Vol.2017-MUS-115 No.38
2017/6/18

無矛盾位相復元を用いた話者間の韻律変換

光本大記[†1]　濱田康弘[†1]　小野順貴[†2]　嵯峨山茂樹[†1]

概要：原音声の声質を保ちつつ言語的な特徴のみをを変化させる音声合成技術は，言語学習支援において効果を上げることが期待されている．我々はこの要件を満たす音声合成手法確立のため，ノンパラメトリックなアプローチを試みてきた．本稿では，操作する特徴量をまず発話リズム・イントネーションに絞り，無矛盾位相復元を用いて音声合成を行う手法について述べる．

キーワード：無矛盾位相復元，韻律変換

Phase reconstruction for cross-speaker prosody transformation

DAIKI MITSUMOTO[†1]　YASUHIRO HAMDA[†1]　NOBUTAKA ONO[†2]
SHIGEKI SAGAYAMA[†1]

1. はじめに

非日本語母語話者にとっては日本語の特殊拍の知覚・発声が困難であり，音素の脱落や挿入が起こることが知られている[1]．逆に，日本語を母語とする英語学習者の多くにとって，/l/と/r/の知覚・発声は難しい．また，日本語の 5 種類の母音の中間に位置する発声も習得し難い．

このような困難性は言語間の差異によるものであり，言語学習の場では差を緩和するために様々な工夫がなされている．例えば母語話者の発話に続けて学習者が復唱する，母語話者とのオーラルコミュニケーションの機会を増やすなどが挙げられる．これらの方法により学習者は母語話者の発音と自らの発音との差異を発見し，より正確な発音に近づくよう修正することができる．これを発展させ，母語話者の音声そのものではなく，声質は学習者本人のものであるが，発音・発話リズム・イントネーションは母語話者の特徴を持つ合成音声を学習者に聴取させることが，新しい学習法として効果を上げることが期待されている．

発話リズムを置換する関連研究として，非負値時空間分解を用いた手法がある[2]．この手法では非負値時空間分解を用いてLSPパラメータの系列を周波数情報と時間情報に分解し，入力音声と参照音声とで組み合わせを入れ替えることによるパラメトリックな発話リズム変換を実現している．

ノンパラメトリックな時間伸縮・ピッチ変換の手法として，無矛盾位相復元[3]を用いる手法がある[4]．

我々は声質を変えずに発音・韻律を置換する音声合成手法を確立することを目的とし，手始めに入力声質を保ったまま発話リズムのみを置換する手法を提案してきた[5]．本研究では，声質を変化させずに発話リズム・イントネーション双方を置換する手法を述べる．

2. 無矛盾位相復元による韻律変換

本手法では，参照音声との時間対応関係に従って発話リズム変換・イントネーション変換がなされたスペクトログラムを生成し，無矛盾位相復元を行うことで目的の音声を出力する(図 1)

2.1 発話リズム変換

我々はこれまでに，無矛盾位相復元によって入力音声に参照音声の発話リズムを与える手法を提案してきた．これは，DP マッチングにより得られた時間対応を用いて発話リズム変換された音声を表すスペクトログラムを生成し，無矛盾位相復元によって音声波形を合成する手法である(図 1).

音声の発話リズムをスペクトログラムの時間構造と考えると，発話リズム変換はスペクトログラムの時間伸縮により行われる．参照音声の発話リズムを入力音声に与えええることは，音声同士をもっともよく時間整合させる時間伸縮率を求める非線形伸縮問題であり，これは DP マッチングにより解決される．

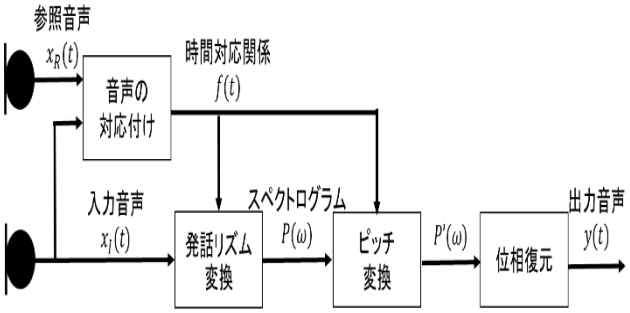

図 1 韻律変換の流れ図

†1 明治大学
Meiji University
†2 国立情報学研究所 / 総合研究大学院大学
National Institute of Informatics
/ SOKENDAI

Vol.2017-MUS-115 No.38
2017/6/18

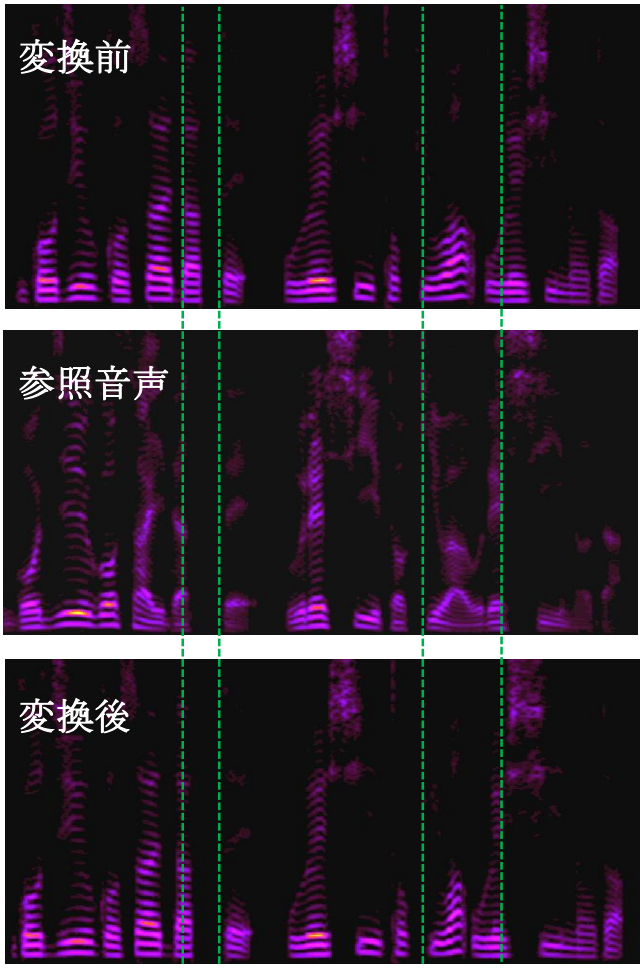

図 3 発話リズム変換前後のスペクトログラム．変換前と参照音声・変換後の差がよく表れている箇所に線を引いた

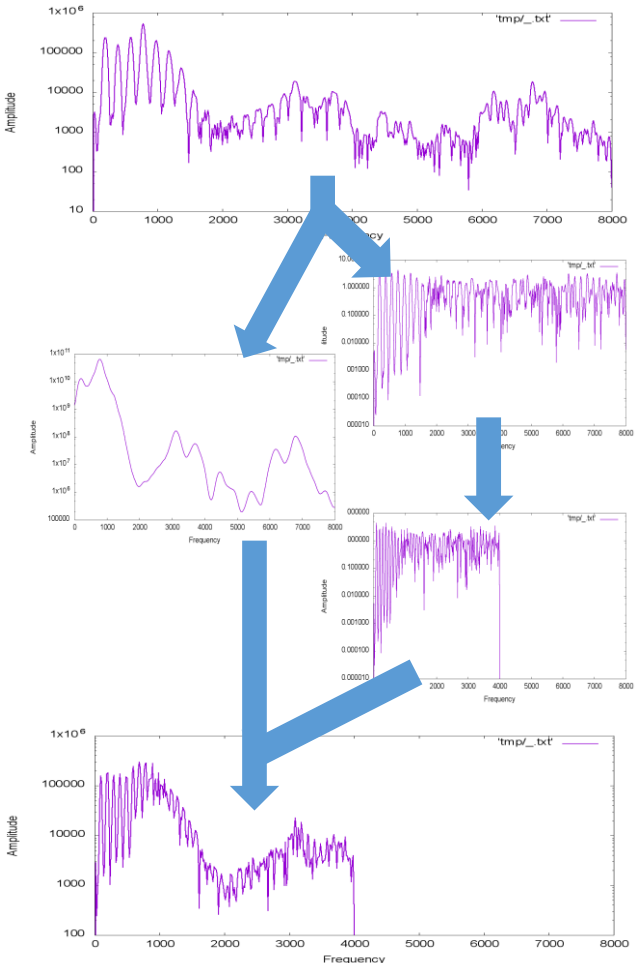

図 3 スペクトルの調波構造に，F_0 を小さくする伸縮処理が施される様子

スペクトログラムの時間伸縮は，分析時と合成時でフレームシフトを別の値に設定することで行われる．例えば，4ms のフレームシフトで分析すると，得られるフレーム数は 8ms で分析した時の 2 倍になる．そのため，フレームシフト 8ms で合成すると信号長が 2 倍になる．これを局所的に利用することで，スペクトログラムを各時刻で自在に伸縮させることができる．

求められた時間伸縮率に基づきこの伸縮法を用いてスペクトログラムを伸縮させることで，発話リズム変換のなされた音声を表すスペクトログラムが得られる．得られたスペクトログラムに無矛盾位相復元を施すことにより音声波形を得ることができる．

2.2 イントネーション変換

イントネーションは F_0 の時間変動であるとすると，参照音声の F_0 の上下動の特徴を再現するよう，スペクトログラムの周波数伸縮が行われれば良いと言える．ただし，音声のスペクトルは声帯由来の調波構造のみならず声道由来の包絡構造から構成されているため，スペクトルの単純な伸縮は個人性が反映される声道特性をも変化させてしまう．したがって，F_0 変換は調波構造と包絡構造の分離がなされた後，前者のみの伸縮処理によって行われなければならない．また声の高さは人により異なるので，変換前後で平均基本周波数が等しくなるよう各時刻のスペクトル伸縮率が決定されなければならない．　スペクトルの分離はラグ窓法により行われる[6]．フーリエ変換により，パワースペクトルと自己相関関数は総合に変換され，自己相関関数の低次の成分はパワースペクトルの包絡構造に相当する．よって，高次の成分を減衰させるラグ窓をかけられた自己相関関数のフーリエ変換により，スペクトルの包絡が得られる．スペクトルは包絡構造と調波構造の積なので，得られた包絡構造でスペクトルを割ることにより調波構造が得られる．この調波構造が伸縮された後再び包絡構造と合成されることにより，スペクトルの F_0 変換がなされる．また，ラグ窓法による調波構造の対数パワースペクトルを逆フーリエ変換することで得られるケプストラムは F_0 に強いピークが現れるため，ピーク位置のケフレンシーから伸縮率計算のための F_0 が高精度で推定される．

各時刻のスペクトルの伸縮率は，参照音声のイントネーションを再現するように決定されなければならない．参照音声と出力音声の時刻tでの基本周波数を$F_{0_R}(t), F_{0_O}(t)$，平

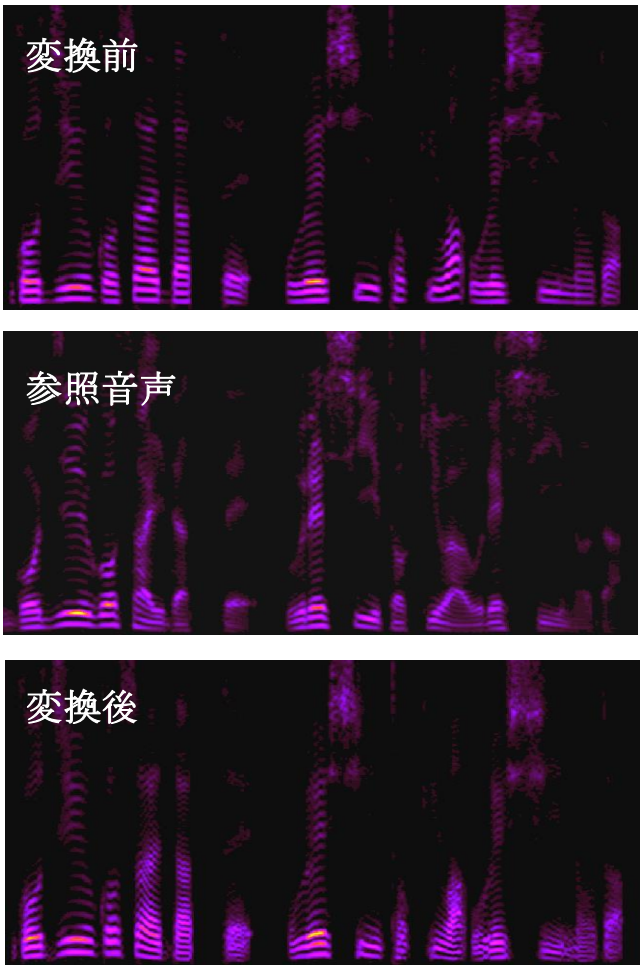

図 4 イントネーション変換前後のスペクトログラム

均基本周波数を$\overline{F_{0_R}}, \overline{F_{0_O}}$としたとき，以下の式を満たすように，時刻tでのスペクトルの伸縮率$\mathrm{k_t}$を決定すれば，参照音声のイントネーションの特徴を与えられる．

$$\frac{F_{0_O}(t)}{\overline{F_{0_O}}} = \frac{F_{0_R}(t)}{\overline{F_{0_R}}} \tag{1}$$

また，入力音声と出力音声の平均基本周波数を等しくするために，以下の条件を設定する．

$$\overline{F_{0_O}} = \overline{F_{0_I}} \tag{2}$$

$\mathrm{F}_{0_O}(t)$は調波構造の伸縮率と入力音声の基本周波数$\mathrm{F}_{0_I}(t)$を用いると

$$F_{0_O}(t) = \mathrm{k_t}\, F_{0_I}(t) \tag{3}$$

(1)〜(3)より，

$$\mathrm{k_t} = \frac{\overline{F_{0_I}} F_{0_R}(t)}{\overline{F_{0_R}} F_{0_I}(t)} \tag{4}$$

以上のようにして求まった各時刻の伸縮率$\mathrm{k_t}$に従って調波構造を伸縮することにより，入力音声の平均基本周波数を保ちつつ参照音声のイントネーションを再現するスペクトログラムが得られる．このスペクトログラムに無矛盾位相ｂ復元が施されることにより，所望の音声が出力される．

3. 実験

本手法の発話リズム・イントネーション変換の性能を調べるため，英語母語話者の日本語音声の韻律を，日本語母語話者のものに置き換える例で実験を行った．

3.1 実験条件

入力音声には留学生による読み上げ日本語音声データベース(UME-JRF)より男性英語母語話者 1 名の日本語音声 50 音を使用し，参照音声には ATR デジタル音声データベースセット B から，入力音声と同一文章の音声を男性話者 1 名分使用した．また，言いよどみなどによって入力音声に 300ms 以上の無音区間がある場合はこれを削除した．スペクトルの分解・ピッチ抽出にはラグ窓法を用いた．調波構造が縮小されることにより生じる空白部分には 0 埋めを行った．

サンプリング周波数は 16000Hz であり，フレーム長は 32ms，合成時フレームシフトは 5ms，位相復元反復回数は 1024 回に設定し，音声の分析及び合成を行った．DP マッチングの特徴量は 12 次元の LPC ケプストラム係数とし，傾斜制限は$\frac{1}{2}$〜2に設定した．被験者は日本語を母語とする大学生・大学院生 6 名であった．

被験者には，原音声と合成音声 98 音をランダムに聞かせ，各音声について，日本語としての発話リズム・イントネーションの自然性を，「非常に悪い」〜「非常に良い」の 5 段階で評価してもらった．

3.2 結果

発話リズム・イントネーションの評価に二要因分散分析を行った．独立変数は変換前/後(2 水準)と使用音声(49 水準)であり，従属変数は韻律の自然性の評価値とした．結果，変換前後で有意差があった($F(1,5) = 6.85, p < 0.05$)．また，使用音声との交互作用が有意であった($F(48,240) = 2.93, p < 0.01$)．使用音声ごとに変換前後の比較をしたところ，3 音声で$p < 0.01$，9 音声で$p < 0.05$，8 音声で$p < 0.10$，29 音声で有意差なしであった(表 1)．

3.3 考察

分散分析の結果は，本手法による合成音声は原音声と比べて，概ね自然な韻律を持ち，音声次第で特に効果を発揮することを示した．

有意差のある音声は文章に特殊拍が含まれないか，ある程度発音されているものが多かった．一方有意差の出ない音声については特殊拍が脱落して発音されたり，言いよどみある音が多く含まれた．特殊拍の脱落や言いよどみの多く発生するような場合には，入力音声と参照音声の時間整合を図るためには局所的に大きい伸縮がなされなければならないが，DP マッチングに$\frac{1}{2}$〜2の傾斜制限が設けてあることにより，十分な伸縮がなされず，韻律の自然性が向上しなかったと考えられる．このような場合には，特徴量に応じて局所的に傾斜制限を緩和する機構を設けることで，結

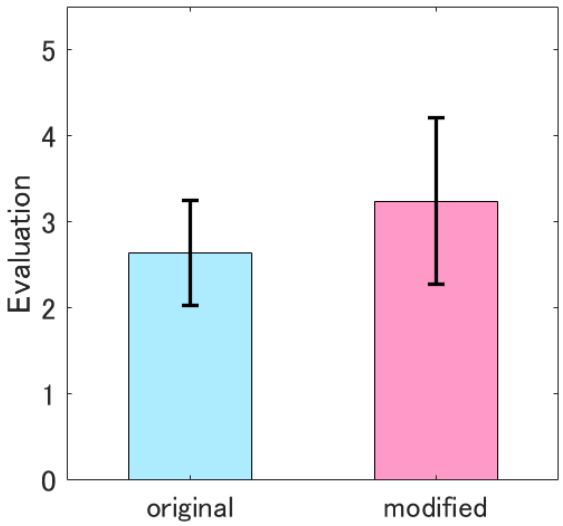

図 5 原音声と合成音声の評価値．サンプル数 49．エラーバーは標準偏差を表す

果の向上が見込まれる．

被験者へのアンケートでは，韻律自然性の評価は難しかったとの感想がいくつか見受けられた．その要因として，合成された音声のノイズが顕著になることがあり，発話に集中しにくかったことや，聴取を重ねるにつれて非日本語母語話者の発話に慣れてしまい評価が不安定になったことが挙げられる．今後は，ノイズの発生しにくいイントネーション変換法を模索すると共に，より厳密な実験を行う必要がある．

4. おわりに

本稿では，原音声の声道特性を保ったまま話者間で韻律を変換する手法を提案した．入力音声のスペクトログラムに対して,参照音声との時間対応付けに基づいて時間伸縮・調波構造の周波数伸縮を行うことで韻律変換を施し，無矛盾位相復元を用いて音声波形を合成した．本手法を用いて合成された非日本語母語話者の日本語音声に対して韻律の自然性についての主観評価実験を行った結果，原音声との間に有意差が現れた．今後は，特殊拍の脱落などにより局所的に大きな伸縮が必要な場合への対応や，イントネーション変換時の音質劣化の改善を行い，より高品質な韻律変換手法を検討していく予定である．

参考文献

[1] 戸田貴子，“外国人学習者の日本語特殊拍の習得,” 日本音声学会，音声研究 第 7 巻 2 号, pp. 70-83, 2003.
[2] 廣谷定男，“非負値時空間分解を用いた発話リズム変換の検討,” 日本音響学会講演論文集 pp. 425-426, Sep. 2014.
[3] D. W. Griffin and J. S. Lim, “Signal Estimation from Modified Short-Time Fourier Transform,” IEEE Trans. Acoustics, Speech, and Signal Processing, vol. 32, no. 2, pp. 236-243, Apr. 1984.
[4] 水野優，小野順貴，西本卓也，嵯峨山茂樹，“パワースペクトログラムの伸縮に基づく多重音信号の再生速度と音高の実時間制御,” 信学技報，聴覚研究会資料, vol. 39, no. 6, pp. 447-452, 2009.
[5] 光本大記，濱田康弘，小野順貴，嵯峨山茂樹，“無矛盾位相復元による発話リズムの話者間変換,” 日本音響学会講演論文集, Sep. 2016.
[6] 嵯峨山茂樹，古井貞熙，“ラグ窓を用いたピッチ抽出の一方法,” 信学会総合大会, 5-263, 1978.