情報処理学会論文誌 Vol.63 No.4 1149–1158 (Apr. 2022)

日本語逐次音声合成における合成単位

柳田 智也[1,a)] サクテイ サクリアニ[2,1,3,b)] 中村 哲[1,3,c)]

受付日 2021年6月18日, 採録日 2022年1月11日

概要：同時音声通訳システムは，話者の発話中に翻訳を行い音声を生成する．その実現のために，文より短いテキストから，音声を生成する逐次音声合成が必要である．本論文は，同時音声通訳システムの実現を目指して，日本語における逐次音声合成の提案を目的とする．先行研究は，逐次音声合成のために使用する言語特徴を制限し，合成範囲を単語としている．しかし，日本語音声合成は，アクセント句と呼ばれる単位が重要であり，単語の逐次音声合成が，音声品質と遅延のトレードオフとして適さない可能性がある．本論文では，日本語逐次音声合成のため，逐次音声合成の言語特徴を提案する．そして，言語特徴の組合せから，遅延と音声品質に最適な合成範囲を決定する．実験結果より，アクセント句から呼気段落の合成範囲が音声の品質を保持するために必要であることを示した．さらに，遅延評価を通して，アクセント句が日本語の逐次音声合成へ適することを示した．

キーワード：同時音声通訳システム，逐次音声合成，アクセント句，日本語

Synthesis Unit for Japanese Incremental Text-to-Speech

Tomoya Yanagita[1,a)] Sakriani Sakti[2,1,3,b)] Satoshi Nakamura[1,3,c)]

Received: June 18, 2021, Accepted: January 11, 2022

Abstract: A simultaneous speech translation system translates while the speaker speaks and generates speech sequentially. To construct the system, an incremental Text-to-speech (iTTS) system which synthesizes a speech in a shorter synthesis unit is required. This work proposes a Japanese iTTS system for the simultaneous speech translation. Most of the researchers used the word unit as the synthesis unit. However, in Japanese speech synthesis, a unit called an accent phrase is important, and word-by-word synthesis may not be suitable. In this paper, we propose a linguistic feature and synthesis unit for Japanese iTTS. Experimental result shows that accent phrase or breath group are essential for a Japanese iTTS system as a trade-off between quality and synthesis units for the Japanese iTTS. Then, an accent phrase is a more appropriate incremental synthesis unit than a breath group through delay analysis.

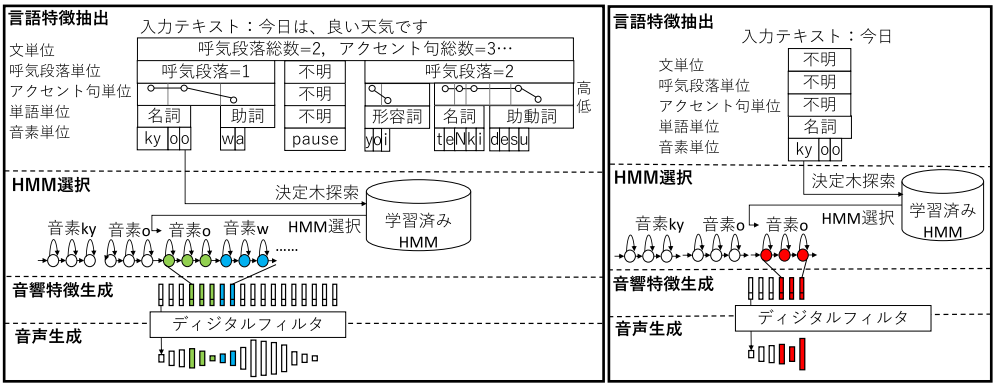
Keywords: simultaneous machine speech translation, incremental TTS, accent phrase, Japanese language

1. はじめに

1.1 研究背景

音声翻訳システムは，原言語の音声を入力しながら目的言語の音声を生成するシステムであり，自動音声認識（ASR: Automatic Speech Recognition）・機械翻訳（MT: Machine Translation）・音声合成（TTS: Text-to-speech）から構築される．従来の音声翻訳システムでは，ASR は発話の終了前に処理を開始するが，MT と TTS は，発話終了後に処理が始まる [1]．しかしながら，入力音声が非常に長い場合，MT と TTS は発話終了を待つ必要があり，非常に長い遅延を生じる．一方で，人間の通訳者は，短い遅延で音声を翻訳するために，入力を短い単位に分割して翻訳する．この特性に着目して，同時音声通訳システムの構築に取り組む研究が存在する [2], [3]．文献 [2] の研究では，

1 奈良先端科学技術大学院大学
NAIST, Ikoma, Nara 630–0101, Japan
2 北陸先端科学技術大学院大学
JAIST, Nomi, Ishikawa 923–1292, Japan
3 理化学研究所観光情報解析チーム
RIKEN AIP, Ikoma, Nara 630–0101, Japan
a) yanagita.tomoya.yo8@is.naist.jp
b) ssakti@jaist.ac.jp
c) s-nakamura@is.naist.jp



（a）通常の音声合成　　（b）逐次音声合成

図 1　文単位の音声合成と逐次音声合成の処理および合成範囲（合成時）
Fig. 1　Process unit of Sentence-Based TTS and iTTS (Synthesizing Step).

ASR の出力を MT の翻訳のために再度分割する方法を提案しており，文献 [3] は，通話における同時通訳システムに取り組む．また，語順が異なる言語に対して，翻訳単位の分割方法が提案されている [4], [5], [6]．これらの研究は，翻訳処理に関わる遅延を最小化しつつ高品質な音声翻訳を目指す．ただ，これらの研究では speech-to-speech 同時通訳に必要な逐次音声合成については触れられていない．本研究は，日本語における逐次音声合成（iTTS: incremental TTS）に焦点を当て，音声の自然性が普通と判断可能な範囲で，最小の遅延となる逐次音声合成の構築が目的である．speech-to-speech 同時通訳システムにおける逐次音声合成の遅延と音声品質の最適性は，ASR および逐次機械翻訳の品質と遅延も考慮する必要があり，その状況における逐次音声合成の最適性は，今後，主観評価を通して検討を考えている．

1.2　隠れマルコフモデルに基づく音声合成と逐次音声合成

初めに，本節では隠れマルコフモデル（Hidden Markov Model: HMM）に基づく音声合成の学習と合成方法について記述し，その後，HMM に基づく逐次音声合成の課題を記述する．

学習時の処理手順を次に示す．まず，音響特徴とそれに対応する音素から，音素ごとに HMM のパラメータが尤度最大化に基づいて学習される．次に，時間的・音韻的な音声の変化を考慮するため，周辺の音素や言語特徴を考慮した HMM（コンテキスト依存 HMM）が学習される．このとき，言語特徴の組合せは膨大であり，コンテキスト依存 HMM の学習不足が生じる．最後に，学習不足を防止する目的で，言語特徴と決定木に基づく分類が行われる．この分類により，類似するコンテキスト依存 HMM のパラメータが共有される．

生成時の処理手順を次に示す．まず，入力テキストから言語特徴が抽出される．次に，言語特徴と音素から，学習時に構築した決定木を探索してコンテキスト依存 HMM が選択される．その後，HMM 系列から音響特徴が生成される．このとき，大域的最適化により，音響特徴の平滑化が行われる [7], [8]．最後に，音響特徴がディジタルフィルタに入力され，音声が生成される．通常の音声合成の場合，合成範囲[*1]は文である．一方で，逐次音声合成の場合，合成範囲は文より短い．したがって，言語特徴の使用と音声生成に関して，制限が生じる．

図 1 に，合成時における通常の音声合成と逐次音声合成との差異を示す．通常の音声合成の場合（図 1 (a)），言語特徴抽出は，音声品質を改善する目的で文単位・句単位・単語単位・音素単位から行われる．さらに，HMM 選択から音声生成の処理において，合成範囲は文であるため，後続の音素や言語特徴が考慮される．たとえば，図 1 (a) の HMM 選択において，選択されるコンテキスト依存 HMM（音素 o）は，先行音素 o や後続音素 w だけでなく，より長い言語特徴（単語単位やアクセント句単位等）が考慮される．さらに，音響特徴生成時について，合成範囲が文単位であり後続の音響特徴の変化が考慮される．

逐次音声合成の場合（図 1 (b)），合成範囲は文より短い．図 1 (b) は，合成範囲を単語とする．この場合，言語特徴抽出は単語単位・音素単位で行われる．使用可能な言語特徴は通常の音声合成より少なく，後続の音素や言語特徴は考慮できない．たとえば，図 1 (b) の HMM 選択において，選択されるコンテキスト依存 HMM（音素 o）は，後続音素 w や単語単位より長い言語特徴（アクセント句単位等）を考慮できない．さらに，音響特徴生成時において合成範囲は単語単位であり，後続の音響特徴の変化も考慮できないため音声波形は合成範囲ごとに不連続となる．前述の要因により，逐次音声合成の音声品質は通常の音声合成より劣化する．逐次音声合成の音声品質を改善するために，最適な言語特徴の使用とその合成範囲の最適化は，解決すべき課題である．

[*1] 通常は合成単位だが，逐次音声合成の場合を考慮して，本論文では「範囲」と呼ぶことにする．

表 1 本論文で使用する言語特徴
Table 1 Utilization of linguistic features in this paper.

合成単位	通常の音声合成	逐次音声合成（下線引きは利用可能な場合に使用）
音素単位	{先行，当該，後続} の音素	{先行，当該，後続} の音素
単語単位	{先行，当該，後続} の品詞タグ情報	{先行，当該，後続} の品詞タグ情報
アクセント句単位	アクセント核と当該モーラの相対位置， {先行，当該，後続} アクセント句内のモーラ数， {先行，当該，後続} アクセント句のアクセント型， 当該アクセント句内の {前方向，後ろ方向} モーラ位置， アクセント句間のポーズの有無	アクセント核と当該モーラの相対位置， {先行，当該，後続} アクセント句内のモーラ数， {先行，当該，後続} アクセント句のアクセント型， 当該アクセント句内の {前方向，後ろ方向} モーラ位置
呼気段落単位	{先行，当該，後続} 呼気段落内のアクセント句の個数， 当該呼気段落のアクセント句位置， 呼気段落の {前方向，後ろ方向} 位置	{先行，当該} 呼気段落内のアクセント句の個数， 当該呼気段落のアクセント句位置， 呼気段落の {前方向} 位置
文単位	文中の {モーラ，アクセント句，呼気段落内} 総数， 文中の {モーラ，アクセント句} による {前方向，後ろ方向} 位置	

1.3 先行研究と日本語における課題

言語特徴の制限に関して先行研究では，合成時に不明な後続の言語特徴が及ぼす影響が調査され，その言語特徴を学習データの最頻値や平均値に置き換える方法が提案されている [9], [10]．さらに，後続の言語特徴が不明という情報を学習に用いる方法や，後続の品詞を予測し言語特徴として使用する方法が提案されている [11], [12]．合成範囲の拡張に関する先行研究は，後続の入力を待ち，後続の合成範囲を考慮して韻律の改善が報告されている [13]．

深層学習に基づく End-to-End 逐次音声合成も提案されている．日英の言語を対象に，合成範囲の後続の入力を待たない逐次音声合成が検討されている [14]．また，英語においては，1 単語もしくは 2 単語の入力待ちを許容することで，End-to-End 音声合成と同程度の品質の逐次音声合成が報告されている [15]．文献 [16] では，文献 [15] による 1-2 単語という固定長の入力待ちを可変にする目的で強化学習が使用されている．文献 [17] は，文献 [15] に影響を受けて，入力待ちの単語による近似について分析し，入力待ち単語の単語長が，品質へ顕著に影響することを示している．この分析から入力単語を待たずに言語モデルで後続の入力を近似する方法が提案されている [18]．また，同様の方法が，文献 [19] においても提案されている．さらに，文献 [19] の方法に基づき，言語モデルの予測にかかる計算量を低減する方法が提案されている [20]．先行研究の多くは，英語のような単語単位の逐次合成を対象とする．

日本語は，モーラ等時性かつ高低アクセントの言語である [21], [22]．モーラは，日本語発声の基本単位として扱われ，おおよそ仮名 1 文字が 1 モーラに相当する．日本語は，モーラ単位で音の高低が変化する高低アクセントを持つ．このアクセントのまとまりをアクセント句と呼び，しばしば文節と一致する [23]．日本語の音声合成において，アクセント句に関する言語特徴は，韻律の改善に重要である [24]．したがって，日本語の言語特徴と合成範囲は，先行研究と異なり単語より長いアクセント句が必要と考えられる．

日本語の逐次音声合成として，HMM と End-to-End に基づく逐次音声合成の初期検討が行われている [14], [25]．本論文では，少量データかつリアルタイムアプリケーションに適する HMM 逐次音声合成について詳細な検討を行った[*2]．今回得られた知見は，日本語における End-to-End 逐次音声合成の初期検討にも応用可能である [14]．

2. 言語特徴と合成範囲の設計

2.1 逐次音声合成のための言語特徴の分類

表 1 に，本論文において，通常の音声合成と逐次音声合成で使用する言語特徴を合成単位別に示す．表 1 において，呼気段落はアクセント句より長い単位であり，音声では発話中のポーズごとに分割され，テキストでは文中の句読点や中点により分割される．逐次音声合成において，下線部引きの言語特徴は，使用可能な状況と使用不可能な状況がある．たとえば，単語を合成範囲とする逐次音声合成の場合，単語末尾の音素以外は後続音素の言語特徴を利用可能である．

通常の音声合成と逐次音声合成における言語特徴の違いは，以下の 2 点である．1 点目は，文単位の言語特徴は，文全体の入力が必要となり使用しない．2 点目は，アクセント句単位のアクセント句間のポーズの有無は今回使用しない．ポーズは，テキストにおいて句点で表現される．事前

[*2] 本研究の一部は，文献 [25] に示している．文献 [25] では，品質評価のみに着目しており，アクセント句より長い単位の呼気段落単位もまた逐次音声合成の候補となる．また，単語単位内における後続音素のように，合成時に遅延を増加させなくても後続の言語特徴が使用可能な状況が存在する．その状況における逐次音声合成について検討していない．本論文では，音声品質のみではなく遅延評価の観点からもアクセント句単位の有効性を示し，後続の言語特徴が使用可能な状況を考慮した追加実験を行っている．

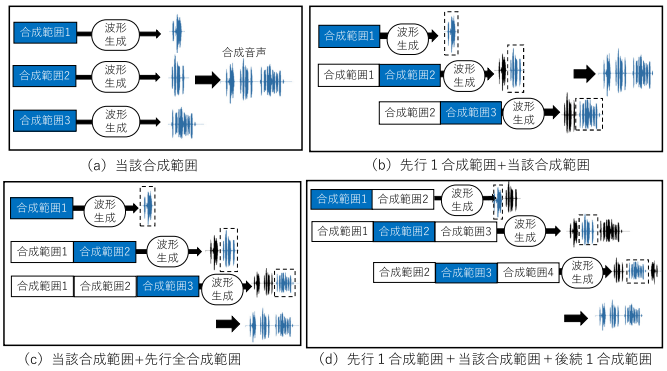


図 2 本論文で提案する合成範囲
Fig. 2 Proposed synthesis unit in this paper.

実験において，ポーズの有無に関する情報を使用して学習し，合成時にポーズの有無が不明な場合，決定木の探索に失敗した HMM が選択され，合成音声が著しく劣化した．決定木の探索の失敗による音声品質への影響を削減するため，本論文では，ポーズの検出を想定せず，ポーズの言語特徴を使用しない．前処理において逐次的なポーズ検出を行い，ポーズの有無を言語特徴として利用可能な場合，合成品質はより改善されると考えられる．

2.2 言語特徴の組合せと合成範囲の設計

逐次音声合成において音声品質を改善するため，より多くの言語特徴の使用が望ましい．しかし，多くの言語特徴を使用するには，合成範囲を長く保つ必要があり，遅延の増加につながる．したがって，音声品質の保持に最適な言語特徴の組合せと，その合成単位を調査する．まず，実験で使用する言語特徴の組合せについて，次に示す．

Pho：音素単位のみ使用
Pho+POS：音素単位と単語単位を使用
Pho+Accphr：音素単位とアクセント句単位を使用
Pho+Bre：音素単位と呼気段落単位を使用
Pho+POS+Accphr：音素単位，単語単位，およびアクセント句単位を使用．
Pho+POS+Bre：音素単位，単語単位，および呼気段落単位を使用．
Pho+Accphr+Bre：音素単位，アクセント句単位，および呼気段落単位を使用．
Pho+POS+Accphr+Bre：音素単位，単語単位，アクセント句単位，および呼気段落単位を使用．
+Next：後続言語特徴（下線引きの言語特徴）が遅延を生じず利用可能な場合に使用．

言語特徴の組合せ例として，Pho+POS の場合，音素単位と単語単位の言語特徴を使用するが，下線引きの言語特徴（後続の言語特徴）は使用しない．一方，Pho+POS+Next の場合，Pho+POS に加えて，後続の音素（音素単位の下線引きの言語特徴）は遅延を生じず利用可能なため使用する．ただし，単語単位の後続の言語特徴については，後続の単語を待つ必要が生じるため使用しない．

次に，主観評価が良好な言語特徴の組合せに対して，逐次音声合成に適した合成範囲を調査する．実験では，合成範囲間の不連続性を改善するために合成範囲を接続する．逐次音声合成の合成範囲は次に示し，合成範囲を接続した場合における波形生成の処理は**図 2** に示す．ただし，入力は言語特徴とし，音声合成の波形生成部（図 1，HMM 選択から音声生成に相当）へ入力され，音声波形を生成する．

当該合成範囲：当該合成範囲ごとに合成する（図 2 (a) 参照）．
先行 1 合成範囲+当該合成範囲：先行 1 合成範囲の言語特徴と，当該合成範囲の言語特徴を接続し，音声全体を合成する．その後，当該合成範囲の音声波形を切り出す（図 2 (b) 参照）．
当該合成範囲+先行全合成範囲：先行全合成範囲の言語特徴と，当該合成範囲の言語特徴を使用し，音声全体を合成する．その後，当該合成範囲の音声を切り出す（図 2 (c) 参照）．
先行 1 合成範囲+当該合成範囲+後続 1 合成範囲：先行 1 合成範囲と当該合成範囲および後続 1 合成範囲の言語特徴を使用し，音声全体を合成する．その後，当該合成範囲の音声波形を切り出す（図 2 (d) 参照）．

3. 実験的評価：後続の言語特徴を使用しない場合

本章では，まず，2.2 節で示した言語特徴の組合せによる音声品質を評価する（3.1 節および 3.2 節）．次に，その結果を受けて，合成範囲を決定するために，遅延評価を行う（3.3 節および 3.4 節）．最後に，合成範囲を変更した場合の音声品質について評価する（3.5 節および 3.6 節）．

3.1 言語特徴の組合せに関する実験条件

言語特徴の組合せを簡略化する目的で，後続の言語特徴（+Next）は使用しない．合成範囲は文として，言語特徴を

制限した場合の音声品質の上限を確認する．

音声と言語特徴は HTS システム [26] に付属しているデモ用の ATR 音素バランス 503 文 [27] を使用する．音声は 16 kHz でサンプリングされており，音響特徴は，39 次元のメルケプストラムと，基本周波数および 5 帯域の非周期成分を使用する．音響特徴抽出は，STRAIGHT [28] を使用する．さらに，音響特徴量を平滑化し改善するために動的特徴量 [7] が使用される．フレームシフト幅は 5 ミリ秒である．音声は，450 文を学習に，53 文を評価に用いる．HMM の学習は HTS を使用する．客観評価は，基本周波数およびメルケプストラムから計算する．客観評価は，文献 [11] と同様に，通常の音声合成から生成される音響特徴を基準とし，基本周波数とメルケプストラム係数に対して求める．基本周波数の客観評価値は，次式で求める．

$$C_{f_0} = \frac{1}{T}\sum_{t=1}^{T} 1200 \log_2 \frac{|f_0^{tar}(t)|}{|f_0^{src}(t)|} \tag{1}$$

式 (1) は．通常の音声合成より得られる F0 系列を基準とした誤差であり，1,200 [cent] で基準から 1 オクターブの差を示す．ただし，T は，動的時間伸縮により系列間の対応をとり，基本周波数が非ゼロの共通句間（有声音区間）である．$f_0^{tar}(t)$ と $f_0^{src}(t)$ は，それぞれ，逐次音声合成を考慮した言語特徴から得られる基本周波数系列と，通常の音声合成から得られる基本周波数系列である．メルケプストラムの客観評価値は，メルケプストラム歪み（MCD: Mel cepstrum distortion）[29] を次式より求める．

$$MCD = \frac{10}{T \log_e(10)} \sum_{t=1}^{T} \sqrt{2 \sum_{d=1}^{D} (y_t^{tar}(d) - y_t^{src}(d))^2} \tag{2}$$

ただし，D はメルケプストラムの次元数，$y_t^{tar}(d)$ および $y_t^{src}(d)$ は，逐次音声合成を考慮した言語特徴のメルケプストラム系列と通常の音声合成のメルケプストラム系列とする．

主観評価は平均オピニオン評点（MOS: Mean Opinion Score）を使用し [30]，音声の自然性に対して 5 段階（1：とても悪い，2：悪い，3：普通，4：良い，5：とても良い）で評価する．評価者は日本語を母語とする日本人 16 名である．実験では，初めに，著者らが 53 文から評価用の 15 文をランダムに選択する．評価では，その 15 文のみを固定的に使用する．つまり，選択されない文は，評価に用いない．評価者は，選択した 15 文すべてを評価する．1 方法あたりの評価サンプルは，$16 \times 15 = 240$ サンプルであり，評価された音声あたりの評価者数に偏りはない．また，合成条件が異なる場合，同じ文章の音声を各方法ごとに生成し評価する．評価は，音声を何度も再生できる状況で行う．

3.2 言語特徴の組合せに関する実検結果

表 2 に，3.1 節の実験条件による C_{f_0}，MCD，および MOS の平均値を示す．MOS に対する検定は，全実験でマン・ホイットニーの U 検定を用いる．表 2 より，音素（Pho）と音素と単語（Pho+POS）の言語特徴を使用した場合とを比較すると，C_{f_0} がおよそ 30 [cent] ほど改善される．音素（Pho）と音素とアクセント句（Pho+Accphr）および音素と単語とアクセント句（Pho+POS+Accphr）の C_{f_0} を比較すると，単語より長いアクセント句単位の言語特徴により，C_{f_0} が大きく減少する．Pho+POS+Accphr と Pho+POS+Accphr+Bre や Pho+Accphr+Bre の結果を比較すると，呼気段落単位とアクセント句単位との言語特徴の併用（Pho+Accphr+Bre や Pho+POS+Accphr+Bre）により，C_{f_0} は，アクセント句のみを使用する場合（Pho+POS+Accphr）より改善される．また，MCD は，3.3 から 3.5 の範囲であり，F0 と比べて品質改善はない．

表 2 の結果より，音素と単語（Pho+POS），音素と単語とアクセント句（Pho+POS+Accphr），音素とアクセント句と呼気段落（Pho+Accphr+Bre），および通常の音声合成による音声を選択し，主観評価を行った．主観評価より，音素と品詞タグを使用した結果（Pho+POS）の自然性は，2（悪い）に近い．したがって，日本語において，単語単位の逐次音声合成は，2.25 以上の改善が望めない．単語単位より長いアクセント句や呼気段落を用いた場合（Pho+POS+Accphr と Pho+AccPhr+Bre）を比較すると，それぞれの自然性は 3（普通）に近く，それぞれ 3.16 と 3.33 である．Pho+POS と Pho+POS+AccPhr とを比較すると，1%未満の水準で有意差が確認できた．したがって，アクセント句までの言語特徴を使用することで，自然性を 3（普通）程度に保ちつつ逐次音声合成が可能と考えられる．また，アクセント句より高品質な逐次音声合成として呼気段落を合成範囲とする方法も考えられる．後続の言語特徴を使用していないにもかかわらず，自然性が 3（普通）程度である理由は，合成範囲が文であり，後続音素の HMM を考慮しているためと考えられる．しかし，実際の

表 2 後続の言語特徴を用いない場合の MCD および C_{f_0} の誤差
Table 2 MCD and f_0 error of cent unit without availavle next linguistic fearures for iTTS.

言語特徴	C_{f_0} [cent]	MCD [dB]	MOS
Pho	242.5	3.5	-
Pho+POS	**211.2**	**3.5**	**2.25**
Pho+Accphr	178.8	3.4	-
Pho+Bre	186.8	3.5	
Pho+POS+Accphr	**141.1**	**3.4**	**3.16**
Pho+POS+Bre	175.3	3.4	-
Pho+Accphr+Bre	**83.9**	**3.3**	**3.33**
Pho+POS+Accphr+Bre	84.2	3.3	-
通常の音声合成	0.0	0.0	3.66

逐次音声合成は，合成範囲が文より短い．その結果，合成範囲間の音声が不連続となり，音声品質の悪化が考えられる．したがって，合成範囲を変更した場合の音声品質も後述の 3.5 節において評価する．

3.3 遅延評価に関する実験条件

本節では，以降の実験で使用する合成範囲を決定するために，合成範囲の変更による遅延分析の実験条件を示す．合成範囲の変更は，合成品質の低下を生じる．しかし，合成の遅延は短くなり，リアルタイムアプリケーションへの応用が期待できる．遅延評価は，評価のコストを低減させるため，客観評価を用いる．評価方法は，合成範囲と使用する言語特徴を変更し，合成に必要な総遅延時間を計測する[*3]．総遅延時間は，次に示す遅延の総和とする．

入力遅延：入力テキストから言語特徴を分析する遅延
合成遅延：言語特徴から合成音声を生成する遅延
再生遅延：合成音声を再生し終えるまでの遅延

また，テストデータに対するテキスト長も客観評価値とする．評価はテストデータすべてを対象とし，テキスト長は，合成範囲内に存在するテキストの文字数とする．評価の合成範囲は，3.2 節の結果より，アクセント句と呼気段落とする．言語特徴は，テキストからそれぞれ Pho+POS+Accphr と Pho+POS+Accphr+Bre とを抽出して使用する．

3.4 遅延評価に関する実験結果

表 3 に，平均テキスト長および平均総遅延時間を示す．表 3 より，文とアクセント句の総遅延時間を比較すると，約 2.7 秒ほど遅延が減少する．したがって，アクセント句の逐次音声合成は，通常の音声合成よりリアルタイムアプリケーションへの応用に適する．

合成範囲を呼気段落とする場合，通常の音声合成と比べて，遅延はおよそ 2 秒ほど短い．したがって，呼気段落の逐次音声合成もまた，通常の音声合成よりリアルタイムアプリケーションに適する．呼気段落の平均テキスト長は，アクセント句と比較しておよそ 2.5 倍となる．いくつかのテストデータは，呼気段落の分割がない．そのような文章の遅延は，通常の音声合成と同様となる．一方で，アクセント句を合成範囲とする逐次音声合成の場合，文中に 2 つ以上のアクセント句を含み，遅延は必ず通常の合成単位より短い．この結果より，2.2 節で表記した合成範囲の変更による音声品質の調査は，合成範囲をアクセント句として行う．

表 3 平均テキスト長と平均総遅延時間の関係
Table 3 Average text length and total delay.

当該合成範囲/ 使用する言語特徴	テキスト長[文字数]	総遅延時間[秒]
文単位/ 通常の音声合成	20.36	4.41
呼気段落単位/ Pho+POS+Accphr+Bre	9.76	2.68
アクセント句単位/ Pho+POS+Accphr	**3.96**	**1.71**

[*3] CPU は，Intel Core(TM) i7-6700K を使用し，言語特徴分析として Open Jtalk（http://open-jtalk.sourceforge.net/）を修正し使用する．

3.5 合成範囲の変更に関する実験条件

本節では，合成範囲を変更した場合における音声品質の評価に関する実験条件を示す．学習および合成に使用する言語特徴は，Pho+POS+Accphr を使用した．当該合成範囲は，アクセント句とし，2.2 節に示した合成範囲を使用する．評価時には，各合成単位の音声波形を接続し，文に相当する音声を構築する．評価は，3.1 節と同様である．

3.6 合成範囲の変更に関する実験結果

合成範囲をアクセント句とした評価結果を**表 4** に示す．表 4 の当該合成範囲と，表 2 の文を合成範囲とする結果（Pho+POS+Accphr）とを比較すると，F0 と MCD の客観評価値は，表 2 の Pho+POS+Accphr より悪い．これは，合成範囲の変更により，後続の合成範囲を考慮できないためと考えられる．**図 3** に，同一のテキストに対する F0 系列を示す．図 3 (a) は，通常の音声合成から生成された F0 系列を，(b) は，表 2 から，文を合成範囲とし，言語特徴として Pho+POS+Accphr を使用した場合の F0 系列を，(c) は，当該合成範囲を用いた場合の F0 系列である．ここで，縦軸は周波数を示し，横軸は時間軸を示す．図上の破線はアクセント句境界を示す．図 3 (a) と (b) を比較すると，合成範囲を文とした場合，アクセント句間で F0 の値は連続的に変化する．しかし，(b) と (c) とを比較すると，アクセント句ごとに合成した場合 (c) はアクセント句境界に短い無音区間が生じ，F0 の値はアクセント句境界間で不連続となる．

表 4 より，過去のアクセント句を接続し合成した場合

表 4 合成範囲をアクセント句とし，Pho+POS+Accphr の言語特徴による評価結果
Table 4 Evaluation of iTTS systems with accent phrase and Pho+POS+Accphr as linguistic feature.

合成範囲	C_{f_0} [cent]	MCD [dB]	MOS
当該合成範囲	**232.6**	**5.2**	**2.68**
先行 1 合成範囲 +当該合成範囲	170.5	4.5	-
当該合成範囲 +先行全合成範囲	160.8	4.2	2.83
先行 1 合成範囲 +当該合成範囲 +後続 1 合成範囲	**157.3**	**4.0**	**3.29**

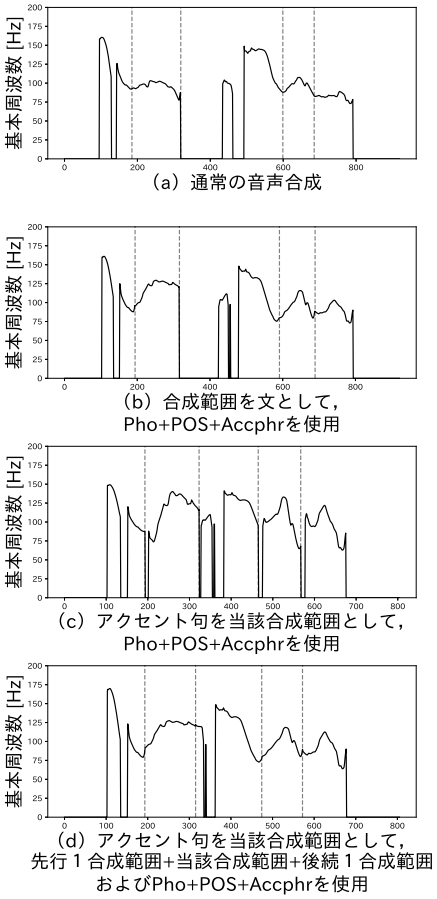


図 3 予測された F0 系列
Fig. 3 Predicted F0 sequences.

(当該合成範囲+先行全合成範囲)，客観評価と主観評価において，改善が確認できた．さらに，後続のアクセント句を待ち，合成した場合（先行 1 合成範囲+当該合成範囲+後続 1 合成範囲）が，最も良く，MOS は 3（普通）以上の品質を実現する．また，先行全合成範囲+当該合成範囲と当該合成範囲+先行 1 合成範囲+後続 1 合成範囲に対して検定を行った．その結果，5%未満の水準で有意差が確認できた．図 3 (d) に，先行 1 合成範囲+当該合成範囲+後続 1 合成範囲の F0 系列を示す．(d) より，後続の入力を待つことで，F0 系列は時系列方向へより連続的に変化し，自然性が改善された．また，先行 1 合成範囲+当該合成範囲+後続 1 合成範囲は，表 2 の通常の音声合成の MOS より改善される．これは評価時に，評価者は通常の音声合成を聴取しておらず，評価にバイアスが生じたと考えられる．

後続のアクセント句を待つ場合（先行 1 合成範囲+当該合成範囲+後続 1 合成範囲)，文末のアクセント句を除き，必ず後続の 1 アクセント句が必要となり，各アクセント句ごとに合成した場合（当該合成範囲）の遅延から，1 アクセント句相当の遅延が増加する．表 3 の合成範囲を当該アクセント句とした遅延（1.71 秒）を考慮すると，後続のアクセント句を待つ場合の総遅延時間は，最大で 1.71 秒の 2 倍程度と考えられる．

表 5 後続の言語特徴を用いる場合の MCD および C_{f_0} の誤差
Table 5 MCD and f_0 error of cent unit with available next linguistic features for iTTS.

言語特徴	C_{f_0} [cent]	MCD [dB]
Pho+POS+Next	207.0(↓4.2)	3.3(↓0.2)
Pho+Accphr+Next	172.4(↓6.4)	3.2(↓0.2)
Pho+Bre+Next	182.5(↓4.3)	3.0(↓0.5)
Pho+POS+Accphr+Next	**142.0(↑0.9)**	**3.1(↓0.3)**
Pho+POS+Bre+Next	177.2(↑1.9)	3.0(↓0.3)
Pho+Accphr+Bre+Next	**74.7(↓9.2)**	**2.8(↓0.5)**
Pho+POS+Accphr+Bre+Next	**73.4(↓10.8)**	**2.8(↓0.5)**

4. 実験的評価：後続の言語特徴を使用する場合

本章では，2.2 節で示した後続の言語特徴を利用する場合（+Next）における実験を行う．評価に関する実験条件を 4.1 節に示す．実験は，まず，合成範囲を文として，後続の言語特徴を使用した場合における音声品質を評価する．次に，後続の言語特徴の有無による音声品質を評価する (4.2 節)．最後に，後続の言語特徴を使用し，合成範囲を変更した場合における音声品質を評価する（4.3 節)．

4.1 実験条件

使用するデータセットおよび客観評価方法は，3.1 節と同様である．主観評価は，日本語を母語とする日本人 10 名が行う．主観評価は，テストセットから 15 音声をランダムに選択し，各方法ごとに同一文章を使用する．評価は音声を何度も再生できる状況で行った．

4.2 合成範囲を文とする場合の評価結果

表 5 に後続の言語特徴を使用した場合における F0 の誤差（C_{f_0}）および，MCD の平均値を示す．括弧内の数値は，表 2 における Next なしとの差分を示しており，下矢印が誤差の減少に対応し，上矢印が誤差の増加に対応する．表 5 の差分から，後続の言語特徴を使用した場合，MCD は 0.2 から 0.5 の範囲で全体的に改善される．C_{f_0} は，呼気段落とアクセント句の組合せを使用した場合（Pho+Accphr+Bre および Pho+POS+Accphr+Bre）に，9.2 および 10.8 と他の組合せより大きく改善される．また，Pho+POS+Accphr と Pho+POS+Bre の場合，F0 誤差が増加しているが，Pho+POS+Accphr では，0.9 程度であり悪化はわずかである．

合成範囲を文とした主観評価は，通常の音声合成と，言語特徴として Pho+POS+Accphr を使用した場合と，Pho+POS+Accphr に後続の言語特徴を使用した場合（Pho+POS+Accphr+Next）とを用いた．**表 6** に，主観評価の結果を示す．表 6 より，後続の言語特徴を使用する場合（Pho+POS+Accphr+Next）と，後続の言語特徴を使

表 6 合成範囲を文とし，後続の言語特徴の有無による主観評価結果
Table 6 MOS score with the presence or absence of next linguistic features.

言語特徴	MOS
Pho+POS+Accphr	**3.60**
Pho+POS+Accphr+Next	**3.63**
通常の音声合成	3.84

表 7 合成範囲をアクセント句とし，後続の言語特徴を使用した場合と通常の音声合成の評価結果
Table 7 Evaluation of iTTS systems with various input chunks and available next feature.

合成範囲/言語特徴/平滑化の有無	C_{f_0} [cent]	MCD [dB]	MOS
文/通常の音声合成/なし	0.0	0.0	3.76
当該合成範囲/ Pho+POS+Accphr+Next/なし	201.7	5.61	2.21
先行 1 合成範囲 +当該合成範囲+後続 1 合成範囲/ Pho+POS+Accphr+Next/なし	**154.8**	**3.71**	**2.81**
先行 1 合成範囲 +当該合成範囲+後続 1 合成範囲/ Pho+POS+Accphr+Next/あり	**152.3**	**3.52**	**2.67**

用しない場合（Pho+POS+Accphr）において，MOS の差はわずか（0.03）である．さらに，Pho+POS+Accphr と Pho+POS+Accphr+Next に対して，p 値による有意差は確認できなかった．したがって，後続の言語特徴を使用しても自然性の改善は確認できない．また，通常の音声合成の MOS は，表 2 の通常の音声合成と比較すると 0.2 ほど高い．評価結果を確認すると，評価者は，通常の音声合成の自然性を高く評価する傾向が確認でき，評価にバイアスが生じたと考えられる．

4.3 合成範囲をアクセント句とする場合の評価結果

本節では，後続の言語特徴（Pho+POS+Accphr+Next）を使用した状況で，合成範囲の長さとその音声品質に関して定量的な評価を行うことを目的とする．合成範囲は，文単位（通常の音声合成）と当該合成範囲および先行 1 合成範囲+当該合成範囲+後続 1 合成範囲を使用する．また，先行 1 合成範囲+当該合成範囲+後続 1 合成範囲を使用した場合，音声波形間の不連続性による雑音が生じ，通常の音声合成と比べて音声品質が低下する．この影響を抑制する目的で，ハニング窓を用いて合成波形間の簡易的な平滑化を試みる．

評価結果を**表 7** に示す．表 7 より，後続のアクセント句を待ち合成した場合（先行 1 合成範囲+当該合成範囲+後続 1 合成範囲）は，当該アクセント句ごとに合成した場合（当該合成範囲）と比べて，客観評価と主観評価両方において改善される．また，検定を行った結果，1%未満の水準で有意差が確認できた．この結果より，後続の言語特徴を使用する場合においても後続の合成範囲が音声品質の改善に必要と考えられる．先行合成範囲＋当該合成範囲＋後続合成範囲の平滑化の有無による評価を比較すると，自然性は悪く，検定による有意差は確認できなかった．また，客観評価値はわずかに改善された．この結果より，平滑化により合成波形間の不連続性は低減したが，評価者に違和感を生じない平滑化の方法は，ハニング窓のような単純な方法では困難であると考えられる．音声品質を改善する平滑化として，深層学習における再帰型ネットワークを，HMM の代わりに使用することが考えられる．しかし，再帰型ネットワークを使用する場合，遅延時間は HMM より増加する傾向がある．再帰型ネットワークと後続の合成範囲を使用した場合における品質や遅延および平滑化の提案については，今後の課題である．

5. おわりに

先行研究による単語の逐次音声合成と異なり，より長い合成範囲を使用する日本語の逐次音声合成の提案について，音声品質のみではなく，遅延評価からその合成範囲の最適性を検討した．

後続の言語特徴を使用しない場合における実験において．まず，言語特徴の組合せにおける品質評価を行い，逐次音声合成の合成範囲として呼気段落とアクセント句を単位として検討した．次に，遅延評価を通して逐次音声合成の合成範囲について分析を行い，遅延の少ない逐次音声合成の合成範囲として，アクセント句単位が適していることを示した．最後に，合成範囲をアクセント句単位とする逐次音声合成において，後続の合成範囲を待つことが音声品質の改善に有効であることを示した．

後続の言語特徴が利用可能な場合においても実験を行った．合成範囲を文として，言語特徴の組合せにおける品質評価を行った．その結果，後続の言語特徴を使用する場合，MCD の改善が確認できた．しかし，主観評価値の改善は確認できなかった．合成範囲をアクセント句単位に変更した場合における音声品質の評価を行い，後続の言語特徴を使用する場合でも後続の合成範囲が音声品質の改善に必要であることを明らかにした．また，合成範囲ごとの音声波形に対して平滑化を行ったが，主観評価値は改善されなかった．

今後の課題としては，高品質化のための波形間の平滑化やより自然な音響特徴量生成のための深層学習に基づく再帰型ネットワークの使用があげられる．再帰型ネットワークの使用により，遅延時間は HMM より増加するため，その遅延が同時通訳システムへ適するか検討する必要があり，こちらも今後の課題である．

謝辞 本研究は科研費（JP21H05054，JP21H03467）の助成を受けたものである．

参考文献

[1] Matusov, E., Mauser, A. and Ney, H.: Automatic sentence segmentation and punctuation prediction for spoken language translation, *Proc. IWSLT*, pp.158–165 (2006).

[2] Fügen, C., Waibel, A. and Kolss, M.: Simultaneous translation of lectures and speeches, *Machine Translation*, Vol.21, No.4, pp.209–252 (2007).

[3] Bangalore, S., Rangarajan Sridhar, V.K., Kolan, P., Golipour, L. and Jimenez, A.: Real-time incremental speech-to-speech translation of dialogs, *Proc. NAACL*, pp.437–445 (2012).

[4] Fujita, T., Neubig, G., Sakti, S., Toda, T. and Nakamura, S.: Simple, lexicalized choice of translation timing for simultaneous speech translation, *Proc. INTERSPEECH*, pp.3487–3491 (2013).

[5] Shimizu, H., Neubig, G., Sakti, S., Toda, T. and Nakamura, S.: Constructing a speech translation system using simultaneous interpretation data, *Proc. IWSLT*, pp.212–218 (2013).

[6] Oda, Y., Neubig, G., Sakti, S., Toda, T. and Nakamura, S.: Optimizing segmentation strategies for simultaneous speech translation, *Proc. ACL*, pp.551–556 (2014).

[7] Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T. and Kitamura, T.: Speech parameter generation algorithms for HMM-based speech synthesis, *Proc. IEEE ICASSP*, Vol.3, pp.1315–1318 (2000).

[8] Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T. and Kitamura, T.: Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis, *Proc. EUROSPEECH*, pp.2347–2350 (1999).

[9] Baumann, T.: Partial representations improve the prosody of incremental speech synthesis, *Proc. INTERSPEECH*, pp.2932–2936 (2014).

[10] Baumann, T.: Decision tree usage for incremental parametric speech synthesis, *Proc. IEEE ICASSP*, pp.3819–3823 (2014).

[11] Pouget, M., Hueber, T., Bailly, G. and Baumann, T.: HMM training strategy for incremental speech synthesis, *Proc. INTERSPEECH*, pp.1201–1205 (2015).

[12] Pouget, M., Nahorna, O., Hueber, T. and Bailly, G.: Adaptive latency for part-of-speech tagging in incremental text-to-speech synthesis, *Proc. INTERSPEECH*, pp.2846–2850 (2016).

[13] Baumann, T. and Schlangen, D.: Evaluating prosodic processing for incremental speech synthesis, *Proc. INTERSPEECH*, pp.438–441 (2012).

[14] Yanagita, T., Sakti, S. and Nakamura, S.: Neural iTTS: Toward synthesizing speech in real-time with end-to-end neural text-to-speech framework, *Proc. SSW*, pp.183–188 (2019).

[15] Ma, M., Zheng, B., Liu, K., Zheng, R., Liu, H., Peng, K., Church, K. and Huang, L.: Incremental text-to-speech synthesis with prefix-to-prefix framework, *Proc. Findings of ENLNP*, pp.3886–3896 (2020).

[16] Mohan, D.S.R., Lenain, R., Foglianti, L., Teh, T.H., Staib, M., Torresquintero, A. and Gao, J.: Incremental text to speech for neural sequence-to-sequence models using reinforcement learning, *Proc. INTERSPEECH*, pp.3186–3190 (2020).

[17] Stephenson, B., Besacier, L., Girin, L. and Hueber, T.: What the future brings: Investigating the impact of lookahead for incremental neural TTS, *Proc. INTERSPEECH*, pp.215–219 (2020).

[18] Stephenson, B., Hueber, T., Girin, L. and Besacier, L.: Alternate endings: Improving prosody for incremental neural TTS with predicted future text input, *Proc. INTERSPEECH*, pp.3865–3869 (2021).

[19] Saeki, T., Takamichi, S. and Saruwatari, H.: Incremental text-to-speech synthesis using pseudo lookahead with large pretrained language model, *IEEE Signal Processing Letters*, Vol.28, pp.857–861 (2021).

[20] Saeki, T., Takamichi, S. and Saruwatari, H.: Low-latency incremental text-to-speech synthesis with distilled context prediction network, *Proc. ASRU* (2021).

[21] Grabe, E. and Low, E.L.: Durational variability in speech and the rhythm class hypothesis, *Papers in Laboratory Phonology*, Vol.7, No.515-546 (2002).

[22] Hirst, D. and Di Cristo, A.: *Intonation systems: A survey of twenty languages*, Cambridge University Press (1998).

[23] 鈴木雅之, 黒岩 龍, 印南圭祐, 小林俊平, 清水信哉, 峯松信明, 広瀬啓吉：条件付き確率場を用いた日本語東京方言のアクセント結合自動推定, 電子情報通信学会論文誌 D, Vol.96, No.3, pp.644–654 (2013).

[24] Yokomizo, S., Nose, T. and Kobayashi, T.: Evaluation of prosodic contextual factors for HMM-based speech synthesis, *Proc. INTERSPEECH*, pp.430–433 (2010).

[25] Yanagita, T., Sakti, S. and Nakamura, S.: Incremental TTS for Japanese language, *Proc. INTERSPEECH*, pp.902–906 (2018).

[26] 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, *Proc. SSW*, pp.294–299 (2007).

[27] 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).

[28] 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).

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

[30] Recommendation ITU-T P.800: Methods for subjective determination of transmission quality, *International Telecommunication Union* (1996).

柳田 智也 （学生会員）

2018 年奈良先端科学技術大学院大学博士前期課程修了．同年同大学院大学博士後期課程に進学．現在，逐次音声合成の研究に従事．

サクティ サクリアニ （正会員）

1999 年インドネシアバンドン工科大学情報学部卒業．2002 年ドイツウルム大学修士課程修了．2003 年音声言語コミュニケーション研究所研究員．2006 年（独）情報通信研究機構専門研究員．2008 年ドイツウルム大学博士課程修了．2009 年インドネシア大学コンピューターサイエンス学部客員教授．2011 年奈良先端科学技術大学院大学助教．2015 年フランス INRIA Paris-Rocquencourt の客員研究員．2018 年奈良先端科学技術大学院大学准教授，理化学研究所革新知能統合センター研究員．現在，北陸先端科学技術大学院大学准教授，奈良先端科学技術大学院大学客員准教授，理化学研究所革新知能統合センター客員研究員．JNS，SFN，ASJ，ISCA，IEICE，IEEE 各会員．

中村 哲 （正会員）

1981 年京都工芸繊維大学工芸学部電子工学科卒業．京都大学博士（工学）．1994 年奈良先端科学技術大学院大学助教授，1996 年米国 Rutgers 大学客員教授（文科省在外研究員），2000 年 ATR 音声言語コミュニケーション研究所室長，2005 年所長，2006 年（独）情報通信研究機構音声コミュニケーション研究グループリーダ，2010 年研究センター長，けいはんな研究所長等を経て，2011 年奈良先端科学技術大学院大学情報科学研究科教授，2017 年データ駆動型サイエンス創造センター長，2017 年理化学研究所革新知能統合研究センター観光情報解析チームリーダー．2003 年からカールスルーエ大学客員教授．音声翻訳，音声対話等の音声言語情報処理，自然言語処理の研究に従事．電子情報通信学会論文賞，AAMT 長尾賞，日本音響学会技術開発賞，人工知能学会研究会優秀賞，情報処理学会喜安記念業績賞，総務大臣表彰，文部科学大臣表彰，Antonio Zampoli 賞受賞，ドコモモバイルサイエンス賞，IBM Faculty Award，Google AI Focused Research Award 等受賞．ISCA 理事，IEEE SLTC 委員歴任．ATR フェロー，IEEE フェロー，ISCA フェロー，本会フェロー．