イントネーション・アクセント　基本周波数パターン　音声分析・合成・変換

音声のイントネーションとアクセントを分析，合成，変換

本稿では，人が話す際の“自然な感じ”を保ったまま声の高低（基本周波数パターン）を調整し，さまざまな話し方の音声に変換することができる音声処理技術を紹介します．この技術により，非母語話者の音声を母語話者のような音声に変えたり，普通の音声をアナウンサーのようにメリハリのある音声に変えることが可能になります．また，入力したテキストをコンピュータに読ませるテキスト音声合成技術にも応用することができます．

かめおか　ひろかず
亀岡　弘和

NTTコミュニケーション科学基礎研究所

音声の「話し方」を変換

普段私たちは会話をするとき言葉を使って相手にメッセージを伝えますが，言葉とともに声の高低を効果的に使いながら声に表情をつけ，調子や意図，その人っぽさなどのさまざまな情報を相手に伝えています．本稿では，人が話す際の“自然な感じ”を保ったまま声の高低を調整し，さまざまな話し方の音声に変換することができる音声処理技術を紹介します．

本技術により，例えば非母語話者の音声を母語話者のような音声に変えたり，標準語の音声を地方の方言の音声に変えたり，普通の音声をアナウンサーのようにメリハリのある音声に変えたりできるようになります．役者や声優の演技における話し方を，撮り直すことなく後から調整する，というような使い方も可能です．また，入力したテキストをコンピュータに読ませるテキスト音声合成技術にも応用することができます．声の高低が一定調となる合成音声は機械っぽい印象を与えますが，人が話すような自然な話し方をする音声合成方法を実現することにも役立ちます．このほかにもプレゼンテーションや語学の指導・学習の支援にも役立つだろうと考えています．また，がんなどで喉頭を摘出した人による無喉頭音声を自然音声に変換する発声障がい者補助への応用も大学と連携しながら取り組んでいます．

基本周波数パターン（イントネーションとアクセント）

声の高低の時間変化を表す基本周波数パターンはイントネーションとアクセントからなります．

イントネーションとは文や句の全体に及ぶ範囲で緩やかに変化する基本周波数パターンのことで，話者の調子や意図，文の区切りや係り受けを表現するのに重要な役割を担います．例えば，「そうですか」という文は普通に読めば納得したような話し方になりますが，語尾が高くなると疑いを示したような話し方になります．また，「これじゃない」という文も読み方によっては全く異なる意図の話し方になります．

一方，アクセントとは各単語の中で急峻に変化する基本周波数パターンのことで，単語の意味や方言の違いに関係します．例えば「はし」という単語は「は」が高い場合と低い場合とでは意味が違いますし，「おいおい」という単語も先頭の「お」が高い場合と低い場合とで意味が違います．特に日本語の場合，単語アクセントと単語の意味の関係は方言によって異なるので，単語内の基本周波数パターンを変えて話せばたちまち異なった方言になります．また，イントネーションとアクセントの強度は，文や句や単語の強弱を表すいわゆるメリハリに相当します．これらの強度が大きい場合と小さい場合とでは話し方の印象は大きく変わり，メリハリをつけることで発話の中で注目すべき文や単語を相手に示すことができるようになります．以上のように，基本周波数パターンはさまざまな情報を持っており，言葉に勝るとも劣らないくらい音声コミュニケーションにおいて大きな役割を果たしています．

基本周波数パターン生成過程モデルのパラメータ推定

■甲状軟骨による基本周波数制御

基本周波数パターンは声帯を伸縮させる甲状軟骨という部位により制御されています．今から40年以上前，甲状軟骨による基本周波数パターンの制御メカニズムを模擬した物理モデルが提案されています(1),(2)（図1）．甲状軟骨の運動がどのような基本周波数パター

ンをもたらすかを明快に説明した方程式で，日本語を含む多言語の音声の基本周波数パターンを極めて良く表現できることが知られています．このモデルでは，甲状軟骨の２つの独立な運動（並進と回転）に伴う声帯の長さの変化の合計と対数基本周波数の変化が比例関係にあり，それぞれの運動がイントネーションとアクセントに関与しているという仮定がベースとなっています．このモデルに基づき，音声から甲状軟骨がどう動いたかを推定することができれば，その声にそっくりの基本周波数パターンを再現することができ，さらにその数値を変えてやれば甲状軟骨が異なる動きをした場合の基本周波数パターンの音声を再合成することができるようになります．ところが，この逆問題は一筋縄ではいかず長らく未解決問題とされていました．例えば７＋３を解くのは簡単でもX+Y=10となるXとYを一意に決められないのと同様で，イントネーションの成分とアクセントの成分から基本周波数パターンを得ることはできても基本周波数パターンのみからイントネーションの成分とアクセントの成分を一意に決めることはできないからです．しかし全く手がかりがないわけではありません．自然音声におけるイントネーションやアクセントのタイミングや強度には統計的な偏りがあります．

■基本周波数パターン生成過程の確率モデル化

筆者は基本周波数パターンをイントネーションとアクセントの成分に分解する問題を音源分離問題と同形ととらえ，統計的信号処理の考え方に基づくユニークなアプローチにより両成分を高精度かつ高速に推定することを可能にしました(3),(4)．音源分離とは文字どおり複数の音の信号の波形が重ね合わさった混合信号の波形から個々の音の信号波形を分離することを指します．X+Yの例のように一度混ざり合ってしまったものを元通りに分離するのは一般的には難しいため，この問題も一見すると解くのが不可能のように思えるかもしれませんが，音の信号波形に関する統計的偏りや統計的性質を活かした分離方法が効果的であることが知られています．そこで音源分離手法の考え方をヒントにして考案したのが今回の方法です（**図２**）．

提案手法によるイントネーションとアクセントの成分の推定例と，推定した成分を変えて再合成した基本周波数パターンおよび音声の例を**図３**(a)に示します．図３(b)ではアクセントの強さを大きくすることでメリハリのある音声に，図３(c)ではアクセントのタイミングを変えることで異なる方言の音声にすることができています．いずれの変換音声もあたかも人が話した自然な感じを保っているところがポイントです．これは，声の高低の制御メカニズムの物理モデルに準拠し，人が発声し得ない基本周波数パターンには決

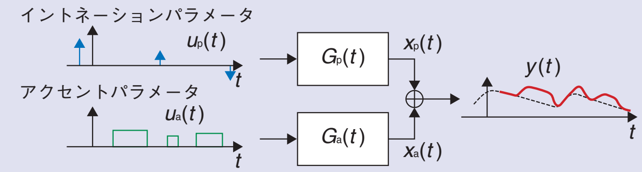


図１　基本周波数パターン生成過程の物理モデル

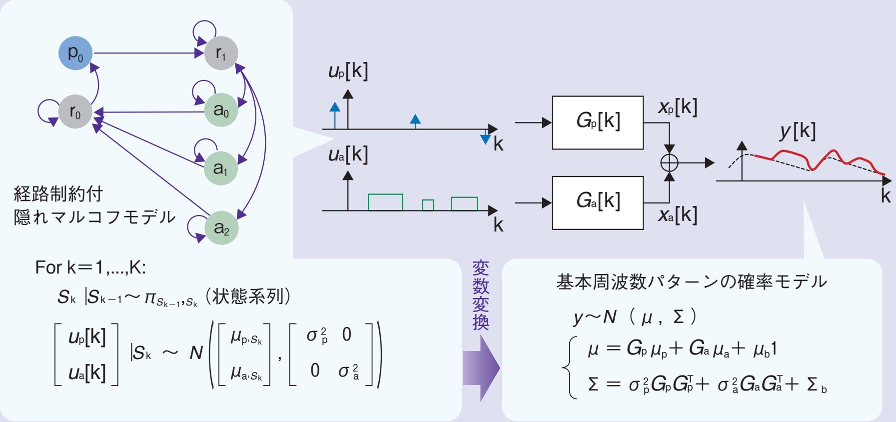
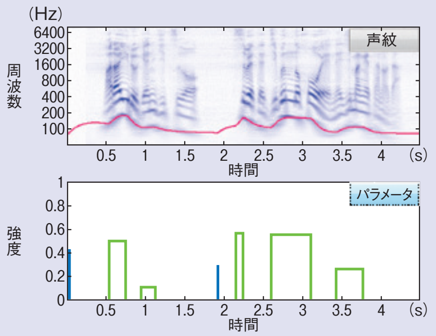
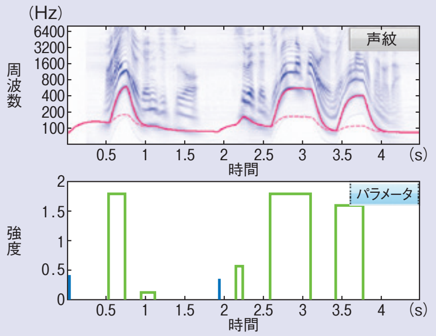


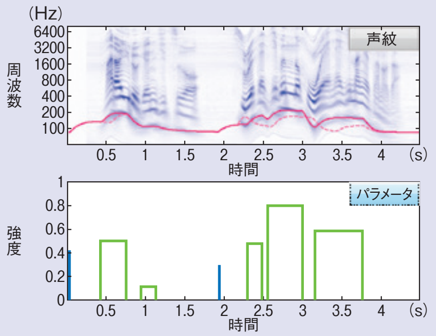
図２　基本周波数パターン生成過程の確率モデル化



(a)　イントネーションとアクセントの成分の推定例



(b)　メリハリのある音声への変換例



(c)　異なる方言の音声への変換例

図3　基本周波数パターンの分析と変換

してならないような範囲で自由に基本周波数パターンを変換することができるからです．

今後の展開

1960年代に登場した線形予測符号化（LPC）技術は近代式の音声分析合成系を誕生させ，携帯電話という新たなコミュニケーション手段と統計的音声情報処理という研究パラダイムをもたらしました．LPCは，言語情報（文字に書き起こせる情報）に関係する音韻的特徴の分析合成系を実現するものであったのに対し，本研究は，非言語情報に関係する韻律的特徴の分析合成系を実現するものです．LPCでは声道の物理モデルによる音声信号生成過程を確率モデル化し，統計的手法により声道パラメータを推定する枠組を与えたのに対し，本研究では甲状軟骨の物理モデルによる基本周波数パターン生成過程を確率モデル化し，統計的手法により韻律パラメータを推定する枠組を確立しました．LPCが音声通信や音声情報処理の分野に大きな発展をもたらしたように，本技術も新たな分野を開拓していける可能性を秘めています．音声の韻律に着目した今までにないさまざまなサービスを生み出せるよう今後も研究を続けていきたいと思っています．

■参考文献

(1) H. Fujisaki and S. Nagashima：“A model for synthesis of pitch contours of connected speech,” Annual Report of Engineering Research Institute, University of Tokyo, Vol.28, pp. 53-60, 1969.

(2) H. Fujisaki：“A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour,” Vocal Physiology: Voice Production, Mechanisms and Functions, pp.347-355, New York, U.S.A., 1988.

(3) H. Kameoka, J. Le Roux, and Y. Ohishi：“A Statistical Model of Speech F_0 Contours,” Proc. of SAPA 2010, pp.43-48, Makuhari, Japan, Sept. 2010.

(4) H. Kameoka, K. Yoshizato, T. Ishihara, K. Kadowaki, Y. Ohishi, and K. Kashino：“Generative Modeling of Voice Fundamental Frequency Contours,” IEEE/ACM Trans. Audio, Speech and Language Processing, Vol.23, No.6, pp.1042-1053, June 2015.

亀岡　弘和

最先端の音声合成・変換システムでは明瞭性の高い音声が生成できるようになってきているものの，人間の肉声と比べるとぎこちなく感じられます．本研究はそのぎこちなさを解消する鍵となる可能性があり，より人間らしく，自然で，表情豊かな音声を生成できる音声合成・変換技術につなげていきたいと考えています．

◆問い合わせ先

NTTコミュニケーション科学基礎研究所
メディア情報研究部
TEL　046-240-3645
FAX　046-240-4708
E-mail　kameoka.hirokazu@lab.ntt.co.jp