Vol.2017-SLP-119 No.24
2017/12/22

日本語インクリメンタル音声合成システム実装のための言語特徴の検討

柳田 智也[1,a)]　Sakriani Sakti[1]　中村 哲[1]

概要：同時音声翻訳システムは，話者が発話を終える前に翻訳を行い音声を生成する．その目的は，高品質な翻訳を生成すると同時に，翻訳処理の待ち時間を最小化することである．そのため，同時音声翻訳のための音声合成（TTS）システムでは，テキストが入力される間に音声を合成する機能が必要である．しかしながら，従来の TTS における音声合成は，テキスト入力が終了するまで待たなければならない．その理由は，文からコンテキストと呼ばれる言語特徴を抽出し，高品質の音声を合成するためである．この制約のために，TTS の合成単位は文で固定されている．一方で，インクリメンタル TTS（ITTS）と呼ばれる TTS が存在する．ITTS は文全体のコンテキストを使用せずに，文より小さい合成単位で音声を合成する．従って，同時音声通訳に応用できると考えられる．しかしながら，ITTS における多くの研究は西欧圏の言語で行われている．高低アクセント及びモーラ単位である日本語において，ITTS は未だ実現されていない．日本語 ITTS 実現のため，本研究では，制限されたコンテキストと合成単位を選択して合成音声の品質を調査する．その結果，音声品質と合成単位のトレードオフとして，アクセント句が適していることが確認できた．

キーワード：インクリメンタル音声合成，HMM based TTS，コンテキスト

1. はじめに

同時音声翻訳システムは，自動音声認識（ASR），機械翻訳（MT），音声合成（TTS）から構成される．従来法では，話者が話している間に ASR が実行され，その後，MT と TTS が文単位毎に実行される．講義のように話者の発話が非常に長い場合，従来法では長い遅延が生じてしまう．この問題を解決するため，話者が話している間に，翻訳を行う同時翻訳に関する研究が行われている．このように，リアルタイムに動作する ASR，MT，TTS の実現は必要である．本研究は，それらの中でも TTS システムの実現に焦点を当てている．

TTS では，音声特徴を推定するために多くのコンテキストが使用される．隠れマルコフモデル（HMM）に基づく TTS では，次の 3 つの処理により実行される．(1) 入力された文を解析し，コンテキスト（音素の位置や，単語の品詞タグ等）を抽出する．(2) コンテキストから文に関する HMM 系列を構築して，音声特徴が滑らかに変化するように動的特徴を用いて大域的最適化を行い，音声特徴を生成する [1], [2]．(3) 生成された音声特徴からディジタルフィルタにより音声を合成する．従来の TTS は，文全体からコンテキストを抽出するために，文単位での合成を行う必要がある．一方で，文が入力され終わる前に，合成を行うインクリメンタル音声合成（ITTS）と呼ばれる方法が提案されている．ITTS は，文が入力され終わる前に音声を合成することを仮定している．従って，文より短い単語等の合成単位で合成を行う．その結果，TTS より速い動作が期待される．しかしながら，テキストが入力されながら，合成を行うことを仮定するため，部分的なコンテキストから音声特徴を生成する必要がある．特に，上記した TTS の動作と比較して，(1) の処理では，現時刻の音声特徴を生成する際に，いくらかのコンテキスト（後続の品詞タグ等）が未知となる．更に，(2) の処理では，制限された HMM 系列のみに対して最適化を行い，音声特徴を生成する必要がある．これらの要因により，より自然な音声特徴の生成が困難となり，合成音声の品質は劣化する．

ITTS の品質を改善するための方法はいくらか提案されている．Baumann らは，英語やドイツ語に対して，初めて未知のコンテキストが，音声特徴に与える影響を調査した [3], [4]．Pouget らは，未知のコンテキストによる HMM に基づく ITTS の学習方法を提案している [5]．更に，次の単語の品詞タグを予測して，コンテキストとして使用する

[1] 奈良先端科学技術大学院大学
[a)] yanagita.tomoya.yo8@is.naist.jp

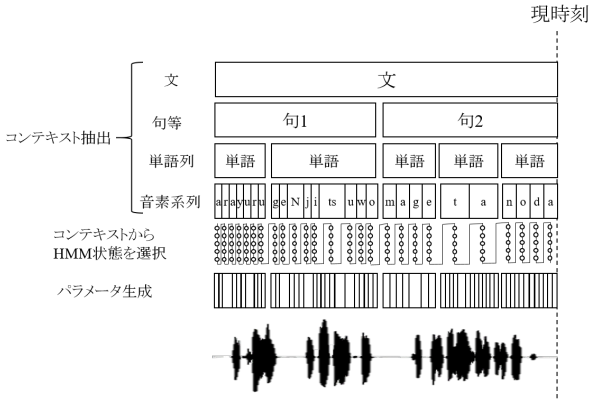

図 1 従来の TTS による音声合成

方法を提案している [6]．上記の方法により，ITTS 品質は改善されている．しかしながら，多くの ITTS の研究は，強勢アクセントのような西欧圏の言語に焦点が当てられている．TTS で使用されるコンテキストは言語により異なる [7]．そのため，モーラ単位・高低アクセントである日本語 ITTS 実現のためには，未知のコンテキストによる合成音声への影響を調査する必要がある．従って，日本語 ITTS において，未知のコンテキストによる合成音声への影響を調査することを目的とし，実験を行う．

2. 日本語 TTS におけるコンテキスト

Yokomizo らは，日本語の TTS において，アクセントに関する情報が，韻律を改善するために重要であると述べている [7]．従って，アクセントに基づいて，未知のコンテキストと合成単位の変更が合成音声へ及ぼす影響を調査する．

2.1 コンテキストと音声合成

図 1 に TTS における処理を示す．まず，文から文及び句や単語，音素に関するコンテキストを抽出する．次に，コンテキストから HMM 状態を選択する．その後，HMM 系列から音声特徴を推定し，ディジタルフィルタにより音声を合成する．今回，TTS において使用するコンテキストを次に示す．

音素: { 先行，当該，後続 } 音素
単語: { 先行，当該，後続 } 単語の品詞情報
アクセント句: { 先行，当該，後続 } アクセント句のモーラ数，{ 先行，当該，後続 } アクセント句のアクセント型，アクセント句間のポーズの有無，
当該アクセント句のモーラ位置
呼気段落: { 先行，当該，後続 } 呼気段落中の { モーラ，アクセント句 } 数，{ 先行，当該，後続 } 呼気段落内の { 文頭，文末 } からの当該アクセント句位置
文: 文全体の { モーラ，アクセント句，呼気段落 } 数，{ 文頭，文末 } からの当該位置

図 1 に示す TTS の処理とは対照的に，ITTS では，まず，文全体が不明な状態で，句やそれ以下の単位からコンテキストを抽出する．その後，限られたコンテキストから HMM 状態を選択する．その後，HMM 系列から音声特徴を，文以下の単位に対して生成する．その結果，生成された音声特徴は滑らかに変化せず，不自然となる可能性がある．そのため，後述する ITTS の実験において，次に示すようにコンテキストを制限して使用する．

音素: { 先行，当該 } 音素
単語: { 先行，当該 } 単語の品詞情報
アクセント句: { 先行，当該 } アクセント句のモーラ数，{ 先行，当該，} アクセント句のアクセント型，
当該アクセント句のモーラ位置
呼気段落: { 先行，当該 } 呼気段落中の { モーラ，アクセント句 } 数，{ 先行，当該 } 呼気段落内の { 文頭 } からの当該アクセント句位置

TTS で使用するコンテキストとの違いは，後続や文末に関するコンテキストを未知として使用しない点である．ここで，先行・後続は，現在生成する音素が属する当該（単語，アクセント句等）部分の先行及び後続の情報を意味する．更に，ポーズの有無といった ITTS において実時間での検出が困難なコンテキストも，今回未知として使用しない．

表 1 ITTS におけるコンテキストの組み合わせ

コンテキスト	Pho	Pho +POS	Pho +Acc	Pho +Bre	Pho +POS +Acc	Pho +POS +Bre	Pho +Acc +Bre	Pho +POS +Acc +Bre
音素	○	○	○	○	○	○	○	○
単語		○			○	○		○
アクセント句			○		○		○	○
呼気段落				○		○	○	○

2.2 コンテキストと合成単位の局所性

ITTS では，多くのコンテキストを使用し，より自然な音声特徴を生成する必要がある．一方で，より多くのコンテキストを抽出するためには，入力される文を待つ必要がある．従って，局所的なコンテキストの使用と，その音声品質を最適化する必要がある．そのために，まず，使用するコンテキストを表 1 に示す組み合わせに分類して実験を行う．但し，制限されたコンテキストでの音声品質の上限を調査するため，合成単位は文単位とする．

次に，アクセント句を合成単位として音声を合成する．コンテキストは，2.1 節で示した ITTS 用のコンテキストから選択し，次に示す組み合わせで実験を行う．

CurAcc: 音素と当該アクセント句のみ使用
CurPos+CurAcc: 音素と当該品詞タグ，当該アクセント句のみ使用
PasPos+CurPos+CurAcc: 音素と { 先行，当該 } 品詞タグ，当該アクセント句のみ使用
PasAcc+CurPos+CurAcc: 音素と当該品詞タグ，{ 先行，当該 } アクセント句のみ使用

PasPos+PasAcc+CurPos+CurAcc: 音素と { 先行，当該 } 品詞タグ，{ 先行，当該 } アクセント句のみ使用

最後に，アクセント句間で音声特徴を滑らかに変化させた場合の影響を調査する．図 2 に示すように，複数のアクセント句のコンテキストラベルを結合して合成し，その後，関連する部分（白色で示すアクセント句）のみの音声を結合する．図 2 において，(a) から (c) は，当該アクセントが判明すれば逐次に合成を行える．しかしながら，(d) は，後続アクセント句を待つ状況を考慮しており，実際に使用する場合，1 アクセント句を待つ遅れを許容しなければならない．図 2 において，使用したコンテキストは，前述の PasPos+PasAcc+CurPos+CurAcc である．

3. 実験条件

使用した音声は HTS デモに付属している ATR 音素バランス 503 文 [8] である．450 文を学習に使用して，53 文をテストに用いる．音声特徴は，39 次元のメルケプストラムと，基本周波数と 5 帯域の非周期成分及び各動的特徴を使用する．HMM の学習は，各々の音声特徴が学習できるように拡張した HTS を用いる [9]．音声特徴は，STRAIGHT[10] により取得した．客観評価は，基本周波数及びメルケプストラムに対して行う．基本周波数は，[5] と同様に，TTS による合成音声を基準とした対数比（C_{f_0}[cent]）で求め，メルケプストラムは [5] と同様にメルケプストラム歪み（MCD）を求める [11]．主観評価は平均評定オピニオンスコア（MOS）により行う．評価者は 16 人の日本人母語者で，53 音声から 15 音声を無作為に選択・再生し，五段階評価を行う．各音声は，評価者が望む限り再生できる状況で行う．

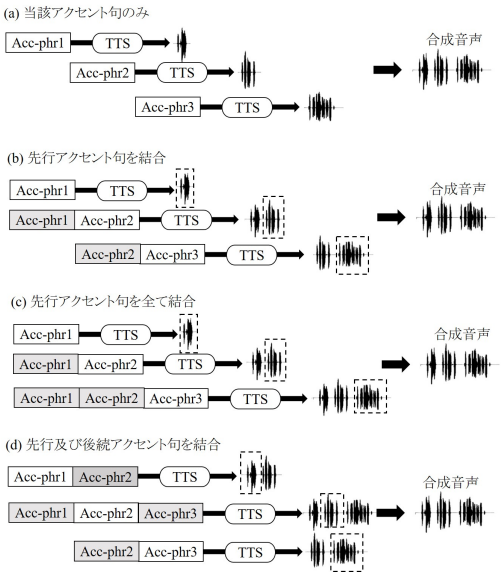

図 2 アクセント句の結合

表 2 制限されたコンテキストによる音声品質への影響

コンテキストセット	C_{f_0}[cent]	MCD[dB]	MOS±95% 信頼区間
Pho	242.5	3.5	-
Pho+POS	211.2	3.5	2.2 ± 0.12
Pho+Acc	178.8	3.4	-
Pho+Bre	186.8	3.5	-
Pho+POS+Acc	141.1	3.4	3.2 ± 0.12
Pho+POS+Bre	175.3	3.4	-
Pho+Acc+Bre	83.9	3.3	3.4 ± 0.14
Pho+POS+Acc+Bre	84.2	3.3	-
Standard TTS	-	-	3.6 ± 0.12

4. 合成音声の主観評価と客観評価

4.1 コンテキストによる影響について

本節で，2.1 節で定めた ITTS 用のコンテキストによる合成音声の品質について調査する．インクリメンタル TTS の音声品質の上限を調査する為に，制限されたコンテキストを使用し，音声合成は文単位で行う．表 2 に，主観評価と客観評価の結果を示す．但し，表 2 に示す値は，平均値である．表 2 より，コンテキストの選択による MCD の著しい悪化は確認できない．更に，Pho 及び Pho+POS と Pho+Acc を比較すると，アクセント句が韻律を著しく改善していることが確認できる．従って，後続アクセント句のコンテキストが不明な場合でも，合成音声の韻律を著しい改善が可能である．また，コンテキストを増加するほど音声品質が改善することも確認できる．

表 2 の客観評価の結果より，3 種のコンテキストセット（Pho+POS，Pho+POS+Acc，Pho+Acc+Bre）と通常の TTS を用いて主観評価を行う．表 2 の Pho+POS の結果より，Pho+POS は，韻律の改善に十分でないことが確認できる．これは，単語単位での日本語 ITTS の困難さを示していると考えられる．更に，Pho+POS+Acc と Pho+Acc+Bre の結果より，両者の差異は Pho+POS と比べると小さい．従って，コンテキストとして最低限アクセント句が必要であると考えられる．また，後続のコンテキストを使用しないにもかかわらず，Pho+POS+Acc と Pho+Acc+Bre の結果は従来の TTS の品質に近づいていることが確認できる．しかしながら，この実験では，文単位で合成を行っており，動的特徴による大域的最適化が音声品質に良い影響を与えた可能性がある．その影響を調査するために，次節でアクセント句を合成単位として実験を行う．

4.2 合成単位変更による影響について

本節では，アクセント句に基づく日本語 ITTS に焦点を当てる．そのため，合成単位をアクセント句として，コンテキストの選択による音声品質への影響を調査する．使用するコンテキストは，2.2 節で示した組み合わせを使用する．合成単位をアクセント句にするため，動的特徴による大域的最適化は，アクセント句内のみを局所的に最適化す

表 3　アクセント句単位の合成による音声品質への影響

コンテキストセット	C_{f_0}[cent]	MCD[dB]
CurAcc	232.6	5.2
CurPos+CurAcc	203.9	5.1
PasPos+CurPos+CurAcc	198.1	5.9
PasAcc+CurPos+CurAcc	198.6	5.1
PasPos+PasAcc+CurPos+CurAcc	195.2	5.7

表 4　アクセント句結合による音声品質への影響

実験条件	C_{f_0}[cent]	MCD[dB]	MOS±95% 信頼区間
(a) 当該アクセント句のみ	195.2	5.7	2.7 ± 0.13
(b) 先行アクセント句を結合	170.5	4.5	-
(c) 先行アクセント句を全て結合	160.8	4.2	2.8 ± 0.12
(d) 先行及び後続アクセント句を結合	157.3	4.0	3.3 ± 0.12

る．その結果，アクセント句間の韻律が崩れる可能性がある．客観評価の平均値を表 3 に示す．表 3 より，コンテキストを増加することで，音声品質の改善が確認できる．しかしながら，改善は僅かである．更に，表 3 の全ての音声品質は，表 2 の Pho+POS+Acc より悪くなっている．これは，アクセント句間の韻律の滑らかな変化を考慮できないことが原因として考えられる．従って，次節でアクセント句間の韻律変化を考慮し，実験を行う．

4.3　アクセント句の結合による影響について

本節では，図 2 で示すアクセント句の接続による音声品質への影響を調査する．使用するコンテキストは，2.2 節に記述している．表 4 に主観評価と客観評価の平均値を示す．表 4 より，先行及び後続アクセント句を接続した結果が最も良い．これは，先行及び後続アクセント句を接続したことで，アクセント句間で韻律の変化が滑らかに推定されたためと考えられる．表 4 より，韻律の改善が良い 2 つの結果（先行及び後続アクセント句を結合，先行アクセント句を全て結合）と，前述した PasPos+PasAcc+CurPos+CurAcc の結果を使用して主観評価を行う．表 4 より，先行アクセント句を全て結合することでも音声の自然性は改善している．しかしながら，後続アクセント句を使用する結果が最も良い．従って，日本語 ITTS では，後続アクセント句を 1 つ待つ戦略が有効である．

5.　おわりに

日本語 ITTS を実現するため，コンテキストと合成単位について調査を行った．特に，後続コンテキストが未知の場合における合成音声への影響と，合成単位をアクセント句とした場合の合成音声への影響を調査した．実験結果より，単語単位による日本語 ITTS は，低品質な音声となる可能性が示唆された．また，後続コンテキストが未知な場合においてもアクセント句が韻律を著しく改善することが確認された．従って，日本語 ITTS は，アクセント句に基づいて合成を行う必要性があると考えられる．更に，日本語 ITTS において，合成音声の品質を保つために，後続アクセント句を 1 つ待つ戦略が有効であることを確認した．今後の課題は，深層学習による日本語 ITTS の実現可能性の検討である．

謝辞　本研究の一部は JSPS 科研費 JP17H06101 および JP17K00237 の助成を受けたものである．

参考文献

[1] Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T. and Kitamura, T.: Speech parameter generation algorithms for HMM-based speech synthesis, *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, Vol. 3, IEEE, pp. 1315–1318 (2000).

[2] Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T. and Kitamura, T.: Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis, *Sixth European Conference on Speech Communication and Technology* (1999).

[3] Baumann, T.: Partial representations improve the prosody of incremental speech synthesis, *Fifteenth Annual Conference of the International Speech Communication Association* (2014).

[4] Baumann, T.: Decision tree usage for incremental parametric speech synthesis, *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, IEEE, pp. 3819–3823 (2014).

[5] Pouget, M., Hueber, T., Bailly, G. and Baumann, T.: HMM training strategy for incremental speech synthesis, *16th Annual Conference of the International Speech Communication Association (Interspeech 2015)*, pp. 1201–1205 (2015).

[6] Pouget, M., Nahorna, O., Hueber, T. and Bailly, G.: Adaptive Latency for Part-of-Speech Tagging in Incremental Text-to-Speech Synthesis, *Interspeech 2016*, pp. 2846–2850 (2016).

[7] Yokomizo, S., Nose, T. and Kobayashi, T.: Evaluation of prosodic contextual factors for HMM-based speech synthesis, *Eleventh Annual Conference of the International Speech Communication Association* (2010).

[8] Kurematsu, A., Takeda, K., Sagisaka, Y., Katagiri, S., Kuwabara, H. and Shikano, K.: ATR Japanese speech database as a tool of speech recognition and synthesis, *Speech Communication*, Vol. 9, No. 4, pp. 357–363 (1990).

[9] Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A. W. and Tokuda, K.: The HMM-based speech synthesis system (HTS) version 2.0., *SSW*, pp. 294–299 (2007).

[10] Kawahara, H., Masuda-Katsuse, I. and De Cheveigne, A.: Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds, *Speech communication*, Vol. 27, No. 3, pp. 187–207 (1999).

[11] Kubichek, R.: Mel-cepstral distance measure for objective speech quality assessment, *Communications, Computers and Signal Processing, 1993., IEEE Pacific Rim Conference on*, Vol. 1, IEEE, pp. 125–128 (1993).