HMM 音声合成手法による早口音声合成の検討*

◎酒向 慎司, 西本 卓也, 嵯峨山 茂樹（東大・情報理工）

1 はじめに

現在，音声合成システムが積極的に利用されている場の一つに，視聴覚障害者が計算機を利用する際に用いられる「スクリーンリーダー」と呼ばれるソフトウエアがある．これは，計算機上のテキスト情報を合成音声として出力することで，他者によるサポートを介さず，即時に，さまざまな情報へのアクセスを可能にするものである．そのため，視覚障害者にとって音声合成システムへの期待は，その出現当初から非常に大きく，また，性能や実用的な側面からの要求も様々である．本研究では，現在のスクリーンリーダーでは，高速な発話速度の合成音声が必要とされている点に着目し，HMM 音声合成手法において，高速な発話速度でも聞き取りやすい合成音声を実現するための手法について検討する．

2 スクリーンリーダー

スクリーンリーダーでは，音声合成システムが必要不可欠なものであるが，汎用的な音声合成システムをベースとしたものが多く，実用上の様々な観点からの要求もある．その中の一つとして，柔軟な発話速度の制御，特に速い話速での読み上げがある．その背景には，合成音声から画面上の情報を得ているため，「斜め読み」のような時間を飛ばした操作が出来ないことから，取得できる情報の量は発話速度に依存することになる．スクリーンリーダーの利用者は，より速い合成音声が得られるよう音声合成エンジンの発話速度を調整する傾向があり，習熟した利用者ほどその高速性の要求が強く，実装されている以上の発話速度までもが求められている．

このような状況を踏まえ，発話速度の速い合成音声における習熟度合いについての研究も進められている [1]．なお，文章要約によって情報を簡約化するアプローチも考えられるが，ここでは合成音声の発話速度を精度よく向上させることを主眼に置く．

3 音声合成における発話速度の制御

発話速度を柔軟に制御することは，自然な対話音声を実現する上でも，一般的に音声合成の重要な課題であり，これまでにも自然な発話速度の変化を実現するための試みがいくつか報告されている [2], [3].

これらの研究でも言及されているが，発話速度の変化による文章内の時間構造の変化は，一様な時間伸縮では表現できないことから，実際に早口音声を収録することによって，発話速度の違いを反映したモデル学習が可能になると期待される．そこで，実験用のデータベースとして，通常な発話速度に加えて，

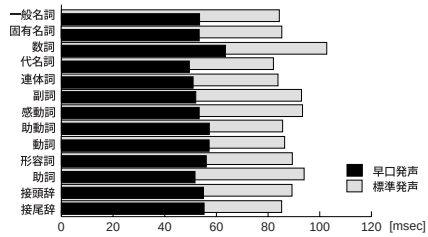

Fig. 1　品詞別にみた発話速度の比較

早口で発話した音声を収録した．ただし，データの収録にあたって，厳密な発話速度を定めて収録することは難しいため，通常発話の 2 倍程度を目標値とした指示を与えるだけに留め，あとは発話者の裁量によって目標の発話速度になるように読み上げた音声を収録した．発話された音声データから得られたポーズおよび無音を除いた区間を比較したところ，標準発声の約 1.58 倍の発話速度の早口音声が得られた．

音素セグメンテーションによって得られた音素境界に基づいて，標準音声と早口音声の品詞ごとに分類した平均音素長の差異を Fig.1 にまとめた．図からは，標準音声では，品詞の分類に応じて継続長の変動が大きく，早口音声では品詞ごとの差が少なくなり，またその分散も少なくなっていることが分かる．つまり，標準音声では局所的な発話速度は，文章の構造などの周辺情報によって何らかの影響を受けていると考えられる一方，早口音声ではそれが少なくなり，一定調になりがちといえる．また，早口音声では文中のポーズも極端に短くなっていることも早口音声の特徴である．

3.1 HMM 音声合成における話速制御

本研究では，隠れマルコフモデル (Hidden Markov Model, HMM) に基づいた音声合成手法を用いる [4]. この手法では，音声データから得られるメルケプストラム係数と基本周波数 F_0 の系列を特徴量として HMM を学習し，HMM から上記の音声パラメータを合成するものである．時間構造のモデル化は，HMM の学習過程で得られる，各状態における滞在確率を多次元ガウス分布によってモデル化することで，学習データに基づいた継続長モデルを得ることができる．

合成時には，与えられたテキストから状態継続長分布に基づいて HMM の各状態の時間構造を決定し，それに従って音声パラメータを合成する．発話速度を制御する場合，学習データにおける平均的な発話速度から，継続長モデルの分散に基づいて各音素の状態を伸縮させることによって，文章全体を一様に伸縮させるのではなく，学習データにおける様々な要因に基づいた話速の制御を行うことができる．

しかし，学習データの発話速度から極端に発話速

* A study on rapid speech synthesis using HMM-based speech synthesis technique. by Sako Shinji, Nishimoto Takuya, Sagayama Shigeki (Graduate School of Information Science and Technology, The University of Tokyo).

度を縮めた場合には，状態継続長分布の伸縮によって音素の一部，あるいは音素全体が欠落してしまう可能性もある．また，発話速度の違いによって，それに応じた発話様式があり，通常の発声速度とは音韻的な，また時間的なモデル構造が異なることも考えられる．つまり，標準の発話速度の音声データだけでは，話速の変化に対応したモデルを充分に表現できないことになる．

4 発話速度に依存したモデル学習

異なる発話速度の音声データから，発話速度に応じた継続長モデルを得るため，発話速度をコンテキストの要因として加える．これにより，発話速度の変化における時間構造の違いを明示的に反映した継続長モデルを構成することが可能となる．本研究では，発話速度に関して文章内の局所性の有無を考慮して，文章単位と形態素単位の単位時間当たりのモーラ数を発話速度情報として付与した．

各音素ごとに該当する周辺区間を含んだ発話速度をコンテキストとして与えることで，異なる発話速度の混在した学習データに対してコンテキストクラスタリングを行う．これにより，時間構造だけでなく，スペクトルや F_0 の各分布について，発話速度の異なるデータ間での共有構造を自動学習することができる．とくに早口音声では，音素あたりのフレーム数が減少するため，発話速度ごとに単独で学習するよりも統計モデルを学習する上で有利であると考えられる．

4.1 実験

早口発声と標準発声の音声データをそれぞれ 100 文章を用いて F_0 抽出，18 次メルケプストラム分析を行い，5 状態 left-to-right 型 HMM の学習を行った．メルケプストラム，F_0，状態継続長の各分布に対して，[5] にて用いられているコンテキスト要因に発話速度を加え，コンテキストクラスタリングを行った．なお，発話速度に関する要因として，形態素区間の平均モーラ速度（提案法 A）と文章全体の平均モーラ速度（提案法 B）をもちいた．

上記の二つのモデルと，同様の条件で標準話速と早口話速の音声データ単独から学習されたモデルの合計 4 種類を用いて，ポーズを除いた発話速度が標準発話速度の 2 倍となるように音声を合成し，対比較試験によって評価を行った．

4.2 結果と考察

クラスタリングの結果から，状態継続長分布では全体の 1/4 が発話速度に関する質問を占め，メルケプストラムおよび F_0 の分布でも 10%前後の発話速度に関する質問が適用されていることから，発話速度に応じた時間構造を反映したモデルが得られていることが確認できる．合成された音声の音素時間長の比較を Fig.2 に，8 名の被験者で行った主観評価実験の結果を Fig.3 に示す．標準速度のデータから時間を伸縮させて合成された音声は，他のモデルと比べて，音素

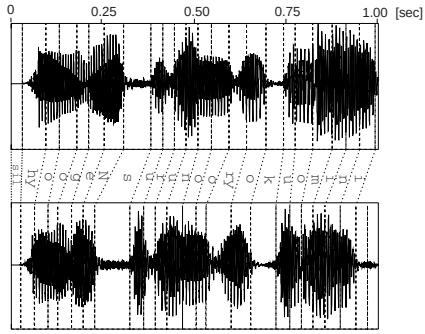

Fig. 2　合成された音声の時間構造の比較（上:標準話速モデル，下:提案法 B）

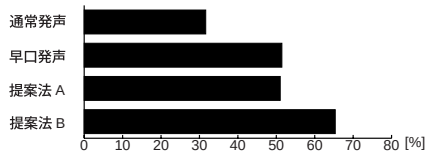

Fig. 3　主観評価実験の結果

系列の時間構造が異なるものが合成されており，大幅な時間伸縮に対して充分に対応できていないことが確認できた．また，発話速度を反映させたモデルが，標準速度のモデルを伸縮した合成音声に対して良い評価を得ていることが確認できる．

5 むすび

本研究では，通常発声から高速な発話速度まで柔軟に対応した音声合成システムを実現するための検討として，局所的な発話速度の変化を考慮したモデル化を試みた．主観評価実験の結果から，標準話速の音声から学習されたモデルと比較して，速い発話速度において良い評価が得られた．本手法は，特定の早口発声の場合に限定したものではなく，発話速度に応じた時間構造のモデル化を主眼としており，様々な発話速度の音声データからモデルを学習することによって，柔軟な発話速度の制御が期待できる．

謝辞　本研究の一部は，科研費特定研究「情報副詞の基礎」Kiki 班および View-Net 神奈川，神奈川工科大学，(株）プロシードとの共同研究の成果である．関係者各位に感謝する．

参考文献

[1] 西本卓也, 酒向慎司, 嵯峨山茂樹, 早口合成音声の聴取実験によるテキスト音声合成の評価, 信学技報, WIT2005-5, pp.23-28, 2005.
[2] 外川太郎, 他, HMM 音声合成における数量化 I 類を用いた発話速度制御法, 音響論（春）, pp.345–347, 2002.
[3] 舛田剛志, 他, 発話速度の異なるデータベースを用いた音声合成手法の検討, 信学技報, vol.101, no.122, pp.61–68, 2002.
[4] 益子貴史, 他, 動的特徴を用いた HMM に基づく音声合成, 信学論 (D-II), vol.J79-D-II, no.12, pp.2184–2190, 1996.
[5] 吉村 貴克, 他, HMM に基づく音声合成のためのスペクトル，ピッチ，状態継続長のモデル化, 信学技報, vol.99, no.255, pp.33–38, 1999.