End-to-End モデルに基づくインクリメンタル音声合成*

☆柳田 智也（NAIST）, サクティ サクリアニ，中村 哲（NAIST/RIKEN AIP）

1 はじめに

同時音声翻訳システムは，元言語の音声を目標言語の音声へ逐次翻訳し，音声として逐次出力する．このシステムは，音声認識・機械翻訳・音声合成により構築される．通常，機械翻訳および音声合成は，元言語の文全体を入力後に処理を行うため，出力に深刻な遅延が発生する．講義のように一発話が長い状況で本システムを使用する場合，聴衆者は講義の理解に大きな支障を生じてしまう．従って，各要素は入力を逐次処理し，出力する機能が必要である．逐次的な音声合成として，Hidden Markov Model（HMM）に基づくインクリメンタル音声合成が提案されている．HMM インクリメンタル音声合成では，未だ十分な音声品質を確保できず，言語依存の処理を持つため他言語適応時に負担が増加する．本研究では，End-to-End 音声合成を用いた，高品質かつ言語依存の少ないインクリメンタル音声合成の実現を目指す．

2 関連研究

2.1 HMM に基づくインクリメンタル音声合成

通常の HMM 音声合成では，まず，入力された文を解析し言語特徴（音素表記や単語の品詞タグ，それら位置関係等）を抽出する．次に，言語特徴から HMM 系列を構築し音響特徴を生成する．その後，音響特徴からボコーダにより音声を合成する．インクリメンタル音声合成は，文の入力終了前に出力を得るため，文より短い合成単位で処理を行う．その結果，一部言語特徴（後続の品詞タグ等）が未知となる．更に，音響特徴は後続音声の変化を考慮不可である．これら要因により，HMM インクリメンタル音声合成の品質は通常の HMM 音声合成と比較して劣化する．品質の改善方法としては，未知の言語特徴の置換や，未知の言語特徴が存在する場合の学習方法の提案，言語特徴を予測し使用する方法がある [1, 2, 3].

上記の HMM インクリメンタル音声合成は，以下の問題を持つ.（1）言語特徴抽出・継続長モデル・音響モデル・ボコーダ各要素の誤差が伝搬し音声品質が低下する．また，通常の HMM 音声合成品質を上限と仮定するため，より高品質な音声が生成できない.（2）言語特徴抽出が言語依存処理であり他言語適応の負担が増加する．

同時音声通訳システムは，人対人のコミュニケーションを想定しており，より高品質なインクリメンタル音声合成が求められ，更に，同時通訳システムの他言語適用コスト低減のため，インクリメンタル音声合成の言語特徴設計コストの削減が求められる．

2.2 End-to-End 音声合成

言語依存の影響を逓減し高品質な音声を生成するため，深層学習に基づく End-to-end 音声合成が近年提案されている [4, 5, 6]．これらのモデルは深層学習によるエンコーダデコーダモデルに基づいており，入力は表層単語を用いるため言語特徴抽出を行わない．従って，言語依存部分の設計負担が低下する．更に，言語特徴抽出・継続長モデル・音響モデルを一つのエンコーダデコーダで表現し，各モデルでの誤差伝搬を逓減させ，より高品質な音声を生成可能とした．

3 End-to-End インクリメンタル音声合成の課題

End-to-End 音声合成によるインクリメンタル音声合成は，未だ実現されていない．実現のため，次の課題に取り組む必要がある．まず，音声品質を保持可能な入力長が不明であり，End-to-End インクリメンタル音声合成において，適切な合成単位の調査が必要である．次に，End-to-End 音声合成では，一部後続音声の変化を HMM モデル程容易に考慮できない．HMM インクリメンタル音声合成では，後続音声を考慮するため，現入力と後続入力とを結合して合成する方法が提案されている [7]．HMM インクリメンタル音声合成の場合，継続長モデルから現入力部分を判別可能である．しかし，End-to-End 音声合成は，モデル自体が自動的に音響特徴と継続長とを対応付けるため，所望の音声区間を容易に出力できない．

本論文は，上記の課題の内，音声品質を保持可能な合成単位について検討する．特に，日本語を対象とする．筆者らは，HMM インクリメンタル音声合成において，日本語ではアクセント句単位が言語特徴及び合成単位として有効であることを示した [8]．上記を踏まえて，本研究では，日本語における End-to-End インクリメンタル音声合成のため，アクセント句単位のインクリメンタル音声合成を行い，評価実験から検討事項を明確化する．

---

*Incremental speech synthesis based on End-to-End model, by YANAGITA, Tomoya, SAKTI, Sakriani, NAKAMURA, Satoshi (Nara Institute of Science and Technology/RIKEN AIP).

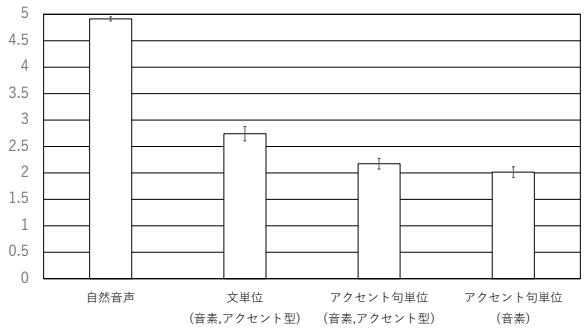

Fig. 1　入力単位と言語特徴による主観評価結果

4　End-to-End インクリメンタル音声合成の評価実験

4.1　実験条件

コーパスとして，JSUT を用いる [9]．音声とテキストからフォースアライメントにより音素継続長を取得し，文頭と文末の無音区間を除去する．フォースアライメントにより継続長が取得できた 7530 文を学習用に，71 文を開発用とし，71 文をテスト用として使用し，End-to-End 音声合成として Tacotron を用いる [4]．モデルには [5] と同様の機構として，コンテキストベクトルを入力とし，音響特徴の出力停止を予測する一層のフィードフォワード層をデコーダに追加する．出力層の活性化関数は Sigmoid 関数を用いて，Binary Cross Entropy を損失関数に用いて学習する．学習時のバッチサイズは 16 である．モデルへの入力は，音素キャラクタ (ポーズ・未知語・文頭文末記号を含む 46 個) を用いる．韻律を考慮するためアクセント句のアクセント型も用いる．その他の言語特徴は，言語依存を逓減するため使用しない．学習は文単位で行い，合成時は文単位及びアクセント句単位で合成する．アクセント句単位とアクセント型は，テキストから OpenJTalk により抽出する．主観評価は以下の条件と自然音声とを用いる．

・文単位合成，音素とアクセント型を入力
・アクセント句単位合成，音素とアクセント型を入力
・アクセント句単位合成，音素を入力

主観評価として自然性に関する MOS テストを行う．評価者は日本語母語者 16 名で，1 評価者 1 条件辺り 15 音声を使用し，音声は再度聴取可能とした．

4.2　実験結果

実験結果を Fig. 1 に示す．自然音声が最も高い評価を得ており，文単位の End-to-End 音声合成の自然性は，自然音声よりおよそ 2 から 2.5 ポイント悪い．この結果より，ベースラインとなる文単位の日本語 End-to-End 音声合成の性能向上に取り組む必要がある．End-to-End 音声合成の性能が悪化した理由としては，データセットは約 10 時間の音声で構築され，モデル学習に対して不十分な可能性が考えられる．更に，文単位合成とアクセント句単位合成（音素とアクセント型を入力）を比較すると，アクセント句単位への変更による自然性の劣化は約 0.56 であり，自然音声と文単位の場合と比較して差は小さい．この原因として，アクセント句間の音声が不平滑となり自然性を低下させた可能性がある．評価結果より，これら要因を検討する必要があることが明確となった．

5　おわりに

同時通訳システム実現のため，高品質かつ言語依存の少ない End-to-End インクリメンタル音声合成の実現を目指す．実現に向け検討事項を明確化するため，End-to-End インクリメンタル音声合成の品質を評価した．今後，End-to-End 音声合成の品質向上を検討し，その後，合成単位間の音響特徴の時間変動を考慮し，自然性の改善を検討する．

参考文献

[1] Baumann Timo, ”Decision tree usage for incremental parametric speech synthesis.” Proc. ICASSP, pp. 3819-3823, 2014.

[2] Pouget, *et al.* ”HMM training strategy for incremental speech synthesis,” Proc. Interspeech, pp. 1201-1205, 2015.

[3] Pouget, *et al.*, ”Adaptive Latency for Part-of-Speech Tagging in Incremental Text-to-Speech Synthesis,” Proc. Interspeech, pp. 2846-2850, 2016.

[4] Wang *et al.*, ”Tacotron: Towards End-to-End Speech Synthesis,” Proc. Interspeech, pp. 4006-4010, 2017.

[5] Shen *et al.*, ”Natural tts synthesis by conditioning wavenet on mel spectrogram predictions,” Proc. ICASSP, pp. 4779-4783, 2018.

[6] Tachibana *et al.*, ”Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention,” Proc. ICASSP, pp. 4784-4788, 2018.

[7] Baumman Timo, SCHLANGEN David, ”Evaluating prosodic processing for incremental speech synthesis,” Proc. Interspeech, pp. 438-441, 2012.

[8] Yanagita *et al.*, ”Incremental TTS for Japanese Language,” Proc. Interspeech, pp. 902-906, 2018.

[9] 園部 他，”JSUT コーパス：End-to-End 音声合成に向けたフリーの大規模日本語音声コーパス,” 日本音響学会 2018 年春季研究発表会講演論文集, 1-Q-37, 2018.