社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－MUS－74 (8)
2008－SLP－70 (8)
2008／2／8

テンポの変化による影響を考慮した歌唱音声合成に関する検討

川添 正人，坂野 秀樹，板倉 文忠
名城大学大学院理工学研究科 〒468-5802 愛知県名古屋市天白区塩釜口 1-501
E-mail: m0632008@ccmailg.meijo-u.ac.jp

子音を含む歌唱音声合成の品質を向上させる方法について検討している．実際の歌唱音声を収録して観察した結果，楽曲のテンポが速くなるにつれ，スペクトルの変動が小さくなる現象があることが明らかとなった．これは，声道の形状が急に変化できないことに起因すると考えられる．既存の歌唱音声合成手法では，このことがあまり考慮されておらず，これが合成音が不自然となる原因の 1 つとなっている．そこで，本稿では，発声タイミングから推定したスペクトル変動の減少量を表す尺度を用い，線形予測分析により推定された声道断面積関数に対して時間軸に関する平滑化を行うことで，実際の音声を模擬する方法を提案する．これにより，実際の音声のテンポによるスペクトル変動特性の変化を良く近似できることが分かった．

Singing Voice Synthesis Taking Account of Tempo Effect

Masato KAWAZOE, Hideki BANNO, Fumitada ITAKURA
Graduate School of Science and Technology, Meijo University Siogamaguchi 1-501, Tempaku-ku, Nagoya-shi, Aichi 468-5802 Japan
E-mail: m0632008@ccmailg.meijo-u.ac.jp

This paper describes a method to improve quality of singing voice synthesis system including consonants. Analysis of recorded signal of real singing voice including consonants indicated that a faster tempo reduces the spectral variance of singing voice, because the vocal tract shape can only change at a limited speed. Since conventional synthesis systems do not take the spectral variance reduction into account, degradation of the synthetic singing voice is caused. We have developed a new measure representing the spectral variance reduction, that is estimated from note-on events of the MIDI signal. Spectral smoothing based on this measure in the domain of the vocal tract area function with respect to time is then applied to the conventional synthesis system. It is found that a generated singing voice by the proposed method successfully simulates the spectral variance reduction of real singing voice.

1 はじめに

現在，音声合成技術の進歩によりそれまで困難とされてきた肉声らしい歌唱音声の合成が可能となり，様々な歌唱音声合成システムが開

発され，実用化に至っている．しかしながら，歌唱音声の合成は，楽器とは異なり，調音様式の複雑さや，歌詞情報の添加に伴う韻律の制御，唱法の制御など，考慮すべき要素が膨大になるため，改善が必要な点は依然として多い．

本研究では，より高品質な歌唱音声合成システムの実現を目的とした歌唱音声合成手法について検討している．特に，速く歌わせた部分において，音声を切り貼りしたように聞こえる，あるいは，実際より滑舌が良すぎるように聞こえるなど，現状の歌唱音声合成システムの合成音が子音部において不自然になる場合があることに着目し，楽曲のテンポの変化による影響を考慮した歌唱音声合成手法の確立を目指す．

2　テンポの違いによる歌唱音声への影響

音声の発声においては，速く話す場合，声道の形が急に変化できないために，完全な調音が行われない場合がある．同様に歌唱音声においてもテンポの変化に伴い声道形状の変動特性が変化することが予想される．また，声道形状の変化より声道の周波数特性も変化する可能性もあることから，ここではまず，テンポの異なる実際の歌唱音声を収録し，スペクトルの変動特性を調査する．

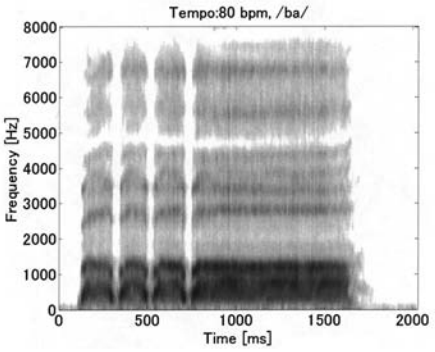
2.1 収録条件

歌い手にヘッドホンを装着し，収録用に作成した midi 音を再生する．歌い手は，それに合わせ，特定の音節のみで構成されたフレーズをマイクロホンに向かって発声する．このとき，midi 音のテンポを 80 bpm, 90 bpm, 100 bpm, 110 bpm, 120bpm, 130 bpm, 140 bpm の 7 種類の場合について収録する．midi 音は，16 分音符の時間間隔で G3(392 Hz) の音程のピアノ音が 4 回発生するものである(Fig. 1)．なお，歌ってもらう音節は，母音を/a/のみとする/ba/, /wa/, /na/, /ma/, /ya/, /da/, /ga/, /za/, /ka/, /sa/, /ta/, /ha/, /ba/, /pa/, /sha/, /cha/の 16 種類である．

Fig. 1　midi 音の譜面

2.2 スペクトル変動の分析

収録音声をスペクトログラムで表示しテンポによる歌唱音声への影響を調査する．なお，収録音声はサンプリング周波数 44.1 kHz, 量子化ビット数 16 bit であり,示すスペクトルは次数を60次(1.36 ms)としてリフタリングを施したものである．Fig. 2 は音節を/ba/とした場合のスペクトログラムであり，上段はテンポを 80 bpm, 中段は 110 bpm, 下段は 140 bpm とした場合のものである．テンポ 80 bpm において母音と思われる部分ではフォルマントのピークが時間と共に弧を描くように変化しているが，テンポが速くなるにつれ弧が水平に近づいていくことが観測できる．また，同時に振幅の変化も滑らかになることが観測できる．



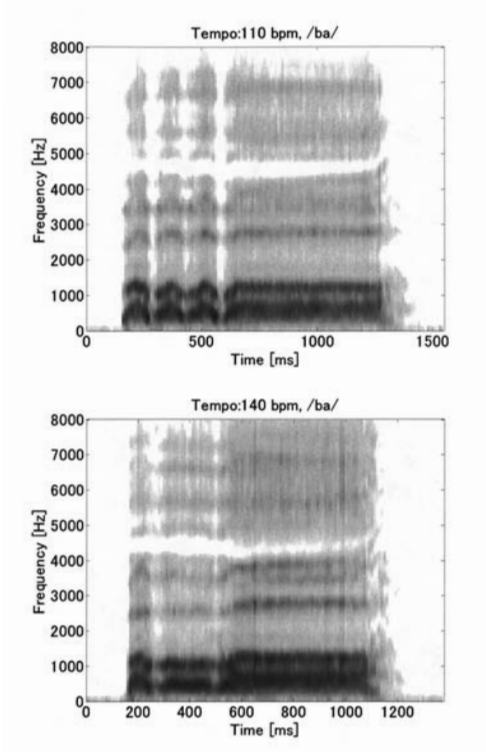


Fig. 2 収録音声のスペクトログラム

2.3 スペクトル変動の定量化

スペクトログラムにより観察されたスペクトル変動の程度を定量的に調査するため，ケプストラムを用いた分析を行った．ここでは，スペクトルの動的特徴量 $D_{\Delta c}(t)$ を以下のように定義する．

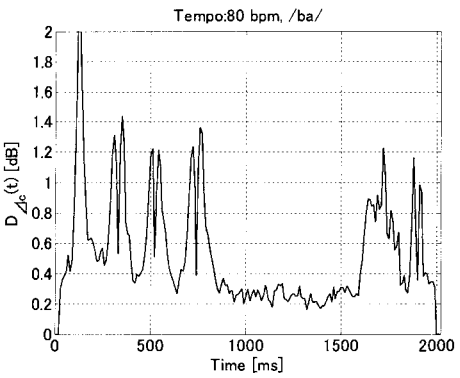
$$\Delta c_n(t) = \frac{\sum_{k=-K}^{K} k c_n(t+k)}{\sum_{k=-K}^{K} k^2}$$

$$D_{\Delta c}(t) = \sqrt{2\sum_{n}(\Delta c_n(t))^2}$$

Eq. 1 スペクトルの動的特徴量 $D_{\Delta c}(t)$ の定義

ここで，$c_n(t)$ はケプストラム係数であり，$\Delta c_n(t)$ は線形回帰係数によって平滑化したケプストラム時系列の微係数である．$D_{\Delta c}(t)$ は時刻 t における両側 $\Delta c_n(t)$ のノルムに相当する．これにより，スペクトルが時間的に大きく変動している時点では $D_{\Delta c}(t)$ の値は大きくなり，反対にスペクトルが安定している時点では $D_{\Delta c}(t)$ の値は小さくなると考えられる．なお，実際の計算においては1次から60次(1.36 ms)のケプストラム係数を使用した．また，K の値は平滑化時間窓長が 50 ms となるよう設定した．

Fig. 3 は音節を/ba/とした場合の分析結果である．上段はテンポを 80 bpm，中段は 110 bpm，下段は 140 bpm とした場合のものである．これは，2.2 節で示したスペクトルの分析に用いた音声と同じものである．母音部では変動が小さく，子音部付近では変動が大きいことが確認できる．子音部付近にピークが2つ見られることは，母音の終了時と直後の子音の開始時において変動が大きいことを示している．テンポ 80 bpm では子音部付近における変動が大きいが，テンポが速くなるにつれ変動が減少していくことが分かる．また，4回連続発声中，特に2・3 回目の発声時における変動が大きく減少していることが分かる．音節の違いにより変動の減少度合は様々であったが，分析に用いた全ての音節について，ほぼ同様の結果が得られた．



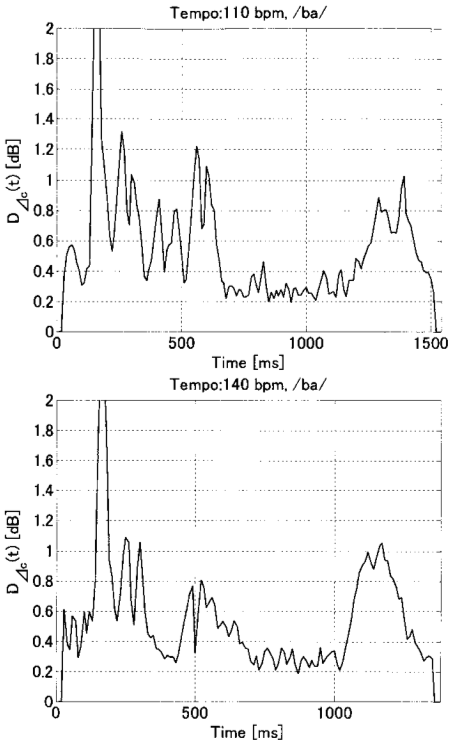


Fig. 3 収録音声のスペクトル動的特徴量 $D_{\Delta c}(t)$

3　スペクトル変動のモデル化

テンポが速くなるにつれ，スペクトルの変動が小さくなる現象が観測された．これは，声道の形状が急に変化できないために，調音が不完全になることに起因すると考えられる．既存の歌唱音声合成手法による合成音では，このことがあまり考慮されておらず，これが不自然に聞こえる原因の 1 つとなっている．そこで，既存の歌唱音声合成手法による合成音に対し，スペクトル変動を減少させる処理を行うことを考える．これにより，テンポによるスペクトル変動特性の変化が再現され，自然性の高い歌唱音声が合成できると考えられる．そのためには，まずスペクトル変動の減少量をモデル化する必要がある．前節の分析結果から，発声タイミングの間隔が密になっている部分において変動の減少量が大きいことが分かった．よって，ある発声タイミングにおいて，その前後の発声タイミングが近い位置にある場合のスペクトル変動の減少量を増加させれば良い．すなわち，発声タイミングが集中している箇所に対してはスペクトルを時間方向に平滑化し，スペクトル変動を小さくする処理を行えば良いと考えられる．そこで，スペクトル変動の減少量を表す尺度を定義する．まず，発声タイミングにおいて振幅が 1 となるパルスが立つような時間関数を作成した．これに，両側に重みを持つ窓関数を畳み込み，これを尺度関数とした．定義式を Eq. 2 に示す．ここで $\delta_{n_i,n-\tau}$ はクロネッカのデルタ関数であり，$n_i = n-\tau$ の場合に 1，それ以外の場合に 0 となる関数である．また，窓関数には，2 周期分のハニング窓を使用し，窓長を 500 ms とした．結果を Fig. 4 に示す．破線で示すものは発声タイミングを表す時間関数であり，これに窓関数を畳み込んだ関数を実線で示している．なお，以下これを発声密度関数と呼ぶ．上段はテンポを 80 bpm，中段は 110 bpm，下段は 140 bpm とした場合のものである．発声タイミング付近における発声密度関数に注目すると，テンポ 80 bpm では，1 を超える時間はわずかであるが，テンポが速くなるにつれ，1 を超える区間が長くなるとともに，値も大きくなっていることが分かる．また，特に 2・3 回目の発声タイミングにおける発声密度関数の値が大きく増加することが分かる．このことから，発声密度関数はスペクトル変動の減少量を適切に表現していると言える．

$$W_N(n)=\begin{cases}0.5-0.5\cos\left(\dfrac{4n\pi}{N-1}\right) & (0\le n\le N-1)\\ 0 & (\text{otherwise})\end{cases}$$

$$E(n)=\sum_{\tau=-\infty}^{\infty}\delta_{n_i,n-\tau}W_N(\tau) \qquad (i=1,2,...)$$

n_i : note - on time

Eq. 2 発声密度関数 $E(n)$ の定義

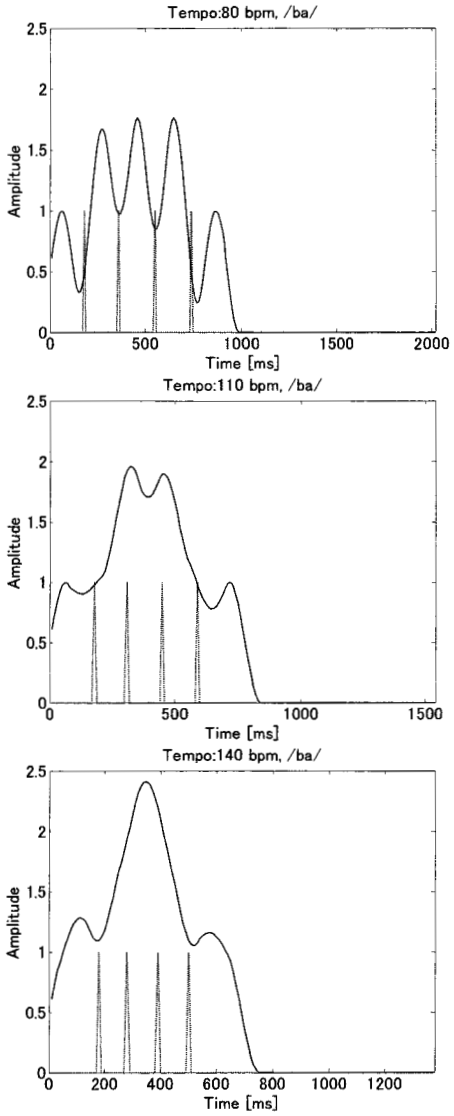


Fig. 4 発声密度関数 $E(n)$

4 合成音の作成

発声密度関数に従い，既存の歌唱音声合成手法による合成音のスペクトル変動を減少させることで，より自然性の高い歌唱音声が合成できると予想される．ここでは既存の手法として波形処理に基づく高品質な時間長伸縮法である PICOLA を使用した．2.1 節で示した収録音声のうち，テンポ 80 bpm の音声を基本とし，これを PICOLA によりテンポを 90 bpm, 100 bpm, 110 bpm, 120 bpm, 130 bpm, 140 bpm となるよう時間長圧縮し，既存の手法による合成音と仮定した．これらを試聴した結果，テンポが速いものに関しては実際よりも滑舌が良すぎるように聞こえ，不自然な印象を受けた．これは，声道の形状が実際よりも急に変化していることに起因すると考えられる．そこで，音声から声道の形状を推定し，その時間変化を滑らかにすることで実際の音声を模擬することを考える．声道の形状は，音声に対し線形予測分析[1][2]を行い，得られた PARCOR 係数から声道断面積関数[3][4]として推定することができる．これを時間毎に求め，声道断面積関数の時系列を作成した．次に，これに対して時間軸に関する平滑化を行った．平滑化には，窓による重み付き移動平均を用いた．その際，窓長を固定ではなく，発声密度関数の値が大きい時点では窓長を長く，値が小さい時点では窓長を短くなるよう可変にした．こうすることで，発声密度が大きい部分，言い換えるとテンポが速い部分では変動が大きく減少し，テンポが遅い部分では変動があまり減少しないため，実際の変動特性に近づけることができる．この提案手法の効果を調査するため，PICOLA による音声と，PICOLA による音声に提案手法を適用した音声に対して 2.2 節で示したケプストラムによる分析を行った．結果を Fig. 5 に示す．破線で示しているものが既存の手法である PICOLA による音声の分析結果であり，実線で示すものが PICOLA による音声に提案手法を適用した音声の分析結果である．既存の手法では，テンポの変化によるスペクトルの変動特性の変化がほとんど観測されないが，提案手法ではテンポが速くなるにつれスペクトルの変動が減少していることが確認できる．また，4 回連続発声中，特に 2・3 回目の発声時におけ

る変動が大きく減少していることが確認できる．これは，2.3節で示した実際の音声における変動特性と近い特性を持つと言える．提案手法による音声を試聴した結果，既存の手法による音声に比べ，極めて人間らしい音声となり，自然性が向上したという印象を受けた．

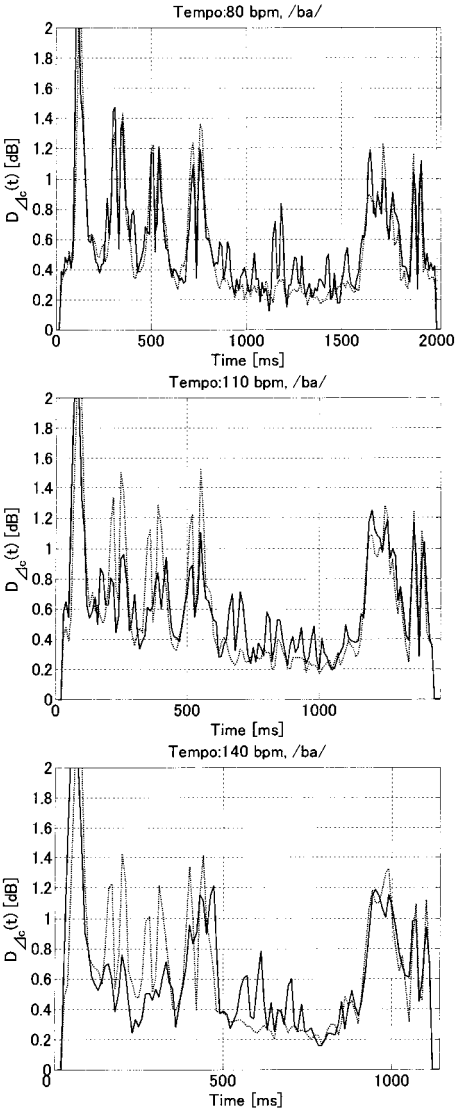


Fig. 5 既存の手法による音声と提案手法による音声のスペクトル動的特徴量 $D_{\Delta c}(t)$

6 まとめ

既存の歌唱音声合成システムでは，速く歌わせた部分において実際よりも滑舌が良すぎるように聞こえる場合があるという問題がある．そこで，異なるテンポで歌った実際の歌唱音声を収録し，そのスペクトル変動の程度を分析した．その結果，速く歌った場合ではゆっくり歌った場合に比べスペクトル変動の程度が減少することが分かった．発声タイミングが密になる部分において変動が減少しやすいため，発声タイミングにおいて振幅が 1 となるパルスが立つような時間関数を作成し，これに両側に重みを持つ窓関数を畳み込む手法でスペクトル変動の減少量を表す発声密度関数を定義した．これに基づき，既存の手法による音声から推定した声道断面積関数に対し，時間軸に関する平滑化を行うことで実際の歌唱音声に近い合成音声を生成することが可能となった．今回は4回連続発声という簡単な発声パターンの音声に対して分析と合成を行ったが，より複雑な発声をした場合についても検討する必要がある．また，提案手法の効果を明らかにするため，主観評価実験を行う予定である．

参考文献

[1] 板倉，斉藤：“統計的手法による音声スペクトル密度とホルマント周波数の推定”，電子通信学会論文誌，53-A，1，pp.35-42，1970.

[2] Atal, B.S. and Hanauee, S. L.：“Speech analysis and synthesis by linear prediction of the speech wave”, j.Acoust.Soc.Amer., 50, 2(pt.2), pp. 637-655, 1971.

[3] 中島隆之 他：“デコンボルーションによる声道形の推定と適応型音声分析システム”，日本音響学会誌，34巻，3号，pp.157-166，1978.

[4] Wakita,H： “Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms” ,IEEE Trans. Audio, Electroacoust., AU-21, 5, pp.417-427, 1973.