Vol.2023-MUS-137 No.17
Vol.2023-SLP-147 No.17
2023/6/23

モーラ単位で高さを制御可能な音声デザインを前提とした日本語テキスト音声合成システムの試作

森勢 将雅[1,a)]

概要：本研究では，人間が大雑把なリクエストを与え，与えられた条件をある程度満たしつつ自然な音声を合成する音声デザイン（本プロジェクトでは，このようなデザイン法を「ビスポーク音声デザイン」として実現を目指している）に向けた取り組みを進めている．現在のテキスト音声合成（Text-to-Speech; TTS）技術は，Tacotron 2 等ですでに人間と等価な品質を実現しており，現在では表情豊かな発話や計算コストの削減など様々な方向で発展的な研究が進められている．本稿では，VOICEROID や VOICEVOX などの日本語 TTS システムにはモーラ単位でのピッチ操作機能が備わっていることに着目し，ピッチの制御性を重視した日本語 TTS システム「サーフィス」を提案する．具体的には，点ピッチパターンの考え方に着目し，モーラ単位で 7 段階のピッチ情報を与えることにより，大雑把なピッチ情報をリクエストできるシステムを試作した．本稿では実装の概要，および簡単に音質の検証をした結果を述べる．

1. はじめに

入力されたテキストを朗読する合成音声技術であるテキスト音声合成（Text-to-Speech; TTS）は，ニコニコ動画や YouTube などのコンテンツ制作において，幅広く利用されるツールとなった．合成音声そのものはスマートスピーカでの利用や駅などでのアナウンスなどでも利用されており，品質は Tacotron 2 [1] が提案された時点で人間の音声と等価な水準に達している．人間の音声と等価な品質を達成した昨今，より表情豊かな音声合成を目指し話速の制御ができる TTS [2] や，感情の制御 [3] 等が研究のターゲットにシフトしつつある．

スマートスピーカや自動応答等のアナウンスへの利用の場合，ユーザが感情表現や細かいイントネーションを制御する機能は必ずしも必要なく，人間と等価な音声を発話できることが重要となる．一方，動画コンテンツでの解説や実況へ利用する場合，ユーザはコンテンツの内容に合わせて細かく大きさやイントネーション，感情情報などを制御することで，コンテンツの完成度を高めている．筆者らは，このユーザが操作する「音声デザイン」を前提とした TTS について Human-in-the-loop speech design として検討を進めてきた [4].

現在は，この考え方を拡張し，ユーザが詳細まで制御するのではなく，大雑把な情報を与えることで計算機がそのリクエストに応える音声を生成するビスポーク音声デザインを目指している．本稿では，この機能を備えたプロトタイプである日本語 TTS システム「サーフィス」について報告する．

2. 関連研究と本研究の位置づけ

TTS の研究は言語に依存する側面があるため，英語と日本語とではアプローチが異なる．書籍 [5] でも示されているが，英語版 Tacotron 2 ではアルファベットをそのまま入力とする一方，日本語版 Tacotron 2 では音素系列とアクセント情報を入力する必要がある．加えて，Deep neural network (DNN) を用いた初期の TTS [6] では Vocoder を利用し波形生成するため，基本周波数などの音声パラメータを出力としてきた．一方現在は，DNN の出力はメルスペクトログラムとなり，Neural vocoder により波形生成する手法が主流である．Neural vocoder は日進月歩で数多く提案されており，HiFi-GAN [7] など 2020 年代に提案された Neural vocoder の多くは，すでに人間の音声と等価な品質での波形生成を実現している．

現状の TTS は End-to-End と呼ばれることもあるが，日本語の場合は入力テキストからアクセント情報を含む音素列へ変換する処理が必要である．また，音素系列から直接波形を生成するのではなく，Neural vocoder も必要不可欠である．日本語 TTS に関する情報を整理すると，

- 入力テキストからアクセント情報を含む音素系列を出力するシステム

[1] 明治大学
Meiji University, Nakano, Tokyo 164–8525, Japan
[a)] mmorise@meiji.ac.jp

- アクセント情報を含む音素系列からメルスペクトログラムを出力するシステム
- メルスペクトログラムから波形を生成するシステム

の 3 つを構築することが求められる．End-to-End を目指した方法では，文字列を画像として入力し波形を生成する方法なども検討されつつある [8]．

本研究では，この中で 3 番目以外について独自の機能を持つ日本語 TTS システムの構築を目指している．具体的に 2 番目については，アクセント情報の代わりに点ピッチパターン [9], [10] の考え方を取り入れる．日本語音声のイントネーションについては藤崎モデル [11] が有名であるが，点ピッチパターンは，単語音声の各母音のエネルギー重心点におけるピッチの系列がアクセントの本質を示しているという考え方である．点ピッチパターンの制御は，VOICEROID や VOICEVOX など既存の TTS ソフトウェアにおいて標準的に実装されているインタフェースとの親和性が高い．DNN を用いない信号処理的なピッチ制御ではなく，各モーラのピッチ情報を入力可能な TTS を実現することができれば，日本語 TTS ソフトウェアとして品質を高められる可能性がある．

筆者らは，音声デザインの考え方として詳細なピッチ軌跡を入力できる Human-in-the-loop speech design [4] を提案してきた．しかしながら，厳密なピッチ軌跡をデザインすることよりも，大雑把な点ピッチで入力した結果から自然な音声を生成できる方が，ユーザの作業コストを低減できる可能性がある．そこで，日本語版の Tacotron 2 をベースにモーラ単位で点ピッチに相当する情報を入力することにより，厳密なピッチ情報の入力を必要としない日本語 TTS システムの構築を目指す．以下では，このシステムを「サーフィス」と命名し，プロトタイプを実装した結果について説明する．

3. 提案する日本語 TTS システム「サーフィス」

本研究では，前節で述べたとおりモーラ単位でのピッチ情報を入力できる日本語 TTS システムの確立を目指している．本稿では，そのプロトタイプとして点ピッチパターンの特徴に着目した日本語 TTS システム「サーフィス」のコンセプトを説明し，プロトタイプ実装した結果について説明する．

3.1 システム全体の構成

図 1 に，提案するサーフィスの全体像を示す．サーフィスでは，入力テキストとして漢字混じりの全角文字を想定している．ただし，Text analysis には新規性は無く，メルスペクトログラムからの波形生成も同様で，Kong らが提案した HiFi-GAN [7] をそのまま利用している．サーフィスのオリジナル部分は，

(1) 入力テキストと音素系列からピッチ情報を予測する Surface Pitch
(2) 音素系列とピッチ情報からメルスペクトログラムを予測する Surface Core

の 2 点となる．Surface Pitch で予測対象となるピッチ情報は，学習に用いる音声データベース（DB）を対象に，以下の手順で求めることとした．

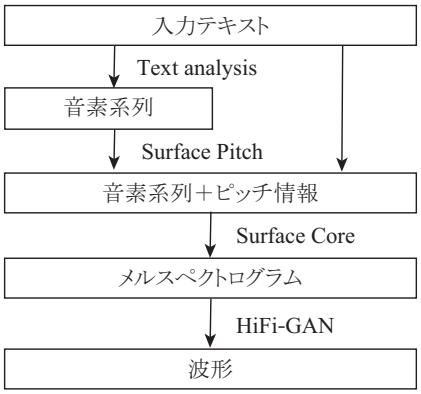

図 1 提案するサーフィスの構造．本稿では Surface Pitch についての詳細は割愛する．

はじめに，音声 DB 全体に対し Harvest [12] により基本周波数を推定する．ここで，基本周波数 f は以下の式によりメル尺度に変換する．

$$m(f) = 1127.01048 \log\left(\frac{f}{700} + 1\right). \tag{1}$$

次いで，音声 DB のテキスト情報に基づいて自動ラベル付けを実施する．自動ラベル付けされたモノフォンラベルから母音と「ん」に相当する音素を検索し，各音素区間の中央の時刻の基本周波数をそのモーラにおけるピッチ情報とする．ラベル情報の推定には多少の誤差が含まれることを勘案し，点ピッチパターンのように母音の重心となる時刻を厳密には算出しない．モノフォンラベルにおいて母音の開始・終了時刻から中央となる時刻を算出し，その時刻の基本周波数をモーラの点ピッチとしている．上記の処理により，各モーラに対する基本周波数の分布が得られるため，メル軸に変換した分布に対し平均が 0，標準偏差が 1 となるよう標準化する．

ここまでの処理により得られた結果が，Surface Pitch においてモデル学習する対象となる．実際に Surface Core に入力する際には，推定されたメル基本周波数をヒストグラムに基づいて 7 段階に量子化した結果を用いる．

3.2 ピッチ情報を予測する Surface Pitch

本稿では詳細を割愛するが，サーフィスは日本語に対応した Tacotron 2 [5] の実装を参考に実装している．ピッチ情報の予測では，入力するモーラ数と出力するモーラ数を揃えるため，Tacotron 2 の中でもエンコーダと Post-Net に相当する部分のみ利用することとした．

Surface Pitch ではモーラ毎のピッチ情報を推定する必

要があることから，音素列をモーラ列に変換している．具体的に，1 モーラにつき 48 次元のベクトルとして，以下のように 0 か 1 のフラグを立てるようにした．48 次元[*1]の内訳は以下のとおりである．

- 1-43 次元：音素情報
- 44 次元：子音無しに相当する情報
- 45 次元：疑問文に対応する情報．文末のモーラに 1 を与える
- 46-48 次元：そのモーラがひらがな（46 次元），カタカナ（47 次元），漢字・その他記号（48 次元）のどれかの分類

例えば「夏季」の「か」であれば，/k/と/a/に相当する次元と 48 次元が 1 となる．「いくら」の「い」の場合は，/i/に相当する次元と 44，46 次元が 1 となる．Surface Pitch の特色は，モーラが平仮名・カタカナ・漢字や記号の何に由来するかを 46-48 次元の入力情報として扱うことである．

このように得られた入力ベクトルに対し，以下で構成された DNN により学習する．

- フィルタサイズが 3 でフィルタ数が 512 の 1 次元畳み込み層・バッチ正規化層・ReLU 層・ドロップアウト層のセット× 3
- 隠れユニット数 512 の双方向 LSTM 層・バッチ正規化層・ReLU 層
- フィルタサイズが 3 でフィルタ数が 512 の 1 次元畳み込み層× 4
- フィルタサイズが 3 でフィルタ数が 1 の 1 次元畳み込み層

Surface Pitch のモデル構造は，双方向 LSTM までは Tacotron 2 におけるエンコーダに相当する処理，および双方向 LSTM より先は Post-Net をベースに決定している．ただし，双方向 LSTM 以降ではバッチ正規化層，ReLU 層，ドロップアウト層を含めていない点で違いがある．

Tacotron 2 では音素単位で入力し畳込み層のフィルタサイズを 5 に設定しているが，Surface Pitch ではモーラ単位で入力しているため，フィルタサイズは 3 とした．同様に，各フレームにつき 80 次元のメルスペクトログラムが Tacotron 2 の推定対象であるが，Surface Pitch では点ピッチに相当する基本周波数が推定対象のため，最後の層のフィルタ数は 1 にしている．「っ」やポーズへの対応など，今回の実装における細かい注意点は 3.4 節で説明する．

3.3 音素系列とピッチ情報からメルスペクトログラムを予測する Surface Core

Surface Core のモデル構造は，入力の音素系列情報の内訳が異なること以外，日本語版 Tacotron 2 と同一である．ただし，Surface Core では 7 段階のピッチ情報を各モーラに対して入力するため，日本語版 Tacotron 2 で用いていたピッチの上がり位置，下がり位置（アクセント核），アクセント句境界，疑問形の文末表現が不要となる．書籍 [5] によると，日本語版 Tacotron 2 の語彙数は 52 であるが，Surface Core の語彙数はモーラ数やアクセント情報等の違いがあり語彙数は 53 となる．

日本語版 Tacotron 2 における語彙数の内訳は，音素数が 44 (空白を示す pau と sil を含む）と，それに特殊記号として文頭，文末，文末（疑問系），ポーズ，アクセント句境界，ピッチの上がり位置，ピッチの下がり位置，パディングの 8 種類からなる．Surface Core では，音素数が 43，パディングと文頭・文末，および 7 段階のピッチ情報の 10 種類を足した 53 が語彙数となる．音素数が一致しない理由は，後述する音声 DB で出現する音素から音素数を決定したため，出現しない音素を除いたことや「ふゅ」の子音に相当する音素/fy/などを加えたことによる．各モーラについて，有声音であれば 7 段階のピッチ情報を母音の次に挿入し，無声音・無音の場合は何も挿入せず次のモーラの情報に繋ぐこととした．

3.4 プロトタイプ実装

ここでは，実装に関する細かい点について説明する．音声 DB については，プロ声優が ROHAN コーパス [13] の 4,600 文を朗読した No.7 音声 DB [14] の発話から，Normal (通常の発話) を利用した．本音声 DB のサンプリング周波数は 96 kHz であるが，Surface では HiFi-GAN の実装を利用するため 22.05 kHz にダウンサンプリングしている．

Surface Pitch ではミニバッチサイズを 32 とし，各ミニバッチを構成する入力テキストのモーラ数は統一した．学習に用いた文の数は，各モーラ数の文章がミニバッチサイズの倍数に設定する条件を満たす 4,128 に設定した．また，エポック毎に同一モーラ数の文についてシャッフルし，出現するモーラ数の順序もシャッフルした上で学習した．入力データは，各ミニバッチの文末に 32 モーラ分のパディングを加えている．エポック数は 800 としたため 1 エポックにつき 129 ステップとなり，合計 103,200 ステップの学習となる．最適化手法には，ADAM を学習率 0.001，β_1 が 0.9，β_2 を 0.99 に設定して用いた．学習率は，開始から終了まで変化させていない．

学習データには，ポーズを示す pau や無声化した母音など，基本周波数を持たない場合もある．これらについては，前後のモーラの基本周波数を線形補間することで与えることとした．これは，無声音に相当する特殊な数字，例えば STRAIGHT や WORLD などの Vocoder で無声音に対応する 0 にした場合は，有声音・無声音の切り替え前後で急峻な変化を生み，推定誤差の拡大に繋がると考えられるためである．推定された無声音・無音の基本周波数は，Surface Core の入力データの際には無視している．

[*1] Surface Pitch は MATLAB で実装したため，1 から 48 次元で表現している．

Surface Core のハイパーパラメータは，以下を除いて日本語版 Tacotron 2 [5] と同一である．学習に用いた音声 DB が違うため，学習データを 4,000 文，検証データを 400 文，テストデータを 200 文とした．ミニバッチサイズを 32 としエポック数を 800 としたため，合計ステップ数は 100,000 ステップとなる．入力音声のサンプリング周波数は 22.05 kHz とし，メルスペクトログラムを算出するフィルタバンクにおける中心周波数の下限を 125 Hz，上限を 10.5 kHz に設定した．次元数は，書籍と同様で 80 次元としている．中心周波数の下限は入力音声の基本周波数の下限に基づいており，上限は最高域のフィルタバンクの上限がナイキスト周波数に届かない数値として設定している．フレームシフトは，後述する HiFi-GAN が 256 サンプルのメルスペクトログラムを入力とすることから，Surface Core も同様に 256 サンプル（11.61 ms）とした．

最後に，HiFi-GAN については，以下の条件を除き公開されている実装 [15] の V1 の条件を利用することにした．具体的に，メルスペクトログラムの上限・下限周波数は Surface Core の条件に併せるため，Hifi-GAN の実装（0 Hz，8,000 Hz[*2]）とは異なっている．学習データ数も論文中の実装より少なく，Surface Core と同様に 4,000 文に対して 3,100 エポック学習することとした．ミニバッチサイズが 16 のため，合計 775,000 回の学習ステップ数となる．これは，論文 [15] で示された学習ステップ数が約 2,500,000 回であることに対し，約 30%の回数である．

4. 有効性に関する議論

ここでは，サーフィスの有効性について議論する．現時点では効果的な評価法について検討中のため，筆者が聴取した範囲での定性的な解析結果を述べる．

4.1 HiFi-GAN 単独の音質

HiFi-GAN については，論文の著者らが学習に利用したデータよりも学習データ量が少ない点が問題であった．しかしながら，学習中のロスを観察すると，論文の実装より学習ステップ数が 1/3 程度でありながら，著者らが Web で公開しているロスの値よりも 1 割程度小さい値になっていることが確認できた．これは，メルスペクトログラムの帯域を変更したことの影響と考えられる．論文中の実装では，人間の音声の基本周波数では一般に想定しなくても良い低域までカバーしており，その帯域の推定誤差が悪影響を及ぼしている可能性が示唆される．

出力された音声を聴取したところ，22.05 kHz の元音声と比較して概ね劣化が知覚されない程度の音質が達成できたと判断している．この原因は，メルスペクトログラムを算出する中心周波数の上限を論文で用いた 8000 Hz より高い周波数に設定していることが影響していると考えられる．下限についても，事前に話者の発話から基本周波数を算出し，最低となる基本周波数より低い 125 Hz に設定していることも寄与している可能性がある．上記については，ランダムに数サンプル切り出して聴取した定性的な評価であるため，今後は主観評価による品質評価を実施する必要がある．

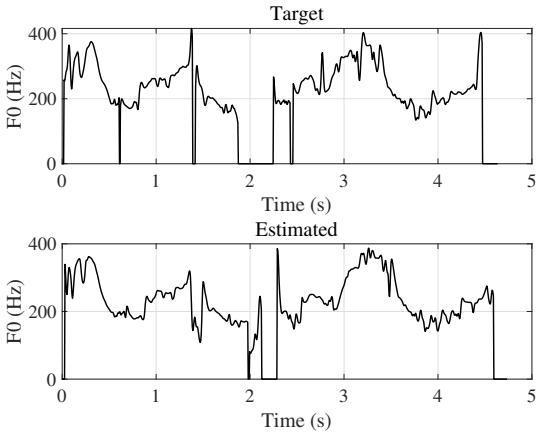

図 2 元音声の基本周波数（上段）と Surface Core で得られたメルスペクトログラムから生成した合成音声の基本周波数（下段）の軌跡.

4.2 生成されたメルスペクトログラムからの波形生成

Surface Core により生成されたメルスペクトログラムから HiFi-GAN により生成した波形を聴取した結果，元音声のイントネーションを概ね再現できていることが確認できた．一方，この評価を客観的な指標で評価することは容易ではない．図 2 は，元音声の基本周波数と Surface Core と HiFi-GAN により生成された合成音声の基本周波数の軌跡を示している．

目視では大局的な構造がある程度類似しているように見えるが，有声区間が一致しないことや，基本周波数推定では有声・無声の境界付近での値が大きく変動することもあるため，単純な RMS 誤差で評価することが適切とは言い難い．メルスペクトログラムの誤差を評価する場合でも同様のため，客観評価法については今後検討する必要がある．現状ではテスト音源について定性的な評価をしているにとどまるが，発話の類似性に関する主観評価についても同様に必要となる．

4.3 残された課題

ここまでの議論は，Surface Core と HiFi-GAN に関するものであり，Surface Pitch については，現時点では評価対象とはしていない．Surface Pitch の評価は，音声デザインを前提としている以上完璧であることは必要ではないこと

[*2] 0 Hz は中心周波数が最低となるフィルタの下限，8,000 Hz は中心周波数が最高となるフィルタの上限であり，中心周波数の上限・下限ではない.

から，目標設定を別途考える必要があるためである．入力テキストに対しアクセントに相当する情報を必要としないことは大きなアドバンテージであるが，その分単語レベルで全く異なるイントネーションを生成する可能がある．この場合，発話全体のイントネーションには大きな破綻はなくとも，特定単語部分での違和感が強調させるなどの問題が想定される．

現時点では，学習データに対する 7 段階のピッチ区分の正答率は約 85%程度であったが，前述の理由によりテストデータに対する正答率は低い状況にある．ただし，これは一部単語でイントネーションが根本的に違うなどの問題があるため，最終的に出力される合成音声の品質が低いことを必ずしも意味しない．加えて，音声デザインが前提であることから，「初期値」として誤差があることそのものが問題であるかについても議論する余地がある．最終的な評価は音声デザインとして包括的に実施する必要があり，この具体的な実験法の確立についても今後の重要な課題となる．

5. おわりに

本稿では，TTS のコンテンツ利用においてピッチ操作が行われることに着目し，ピッチ操作を前提とした日本語 TTS システム「サーフィス」の試作結果を報告した．Surface Core により，音素系列＋ピッチ情報からメルスペクトログラムを推定し，得られた結果から HiFi-GAN で生成した波形は，アンオフィシャルな評価では良好な品質であることを確認した．

今後の課題には，上記の品質評価を公式に実施し，どの程度元音声の発話のイントネーションを再現できているかの検証が第一に挙げられる．また，入力テキストと音素系列からピッチ情報を予測する Surface Pitch を含む全体の評価が未実施であることから，音声デザインの機能まで含めたサーフィス全体の評価法の設計と実施が重要な検討項目である．

謝辞 本研究の一部は，JSPS 科研費 JP21H04900, JP21K19794 の支援を受けました．

参考文献

[1] J. Shen et al., “Natural TTS synthesis by conditioning WaveNet on Mel spectrogram predictions,” in Proc. ICASSP 2018, pp. 4779–4783, 2018.
[2] Y. Ren et al., “FastSpeech: Fast, robust and controllable text to speech,” in Proc. NIPS’19, pp. 3171–3180, 2019.
[3] X. Luo et al., “Emotion-controllable speech synthesis using emotion soft labels and fine-grained prosody factors,” in Proc. APSIPA ASC 2021, pp. 794–799, 2021.
[4] D. Kondo and M. Morise, “Human-in-the-loop speech-design system and its evaluation,” in Proc. APSIPA ASC 2019, pp. 608-612, 2019.
[5] 山本龍一，高道慎之介，“Python で学ぶ音声合成，” 株式会社インプレス，2021.
[6] H. Zen et al., “Statistical parametric speech synthesis using deep neural networks,” in Proc. ICASSP 2013, pp. 7962–7966, 2013.
[7] J. Kong et al., “HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis,” in Proc. NeurIPS 2020, pp. 1–12, 2020.
[8] 中野他，“画像文字からの音声合成，” 言語処理学会第 28 回年次大会, 2022.
[9] 橋本新一郎，“日本語単語アクセントの諸性質，” 信学論 (D)，vol. J56-D, no. 11, pp. 654–661, 1973.
[10] 加藤他，“母音部エネルギー重心点に着目した日本語リズム規則，” 音響誌，vol. 50, no. 11, pp. 888–896, 1994.
[11] 藤崎博也，須藤寛，“日本語単語アクセントの基本周波数パタンとその生成機構のモデル，” 音響誌，vol. 27, no. 9, pp. 445–453, 1971.
[12] M. Morise, “Harvest: A high-performance fundamental frequency estimator from speech signals,” in Proc. INTERSPEECH 2017, pp. 2321–2325, 2017.
[13] 森勢将雅，“ROHAN：テキスト音声合成に向けたモーラバランス型日本語コーパス，” 音響誌, vol. 79, no. 1, pp. 9–17, 2023.
[14] https://voiceseven.com/7rdev/login.php
[15] https://github.com/jik876/hifi-gan