Vol.2017-MUS-115 No.55
2017/6/18

誇張した時間的揺らぎが歌声の人間性知覚に与える影響

森勢 将雅[1,a)]　豊田 裕一[1]　小澤 賢司[1]

概要：本稿では，人間の歌声を対象とした人間らしさの判断基準を明らかにすることを狙い，歌声の時間的な揺らぎが人間らしさに与える影響について調査した結果を報告する．音声を基本周波数（F0）・スペクトル包絡に分離する高品質ボコーダーを利用し，それぞれの時間的揺らぎを独立して制御することで音声刺激を生成した．時間的揺らぎの制御は，人間の歌声と，時間方向に 2 つのパラメータを平滑化した歌声をモーフィングすることで実現した．主観評価実験には MUSHRA 法を用い，分析合成音を基準として時間的揺らぎの程度に相当するモーフィング率と知覚する人間性との関係性について評価させた．実験の結果，F0 は元音声の 1.5 倍程度まで誇張しても人間性への影響が有意ではない一方，スペクトル包絡は 1.5 倍から有意に人間性が低下することを確認した．F0 とスペクトル包絡とを同時に制御した場合，揺らぎを低減する場合は F0 やスペクトル包絡を単独で制御するよりも効果が大きいが，誇張する場合はスペクトル包絡と同程度の影響であることを確認した．

キーワード：歌声情報処理，音声分析合成，基本周波数，スペクトル包絡，人間性の知覚

1. はじめに

テキストから波形を生成する音声・歌声の合成技術は，いまやコンテンツ制作において必要不可欠な中心的存在となりつつある．初期の歌声合成は初音ミク [1] が火付け役であり，歌声を人間に近づけるための調音技術についてクリエイターが競い合う状況にあった．現在では，統計的音声合成技術 [2] を歌声合成に活用した統計的歌声合成技術 [3] が普及している．テクノスピーチ社の CeVIO[*1]や，HOYA 株式会社の VoiceText[*2]などの製品も存在し，クリエイターは，テキストを入力するだけで自然な音声・歌声を得ることが可能になりつつある．

統計的歌声合成には，HMM (Hidden Markov Model) や DNN (Deep Neural Network) を用いた方法 [4] が実現されているが，共通することとして，多くの歌声データから学習し，統計的にもっともらしい音声パラメータを生成することがある．従来ではボコーダー [5] の考えを利用し，音声の構成要素である高さ（F0），音色（スペクトル包絡），掠れの程度（非周期性指標）を生成するが，近年では波形そのものを出力する WaveNet [6] も提案されている．様々な技術が提案され，そのたびに合成結果の品質は向上しつつあるが，生成された音声を目的に沿うように加工することは容易ではない．統計的歌声合成では自然な歌声が出力されるが，人間が加工することで人間らしさが損なわれ，不自然な結果となることが，調音の難しさとして知られている．

加工による音質劣化にはいくつかの要因があり，音声を対象とした合成技術では，イントネーションに関する研究事例が存在する [7], [8]．本研究では，人間が発声する音声には時間的な揺らぎ成分が存在することに着目し，この問題に対し，統計的な性質ではなく主観評価による人間の知覚側からアプローチする．具体的な目的は，揺らぎに関する音声や歌声の自然性を計測し，揺らぎをモデルとして与えることで，揺らぎに起因する劣化を原理的に排除することである．本稿では，その第一歩として，歌声を対象に時間的揺らぎが人間性の知覚に与える影響について，主観評価実験を実施した結果について述べる．

本稿は，以下の流れで構成される．まず 2 章では，時間的揺らぎに関する関連研究について述べ，本研究で行う実験について位置付けを行う．3 章では，本研究で実施する実験に必要となる音声分析合成技術や音声パラメータについて説明し，実験のコンセプトと具体的な手順，および得られた結果について示す．4 章では，先行研究を含めた，本研究に関する考察について述べる．最後に 5 章で，本研究のまとめを示す．

[1] 山梨大学
〒 400-8511，山梨県甲府市武田 4-3-11
[a)] mmorise@yamanashi.ac.jp

[*1] http://cevio.jp/
[*2] http://voicetext.jp/voiceactor/

2. 関連研究と本研究の位置付け

2.1 統計的音声合成に関する時間揺らぎの研究事例

2000 年代から急速に普及しはじめた統計的パラメトリック音声合成は，前述のとおり，様々な分野において広く研究がなされている．しかしながら，現在においても，合成音声の品質は肉声よりも有意に低いという問題がある [9]．その原因として発話のイントネーションや，音声の時間的揺らぎに関する研究が不十分であることが考えられる．発話のイントネーションに関しての研究は多い [7], [8] が，一方で音声の時間的揺らぎに着目した研究は少ない．いくつかの研究事例はあるが [10], [11]，音声の明瞭性向上を目的にしており，人間性知覚にターゲットを絞った研究は少ないのが現状である．ここでは，人間が時間的揺らぎをどのように知覚しているかという観点での検討を目指す．人間の知覚を考慮することで，より簡易的な揺らぎのモデルを実現できる可能性が考えられる．

2.2 時間揺らぎが人間性の知覚に与える影響の分析例

横森らは，人間の知覚特性に着目した時間的揺らぎについて調査した [12]．文献では，F0 を固定して発声した持続母音を高品質ボコーダーにより分析し，得られた F0 とスペクトル包絡を時間方向に平滑化することで，時間揺らぎを除去した音声を合成した．分析された F0，スペクトル包絡をそのまま用いた音声と，時間揺らぎを除去した音声から，音声モーフィングにより段階的に時間的揺らぎを除去した．また，声帯振動と声道情報は完全に独立な関係（F0 を高くするために力むことで声道形状も変化するなど）ではなくインタラクションが存在すると考えられる．この疑問に対し，負のモーフィング率を与えることで，正のモーフィング率とは逆の変動をする音声を生成し，実験に利用した．

実験の結果，モーフィング率が 0，つまり時間揺らぎを除去するにしたがい，音声は人間的ではなくなるという傾向が認められた．負のモーフィング率については，モーフィング率が 0 の場合と比較すると有意に人間的だと知覚されることも示された．一方，モーフィング率が 1 と −1 との場合を比較すると，1 の場合の音声がより人間的であると示されていることから，F0 とスペクトル包絡にはある程度のインタラクションが存在することを示唆する．

2.3 本研究の位置付け

このように，先行研究 [12] では，音声の時間的揺らぎを平坦化，反転したものまでの調査がなされている．一方で時間的揺らぎを本来の人間のものより誇張した場合の人間性知覚については検討されていない．揺らぎのない音声に揺らぎを付与することで人間らしさを向上させる場合，人間的だと知覚する範囲を明確することが必要となる．本研究では，先行研究と同様の実験を行い，1 以上のモーフィング率を与えることで，時間的揺らぎを誇張した場合の人間性知覚について検討することとした．本実験により，人間にとって人間的だと知覚される時間的揺らぎの範囲を定めることが可能となる．

表 1　実験に使用した音声の収録条件

発話者	4 名 (男女各 2 名)
発話内容	持続母音 (/a/, /i/)
音域	130 Hz から 247 Hz の範囲
発話時間	約 1.4 秒
A/D 変換	96 kHz/24 bit
収録マイクロホン	NEUMANN U87Ai
収録環境	レコーディングスタジオ (NC-20)

3. 主観評価実験

3.1 実験のコンセプト

本実験では，合成された音声が人間的であると判断される時間的揺らぎの範囲を明確にすることを目指す．先行研究 [12] と同様に，時間的揺らぎを除去した音声と揺らぎを含む音声をモーフィングし，段階的に時間的揺らぎを変化させる．この際，モーフィング率を 1 以上に設定することで，誇張された時間的揺らぎを有する音声を合成することが可能となる．

3.2 使用音声

本実験では，4 名の歌手から構成される歌声データベース [13] から用いる音声を決定した．音声合成に使用した音源の条件を表 1 に示す．音声は，発話者男女 2 名ずつの計 4 名とし，2 種類の母音/a/および/i/から構成される 1.4 秒程度の持続母音とした．使用音声の音域は，各発話者に依存して設定することとして，130 Hz から 247 Hz までの音域とした．

3.3 制御する音声パラメータと実験に用いる音声分析合成技術

本研究では，音声や歌声の高さ・音色を加工する技術開発が目的となるため，音声の F0 とスペクトル包絡を制御の対象とする．音声からこれらのパラメータを推定する方法には，STRAIGHT [14] や TANDEM-STRAIGHT [15], [16] が利用されている．本稿では，現在統計的歌声合成の領域で利用されつつある，筆者らが開発した WORLD [17] (D4C edition [18]) を用いる．F0 推定法として Harvest [19] を利用し，スペクトル包絡推定法には CheapTrick [20], [21]，非周期性指標推定には D4C [18] を用いることとした．

3.4 音声加工の手順と実験の条件

加工するパラメータは，先行研究と同様に，F0 とスペク

トル包絡のインタラクションが人間性知覚に与える影響を確認するため，F0 のみ，スペクトル包絡のみ，F0 とスペクトル包絡の両方を加工したものの 3 種類とした．時間的揺らぎの除去については，以下の 3 ステップで実施した．

(1) 音声を WORLD により分析し F0，スペクトル包絡，非周期性指標を推定
(2) 音声の開始・終了部を含めないように音声の有声区間を抽出
(3) 切り出された区間の F0 を平均し，全フレームの値を平均 F0 に更新

上記は F0 についてである．他パラメータはスペクトル形状であるため，各周波数における時系列について同様の処理を実施した．2 番目のステップは，音の開始・終了部では F0 が不安定になりやすく，時間平滑化することで全体に悪影響が生じることを避けるために行う必要がある．ただし，人間の知覚特性に合わせるため，平滑化は対数化した値を対象に実施している．

本実験における音声は，持続発声された単母音を F0 を固定して発声したものであるため，モーフィングにおいて重要となる音素ラベルの対応付けは不要である．同一話者の時間的揺らぎの有無の差であるため，フォルマントの対応付けも必要としない．よって，F0 のモーフィングは式 (1) を用いて実施した．

$$f_0(n) = \alpha f_a(n) + (1-\alpha) f_b(n), \tag{1}$$

ここで，$f_0(n)$ はモーフィング後の F0 軌跡，α はモーフィング率を表し，n は離散時間を示す．$f_a(n)$，$f_b(n)$ はそれぞれモーフィングを行う 2 つの F0 軌跡で，$f_a(n)$ が人間の音声，$f_b(n)$ は時間的揺らぎを除去した F0 軌跡である．α=0 の場合が時間的揺らぎを除去した F0 軌跡となり，α=1 の場合が人間音声の F0 軌跡となる．また，スペクトル包絡の加工に関しては，各離散周波数番号について，時間軸方向に対する F0 のモーフィングと同様の処理を行う式 (2) を用いた．

$$S(k,n) = \alpha S_a(k,n) + (1-\alpha) S_b(k,n), \tag{2}$$

$S(k,n)$ はモーフィング後のスペクトル包絡を示し，k は離散周波数番号を示す．$S_a(k,n)$，$S_b(k,n)$ はそれぞれモーフィングを行う 2 つのスペクトル包絡で，前者が人間の音声，後者は時間的揺らぎを除去したスペクトル包絡である．

図 1 に，男性の持続母音/a/についての F0 の加工結果を示す．図 1 の横軸と縦軸は，それぞれ時間と F0 に対応する．モーフィング率 1.0 の F0 軌跡と比較すると，モーフィング率 0 の F0 軌跡は時間的な揺らぎが除去され，モーフィング率 3.0 の F0 軌跡はより大きな揺らぎが付加されていることが確認できる．同様に，スペクトル包絡について加工したものの例を図 2 に示す．各図はスペクトログラムを表すため，横軸は時間，縦軸は周波数の対応する．モーフィング率 1.0 のスペクトログラムと比較すると，モーフィング率 0 のスペクトログラムは，時間軸方向への変動が除去されていることが確認できる．また，モーフィング率 3.0 のスペクトログラムは，周波数ごとのパワーの変化が大きくなり，揺らぎが強調されていることが確認できる．

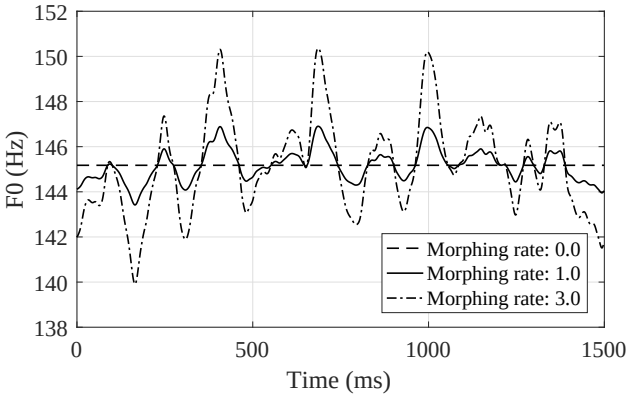

図 1 男性話者の持続母音/a/を分析して得られた F0 軌跡とモーフィング結果

本実験では，F0 のみのモーフィング，スペクトル包絡のみのモーフィング，および F0 とスペクトル包絡を同時にモーフィングの 3 種類の加工法により音声を合成した．これらの音声を用いることで，各パラメータが人間性知覚に与える影響の差，および 2 パラメータを同時にモーフィングすることによる相乗効果の確認が期待される．なお，非周期性指標については，知覚への影響が他 2 つと比較すると小さいため，本実験では加工しないこととした．

実験に用いたモーフィング率のうち，誇張部分の範囲については，予備実験により効果を明確に知覚できる範囲を算出することで決定した．予備実験の結果より，モーフィング率 0 から 1.0 までは 0.25 刻み，1.0 から 3.0 までは 0.5 刻みで制御した 9 段階で生成した．また，音声発声時の音色変化に影響されないように，本実験では，全ての音声について，音声の始点と終点に対して 5 ms ずつのテーパー窓による窓かけを行った．以上の処理に加え，実験に用いる音声に対して音圧の強弱が評価実験に影響しないように，A 特性音圧レベル [22] が均一となるよう音圧を調整した．被験者には 8 種類の条件 (表 1)，3 種類の加工方法，9 種類のモーフィング率により，合計 216 種類の音声を呈示する．

3.5 実験条件

実験条件を表 2 に示す．評価方法は MUSHRA 法 [23] を用い，評価用 GUI として MUSHRAM を利用した．図 3 に MUSHRA 法で評価するための GUI を示す．被験者は，GUI の左下にある 1 つのリファレンスを基準として，9 個のボタンにランダムに配置された音声について 0–100 の 101 段階で点数を付ける．MUSHRA では，9 個の音声の

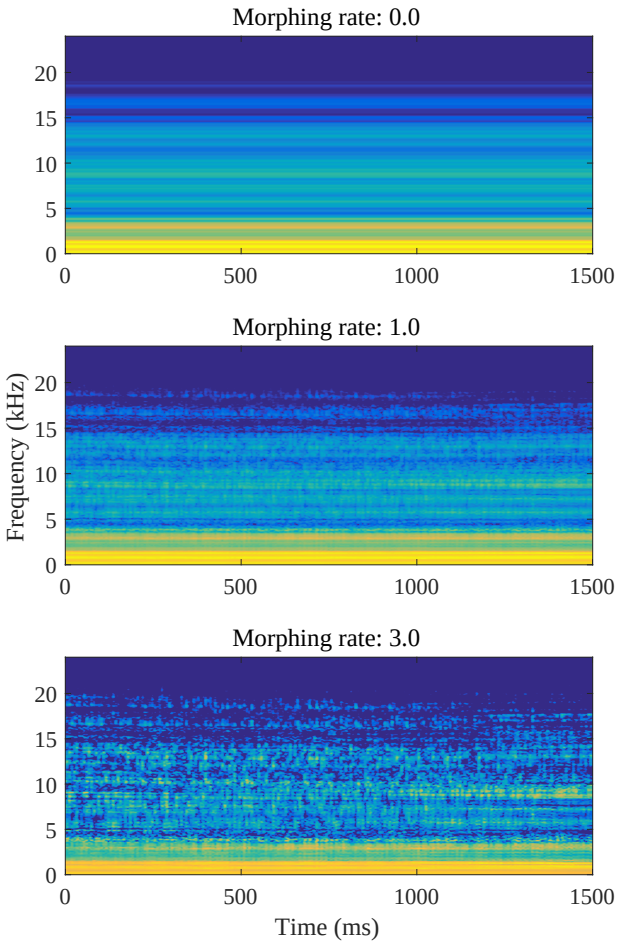

図 2 男性話者の持続母音/a/を分析して得られたスペクトログラムとモーフィング結果（上段からモーフィング率：0.0, 1.0, 3.0）.

うち 1 つはリファレンスと同じものが含まれるため，必ず 1 つの音声には 100 点を付ける制約条件が与えられる．また，被験者は同じ音声を何回でも聴くことが許されており，MOS 評価と比較すると，より精密にスコアを付けることが可能である．

MUSHRA 法は，本来は原音を対象に，符号化等を行って音質が変化した音声の相対的な違いを調査するために用いられる手法である．一方，ここではモーフィング率 1.0 を基準としたときに，音声の時間的揺らぎの影響によってどの程度人間らしさのスコアが変化するか調査した．本実験ではモーフィング率 1.0 の音声をリファレンス音として呈示し，その人間らしさを 100 としたときに，ほかの呈示音の人間性を評価するよう教示した．1 つの GUI で 9 種類の音声を評価し，音声の種類は 24 種類（4 名の話者× 2 種類の母音× 3 種類の加工パラメータ）に設定した．評価に関する操作に慣れず初期の回答がばらつくことを防ぐため，練習タスクとして，被験者にモーフィング率を 5 段階 (0, 0.25, 0.5, 0.75, 1.0) で生成した練習用音声を評価させた．この練習用音声は，実験本番に含まれる音声とは別のものである．

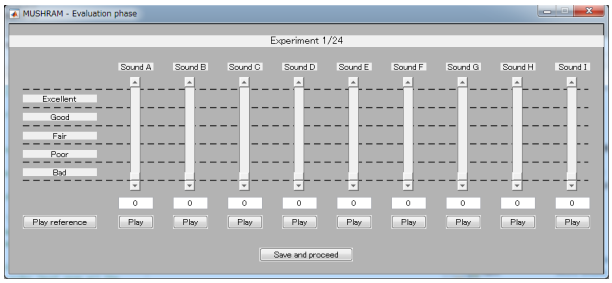

図 3 MUSHRA 評価のサポート GUI である MUSHRAM のスナップショット

表 2 実験条件

実験方法	MUSHRA
被験者数	10 名（男性 9 名，女性 1 名）
評価環境	防音室
騒音レベル	17 dB (A-weighted SPL)
再生機材	Roland, QUAD–CAPTURE SENNHEISER, HD650

3.6 実験結果

実験結果を図 4 に示す．横軸はモーフィング率を示し，縦軸は全被験者のスコアの平均値を表す．図中の黒丸，白丸，三角は，それぞれ F0 のみ，スペクトル包絡のみ (SP)，F0 とスペクトル包絡の両方 (F0+SP) を加工した場合の結果を表す．また，図示にあたり，F0 の結果，F0+SP の結果は左右にずらして表示した．各点に付与された誤差棒は，標準誤差を表す．図 4 より，モーフィング率が 1 から外れると，知覚する人間性が低下する傾向が確認できる．モーフィング率が 0 から 1 については横森らの研究と同様の傾向が認められる．

これらの実験結果から，モーフィング率がどの程度変化すると有意に人間性が低下するかを調査するため，3 種類の加工手法と，9 段階のモーフィング率に関して 2 元配置分散分析を行った．その結果，2 つの主効果，交互作用すべてに有意差 ($p < 0.01$) が認められた．次いで，モーフィング率 1 の音声を基準とし，有意差が認められない範囲を明らかにするため，Bonferroni 法による多重比較検定を行った．図中の「*」，「***」は，多重比較検定の結果，それぞれ $p < 0.05$，$p < 0.001$ で有意であったことを示す．なお，図 4 では有意な差が認められた箇所のうち，以下の主張に必要なもののみを図示した．

図 4 より，モーフィング率が 1.0 から離れるほど人間らしさのスコアが低下していることが確認できる．抑圧した時間的揺らぎにおいて，モーフィング率 1.0 との間に有意差が確認できたのは F0 ではモーフィング率 0.25 以下であったのに対し，F0+SP ではモーフィング率 0.75 であった．同様に，誇張した時間的揺らぎでは，F0 はモーフィング率

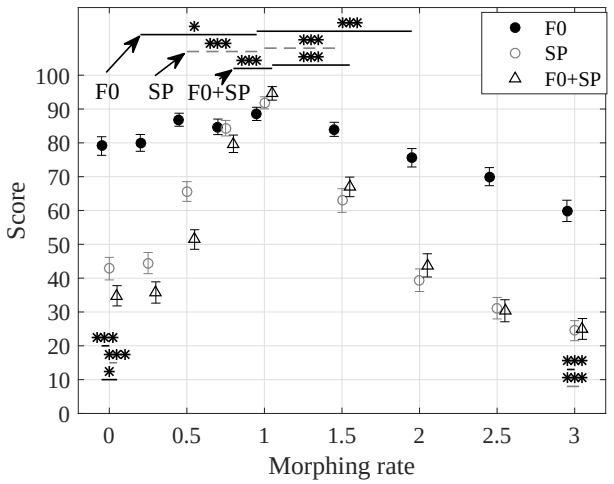

図 4　主観評価実験の結果．図中の*, ***はそれぞれ $p < 0.05$, $p < 0.001$ で有意であってことを示す．なお，全ての組み合わせについて多重比較検定は実施せず，本稿で議論するために必要最低限の組み合わせでのみ多重比較検定を実施することとした．

2.0 以上で有意差が確認でき，SP，F0+SP はそれぞれモーフィング率 1.5 以上で有意差が認められた．時間的揺らぎにおいて，F0 がモーフィング率 1.0 との間に有意差が確認できるまでの変化が最も大きく，次いで SP，F0+SP という結果が得られた．F0，SP，F0+SP の差について議論するため，同一のモーフィング率について 3 種のパラメータについて多重比較検定を実施した．検定の結果，モーフィング率 0 の条件では全ての加工手法の間に有意差が認められたが，モーフィング率 3.0 の条件では加工手法のうち SP と F0+SP の間に有意差が認められなかった．

本実験では，音声の時間的揺らぎが自然発話から離れるほど知覚される人間性が低下するという，従来の結果 [12] を支持する結果が得られた．一方，モーフィング率が 0 に近づく場合と，3 に近づく場合とで，3 種の加工パラメータの結果には差が生じていることが新たな知見として得られた．これらの原因については，次節で考察することとする．

4. 考察

4.1 時間的揺らぎと人間性との関係について

本実験の $\alpha \leq 1$ の条件について横森らの結果 [12] と比較すると，抑圧した時間的揺らぎに関して，同様の傾向を再現できていることが確認できる．F0，スペクトル包絡，F0 とスペクトル包絡の両方の加工手法について結果を比較すると，揺らぎを制御することの影響は，F0 がもっとも小さいという結果が得られた．これは，高さよりも音色のほうが人間性知覚に影響することを示唆する．今回の実験に用いた音声について，ボコーダーによる出力されるパラメータの情報量が，F0 に対してスペクトル包絡は 1025 倍であり，変化に対する知覚的な影響が大きいことが原因と考えられる．

モーフィング率 0 のときに 3 つの加工手法間で有意差が認められたことに対し，モーフィング率 3.0 のときにスペクトル包絡と F0 とスペクトル包絡の両方の間には，有意差が認められなかった．これは，主観評価後に実施した内省報告の結果から，一部の歌声について，合成結果の品質そのものが劣化していたことに起因すると考えられる．対数パワースペクトルに対してモーフィングを実施した理由として，パワースペクトルが非負であることを保証する意図があったが，1 以上のモーフィング率を設定することで，局所的に大きなパワーが発生し，音質の劣化に繋がった可能性が考えられる．人間性の知覚と自然性の知覚については，被験者間での解釈に相違があり，結果的に自然性と人間性の知覚には関連性が強いことを示唆する結果であるといえる．

全体として，時間的揺らぎについては，F0 とスペクトル包絡の相互作用の影響はあるものの，特定パラメータの時間的揺らぎを除去する場合よりも人間的であることが示された．すなわち，揺らぎを含まないパラメータが与えられた際には，相互作用を勘案せず，適切な量の時間的揺らぎを与えることで，人間性を高めることが可能になることを示唆する．

4.2 歌声合成への活用に関して

本実験結果から，人間的であると判断される時間的揺らぎの範囲が示されたといえる．今後，歌声合成の自然性向上に貢献するための検討として，入力音声に対する時間的揺らぎを計測し，人間らしさについて判断する時間的揺らぎの計測モデルを構築することが挙げられる．現在も時間的揺らぎの制御に関する研究は存在しているため，それらの研究についても，歌声知覚の観点から指標を提供することには一定の価値があると考えられる．

本実験の結果は，時間的揺らぎを人工的に付与することにも応用できると考えられる．例えば，人間的な時間的揺らぎを与えるモデルとして Klatt のモデル [24] がすでに提案されている．このモデルにはいくつかのパラメータが存在するため，本実験の結果をパラメータの最適化に利用することが可能であると考えられる．Klatt のモデルは，3 つの周波数が異なる正弦波の和で表されるシンプルなものである．本モデルをそのまま適用するのではなく，今回の実験により合成された音声の時間的揺らぎを周波数解析することで，どの程度の周波数までモデルに組み込めば良いかの指針を得ることが期待される．これらの検討を進めることで，高品質歌声合成をターゲットとした場合の時間的揺らぎモデルについても実現可能であると考えている．

今回の実験では，人間性と自然性の関係性についてまで調査していないが，人間性と自然性が完全に単調の関係にあるか（人間性が損なわれると自然性が下がる），実験する

ことも検討している．歌声合成では，人間には発声不可能な様々な表現が可能であるため，非人間的であるが自然な歌声の合成など，歌声合成の表現力の向上についても引き続き検討していきたい．

5. おわりに

本稿では，時間的な揺らぎに着目し，人間的であると知覚可能な時間的揺らぎの範囲を示すための実験を行った．高品質ボコーダーを活用し，F0，スペクトル包絡，および両方について時間的揺らぎを制御し，時間的揺らぎが無い音声から，本来の音声が有する揺らぎを誇張した音声を合成し，実験に用いた．実験の結果，F0 よりもスペクトル包絡のほうが人間的知覚に与える影響が大きいことを確認した．F0 とスペクトル包絡の両方の時間的揺らぎを制御した場合，揺らぎを除去する方向へのモーフィング率の変化については，単独で制御するよりも効果があることを確認した．一方，時間的揺らぎを誇張する方向へのモーフィング率の変化については，スペクトル包絡単独と同程度の影響となることも確認できた．

今後は，本実験結果を活かした F0 やスペクトル包絡の揺らぎモデルを与えることで，主にテキスト音声合成における自然性向上に繋げるための検討を実施する予定である．F0 については Klatt のモデル [24] が存在するため，今回の実験結果と Klatt のモデルとの整合性について調査する．加えて，スペクトル包絡の時間的揺らぎを与えるモデルについても，Klatt のモデルをベースに構築する予定である．

謝辞 本研究の一部は，科研費 JP15H02726, JP16H05899, JP16K12511, 16H01734 の支援を受けて行われた．

参考文献

[1] H. Kenmochi, “Vocaloid and Hatsune Miku phenomenon in Japan,” in Proc. INTERSINGING2010, pp. 1–4, 2010.
[2] H. Zen, K. Tokuda, and A. W. Black, “Statistical parametric speech synthesis,” Speech Communication, vol. 51, pp. 1039–1064, 2009.
[3] K. Oura, A. Mase, T. Yamada, S. Muto, Y. Nankaku, and K. Tokuda, “Recent development of the HMM-based singing voice synthesis system - Sinsy,” in Proc. SSW7, pp. 211–216, 2010.
[4] M. Blaauw and J. Bonada, “A neural parametric singing synthesizer,” arXiv preprint arXiv: 1704.03809, 2017.
[5] H. Dudley, “Remaking speech,” J. Acoust. Soc. Am., vol. 11, pp. 169–177, 1939.
[6] A. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, “WaveNet: A generative model for raw audio,” arXiv preprint arXiv: 1609.03499, 2016.
[7] 中嶋秀治, 青野裕司, “テキストからの音声合成の為の強調アクセント句内の強調単語の予測, ” 日本音響学会秋季研究発表会講演論文集, pp. 163–164, 2016.
[8] 前野悠, 能勢隆, 小林隆夫, 井島勇祐, 中嶋秀治, 水野秀之, 吉岡理, “多様な発話様式による HMM 音声合成のための韻律コンテキストの検討, ” 日本音響学会秋季研究発表会講演論文集, pp. 385–386, 2011.
[9] Y. Stylianou, “Voice transformation: A survey,” in Proc. ICASSP 2009, pp. 3585–3588, 2009.
[10] T. Toda and K. Tokuda, “A speech parameter generation algorithm considering global variance for HMM-based speech synthesis,” IEICE transactions on information and systems, vol. 90, no. 5, pp. 816–824, 2007
[11] S. Takamichi, T. Toda, A. Black, S. Sakti, G. Neubig, and S. Nakamura, “Post-filters to modify the modulation spectrum for statistical parametric speech synthesis,” IEEE/ACM Transactions on Audio, Speech and Language Processing, vol. 24, no. 4, pp. 755–767, 2016.
[12] 横森文哉, 森勢将雅, 小澤賢司 “基本周波数とスペクトル包絡の揺らぎに着目した合成音声の人間性知覚に関する検討, ” 日本音響学会秋季研究発表会講演論文集, pp. 609–610, 2016.
[13] 右田尚人, 森勢将雅, 西浦敬信, “歌唱データベースを用いたヴィブラートの個人性の制御に有効な特徴量の検討, ” 情報処理学会論文誌, vol. 52, no. 5, pp. 1910–1922, 2011.
[14] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigné, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction,” Speech Communication, vol. 27, no. 3-4, pp. 187–207, 1999.
[15] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno, “TANDEM-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, f0, and aperiodicity estimation,” in Proc. ICASSP 2008, pp. 3933–3936, 2008.
[16] H. Kawahara and M. Morise, “Technical foundations of TANDEM-STRAIGHT, a speech analysis, modification and synthesis framework,” SADHANA - Academy Proceedings in Engineering Sciences, vol. 36, pp. 713–728, 2011.
[17] M. Morise, F. Yokomori, and K. Ozawa “WORLD: a vocoder-based high-quality speech synthesis system for real-time applications,” IEICE Trans. Inf. & Syst., vol. E99–D, pp. 1877–1884, 2016.
[18] M. Morise, “D4C, a band-aperiodicity estimator for high–quality speech synthesis,” Speech Communication, vol. 84, pp. 57–65, 2016.
[19] M. Morise “Harvest: A high-performance fundamental frequency estimator from speech signals,” in Proc. INTERSPEECH 2017, 2017.
[20] M. Morise, “Cheaptrick, a spectral envelope estimator for high-quality speech synthesis,” Speech Communication, vol. 67, pp. 1–7, 2015.
[21] M. Morise, “Error evaluation of an f0-adaptive spectral envelope estimator in robustness against the additive noise and f0 error,” IEICE Trans. Inf. & Syst., vol. E98-D, pp. 1405–1408, 2015.
[22] “Electroacoustics–Sound level meters–Part1,” IEC61672-1, Specifications, 2013.
[23] “Method for the subjective assessment of intermediate quality levels of coding systems,” ITU (International Telecommunication Union) Recommendation BS.1534–1, Jan. 2003.
[24] D. Klatt, and L. Klatt, “Analysis, ynthesis, and perception of voice quality variations among female and male talkers,” J. Acoust. Soc. Am., vol. 82, no. 2, pp. 820–857, 1990.