Vol.2018-MUS-119 No.52
Vol.2018-SLP-122 No.52
2018/6/17

Human-in-the-loop型音声デザインのための DNNを活用した音声パラメータ生成の検討

近藤 大地[1,a)]　森勢 将雅[1]

概要：ユーザが所望する音声を生成したいという需要から，音声デザインを補助する音声合成システムが存在する．これらのシステムでは，加工の程度に伴い品質が劣化することから，加工できる範囲に縛りを設けることが多い．そのため，システムの許容範囲を超える加工をしたいという要求を持つユーザは，システム内で一度生成した合成音声を，システム外でさらに加工するが，やはり加工の程度に伴い品質が劣化する．本研究では，人間がデザインした結果に基づき，DNN により音声パラメータを生成するHuman-in-the-loop 型音声デザインを提案する．提案手法では，品質劣化が目立つような大きな加工を許容し，DNN が品質劣化を抑えるような形で音声パラメータを生成することで，問題の解決を図った．本稿では，提案法と従来手法によって生成された音声の品質について主観評価実験を行い，有用性を検証した．主観評価実験には MOS 評価を用い，音声の品質について評価させた．実験の結果，提案手法の品質が有意に高いことを確認した．

1. はじめに

テキストから音声波形を生成するテキスト音声合成技術は，動画共有サービス YouTube[*1]の実況動画や施設の情報案内などといった，日常生活の様々な場面に見られる．このように，テキスト音声合成は，現代社会に広く浸透している技術であることが窺える．テキスト音声合成が普及した理由としては，テキストを入力するだけで自然な音声を得ることが可能であり，音声生成の容易性や変更への柔軟性などが評価されていることが挙げられる．

テキスト音声合成では，人間が加工不要なほど高品質な合成音声の生成を目的に，人間を極力介在させないようなデザインがなされていた．昨今では，既存のテキスト音声合成手法に対して，音声生成の全てをニューラルネットワークに任せた End-to-End の音声合成手法が標準になりつつある．Google Home[*2] では，WaveNet[1], [2] と呼ばれる End-to-End 音声合成が用いられている．一方で，人間が音声や歌声をデザインしたいという需要から，CeVIO[*3]やVOICEROID[*4]のような，合成音声の加工によって，音声デザインを補助するシステムが存在する．しかしこれらのシステムでは，品質劣化が目立つような大きな加工が許容されていない．許容範囲を超えた加工がしたいという要求を持つユーザは，システム内で生成した合成音声を，システム外でさらに加工する作業が必要となるが，加工の程度に伴って信号処理による品質劣化が生じてしまう．合成結果が劣化する問題は，人間が加工した音声をコンピュータが調整し，再生成することで解決できると考えられる．また，品質劣化が目立つような大きな加工に対してこれを適用することで，製品外での加工を経由せずにユーザーが満足する合成音声を生成することが可能であると考えられる．

本研究では，既存のテキスト音声合成技術を音声デザインの補助として活用することを目指した，Human-in-the-loop 型音声デザイン技術の開発を目的とする．Human-in-the-loop 型音声デザインとは，音声合成の一部に人間が介在し，ユーザがデザインした音声に基づいて，音声を合成することを意味する．本稿では，Human-in-the-loop 型音声デザインによって，ユーザの音声デザインを補助し，デザインした音声をより自然なものへとコンピュータが調整する手法を提案する．

本稿は，以下の流れで構成される．まず 2 章は，音声合成の手法について述べる．3 章では，本研究で提案する音声合成の手法について説明し，4 章では，提案手法により生成された合成音声に対する主観評価実験とその結果および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
University of Yamanashi
[a)] g18tk007@yamanashi.ac.jp
[*1] https://youtube.com/
[*2] https://store.google.com/product/google_home
[*3] http://cevio.jp/
[*4] http://ah-soft.com/voiceroid/

Vol.2018-MUS-119 No.52
Vol.2018-SLP-122 No.52
2018/6/17

2. 音声合成手法

2.1 統計的音声合成

統計的音声合成とは，テキストと音声の関係を統計的なモデルによって表現することで入力テキストに対応する音声を生成する手法である [3], [4]．モデルの生成は，音声コーパスを構成する音声波形やテキストの特徴を HMM (hidden Markov model) や DNN (deep neural network) などの機械学習または深層学習の手法を用いることで行う．

統計的音声合成の手順を図 1 に示す．前処理部では，発話テキストと音声波形から構成される学習データのラベル付けを行う．ラベルには，ある音素の発話時間（音素継続長）を区別するために，テキストにおける音素の境界位置を記述する．また，コンテキスト（文脈や前後の音素などの音響的に影響を与える要因）を区別するために，言語の品詞や活用形といった自然言語特有の特徴である，言語特徴量の記述を行う．加えて，各音素の音響特徴量は前後の音素によって変化することから，前後の音素情報を記述する．学習部では，音声波形から音響特徴量や言語特徴量を抽出する．抽出した言語特徴量と音響特徴量およびラベル付けを行ったデータに対するテキスト解析の結果を用いてモデルの学習を行う．

合成部では，入力テキストの解析によって得られる言語特徴量をモデルに適用することで，音響特徴量の生成を行う．加えて，ボコーダによる音声分析合成技術を利用することで，生成された音響特徴量から音声波形を生成する．音声分析合成では，音声を声帯振動からなる音源と声道特性を表す合成フィルタによって表現する．音源は基本周波数（F0）や非周期性指標（ap）に対応し，声道フィルタはスペクトル包絡（sp）に対応する．

2.1.1 HMM 音声合成

統計的音声合成における音響モデルの学習に HMM を用いたものが，HMM 音声合成である [5]．HMM 音声合成は，マルコフ性に基づく時間構造を持つ離散状態からなる状態系列によって，音声の時間変化をモデル化する手法である．HMM 音声合成の学習部では，音声コーパス中の言語特徴量に対応する音響特徴量の尤度が最大となる HMM の学習を行う．

合成部では，入力テキストから抽出された言語特徴量に従って HMM を連結し，出力確率が最大となる音響特徴量を予測する．予測した音響特徴量から音源を生成し，合成フィルタを通すことによって合成波形を生成する．HMM 音声合成の代表的な製品例としては，CeVIO が挙げられる．

2.1.2 DNN 音声合成

統計的音声合成における音響モデルの学習に DNN を用いたものが DNN 音声合成である [6], [7]．DNN 音声合成では，言語特徴量から音響特徴量を予測する回帰問題に DNN を利用する [8], [9]．学習部では，学習用の音声コーパスから抽出した音素単位の言語特徴量と音素継続長の対を利用して，継続長モデルを DNN で学習する．同様に，抽出したフレーム単位の言語特徴量と音響特徴量の系列の対を利用して音響モデルを DNN で学習する．

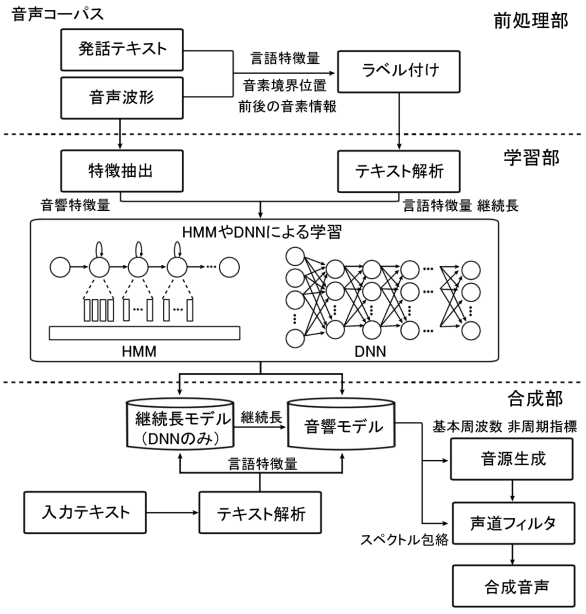

図 1 統計的音声合成の手順

合成部では，入力テキストから音素単位の言語特徴量を求め，これを継続長モデルに適用することで，音素継続長を予測する．この音素継続長に基づいて，フレーム単位の言語特徴量を求め，これを音響モデルに適用することで，音響特徴量を予測する．HMM 音声合成の場合と同様に，予測した音響特徴量から合成波形を生成する．DNN 音声合成は，HMM 音声合成による合成音声の品質を大きく改善したことが示されている [7]．

2.2 End-to-End 音声合成

音声合成の最先端手法として，WaveNet[1], [2] や TACOTRON[10] が挙げられる．これらの研究では，End-to-End 音声合成によって音声波形，あるいはスペクトログラムを直接生成するためのニューラルネットワークが提案されている．End-to-End とは，端から端までという意味の通り，ニューラルネットワークによって，入力と出力の関係を直接モデル化する手法である．

本章でこれまで述べてきたように，テキスト音声合成の合成部では，テキスト解析，音響モデルから音響特徴量の生成および音響特徴量から音声波形の合成といった複数の段階からなる．これに対し，End-to-End 音声合成では，テキスト解析から音声波形を直接生成することで，音声や言語に関する専門知識に基づく処理を軽減できるという利点が存在する．また，従来のテキスト音声合成手法による合成音声の品質を大きく上回り，自然音声に迫る高品質な音声を合成可能なことが示されている [1]．

Vol.2018-MUS-119 No.52
Vol.2018-SLP-122 No.52
2018/6/17

3. Human-in-the-loop 型音声デザイン

3.1 本研究で用いる音声合成手法

WaveNet のような End-to-End 音声合成では，ユーザが操作するための音声パラメータである，音響特徴量が生成されない．音声デザインには，音響特徴量の調整を行う必要があるため，本研究に End-to-End 音声合成は適さないと考えられる．また，2 章で紹介した音声合成手法以外にも，波形接続型音声合成といった手法が存在する [3]．波形接続型音声合成と統計的音声合成は，ともに音声デザインに関する製品が存在し，どちらも音声デザインに適した手法である．

音声をデザインするにあたり，波形接続型音声合成では音声素片の接続と音声パラメータの加工の際に，信号処理の多さに伴う大きな品質劣化が生じてしまうことが懸念される．一方で，統計的音声合成では音声デザインしたパラメータをもとに，モデルから音声を生成することで，信号処理による品質劣化を抑えることができる可能性がある．これらを考慮した結果，統計的音声合成が本研究の目的において最適な音声合成手法であると結論付けた．また，モデルの学習には，HMM と比較して，より良い合成音声の品質が示されている DNN を利用する [7].

3.2 従来の合成音声加工手法

これまで紹介した既存のテキスト音声合成の製品では，加工による音声デザインが可能である．一方で，品質劣化が目立つような大きな加工はシステム上，許容されていない．許容範囲を超えた加工がしたいという要求を持つユーザは，製品内で生成した合成音声を，製品外でさらに加工する処理が必要となるが，加工の程度に伴って，信号処理による品質劣化が生じてしまう．この品質劣化の原因としては，加工に伴う音声パラメータの平滑化や不自然さの強調などが挙げられる．

3.3 提案手法

本稿では，従来手法における合成音声の加工に伴う品質劣化に対して，Human-in-the-loop 型の音声デザインを提案する．提案手法では，品質劣化が目立つような音声パラメータの加工を許容し，DNN が品質劣化を抑えるような形で音声パラメータを再生成する，

人間が加工するパラメータとしては，図 1 の合成部に用いる音響特徴量の F0，sp，ap が候補として挙げられる．これらの音響特徴量は全て時間軸に沿って変化する値であり，自然さを保つような加工そのものが容易ではない．しかし，F0 は 1 次元の値であるため，その他音響特徴量に比べ，人間による加工が容易であると言える．一方で，sp と ap はともに多次元の値であり，人間による加工が困難である．そこで，加工が容易である F0 を人間が加工し，人間が加工困難な sp と ap を DNN が自動調整することで，提案手法を実現できる可能性が考えられる．

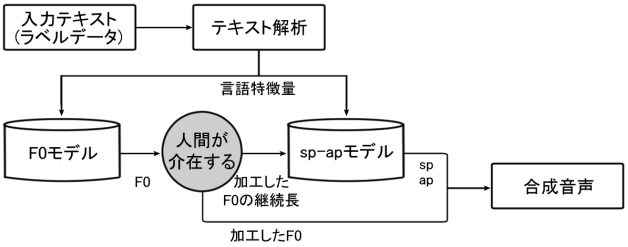

図 2 提案手法における人間の介在箇所

提案手法における人間の介在箇所を図 2 に示す．提案手法では二種類の DNN を用いる．1 つ目は，テキストから F0 を予測するモデルを生成する DNN である（以下，この DNN で学習するモデルを F0 モデルと呼称）．2 つ目は，テキストと F0 の継続長から sp と ap を予測するモデルを生成する DNN である（以下，この DNN で学習するモデルを sp–ap モデルと呼称）．F0 モデルおよび sp–ap モデルともに，継続長モデルおよび音響モデルの学習を行う．図 2 の F0 モデルと人間の介在箇所および sp–ap モデルの入出力部分は，図 1 の継続長モデルおよび音響モデル部分に相当する．F0 モデルによって一度予測した F0 をユーザが加工し，sp–ap モデルによってユーザが加工した F0 の継続長に合わせた sp と ap を予測する．提案手法による利点としては，ユーザが加工した F0 に基づいて，DNN がより自然な sp と ap を生成できる可能性が挙げられる．

3.4 提案手法の構成

前処理部における前後音素を考慮した音素境界位置は，大語彙連続音声認識エンジン Julius[11] を用いて記述する．Julius で用いる音素表記は，学習に用いる音声コーパス中に含まれる音素表記に合わせたものとする．また言語特徴量の記述に際して必要な形態素解析には，汎用日本語形態素解析エンジン Mecab[12] を用いる．加えて形態素解析した結果を日本語音声合成システム Open JTalk[13] を用いて言語特徴量へと変換して，ラベルデータとする．このラベルデータの誤りに関しては，手作業での修正を行う．

音声コーパス中の音声波形から音響特徴量の抽出には，音声分析合成システム WORLD[14], [15] を用いる．また音響特徴量の抽出に際し，ラベルデータの無声区間に関する情報から，音声波形の最初と最後の無声区間の除去を行う．抽出した特徴量は F0，sp，ap である．抽出した F0 は $\log \mathrm{F0}$ へ対数化する．学習には，式 (1) に示す，$\log \mathrm{F0}$ の離散信号，その時間差分 $\Delta \log \mathrm{F0}[n]$ および $\Delta\Delta \log \mathrm{F0}[n]$ を用いる．ここで，$\log \mathrm{F0}[n]$ は $\log \mathrm{F0}$ の離散信号を表す．加えて，$\log \mathrm{F0}$ から有声無声音判定の計算を行う．また，n はパラメータに対する時間方向の値である．

$$\begin{aligned}\log \mathrm{F0}[n] &= \log \mathrm{F0}[n] \\ \Delta \log \mathrm{F0}[n] &= 0.5(\log \mathrm{F0}[n+1] - \log \mathrm{F0}[n-1]) \\ \Delta\Delta \log \mathrm{F0}[n] &= \log \mathrm{F0}[n+1] + \log \mathrm{F0}[n-1] - 2\log \mathrm{F0}[n]\end{aligned} \tag{1}$$

また，抽出した sp はメルケプストラム（mel-frequency cepstrum coefficients; MFCC）へ次元圧縮する．MFCC は，人の声道特性を表すパラメータの一つであり [16]，人の知覚特性を考慮して sp の圧縮を行う．学習には，式 (2) に示す，MFCC の離散信号，その時間差分 $\Delta\mathrm{mfcc}[n]$ および $\Delta\Delta\mathrm{mfcc}[n]$ を用いる．ここで，$\mathrm{mfcc}[n]$ は MFCC の離散時系列を表す．

$$\begin{aligned}\mathrm{mfcc}[n] &= \mathrm{mfcc}[n] \\ \Delta\mathrm{mfcc}[n] &= 0.5(\mathrm{mfcc}[n+1] - \mathrm{mfcc}[n-1]) \\ \Delta\Delta\mathrm{mfcc}[n] &= \mathrm{mfcc}[n+1] + \mathrm{mfcc}[n-1] - 2\mathrm{mfcc}[n]\end{aligned} \tag{2}$$

同様に，抽出した ap は，スペクトル表現の ap に対し，帯域毎の平均値として与えられる bap（band aperiodicity）へ次元圧縮する．学習には，式 (3) に示す，bap の離散信号，その時間差分 $\Delta\mathrm{bap}[n]$ および $\Delta\Delta\mathrm{bap}[n]$ を用いる．ここで，$\mathrm{bap}[n]$ は bap の離散時系列を表す．

$$\begin{aligned}\mathrm{bap}[n] &= \mathrm{bap}[n] \\ \Delta\mathrm{bap}[n] &= 0.5(\mathrm{bap}[n+1] - \mathrm{bap}[n-1]) \\ \Delta\Delta\mathrm{bap}[n] &= \mathrm{bap}[n+1] + \mathrm{bap}[n-1] - 2\mathrm{bap}[n]\end{aligned} \tag{3}$$

ニューラルネットワークで学習するモデルにおける，層毎の次元数の対応を，表 1 に示す．F0 モデルにおける継続長モデルの学習では，予備検討として入力層の次元を，音素単位で抽出した言語特徴量の次元である 1119 次元とした．また中間層は 512 次元とし，出力層は音素継続長の次元である 1 次元とした．F0 モデルにおける音響モデルの学習では，予備検討として入力層の次元を，フレーム単位で抽出した言語特徴量の次元である 1123 次元とした．また中間層は 512 次元とし，出力層は，$\log \mathrm{F0}$ の離散信号およびその時間差分の 3 次元と，有声無声音判定の 1 次元の和である 4 次元とした．

sp–ap モデルにおける継続長モデルの学習では，F0 モデルと同様に，入力層を 1119 次元，中間層を 512 次元，出力層を 1 次元とした．sp–ap モデルにおける音響モデルの学習では，F0 モデルと同様に，入力層を 1123 次元，中間層を 512 次元とした．出力層は，MFCC の離散信号およびその時間差分の 180 次元と，bap の離散信号およびその時間差分の 15 次元の和である 195 次元とした．

合成部では，F0 モデルによって予測した $\log \mathrm{F0}$ を F0 へ変換する．次に，F0 および F0 の継続長の自由な加工を人間が行う．加えて，加工した F0 の継続長に合わせて，sp–ap モデルによる MFCC と bap の予測および sp と ap への復号を行う．最後に，人間が加工した F0，生成した sp，ap を音響特徴量とした，WORLD による音声分析合成によって，音声波形を生成する．

表 1 モデルと次元数の対応

モデル	入力層の次元	中間層の次元	出力層の次元
F0 モデル（継続長モデル）	1119	512	1
F0 モデル（音響モデル）	1123	512	4
sp–ap モデル（継続長モデル）	1119	512	1
sp–ap モデル（音響モデル）	1123	512	195

表 2 音声の収録環境

A/D 変換	48 kHz / 16 bit
マイクロホン	NEUMANN U87Ai
環境	防音室 (A–weighted SPL 18 dB)
テキスト	ATR503 文章から全 503 文章
発話者	20 代男性 1 名

4. 音声デザインに対する主観評価実験

4.1 学習に用いる音声の収録

評価実験の実施にあたり，各モデルの学習に用いる音声の収録を行う．本実験では，ATR503 文章 [17] を用いて音声の収録を行った．音声の収録環境を表 2 に示す．収録音声をラベル付けすることで，音声コーパスの作成をした．

4.2 実験音声の生成

作成した音声コーパスを用いて学習を行い，実験音声を生成する．実験音声の生成方法を図 3 に示す．図中上が提案手法，下が従来手法による音声の生成方法を表す．また，合成条件を表 3 に示す．本実験では，直感的な F0 加工機能を持つインターフェースが未実装なため，別の話者による同一文章の音声における F0，およびその継続長（以下，代替 F0 と呼称）を用いることで，直感的な F0 加工機能の代替処理とする．

提案手法では，入力テキストから一度 F0 を予測する．次に，F0 の加工処理として代替 F0 を与える．そして，sp–ap モデルによって代替 F0 の継続長に合わせた sp と ap を予測する．最後に，代替 F0，sp，ap から音声波形を合成する．

従来手法では，入力テキストから一度 F0 を予測する．次に，sp–ap モデルによって予測した F0 の継続長に合わせた sp と ap を予測する．そして，F0 の加工処理として代替 F0 を与え，sp と ap を代替 F0 の継続長に合わせて線形伸縮する．最後に，代替 F0，線形伸縮した sp および ap から音声波形を合成する．

本実験における，各モデルを生成する DNN のハイパーパラメータは表 4 の通りである．合成するテキストは

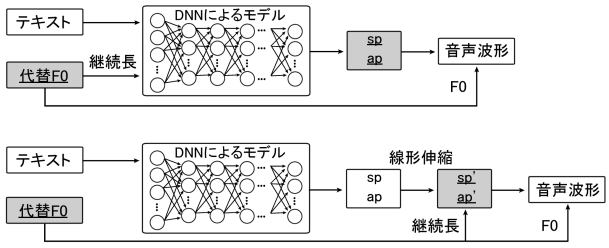

図 3 実験音声の生成方法
図中上は提案手法，下は従来手法

表 3 実験音声の合成条件

提案手法	代替 F0 の継続長に合わせて sp と ap を DNN で予測して合成
従来手法	DNN で予測した sp と ap を 代替 F0 の継続長に線形伸縮して合成

表 4 DNN のハイパーパラメータ

損失関数	最小二乗誤差
活性化関数	tanh
最適化手法	Adam（adaptive moment estimation）[18]
中間層数	10
epoch 数	100
バッチサイズ	256
学習係数	5.0×10^{-5}
重み減衰	1.0×10^{-6}

表 5 実験環境と評価方法

主観評価尺度	MOS 評価
再生機器	SENNHEISER HD650 Roland QUAD–CAPTURE
環境	防音室 (A–weighted SPL 18 dB)
文章数	20 文章
評価音声数	40 音声 （提案手法 20 音声，従来手法 20 音声）
被験者	20 代男性 10 名

ATR503 文章から 20 文章とし，提案手法，従来手法ともに 20 音声の全 40 音声を生成する．また生成された音声に，人間の聴覚に沿った大きさである，等価騒音レベルの正規化を行うことで，実験に用いる音声とする．

4.3 評価方法

音声の評価方法として，MOS（mean opinion score）評価を用いる．音声の評価環境を表 5 に示す．実験では，全 40 音声の中からランダムな順番で実験音声を提示し，「非常に悪い」から「非常に良い」の 5 段階評価を用いて音声の品質を評価する．

4.4 実験結果と考察

主観評価実験の結果を図 4 に示す．図の横軸は MOS を表しており，縦軸の上側が従来手法，下側が提案手法を表す．また，誤差棒は 95%信頼区間を表す．t 検定による p 値は 1.02×10^{-66} となり，提案手法の音声品質が有意に高

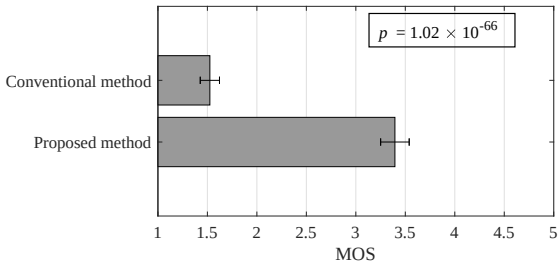

図 4 主観評価実験の結果

いことが示された．

提案手法における評価値の標準偏差が大きいことから，評価値のばらつきが大きいことが分かる．考えられる原因として，評価がモデルの質に左右されている可能性が挙げられる．モデルの質は，学習に用いた音声数や学習方法に依存している．本実験において，学習に用いた音声数は 503 個であり，収録時間の合計としては 1 時間に満たないデータである．WaveNet の主観評価実験において，北米英語では 24.6 時間，中国語では 34.8 時間の，プロの女性発話者による音声が学習に用いられている [1]．本実験の結果より，モデルの質を向上させるうえで，この音声数では不十分であることが示唆された．また，RNN（reccurent neural network）を用いたテキスト音声合成では，DNN より良い音声品質の評価結果が示されている [19]．これらのことから，学習音声を増やすことや，学習方法に RNN を用いることでモデルの質が改善できると考えられる．

図 4 より，従来手法の MOS が低いことが分かる．従来手法では，代替 F0 の継続長に合わせて，sp と ap の継続長を音素毎に補間している．加工前と後で音素毎の継続長の差が大きく，品質が劣化したためであると考えられる．

4.5 実験結果の分析

実験音声の生成の際に，音素毎の継続長の差が大きいと，品質が劣化するという考察から，加工前と後における継続長の差を音声変化率とした，実験音声の分析を行った．従来手法における音声変化率の分析結果を図 5 に，提案手法における音声変化率の分析結果を図 6 に示す．図の横軸は変化率として，発話時間に対する音素継続長差の割合を表しており，縦軸は MOS を表す．また，音素継続長差は，音素毎の継続長の差の絶対値の和により算出した．従来手法と提案手法における変化率と MOS の相関係数はそれぞれ，−0.456, −0.081 であった．従来手法の MOS が低いという考察に対して，分析結果から，従来手法では弱い相関が見られたが，有意差を認めることはできなかった．しかし，テストする音声数を増やすことで有意差が認められる可能性がある．

5. おわりに

本稿では，合成音声の加工に伴う品質劣化に対して，人

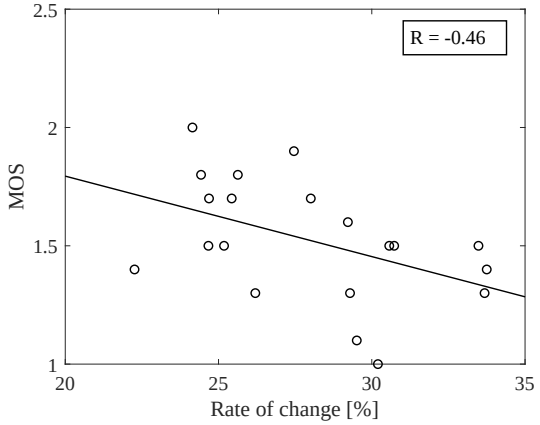

図 5 従来手法における音声変化率の分析

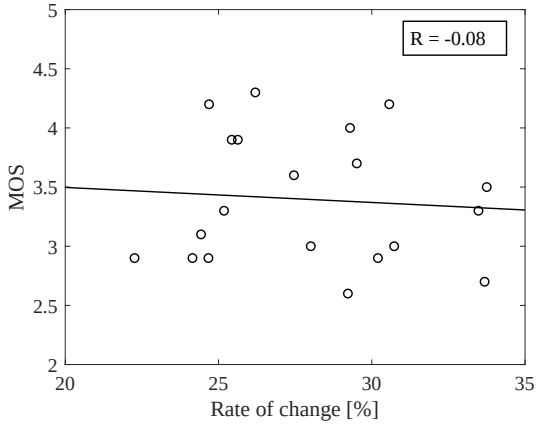

図 6 提案手法における音声変化率の分析

間がデザインした結果に基づいた合成音声の生成に着目し，Human-in-the-loop 型音声デザインの提案を行った．提案手法の音声品質を評価するにあたり，別の話者による同一文章の発話音声における F0 およびその継続長を，加工の代替処理として用いた．提案手法では，加工した F0 の継続長を用いて sp と ap のパラメータ再生成し，音声を生成した．従来手法では，加工した F0 の継続長に合わせた sp と ap の線形補間によって，音声を生成した．主観評価実験の結果，提案手法の音声品質が有意に高いことを確認した．

今後は，学習に用いる音声数の増加および学習方法の改善によって，モデルの改良を行う予定である．また，本稿で F0 の加工に代替処理を用いたことについては，直感な F0 デザイン機能を持つインターフェースの実装を行う予定である．加えて，現状の実装では，デザインした F0 のうち，F0 継続長のみを sp と ap の音声パラメータの生成に利用しており，継続長以外の値を有効活用することができていない．そのため，F0 の値全てを利用して音声パラメータを生成するような実装を検討する必要がある．

謝辞 謝辞本研究は，JSPS 科研費 JP16H05899, JP16H01734 の支援を受けて実施された．

参考文献

[1] van Den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. and Kavukcuoglu, K.: Wavenet: A generative model for raw audio, *arXiv preprint arXiv:1609.03499* (2016).

[2] van Den Oord, A., Li, Y., Babuschkin, I., Simonyan, K., Vinyals, O., Kavukcuoglu, K., Driessche, G. v. d., Lockhart, E., Cobo, L. C., Stimberg, F., Casagrande, N., Grewe, D., Noury, S., Dieleman, S., Elsen, E., Kakchbrenner, N., Zen, H., Graves, A., King, H., Walters, T., Belov, D. and Hassabis, D.: Parallel WaveNet: Fast High-Fidelity Speech Synthesis, *arXiv preprint arXiv:1711.10433* (2017).

[3] Zen, H., Tokuda, K. and Black, A. W.: Statistical parametric speech synthesis, *Speech Communication*, Vol. 51, No. 11, pp. 1039–1064 (2009).

[4] 徳田恵一：統計的機械学習問題としての音声合成，研究報告音楽情報科学 (MUS) 招待講演， Vol. 2013, No. 2, p. 1 (2013).

[5] 徳田恵一：HMM による音声合成の基礎，電子情報通信学会技術研究報告, SP2000-74，p. 43 (2000).

[6] 清山信正：音声合成技術の動向と放送・通信分野における応用展開，NHK 技研 R&D，No. 161, pp. 13–22 (2017).

[7] Zen, H., Senior, A. and Schuster, M.: Statistical parametric speech synthesis using deep neural networks, *in Proc. ICASSP, 2013 IEEE International Conference on*, IEEE, pp. 7962–7966 (2013).

[8] 橋本佳，高木信二：深層学習に基づく統計的音声合成，日本音響学会誌， Vol. 73, No. 1, pp. 55–62 (2017).

[9] Wu, Z., Watts, O. and King, S.: Merlin: An open source neural network speech synthesis system, *in Proc. SSW9, Sunnyvale, USA*, pp. 202–207 (2016).

[10] Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S. et al.: Tacotron: A fully end-to-end text-to-speech synthesis model, *arXiv preprint arXiv:1703.10135* (2017).

[11] 河原達也， 李晃伸：連続音声認識ソフトウェア Julius，Vol. 20, No. 1, pp. 41–49 (2005).

[12] Kudo, T.: Mecab: Yet another part-of-speech and morphological analyzer, *http://mecab. sourceforge. jp* (2006).

[13] 大浦圭一郎：日本語テキスト音声合成システム Open JTalk，日本音響学会春季講論集, 2010， Vol. 1, pp. 343–344 (2010).

[14] Morise, M., Yokomori, F. and Ozawa, K.: WORLD: a vocoder-based high-quality speech synthesis system for real-time applications, *IEICE TRANSACTIONS on Information and Systems*, Vol. 99, No. 7, pp. 1877–1884 (2016).

[15] Morise, M.: D4C, a band-aperiodicity estimator for high-quality speech synthesis, *Speech Communication*, Vol. 84, pp. 57–65 (2016).

[16] 徳田恵一，小林隆夫，深田俊明，斎藤博徳， 今井聖：メルケプストラムをパラメータとする音声のスペクトル推定，電子情報通信学会論文誌 A， Vol. 74, No. 8, pp. 1240–1248 (1991).

[17] 匂坂芳典，浦谷則好：ATR 音声・言語データベース，日本音響学会誌， Vol. 48, No. 12, pp. 878–882 (1992).

[18] Kingma, D. P. and Ba, J.: Adam: A method for stochastic optimization, *arXiv preprint arXiv:1412.6980* (2014).

[19] Fan, Y., Qian, Y., Xie, F.-L. and Soong, F. K.: TTS synthesis with bidirectional LSTM based recurrent neural networks, *Fifteenth Annual Conference of the International Speech Communication Association* (2014).