項で示す．

3.2.1 Vocal fry（creaky）区間の検出

ここでは，先行研究[17] で提案した vocal fry（creaky）区間検出アルゴリズムを使用する．アルゴリズムは vocal fry のパルス性と通常発声よりも低い基本周波数（長いパルス間隔）の特徴を反映するために，通常使用される 25〜32 ms のフレーム長と 5〜10 ms のフレームシフトの短時間処理に対し，5 ms のフレーム長と 2.5 ms のフレームシフトの “超短時間”（“very short-term”）のパワー軌道を用いる．**図 5** のブロック図に示されるように，超短時間パワー軌道から，パワーピークを声帯パルスの候補として検出し，隣り合うピークの周期性と類似性の制約をチェックして，vocal fry による声帯パルスであるかどうかを判断する．検出は主に以下の 3 つのパラメータによって行う．

- パワーピークを検出するためのパワー（*PPw*: Peak Power）
- 自己相関関数に基づいたフレーム内の周期性（*IFP*: Intra-Frame Periodicity）
- ピーク周辺の波形の相互相関に基づいたパルス間の類似性（*IPS*: Inter-Pulse Similarity）

具体的なアルゴリズムとパラメータの詳細や評価については文献 17) を参照のこと．本研究では，$PPw > 7\,\mathrm{dB}$，$IFP < 0.8$，$IPS > 0.6$ と設定した．

3.2.2 非周期・ダブル周期（aperiodicity/double-periodicity）区間の検出

Vocal fry および harsh 発声は，声帯振動の周期性が不規則になる特徴を持っている．この不規則性は，声帯パルスの非周期性またはダブル周期性として現れる．ここでは，先行研究[18] で提案したダブル周期・非周期に関連する音源波形の自己相関関数に基づいたパラメータを使用する．これらのパラメータは本来 creaky（vocal fry）区間を検出するために提案したものであるが，予備的な実験により，harsh 発声による非周期性・ダブル周期性も反映されることが確認できている．ここで検出する非周期性・ダブル周期性区間のうち，3.2.1 項の手法によって vocal fry 区間として検出されない区間を harsh として検出することを試みる．

図 6 に非周期性・ダブル周期性に関連する音響パラメータの推定法の簡単なブロック図を示す．パラメータは，入力音声信号に声道の逆フィルタをかけて求めた音源波形の正規化自己相関関数の，最初の 2 つのピークの関係を表現している．ピーク検出においては，自己相関値が 0.2 以上のもののみピークと見なす．パラメータは以下のものである．

- 最初の 2 つのピークの正規化自己相関値の比率（*NACR*: Normalized Auto-Correlation Ratio）
- 最初の 2 つのピークの正規化自己相関ラグの比率を 2 倍したもの（*TLR*: Time-Lag Ratio）

$NACR > 1$ または $0.8 > TLR > 1.2$ の条件で，ダブル周期性または非周期性をフレームごとに検出する．パラメータの詳細と評価に関しては文献 18) を参照のこと．

3.2.3 気息音（息漏れ雑音：aspiration noise）区間の検出

気息音（息漏れ雑音）とは，breathy 発声や whispery 発声において，声帯振動における声門の不十分な閉鎖，かつ十分な狭めによって生成される気流雑音（turbulent noise）のことを指す．生成メカニズムとしては，breathy と whispery は区別されるが[14]，音

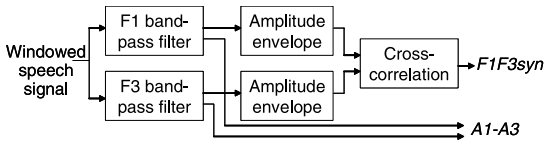


図 7 息漏れ雑音の自動検出における音響パラメータの推定法
Fig. 7 Simplified block diagram of the parameters for aspiration noise detection.

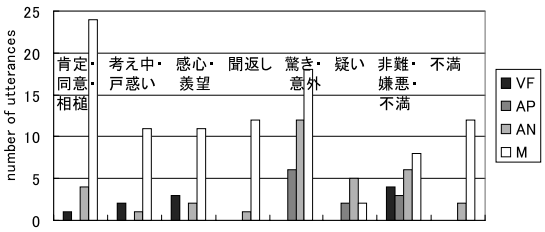


図 8 知覚されたパラ言語情報の項目における自動検出された声質の分布
Fig. 8 Distribution of the detected voice qualities, for each perceived paralinguistic information group.

響的にも知覚的にもその分類は難しい[24]．また，気息音は harsh 発声とともに現れる場合もある（harsh whispery voice[14]）．

気息音を検出する手法として，先行研究[19] で提案したものを使用する．手法は以下の 2 つのパラメータによって検出を行う．

- 第 1 と第 3 のフォルマント（F1，F3）周辺の周波数帯域でフィルタリングした信号の同期性を定量化したもの（*F1F3syn*: F1 and F3 band synchronization）
- F1 と F3 帯域のパワーの差を表すもの（$A1-A3$）

F1F3syn は，F1 と F3 帯域の波形振幅包絡の相互相関によって求める（**図 7** 参照）．気息性がない場合，*F1F3syn* は 1 に近づき，気息性がある場合は 0 に近づく．2 つ目のパラメータの $A1-A3$ は，*F1F3syn* の使用を制限するのに用いられる．$A1-A3$ が比較的大きい場合（つまり F3 帯域のパワーが F1 帯域のパワーと比べて弱い場合）は，F3 帯域の雑音は知覚されていない可能性があり，同期率を図る意味がなくなるのである．F1 帯域は 100〜1,500 Hz，F3 帯域は 1,800〜4,500 Hz に固定した．本手法の詳細およびパラメータの評価に関しては，文献 19) を参照のこと．ここでは $F1F3syn < 0.4$ および $A1 - A3 < 25$ dB の条件でフレームごとに気息音を検出する．

3.2.4 声質パラメータとパラ言語情報との関連および声質ラベルの自動検出の評価

以上のパラメータにより，フレームごと，あるいは区間ごとの情報が得られるが，以下のものを発話ごとのパラメータとして提案する．

- Vocal Fry Rate（$\boldsymbol{VFR}$）：発話全体に対し，vocal fry（creaky）が検出された区間の割合．
- Aperiodicity Rate（$\boldsymbol{APR}$）：発話全体に対し，非周期またはダブル周期が検出され，vocal fry とは検出されなかった区間の割合．
- Aspiration Noise Rate（$\boldsymbol{ANR}$）：発話全体に対し，気息性（息漏れ雑音）が検出された区間の割合．

以上のパラメータにより，$\boldsymbol{VF}$（vocal fry），$\boldsymbol{AP}$（非周期・ダブル周期），$\boldsymbol{AN}$（気息性），$\boldsymbol{M}$（modal 発声）の 4 種類の声質特徴を識別する．$\boldsymbol{VF}$，$\boldsymbol{AP}$，$\boldsymbol{AN}$，$\boldsymbol{M}$ と認識されるカテゴリは，それぞれ 2.3 節および図 3 で示した知覚カテゴリの（$\boldsymbol{pc}$, $\boldsymbol{c}$），（$\boldsymbol{h}$, $\boldsymbol{hw}$），（$\boldsymbol{w}$, $\boldsymbol{a}$），（$\boldsymbol{m}$）に対応する．声質の認識の予備的な実験結果より，これらの発話レベルの声質パラメータの閾値を 0.1 と設定した．したがって，$VFR > 0.1$ の発話は $\boldsymbol{VF}$，$APR > 0.1$ の発話は $\boldsymbol{AP}$，$ANR > 0.1$ の発話は $\boldsymbol{AN}$，それ以外のものは $\boldsymbol{M}$，のように声質のカテゴリの自動識別を行う．自動識別の結果をパラ言語情報ごとに分類して**図 8** に示す．

図 8 の結果より，強い気息性および強い非周期性の声質（$\boldsymbol{AP}$，$\boldsymbol{AN}$）を含む発話は，**驚き**，**意外**，**非難**，**嫌悪**，**疑い**など，比較的強い感情や態度を表す項目を検出するのに有効である可能性を示している．$\boldsymbol{AP}$ と $\boldsymbol{AN}$ の使い分けは明確ではないが，疑いの知覚においては気息性の特徴（$\boldsymbol{AN}$）の方が重要といえる．この結果は，図 3 に示した声質の知覚ラベルの結果と同様の傾向を示す一方で，$\boldsymbol{AP}$ による $\boldsymbol{h}$，$\boldsymbol{hw}$ の検出が不十分であることが分かる．これは，harsh 声質を正しく検出するためには，3.2.2 項で導入した手法が不十分であることを示しており，今後改善が必要である．

また，$\boldsymbol{VF}$ に関しては，図 8 と図 3 に示されているように，強い感情を表す**感心**と**嫌悪**の $\boldsymbol{pc}$（喉頭を力んだ creaky）と，**考え中・肯定**の $\boldsymbol{c}$（柔らかい creaky）が検出できている．力んだ発声を識別するためには，さらなる音響特徴が必要であり，これも今後の課題として残される．

4. 自然対話音声データにおける non-modal な声質とパラ言語情報との関連

2 章ではパラ言語情報の観点からバランスの良いデータを求めるために，パラ言語情報を意図して発声されたものを収録したが，声質のデータとしては，比

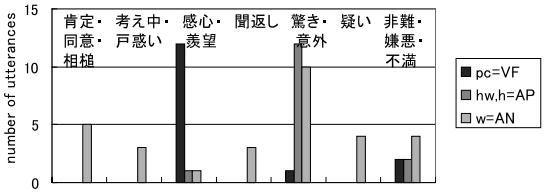


図 9 知覚されたパラ言語情報の項目における自動検出された non-modal な声質の分布（自然発話）

Fig. 9 Distribution of the perceived paralinguistic information groups for each perceived non-modal voice qualities (Natural speech data).

較的強い非周期性や力みを含んだ発声が少なかった．このような non-modal な声質は，自然会話の中では多く現れるが，2 章のように意図して発声する場合には現れにくいという結果となった．なお，自然発話では non-modal な発声は話者の心的状態などにより，無意識に起きる可能性もある．そこで，本章では自然発話から発話「え」および「へ」を抜き出し，その中から non-modal な声質のものを選択し，パラ言語情報との関連を調べる．

自然発話データとしては，CREST/ESP プロジェクトで収録された Expressive Speech Database [20] より，30 代女性話者 1 名（FAN）が長期間（およそ 3 年）にわたって収録した日常会話データを使用する．

データベースの書き起こしデータより，「え」（「ええ」，「えー」，「ええっ」など）および「へ」（「へえ」，「へー」など）を含んだ発話を検索し，被験者 1 名（著者本人）が各発話を聴取し，non-modal な声質が知覚されるものに，2.3 節と同様の基準で声質の知覚ラベルを付与した．発話の大半は**聞返し**を表現した modal 発声であったが，non-modal な声質が知覚されたものが 87 発話そろった．これらの発話に 3.2 節の声質自動検出アルゴリズムを用いて声質の自動ラベリングを行い，手動ラベルと一致した 60 発話 {$\boldsymbol{pc = VF}$ (15), $\boldsymbol{h}$,$\boldsymbol{hw = AP}$ (15), $\boldsymbol{w = AN}$ (30)} をパラ言語情報の分析対象とした．これらの音声サンプルも著者らの用意したホームページ [22] で聞くことができる．

表 3 のリストに基づき，2 章と同じ被験者 4 名が，各発話から印象付けられるパラ言語情報を選択した．ただし，ここでは文脈を考慮し，発話の前後 5 秒を含めて聴取することとした．文脈を考慮することで，被験者間の一致率も高まることを期待したのである．しかしながら，ばらつきが多く，2 名以上一致したものを，パラ言語情報の知覚ラベルとした．その結果，各声質における分布は，**図 9** のようになった．

図 9 の結果は，3 章で意図して発声された音声データについて得られた結果（図 8）と同様に，強い気息性（$\boldsymbol{w = AN}$）および強い非周期性（$\boldsymbol{h}$,$\boldsymbol{hw = AP}$）が，比較的強い感情や態度を表現する項目（**驚き・意外**，**疑い**，**非難・嫌悪・不満**）に多く現れることを示している．力んだ発声（$\boldsymbol{pc = VF}$）に関しては，図 9 の自然発話データでの発話数が多く，**感心**を表現する発話がほとんどであるという結果が得られた．したがって，自然発話データでも non-modal な発声は，比較的強い感情や態度の表現に関連するという結果が得られた．

表 4 知覚された声質と自動検出された声質の混同行列
Table 4 Confusion matrix between perceived and detected voice qualities.

	VF	*AP*	*AN*	*M*
pc	15 (88%)			2
hw, h	11	15 (40%)		11
w		2	30 (91%)	1

最後に，声質の自動検出アルゴリズムと手動ラベルが一致しなかった発話も含んだ混同行列を**表 4** に示す．

この結果より，気息音（$\boldsymbol{w = AN}$）および vocal fry 発声（$\boldsymbol{pc = VF}$）は，9 割近く検出されているものの，harsh 発声の検出（$\boldsymbol{AP}$）では脱落（$\boldsymbol{M}$）および $\boldsymbol{VF}$ として誤検出されたものが多く，検出率が低い．3.2.4 項でも示されたように，3.2.2 項で導入した非周期性検出は harsh 区間の検出にある程度貢献しているものの，十分ではないことを確認した．今後，harsh 区間の表現におけるより適切な音響パラメータを検討する必要がある．

5. 結　　論

さまざまな発話スタイルで発声された，発話「え」および「へ」を分析した結果，韻律特徴は**肯定的な表現**，**聞返し**，**フィラー**，**否定的な表現**のような，機能的なパラ言語情報を識別するのに有効である可能性を示すことができた．また一方，声質特徴（強い気息性，強い非周期性，また喉頭を力んだ発声を含んだ声）は**驚き**，**嫌悪**，**疑い**，**感心**など，比較的強い感情や態度を表すパラ言語情報を検出することに，有効である可能性を示すことができた．さらに，自然発話に現れる non-modal な声質で発声された発話のうち，喉頭を力んだ場合に発する vocal fry は，**感心**，および**嫌悪**を表現する発話に観察され，harsh 発声は比較的気持ちが高ぶりやすい**驚き・意外**，**疑い・嫌悪・非難・不満**で現れた．

今後，主に声質に関連する音響特徴の抽出方法を改善し，韻律特徴との適切な組合せを決定木や SVM などの分類アルゴリズムを用いて抽出するアルゴリズムを実装し，認識システムの識別能力を評価する予定で

ある．

謝辞 本研究は総務省の研究委託により実施したものである．アドバイスもしくは機材のサポートにご協力いただいた，榊原健一氏（NTT），パーハムモクタリ氏（ATR/HIS），北村達也氏（ATR/HIS），IRC の皆様に感謝する．音声収録および知覚実験にご協力いただいた皆様に感謝する．

参 考 文 献

1) Fernandez, R. and Picard, R.W.: Classical and Novel Discriminant Features for Affect Recognition from Speech, *Proc. Interspeech 2005*, pp.473–476 (2005).
2) Schuller, B., Muller, R., Lang, M. and Rigoll, G.: Speaker Independent Emotion Recognition by Early Fusion of Acoustic and Linguistic Features within Ensembles, *Proc. Interspeech 2005*, pp.805–808 (2005).
3) 佐藤信夫，大淵康成：ケプストラムを用いた感情識別手法の検討，日本音響学会 2005 年春季研究発表会講演論文集，Vol.I, pp.211–212 (2005).
4) 野田哲矢，矢野良和，道木慎二，大熊 繁：KL 情報量に基づく音声感情認識に有効な韻律特徴の評価法，日本音響学会 2005 年秋季研究発表会講演論文集，Vol.I, pp.394–395 (2005).
5) 藤江真也，江尻 康，菊池英明，小林哲則：肯定的/否定的発話態度の認識とその音声対話システムへの応用，電子情報通信学会論文誌，Vol.J88-D-II, No.3, pp.489–498 (2005).
6) 田中俊光，柏岡秀紀，ニック・キャンベル：発話機能における音声の非語彙的情報の分析およびその考察，日本音響学会 2004 年春季研究発表会講演論文集，Vol.I, pp.231–232 (2005).
7) 森 大毅，相澤 宏，粕谷英樹：対話音声のパラ言語情報ラベリングの安定性，日本音響学会誌，Vol.61, No.12, pp.690–697 (2005).
8) 藤野真紀，峯松信明，広瀬啓吉：音声の音響的普遍構造に着眼したパラ・非言語情報推定に関する実験的検討，日本音響学会 2005 年春季研究発表会講演論文集，Vol.I, pp.59–60 (2005).
9) Erickson, D.: Expressive speech: production, perception and application to speech synthesis, *Acoust. Sci. & Tech.*, Vol.26, No.4, pp.317–325 (2005).
10) Maekawa, K.: Production and perception of 'Paralinguistic' information, *Proc. Speech Prosody 2004*, pp.367–374 (2004).
11) Klasmeyer, G. and Sendlmeier, W.F.: Voice and Emotional States, *Voice Quality Measurement*, Ch.15, pp.339–358, Singular Thomson Learning (2000).
12) Gobl, C. and Ní Chasaide, A.: The role of voice quality in communicating emotion, mood and attitude, *Speech Communication*, Vol.40, pp.189–212 (2003).
13) Hess, W.: Pitch Determination of Speech Signals, *Vol.3 of Springer Series of Information Sciences*, Springer-Verlag, Berlin, Heidelberg, New York (1983).
14) Laver, J.: Phonatory settings, *The phonetic description of voice quality*, Ch.3, pp.93–135, Cambridge University Press (1980).
15) 森 大毅，相田千尋，粕谷英樹：活性–評価次元に基づくパラ言語情報ラベルの音響関連量，日本音響学会 2005 年春季研究発表会講演論文集，Vol.I, pp.231–232 (2005).
16) Ishi, C.T.: Perceptually-related F0 parameters for Automatic Classification of Phrase Final Tones, *IEICE Trans. Inf. & Syst.*, Vol.E88-D, No.3, pp.481–488 (2005).
17) Ishi, C.T., Ishiguro, H. and Hagita, N.: Proposal of Acoustic Measures for Automatic Detection of Vocal Fry, *Proc. Eurospeech 2005*, pp.481–484 (2005).
18) Ishi, C.T.: Analysis of Autocorrelation-based parameters for Creaky Voice Detection, *Proc. Speech Prosody*, pp.643–646 (2004).
19) Ishi, C.T.: A New Acoustic Measure for Aspiration Noise Detection, *Proc. ICSLP 2004*, Vol.II, pp.941–944 (2004).
20) http://feast.atr.jp/esp/esp-web/
21) Sadanobu, T.: A Natural History of Japanese Pressed Voice, *J. Phonetic Society of Japan*, Vol.8, No.1, pp.29–44 (2004).
22) http://www.irc.atr.jp/ carlos/voicequality/
23) Ito, M.: Politeness and voice quality — The alternative method to measure aspiration noise, *Proc. Speech Prosody 2004*, pp.213–216 (2004).
24) Kreiman, J. and Gerratt, B.: Measuring Vocal Quality, *Voice Quality Measurement*, Ch.7, pp.73–102, Singular Thomson Learning (2000).

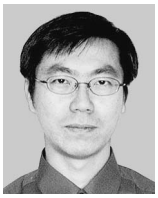
付 録

発話行為の音声収集に用いた台本

A：今日は雨かな?
B：(肯定)，雨だよ．
A：韓国料理は好き?
B：(肯定)，好きだよ．
A：今日は雨やね．
B：(同意)，そうやね．
A：お昼，ファミレス行こうか．
B：(同意)，行こう行こう．
A： 今日は雨みたい．

B：(相槌)，そうやね.
A：今日，また電車遅れてるみたいよ.
B：(相槌)，そうやってね.
A：今日は雨やし，バーベキュー中止しよっか?
B：(戸惑い)，どうしよう.
A：体の調子が悪いから，今日の予定はやめとこか?
B：(戸惑い)，じゃーどうしようかー.
A：今日は rainy だよ.
B：(聞き返し)?なんて?
A：明日の朝，7 時に出発するよ.
B：(聞き返し)？ 何時って?
A：今日は夕食の準備しておいてね.
B：(不満)，なんでよ.
A：この仕事，頼むで.
B：(不満)，なんで.
A：私の趣味は草刈だよ.
B：(意外)，うそ!
A：私，格闘技見るの好きやねん.
B：(意外)，そうなんや!
A：私はブッシュ大統領を支持するよ.
B：(非難)，なんでまたー
A：私，蛇飼ってるんねん.
B：(非難)，なんで蛇なん!?
A：私はゴキブリが好きだよ.
B：(嫌悪)，キモー!
A：満員電車が好きやねん.
B：(嫌悪)，どこがいいん?.
A：今日から 1ヶ月間，海外旅行へ行ってきまーす!
B：(羨望)，いいなー.
A：このネックレス，昨日彼氏が買ってくれてん.
B：(羨望)，ええなー.
A：ロボビーは完璧にしゃべれるようになったよ!
B：(感心)，すごいなー!
A：あの人はどんな曲でもピアノで演奏できるんだって.
B：(感心)，すごいなー!
A：ロボビーは完璧にしゃべれるようになったよ!
B：(疑い)，ありえへん!
A：私，ポルトガル語，ペラペラやねん.
B：(疑い)，うそや〜.
A：今日抽選で当たりました.
B：(驚き)，すごい!
A：昨日空港で中島みゆきに会ってん!
B：(驚き)，ほんまに?
A：もう 3 日も寝ないで仕事してるんだよ.
B：(同情)，大変やんなー..
A：階段から落ちて，骨折してん.
B：(同情)，かわいそうやな.
A：128 + 63 はいくつ?
B：(考え中),...
A：330 を 11 で割ると?
B：(考え中),...

(平成 17 年 10 月 17 日受付)
(平成 18 年 4 月 4 日採録)



石井カルロス寿憲

1996 年 ITA（Instituto Tecnológico de Aeronáutica）電子工学科卒業．1998 年同大学大学院電気通信工学科修士課程修了．1998 年文部省の留学生として東京大学大学院に入学．2001 年東京大学大学院電子情報工学科博士課程修了．工学博士．2002 年 JST/CREST ESP プロジェクトの研究員として，ATR 人間情報科学研究所にて音声情報処理の研究に従事．2005 年 ATR 知能ロボティクス研究所の研究員としてコミュニケーションロボットにおける音声情報処理の研究に従事．日本音響学会会員．

石黒 浩（正会員）

1991 年大阪大学大学院基礎工学研究科物理系専攻修了．工学博士．同年山梨大学工学部情報工学科助手．1992 年大阪大学基礎工学部システム工学科助手．1994 年京都大学大学院工学研究科情報工学専攻助教授，1998 年同大学大学院情報学研究科社会情報学専攻助教授．この間，1998 年より 1 年間，カリフォルニア大学サンディエゴ校客員研究員．2000 年和歌山大学システム工学部情報通信システム学科助教授．2001 年同大学教授．1999 年 ATR 知能映像研究所客員研究員．現在，大阪大学大学院工学研究科知能・機能創成工学専攻教授および ATR 知能ロボティクス研究所客員室長．知能ロボット，アンドロイドロボット，知覚情報基盤の研究に興味を持つ．

萩田　紀博（正会員）

1978年慶應義塾大学大学院工学研究科電気工学専攻修士課程修了．同年電電公社（現NTT）武蔵野電気通信研究所入所．文字認識，画像認識等の研究に従事．NTT基礎研究所，ATRメディア情報科学研究所長等を経て，現在，ATR知能ロボティクス研究所長．工学博士．IEEE，電子情報通信学会，人工知能学会，日本ロボット学会各会員．