TalkingHead: ユーザによる音素継続長と高さの制御を前提としたテキスト音声合成インタフェース

近藤 大地 * 森勢 将雅 †‡

概要. 深層学習に基づくテキスト音声合成技術は，2013 年以降多くの研究事例が蓄積しており，人間の音声と等価な品質の合成音声を生成することが可能になってきた．品質の向上に伴い，合成音声はスマートスピーカなどの幅広い領域で利用されている．一方，音声生成を創作活動の一環と捉えると，音声の品質を劣化することなく自由にデザインする技術には需要があると考えられる．この際，ユーザにとって，音質劣化を抑えた音声パラメータのデザインは困難である．本研究の目的は，深層学習が人間の作業結果の品質向上をサポートする技術の獲得を目指し，音質劣化を抑えた音声デザイン手法を実現することである．この手法を，Hunan-in-the-loop（HITL）型音声デザイン手法として提案してきた．本稿では，このデザイン手法を組み込んだインタフェース TalkingHead について示す．TalkingHead では，テキスト音声合成において各音素のタイミングと基本周波数を人間が操作し，操作結果から相対的に音質の良い音声パラメータを生成して波形を生成する機能を有する．

1 はじめに

テキスト音声合成技術は，合成音声を用いたコンテンツに幅広く利用されている．2013 年以降，深層学習によるテキスト音声合成の手法も多く研究されており，deep neural network（DNN）を用いた統計的音声合成 [1] や，Tacotron [2] といった End-to-End 音声合成が主流となってきている．これらの研究では，音声の加工が必要ない程高品質な音声を生成することを目的としており，既に人間の声と遜色ない品質で音声生成がなされている．一方で，動画共有サービスでは合成音声を加工したユーザ独自の表現が数多くみられることから，音声を自由に加工することで個性を演出したいと考えるユーザも一定数いることが考えられる．しかし，ユーザにとって，音質劣化を抑えた音声パラメータのデザインは困難である．例えば，音声を合成後にイントネーションを局所的に変換すると，その部分において音質が劣化しやすい．

本研究の目的は，そのような劣化を防ぐために，ユーザの音声デザイン結果をもとに DNN が劣化の少ない音声パラメータを生成する手法を実現することである．本研究の新規性は，人間のデザイン作業を DNN がサポートすることによる，音質劣化を抑えた音声デザイン手法を提案する部分にある．この手法および開発したインタフェースを Human-in-the-loop（HITL）型音声デザイン手法と定義する．

Copyright is held by the author(s).

* 山梨大学
† 明治大学
‡ JST さきがけ

2 提案手法

音声パラメータをデザインする際に音質劣化が発生する原因の 1 つとして，声の高さを表す基本周波数（F0）と音色を表すスペクトル包絡の間における相互作用が挙げられる．スペクトル包絡が変更されることなく F0 軌跡のみがデザインされた結果，F0 とスペクトル包絡間の相互作用が消失し，品質が劣化すると考えられる．

上記の問題に対し，過去の検討 [3] では，デザインした F0 軌跡から再度 F0 軌跡を生成する手法として，HITL 型 F0 デザイン手法を提案した．DNN 音声合成では，アクセント依存の音響モデルを用いてテキストから F0 を生成する．HITL 型 F0 デザイン手法の実装にあたり，上記の手法ではデザインした F0 軌跡とアクセント情報が矛盾してしまうという問題があった．これに対し，提案手法では図 1 に示すように，アクセント非依存の言語特徴量を用いて音響モデルを生成した．これにより，上記の矛盾を防ぎつつ，F0 とスペクトル包絡間の相互作用を保つような音響モデルの学習を行った．本研究では，この機能を持つ音声デザインインタフェースを提案する．

3 インタフェース

提案手法の実装にあたり，必要な機能は以下のとおりである．

- テキストから合成音声を生成可能であること
- 発話時間および F0 軌跡を直感的にデザイン可能であること
- HITL 型音声デザイン手法によって，デザイン結果をもとに音声を生成可能であること

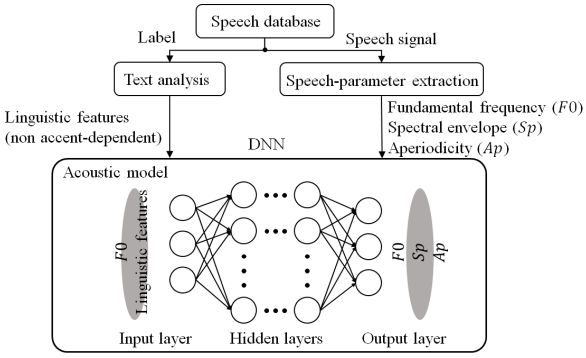

図 1. 提案手法における音響モデルの生成方法

提案手法をもとに，音声デザインインタフェース TalkingHead を開発した．TalkingHead の画面を図 2 に示す．図 2 の上部はテキスト音声合成部であり，下部は音声デザイン部である．

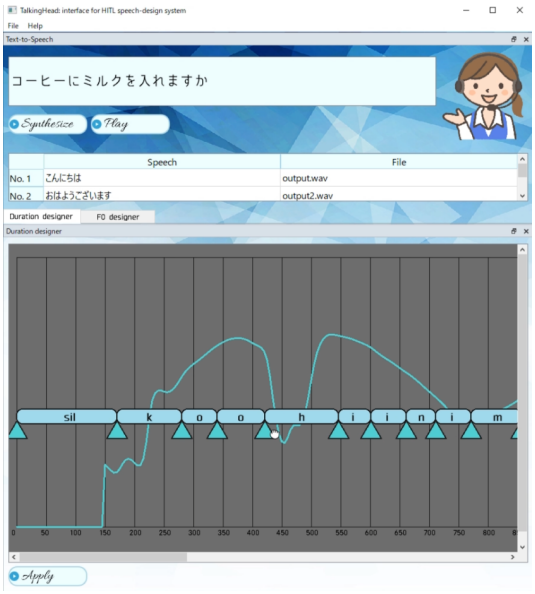

図 2. TalkingHead のインタフェース画面

3.1 テキスト音声合成部

テキスト音声合成部では，テキストボックス内に任意のテキストを入力した後に Synthesize ボタンを押すことで，テキストに対応した合成音声が生成される．このとき，システム内ではテキストからラベルが生成され，ラベルをもとに継続長モデルと音響モデルから音声が生成される．これらのモデルは一般的な DNN 音声合成のモデルと同等の学習を行ったものである．

合成音声が生成されると同時に，音素継続長と F0 軌跡が音声デザイン部に表示される．Play ボタンを押すことで，生成された音声を再生することができる．これまでに生成された合成音声の発話テキストとファイル名は，ボタンの下部に表示される．

3.2 音声デザイン部

音声デザイン部では，音素継続長と F0 をそれぞれデザインするためのタブが実装されている．

3.2.1 音素継続長デザインタブ

音素継続長デザインタブでは，合成音声が持つ各音素の開始時間と終了時間に対応するカーソルおよび F0 軌跡が表示される．デザイン部の横軸と縦軸はそれぞれ時間と周波数を表す．デザインしたい合成音声に対して，表示されたカーソルを横方向にドラッグ操作をする．これにより音素の継続長を操作することで，発話時間をデザインすることができる．

3.2.2 F0 デザインタブ

F0 デザインタブでは，音素継続長デザインタブと同様に F0 軌跡が表示される．表示された F0 軌跡をペンやマウスでドラッグ操作することで，ユーザは F0 軌跡を描くことができる．

3.3 TalkingHead の有効性に関する考察

過去に実施した主観評価実験 [3] から，単に F0 を変更するだけの場合と比べ音質は改善できることが想定される．また，ペンやマウスを用いて視覚的に音声デザインが可能な本インタフェースでは，数値の入力やスライダーによる制御を用いる場合と比べて，直感的なデザインが可能であると考えられる．一方，極端な値となるデザイン結果では，逆に品質が劣化してしまう可能性がある．そのため，極端な値とならないよう自動調整する機能の実装や，極端なパラメータからでも劣化が生じないような音響モデルの改良が必要であると考えられる．

4 おわりに

本稿では，人間がデザイン過程に介入することを前提とした HITL 型音声デザインを組み込んだインタフェースとして TalkingHead を提案し，有効性について論じた．今後の課題として，インタフェースや音響モデルの改良および，ユーザビリティ評価をすることが挙げられる．

謝辞

本研究は JST さきがけ JPMJPR18J8 の支援を受けて実施された．

参考文献

[1] H. Zen et al. Statistical parametric speech synthesis using deep neural networks. in Proc. ICASSP 2013, pp. 7962–7966, 2013.

[2] Y. Wang et al. Tacotron: Towards end-to-end speech synthesis. arXiv preprint arXiv:1703.10135, 2017.

[3] D. Kondo and M. Morise. Human-in-the-loop speech-design system and its evaluation. in Proc. APSIPA ASC 2019, 2019, [Accepted].