2. 声質

―パラ言語情報を持つ第四の韻律パラメータ―

ニック キャンベル

要旨

本稿では、大規模自然対話音声コーパスを分析することにより得られた知見を報告する。パラ言語情報の伝達する音響パラメータとしては、基本周波数が明らかになっているが、今回の研究では、連続した音声から気息性／緊張度の評価値として測定した正規化 AQ パラメータ normalized amplitude quotient(NAQ) を用いて声質評価を行った。その結果、気息性もパラ言語情報を伝達する重要な音響的要素であることが確認された。本稿では、対話者、発話スタイル、および、発話行為の全てが NAQ と関連があり、声質がピッチ、パワー、持続時間と並んで、第四の韻律パラメータであるという主張を展開する。

1. はじめに

韻律研究は音声処理技術の進歩とともに発展を遂げてきた。ピッチ情報(より正確には音声の基本周波数、以下、F0)、および、振幅(または RMS 信号パワー)は比較的音声信号からの抽出が容易であった。しかしながら、発話タイミングや音声の持続時間の学術的研究は、音声データベースの規模、コンピュータの処理速度、ラベリングソフトウェアの信頼性による研究が可能になった 1980 年代に入るまで待たれることとなった。

21 世紀に入り、音声処理技術は新たな進歩を見せつつある、その中でも声質情報の抽出は信頼のおける水準に達してきた。Alku and Vilkman (1996) によって提案された手法を、Mokhtari and Campbell (2003) が自然に行われる対話音声に対応できるように改良したことにより、我々は日常会話の気息性を測定することが可能になった。

本稿では、信号処理の進歩に伴い、我々も第四の韻律パラメータとして、声質を含めることを提案する。そして、音声のF0、振幅、局所的話速同様、音声の中の気息性と緊張がどのように社会的、パラ言語的情報の伝達に用いられているかを示す。

2. 声質のAQパラメータ

声門での発声様式(または“声質”)は喉頭原音波形導関数の予測(声道関数の影響を取り除くための可変長の最適化されたフオルマント(Mokhtari and Campbell 2003)の逆フィルタリングの結果)から計測が可能であり、Alku and Vilkman (1996)により提案された喉頭音源指数(Amplitude Quotient, AQ)により、ピーク間の最大振幅比率、および、サイクル間の最小導関数の最大関数を測定することによって得られる。

実測値のままでは、音声波形の基本周期にわずかながらも影響を受けるが、F0に基づく正規化を行うことで、この影響を最小に抑えることができるため、本研究では、この正規化されたAQパラメータ(以下“NAQ”、Alku, Backstrom, and Vilkman (2002))を用いる。

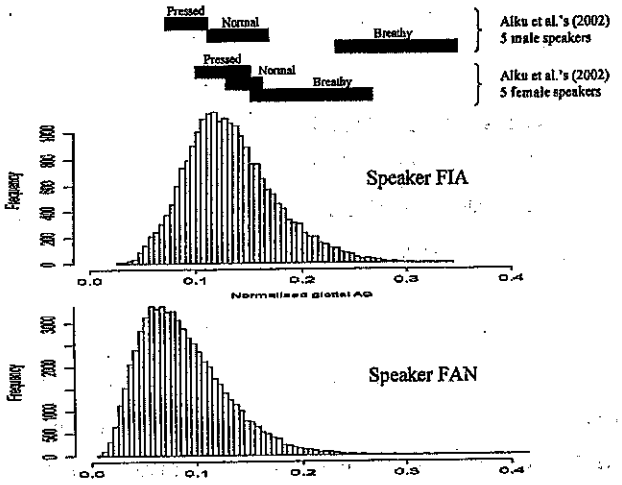


図1：JST/CREST　ESPプロジェクトで収集した2話者のコーパスのNAQ分布
Alkuらの発表した分布と比較し、差がないと考えられる。

図1に本研究で用いた日本語を母国語とする女性話者2名(FIA & FAN)のNAQのヒストグラムを示す。同図の上部には比較のためAlku, Backstrom, and Vilkman (2002)で報告された測定値を示している。この図より、個人差も認められるが、全体的な分布の形状は似ており、本研究の両データも緊張音(pressed)平静音(normal)気息音(breathy)の範囲に納まっている。発話者FANのデータに見られる傾斜部分は第4節で説明するように、気取らない(緊張音)発話スタイルの優位性に起因すると思われる。

本稿では、このNAQの多様性は偶然に生じたものではなく発話者の対話者との関係・発話者の意図・発話態度などのパラ言語情報によって最も効果的に説明できること、それ故にNAQは韻律パラメータであること、を実際のデータの分析結果を示しながら述べる。

3. 対話音声データ

JST／CREST発話様式処理プロジェクトは1,000時間を目標に、高品質な収録機材を用いて、三カ国語(日本語、中国語、英語)の大規模な自然日常対話音声の収集に取り組んでいる(Campbell 2002)。この対話音声の一部(約250時間)は書き起こされ、その一部(約100時間)は発話スタイル、および、発話行動のラベル付与が終了している。ラベル付与された音声の音響的特徴が抽出され、聞き手の聴覚的印象と物理的特性との相関分析が行われている。

本稿では女性話者一名(FAN)の音声データの分析結果について報告する。発話者はヘッドフォン型の高性能マイクロフォンを装着し、日常の発話行動をミニディスクに録音した(Campbell and Mokhtari 2002)。分析は、被験者の音声のみ行われた。ラベラーたちは必要に応じて聞こえる対話者の音声も参考にしながら、聴覚的印象をもとにラベル付与を行った。

音響的、および、聴覚的印象に基づくラベル付与が終了した音声データは13,604発話に及んだ。ここでの「発話」とは、聞こえうる間を含んでいない最短な音声と定義され、恐らくは「イントネーショナルフレーズ」という表現が最も適切な単位である(書き起こし従事者には、発話単位とは最小意味単位であると指導した)。従って、発話長は、1音節から最長では35音節に及んだ。

音声データは公開統計ソフトウェアであるCRANの「R」を用いて統計分析を行った(http://cran.r-project.org/)。各発話単位には対話者(who)，発話スタイル(how)，発話行為(what)のラベルが付与され、NAQ、および、F0との相関について分析を行った。

ラベラーの聴覚的印象に基づき、対話者は表1に示すように分類された。発話スタイルラベルは、本研究では複雑なものは用いず、丁寧(polite)，親しげ(friendly)，および気取らない(casual)と分類し，家族(family)，友人(friend)，および他人(others)の発話者に対して付与した。発話行為カテゴリーは合計24設定されたが、本稿ではそのうちの情報提供(giving information)，感嘆(exclamations)，情報要求(requesting information)，つぶやき(muttering)，および反復要求(requesting repeats)の5カテゴリーに焦点を当てる。

表1：対話者ごとの発話数

子供	家族	友人	他者	独り言
139	3,623	9,044	632	116

4. 分析結果

正規化前、音節ごとのAQとF0はr＝－0.406の相関を示していた。正規化(NAQ＝log(AQ)＋log(F0))後は，図2に示すように両者間にはr＝0.182と弱い相関しか見られなかった。図中、F0レンジの中央に広範囲でNAQの分布が認められる。そして、両者は最も低いF0を除いて、全ての四分円で独立して機能しているが、上で述べた聴覚的印象による分類と有意に相関が認められる。

図3は対話者別のNAQ、および、F0の中点を示したものである。NAQは他人(others，丁寧な発話様式)に話しかけている時に最も高く(最も気息性を伴う音声)、次に子供に話しかけている時に高くなっている(優しく)。独り言が最も低いNAQ値を示した。

家族との発話では友人との発話より、高頻度で気息音(つまり高NAQ値)が認められた。F0は子供に向けられた発話で最も高く、家族(子供を除く)に

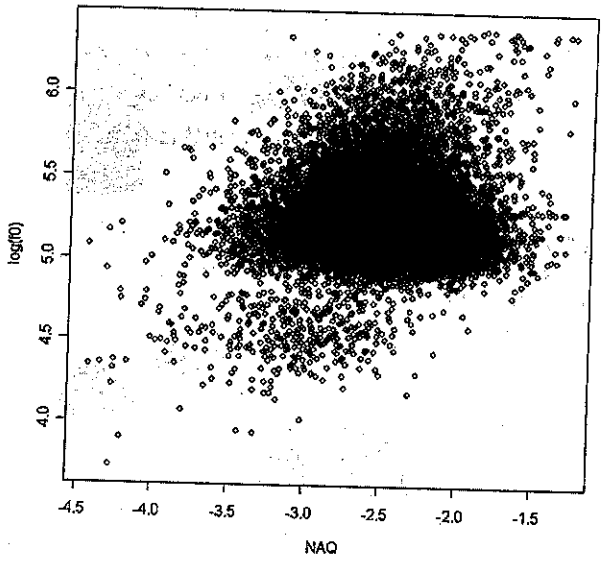


図2：話者FANのF0とNAQ
独立に変化し、依存関係がないことを示す。

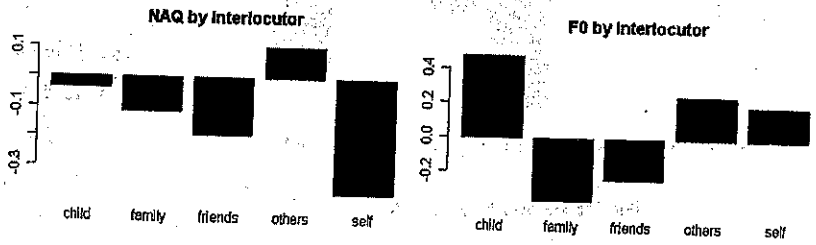


図3：対話者ごとのNAQおよびF0中点
データはz値で表され、0は分布の平均を表す。相手は、子供、家族、友人、他人、自分である。F0だけではなく、NAQも相手によって変化することが見える。＋方向はやわらかい声（裏声）、－方向は硬い声（地声）を示す。

向けられた発話で最も低かった。図からも明らかなように、それぞれの対話者への発話において、F0と気息性は独立してコントロールされている(しかしながら、それが意識的か否か、また、どのようなメカニズムを通してコントロールされているかの議論は別稿に譲る)。

表２は図３でプロットされたNAQデータのペアごとのt-testによる多重比較の結果である。過誤確率(棄却された仮説のうちで第一種の過誤が起こる確率)はFDR法によりコントロールされた(Benjamini and Hochberg 1995)。反復t-testによる子供に向けられた発話(n = 139)以外では声質は有意に異なることが確認された。

表２：ペアごとのt-test結果のp値
子供に向けられた発話は最小発話数と最大分布を持ち、有意差は認められない。
他の対話者に向けられた発話声質は有意に異なる。

	子供	家族	友人	他者
家族	0.58(ns)	-	-	-
友人	0.10(ns)	2.7e-05	-	-
他者	0.16(ns)	0.00042	1.8e-08	
子供	0.00143	0.00042	0.00656	2.0e-06

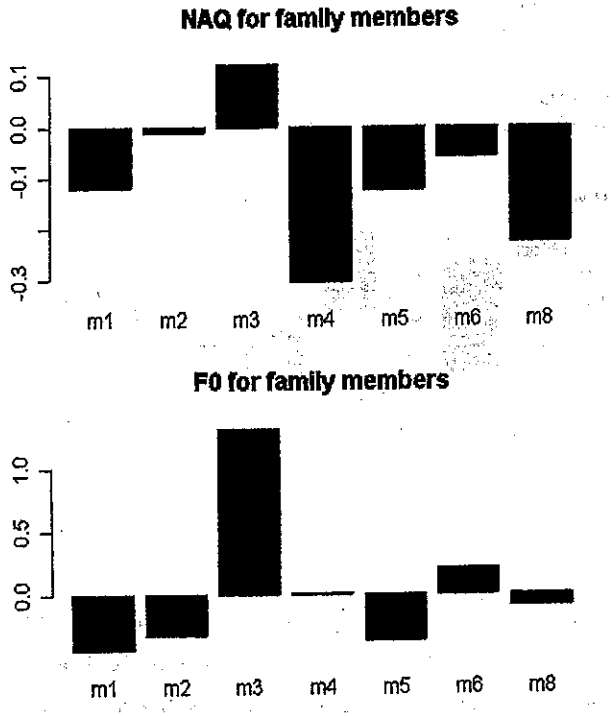


図４：家族構成員ごとのNAQおよびF0中点
m１：母、m２：父、m３：実娘、m４：夫、m５：姉、m６：甥、m８：叔母。
ここでNAQの『意味』が見えてくる。つまり「気遣い度」が示される。

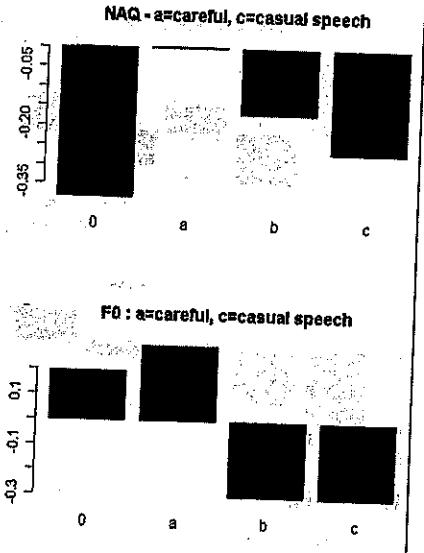
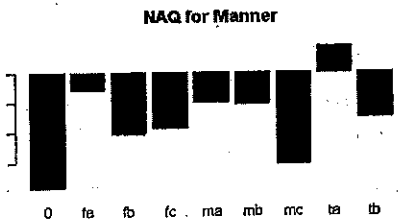
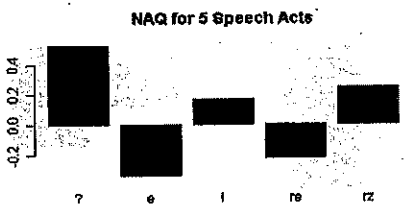


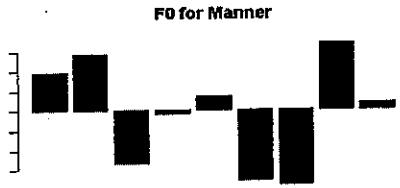
図５：発話様式(Manner)
４種類の発話スタイルのNAQおよびF0
F0よりも、NAQで発話様式や気遣いの違いを示す。

図４では、対話者が“家族”の場合の発話をより詳細に表示し、興味深い傾向を明らかにしている。一歳半の娘に対しての発話者の発話はF0、および、気息性において、最も高い。家族構成員を気息性の高い順に並べると以下のようになる。娘＞甥＞父＞母＝姉＞叔母＞夫。この順序はそれぞれの家族構成員にどの程度「気配り(care)」を払ったかを反映しているようである。これはラベラーたちの経験に基づく聴覚的印象とも一致する。発話行為と発話スタイルの分析結果もこの解釈と一致した。「発話スタイル(Manner)」ごとにデータを表示した図５を見ると、F0については、４カテゴリーにおける違いは３段階に留まっているが、NAQでは４段階見られる。カテゴリー「0」(独り言)は最も低いNAQ値を示しているが、カテゴリー「a」(慎重な発話(careful speech))は最も高いNAQ値を示している。

カテゴリー「c」(気取らない発話(casual speech))は、「b」(親しげな発話)よりも低いNAQを示している。これらは一貫してNAQは慎重深さ(degree of care)を表しているという見解を支持するデータとなっている。これらのカテゴリーのF0値は家族について、NAQと同じ順序は示さなかったが、慎重な発話は親しげな発話や気取らない発話よりも高いF0を示した。図６は対話者カテゴリー毎で分けた同値を示す。この図において、「f」は友人、「m」は家族、「t」はその他を示す。ここで興味深いのは、友人に対しては慎重な発話(fa)のNAQ値の方が親しげな発話(fb)や気取らない発話(fc)よりも高く、後者２カテゴリー間には差が見られないのに対し、家族構成員については、逆の傾向を示していることである。すなわち、慎重な発話(ma)と親しげな発話







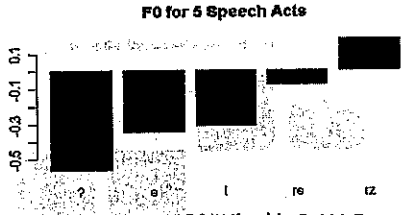


図６：対話者ごとの NAQ および F0 中点
"f" は友人、"m" は家族、
"t" は他人関係を示す。

図７：５種類の発話行為ごとの NAQ および F0

(mb) の間には差が見られないが、気取らない発話 (mc) では NAQ 値は低い。他者へ向けられた音声では、気取らない発話は見られなかったが、慎重な発話と親しげな発話は NAQ については予想された差を示した。

最後に発話行為に見られる差について報告する。前項までの報告から、慎重な発話行動の方がソーシャル・コストの低い発話行動よりも高い NAQ 値を示すのではと予想でき、このことは図７において確認できる。ここでは、つぶやき「？」、間投詞「l」、情報提供「e」、情報要求「re」、反復要求「rz」の５種類の発話行為について報告する。情報提供の NAQ 値の方が情報要求のそれよりも低く、統計分析の結果有意な差が表れている ($t = 3.2805$, $df = 1453.04$, $P = 0.001061$)。また反復要求 (対話者には最も高い負荷を要求する発話行為) は最も高い NAQ 値を示している。きわだって低い F0 値と高い NAQ 値を示したつぶやきは発話行為とは別のカテゴリーと見なした方がよいという聴覚的印象と一致する。

5. 今後の課題

本稿で報告したこの新しい知見は自然な対話音声の中の気息性の測定を可

能にした音声信号処理技術の進歩がもたらしたものである。図８は第４節の図３のデータを箱ひげ図で表したものであり、各カテゴリーの NAQ、および、F0 値の分布や両パラメータの重なり具合を、より視覚的に表示している。今回の報告 (第４節) では、両パラメータの中央値を中心に議論を展開した。中点は、各々 t-test でも確認できたように両パラメータの傾向を示すことは可能だが、図８に示されるように、それぞれのカテゴリーにおいても声質の分布は広く、今後はより細かいカテゴリーに分類して、より詳しい分析を行うことが必要だと考える。

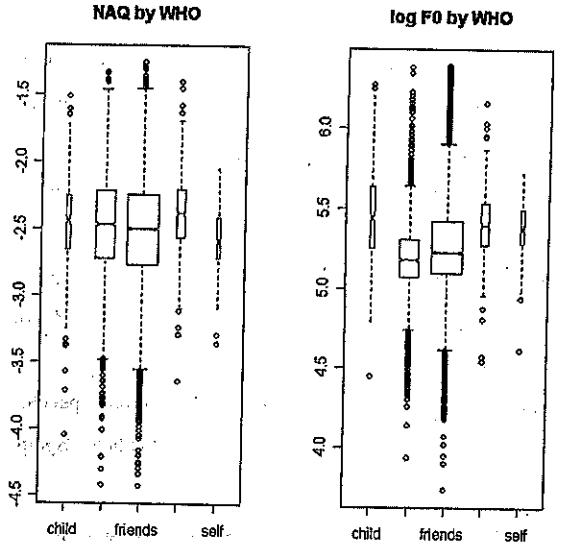


図８：対話者別 NAQ および F0 の箱ひげ図

6. 結論

本稿では、正規化 AQ パラメータ (NAQ) によって測定された声質が対話者、発話スタイル、発話行為と強い相関があることを述べた。本研究で我々は声質は (It) は、F0 とは独立に、音声の中の慎重深さ (degree of care) によって絶えず変化することを確認した。このことより、声質は意味的に異なるパラ言語情報を伝達するために、音声生成の過程において、コントロールされるものであり、F0、持続時間、振幅と並んで、韻律特性に含めるべきであると結

論する。

謝辞

本研究は科学技術振興事業団の助成を受けており、同事業団に感謝致します。また、本稿はICPhS2003で発表したものの翻訳である。飯田明美氏の翻訳結果、Parham Mokhtari氏の研究結果、木村みなこ氏のラベル結果をはじめ本プロジェクトの研究員やラベラー諸氏に感謝します。

参考文献

Alku, P. and Vilkman, E. (1996) "Amplitude domain quotient For characterization of the glottal volume velocity waveform estimated by inverse filtering," *Speech Comm.* 18-2, pp. 131-138.

Alku, P., Backstrom, T. and Vilkman, E. (2002) "Normalized amplitude quotient for parametrization of the glottal flow," *J. Acoust. Soc. Am.* 112-2, pp. 70-710.

Benjamini, Y. and Hochberg, Y. (1995) "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society Series B* 57, pp. 289-300.

Campbell, N. (2002) "Recording Techniques for capturing natural everyday speech," In Proc. *Language Resources and Evaluation Conference* (*LREC*-2002). Spain: Las Palmas.

Campbell, N. and Mokhtari, P. (2002) "DAT vs. Minidisc: Is MD recording quality good enough for prosodic analysis?," In Proc. *Acoustical Society of Japan Spring Mtg* 1-P-27, pp. 405-406.

Comprehensive R Archive Network: http://cran.r-project.org/

Mokhtari, P. and Campbell, N. (2003) "Automatic measurement of presses / breathy phonation at acoustic centres of reliability in continuous speech," In Special Issue on Speech Information Processing of the *IEICE Transactions on Information and Systems*, E-86-D, No. 3 (March), pp. 574-582.