発話リズムを抽出・制御する音声信号処理

日本語母語話者が英語を用いた音声コミュニケーションを円滑に行うためには，発話リズムが重要な役割を果たすと考えられます．本稿では，音声信号から発話リズムを抽出・制御するための技術である非負値時空間分解法について紹介します．

廣谷（ひろや）　定男（さだお）

NTTコミュニケーション科学基礎研究所

外国語を用いた音声コミュニケーション

外国語を用いての音声コミュニケーションの難しさは，しゃべること，聞くことの両方において存在します．とりわけ多くの日本語母語話者は，英語母語話者の話が聞き取れない，自分のしゃべった言葉が英語母語話者に伝わらず聞き返される，などの問題を抱えています．日本語と英語の違いは大きく分けて，母音の数，RとLの区別，アクセントやイントネーションなどに代表される「発音」と，文章発声にみられる大局的な時間構造としての「リズム」の2つがあります*．「Coffee, please.」や「Where is the toilet?」などの単語や短い文章を伝える場合には，正しい発音が不可欠です．ところが，日本の英語教育では発音を重視する傾向があるため，正しいリズムが身につきにくく，長い文章が英語母語話者に伝わりにくいという問題が発生する場合があります．つまり，長い文章の場合には，正しい発音よりも，英語独特の正しいリズムのほうが重要であるということです．

NTTコミュニケーション科学基礎研究所では，「発音に多少の間違いがあってもリズムが正しいと理解されやすい」ことに着目し，日本語母語話者の音声信号から発話リズムを抽出し，英語母語話者らしい発話リズムに変換する技術の研究を行っています（**図1**）．

発話リズムとは

リズムとは大局的な時間構造のことを指します．音声言語の場合，強勢拍リズムと音節拍リズムの2つのリズムがあり，英語は強勢拍リズム，日本語は音節拍リズムといわれ，それぞれの言語で異なっています．強勢拍リズムとは，強勢（ストレス）と強勢が等間隔で出現するリズムで，音節拍リズム

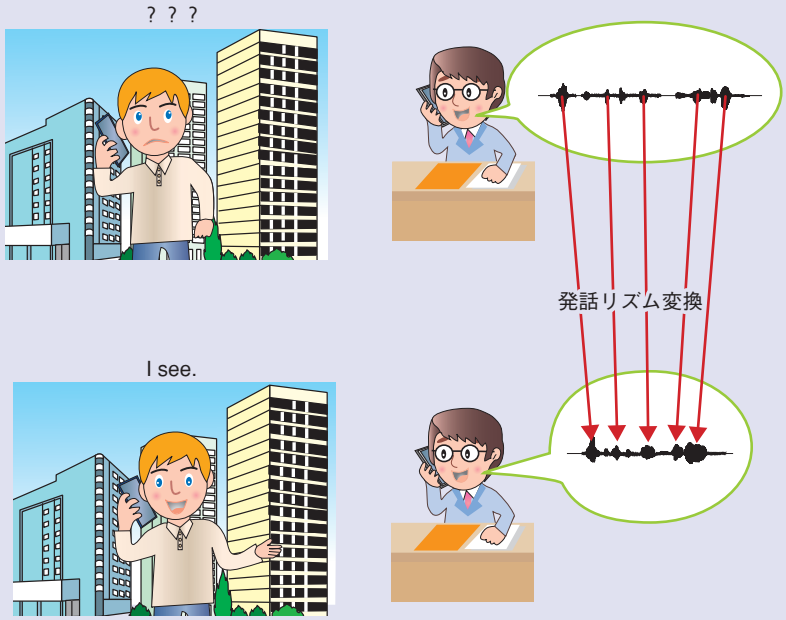


図1　発話リズム変換

＊アクセントやイントネーションは発音に含まれないことがありますが，本稿では，「リズム」以外の特徴として「発音」を定義します．

とは，同じ長さの音節（シラブル）が繰り返されるリズムのことです（**図2**）．音節拍リズムは，強勢拍リズムと比べて単調な時間構造を持っています．

さて，人間は発話以外にも，楽器を演奏する，手を叩く，歩行するなど，さまざまな場面でリズムをとります．つまり，リズムは音の時間構造からだけでなく，音以外の時間構造からも定義できるということです．音声は人間の口唇・舌・軟口蓋などの発話器官の運動（発話運動）により生成されることから，我々はリズムを音の時間構造として定義するのではなく，筋肉によってもたらされる運動の時間構造として定義することとしました．本技術で用いられる発話リズムとは，人間の発話運動の時間構造，つまり言語に依存しない人間固有の口唇や舌の動きの時間変化パタンということになります．ところで，舌などの外からは見ることのできない発話運動をリズムの定義に用いることに疑問を感じる方もいらっしゃるかもしれませんが，我々はこれまでに，磁気センサシステムやMRI（核磁気共鳴画像法）[1]による発話運動の計測や音声信号からの発話運動の推定法[2]の研究を行っており，発話運動の時間構造をリズムとして用いることの有効性を示してきました（**図3**）．例えば，磁気センサシステムを用いて計測した正中矢状面での上下唇・下顎・舌上3点の計6点に貼付したセンサの発話時の時系列データの分析を行ったところ，発話運動は，音素に対応した発話器官の位置情報と，隣り合う音素間での滑らかな補間（時間構造に対応）により表現可能であることを見出しました[3]．この発見は，調音音韻論[4]と呼ばれる学問や最近の発話の脳研究における発見[5]と関連しており，また発話運動が滑らかに変化することと対応しています．

このように人間の音声生成において重要な発話運動をリズムの定義として用いることにより，発話リズムの抽出・制御が容易になると期待されます．

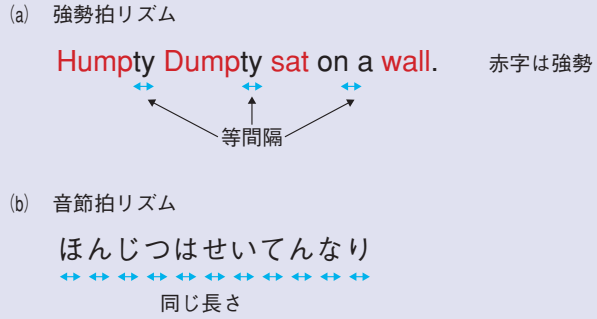
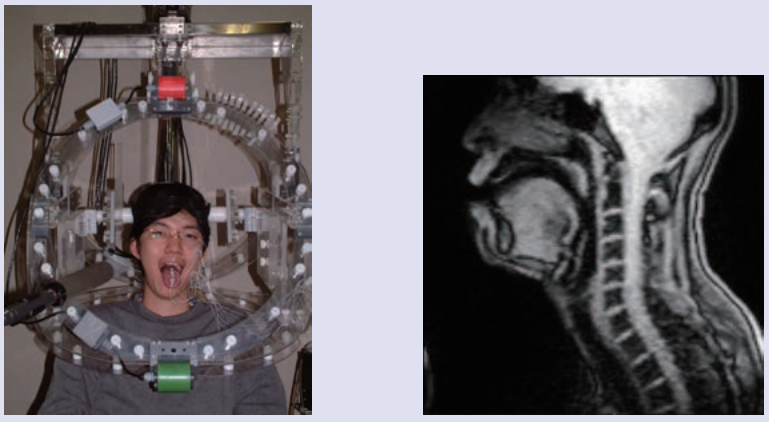


図2　強勢拍リズムと音節拍リズム



(a) 磁気センサシステム　(b) MRI

（甲南大学との共同研究成果）

図3　発話運動の計測

非負値時空間分解法

発話運動により生成される音声信号には，周波数情報と時間情報の両方が含まれています．近年，音響信号処理分野で注目を集めているNMF（Non-negative Matrix Factorization：非負値行列因子分解）[6]では，音信号からの周波数情報と時間情報の分解を可能としますが，発話運動を考慮した時間情報への制約が導入されていないため，時間情報の速度曲線が人間の運動の特徴であるベル型（運動開始から徐々に速度が上がっていき，運動終了に近づくにつれ徐々に速度が下がっていく）を示さない，隣り合う音素以外からも時間情報が影響を受けてしまうなどの問題がありました．

そこで我々は，NMFにはない発話運動に特化した制約を考慮しながら，高い精度で発話リズム（時間情報）を抽出するNTD（Non-negative Temporal Decomposition：非負値時空間分解法）[3]という技術を考案しました．

NTDは，発話器官の位置情報に対

応する特徴パラメータである声道スペクトル（例えば，線スペクトル周波数）を，周波数情報Aと時間情報Fに分解します（**図4**）．時間情報Fの中には，発話リズムを音素単位で表現するモデルが導入されており，また隣り合う音素のみが時間情報に影響を及ぼすことが考慮されています．NTDの処理手順は，まず音声信号から声道スペクトルを抽出します．次に，NMFで用いられている乗算型更新式の考えを応用したアルゴリズムにより，声道スペクトルの値と推定値との間の二乗誤差を最小にする時間関数（正規化された［0，1］の範囲の値）を求めます．この際，時間関数の値が0以上となる制約や，時間関数が滑らかに変化する（音素に対応した単峰性の時間関数）という制約は，非負値に収束する乗算型更新式の仕組みを利用することで実現されます[3]．NMFでも滑らかな時間関数を得るための方法（ペナルティ関数）が提案されていますが，ペナルティ関数を用いたNMFでは時間関数が直線になりやすいのに対して，NTDではベル型の速度曲線を持つ時間関数が得られやすいという特徴があります．

また時間関数の抽出には，音素の中心時点を表す「音素時点」の情報も必要となりますが，最適な音素時点を決定するために，DP（Dynamic Programming：動的計画法）をNTDと組み合わせて用いています[3]．したがって，NTDは，DPかつ発話運動制約付きNMFとみなすことができます．NTDで発話リズムを抽出する際の入力は音声信号のみですが，アルゴリズムの中に発話運動に関する制約が考慮されているため，音声信号から発話運動のリズムの抽出が可能となっています．そのため，NTDは，音声信号からの発話運動の推定法[2]への応用が期待されます．

発話リズムの変換

NTDを用いて日本語母語話者が発声した英語の発話リズムを英語母語話者らしい発話リズムに変換する手法について述べます（**図5**）．まず，日本語母語話者と英語母語話者の両方がそれぞれ同一文章（「Rice is often served in round bowls.」）を発声します．次に，日本語母語話者の声道スペクトルからNTDを用いて周波数情報A_Jと時間情報F_Jを抽出し，英語母語話者の声道スペクトルから周波数情報A_Eと時間情報F_Eを抽出します．最後に，日本語母語話者の時間情報をF_Eに置き換え，A_JとF_Eを掛け合わせることで，発声が日本語母語話者でリズムが英語母語話者となる声道スペクトルを生成し，音源信号をたたみ込むことで音声を合成します．発話リズムを変換した結果，日本語母語話者の発話時間が単に短くなっただけのように見えるかもしれませんが，例えば，「bowls」の部分（図5の赤い四角で囲った部分）の時間情報を日本語母語話者と英語母語話者で比較すると，日本語母語話者の場合は発話リズムが一定なのに対して，英語母語話者の場合は「ow」の部分の時間長が前後と比較して伸びていることが分かります．つまり，日本語母語話者の発話リズムが，英語母語話者らしい発話リズムに変換されることにより，メリハリのある音声になるのです．英語母語話者らしい発話リズムに変換した音声を，英語母語話者に聞かせたところ，聞き取りやすいという感想が得られました．

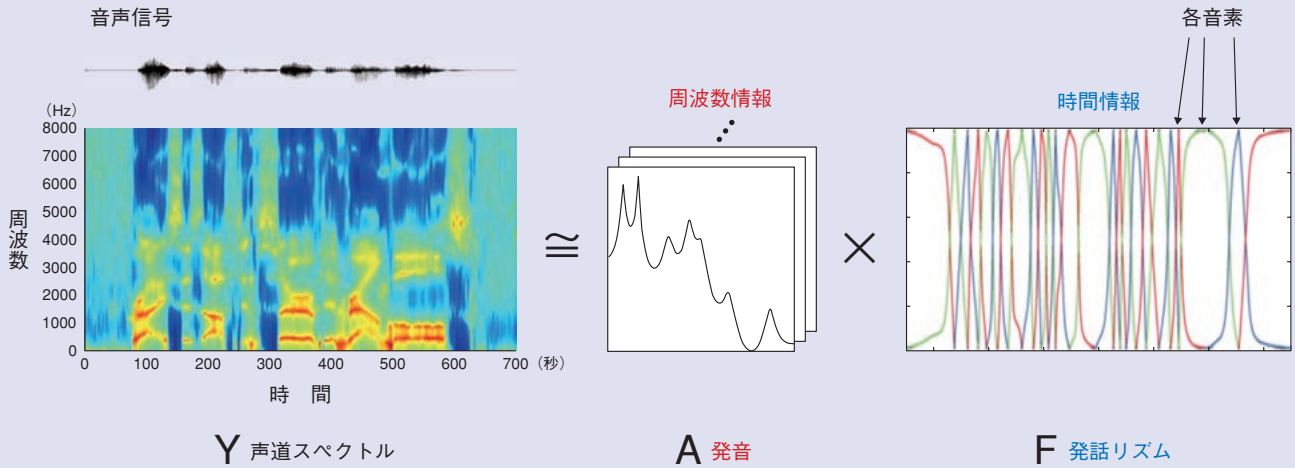


図4　非負値時空間分解法

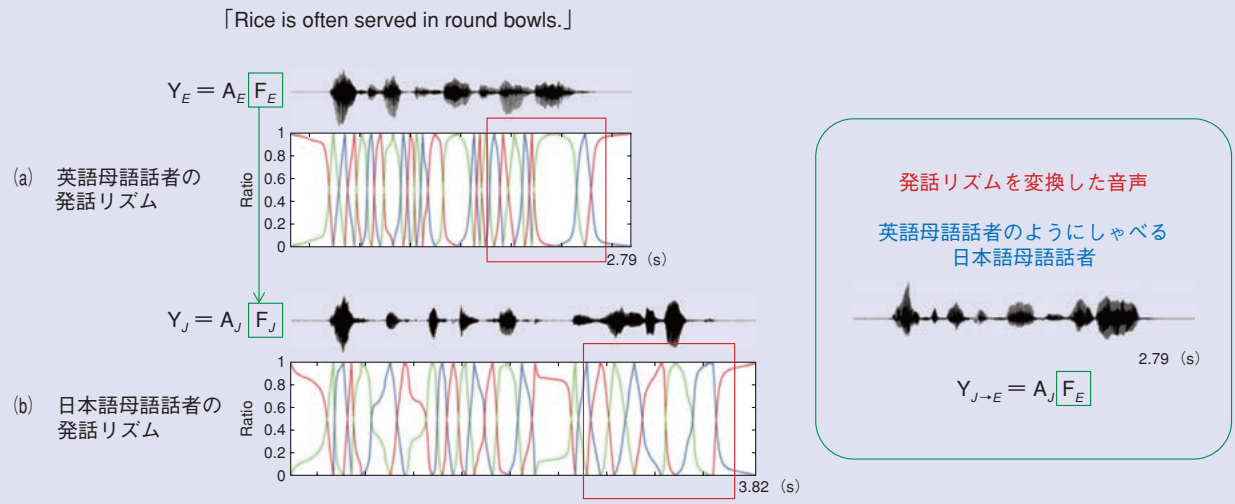


図5　発話リズムの変換

今後の展望

日本語母語話者の英語でのコミュニケーションをアシストする技術としては，ほかにも日本語母語話者がしゃべった日本語を英語に翻訳し，（他人の声で）英語を発声する，あるいは英語母語話者がしゃべった英語を日本語に翻訳し，日本語を発声するという音声翻訳技術があります．しかし，日本では国際化社会での人材育成のために，2011年度から小学校５，６年生の英語の必修化が始まっており，音声翻訳技術に頼らない，自分の声による英語でのコミュニケーションの機会が今後増えると考えられます．

本技術を使えば，声質はそのままで発話リズムのみを変換することができるため，発話リズムの訓練のための効率的な学習法に応用できるのではないかと考えられます．また，携帯端末を通じて，日本語母語話者が発声した英語を英語母語話者の発話リズムに変換する，さらにはTV会議やプレゼンテーションでリアルタイムに発話リズムを変換するなどの応用が考えられます．ただし，現時点では，あらかじめ英語母語話者に同一文章を読み上げさせる必要があるため，任意の発声に対してリズムを変換する技術の構築のためには言語間での発話リズムの違いのモデル化が必要となります．

また本技術は，外国語の聞き取りにおける発話リズムの重要性の検証，およびその脳内メカニズムの解明のために大きく貢献できると考えられます．将来，日本語母語話者の英語の発話リズムを変換することで，英語が英語母語話者に伝わりやすくなり，英語でのコミュニケーションの苦労が軽減されるかもしれません．

■参考文献

(1) S. Hiroya and T. Kitamura: “Generation of a vocal-tract MRI movie based on sparse sampling,” Proc. of ISSP, pp.1-8, Montreal, Canada, June 2011.
(2) S. Hiroya and M. Honda: “Estimation of articulatory movements from speech acoustics using an HMM-based speech production model,” IEEE Trans. SAP, Vol.12, No.2, pp.175-185, 2004.
(3) S. Hiroya: “Non-negative temporal decomposition of speech parameters by multiplicative update rules,” IEEE Trans. ASLP, Vol.21, No.10, pp.2108-2117, 2013.
(4) C.P. Browman and L. Goldstein: “Articulatory phonology: An overview,” Phonetica, Vol.49, No.3-4, pp.155-180, 1992.
(5) K.E. Bouchard, N. Mesgarani, K. Johnson, and E.F. Chang: “Functional organization of human sensorimotor cortex for speech articulation,” Nature, Vol.495, No.7441, pp.327-332, 2013.
(6) D.D.Lee and H.S.Seung: “Learning the parts of objects by nonnegative matrix factorization,” Nature, Vol.401, No.6755, pp.788-791, 1999.

廣谷　定男

人間が言葉をしゃべったり，聞いたりする脳内メカニズムを解明し，科学的根拠に基づいた音声コミュニケーションを支援する技術を考案し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
人間情報研究部　感覚運動研究グループ
TEL　046-240-3578
FAX　046-240-4721
E-mail　hiroya.sadao@lab.ntt.co.jp