早口合成音声に対する聴取者の慣れの効果の検討*

○西本卓也, 酒向慎司, 嵯峨山茂樹 (東大),
△大島一恵, △小田浩一, △渡辺隆行 (東京女子大)

1　はじめに

多くの視覚障害者はスクリーンリーダなどの音声化ソフトを使用し，合成音声によって情報を得ている．このような環境では，短い時間で効率よく情報を得ることが重要である．肉声を伸縮変換して作った早口音声を視覚障害者に聴取させた実験 [1] によると，上級者の最適速度（十分な文理解が得られる速度）は 1100-1170 モーラ／分，最高速度（文理解率 50%程度が得られる速度）は 1400-1500 モーラ／分とされる．しかし，このような早口に対応していない音声合成エンジンは多く，音声合成研究の立場からは早口音声合成について十分に検討されているとは言えない．

我々は HMM 音声合成を用いて，早口コーパスから学習した話者モデル（早口モデル）により 1500 モーラ／分（25 モーラ／秒）以上の話速を実現し，4 桁の数字を聴取する予備評価実験を行った [2]．特に，聴取者の慣れの効果に関して，3 人の被験者が短時間で早口合成音声に慣れており，その効果は 3 日後の実験においても持続する，ということを確認した．

本報告では，この予備評価実験を踏まえて，実験の規模を拡大し，慣れの効果に関する詳細な検討を行う．また，複数の話者モデルを聴取率の観点から比較する手法を提案し，特に早口モデルの有効性について検討する．

2　早口合成音声の作成

本研究では音声合成エンジンとして，Galatea Project [3] および音声対話技術コンソーシアム (ISTC) [4] で開発が進められている GalateaTalk と HTS (HMM-Based Speech Synthesis System) [5] を使用する．各音素は 5 状態の HMM（隠れマルコフモデル）で構成される．GalateaTalk からの主な変更点として，フレームシフトを 5ms から 2ms に変更した．

また，発話速度が早い音声コーパスから話者モデルを構築すれば自然な早口音声を合成できると考え，ATR 音素バランス 503 文のうち，短くて発話が容易な 100 文を選び，この同一の文セットの読み上げ音声として，プロの男性ナレーター 1 名の標準的な速度の音声（標準コーパス）とその約 2 倍の速さで発話させた音声（早口コーパス）を収集した．これらを用いて，標準モデル (Normal) および早口モデル (Fast) の話者モデルを構築した．早口モデルを用いることで標準モデルの約 1.6 倍の話速の合成音声を生成できる．なお，2 つの話者モデルはパワーの平均値が同じになるように補正している．

3　聴取率による評価実験

本実験では「番号は 1234（イチニーサンヨン）です」のように 4 桁の数字をさまざまな話速で聴取させる．聞き取れない桁があった場合は「1?34」のように「?」を用いて回答させた．被験者が回答した数字を，桁ごとの一致・不一致に基づいて集計した．

実験は 3 週間にわたって行った．第 1 週の実験には被験者 19 人が参加した．そのうち 5 人は第 2 週および第 3 週にも引き続き実験に参加した．データ採取時の不具合などを考慮し，第 1 週では 16 人分，第 2 週および第 3 週では 4 人分のデータを分析対象とした．全員，HMM 音声合成の評価の経験がなく，晴眼者でかつ健聴者であり，日本語を母国語とする大学生および大学院生（女性）である．

実験には 1 人 1 台のノート PC (Microsoft Windows XP) とヘッドフォン (Panasonic RP-H750-S) を使用し，ウェブブラウザで音声を提示して回答をキーボードで入力させた．

各週の実験の構成は，試行 T1（200 課題）に続いて，訓練（25 課題）を行い，試行 T2（200 課題）を行う．試行 T1 は実際には 50 種類の音声（数字・話速・話者モデルの出現回数のバランスを考慮）を 4 回反復し，試行 T2 では同じ課題を逆順で 4 回反復する．

標準モデルおよび早口モデルを用いて課題用の音声ファイル 50 個を作成した．早口モデルのデフォルトの発話時間の 0.7-0.3 程度を話速の目標値とした．しかし，各状態のフレーム数が 1 以上になるように制限を設けたため，話速の目標値と実現値が一致しない．後述する評価においては，話速の実現値と聴取率の関係を検討する．

訓練では提示する音声は試行と同じく「番号は XXXX です」という文型であるが，被験者が回答を行った後に，一定のゆっくりした速度で「答えは XXXX です」という音声を提示し，フィードバックを行う．

---

*Listener's familiarizations to text-to-speech synthesizers at fast speaking rates. by Nishimoto Takuya, Sako Shinji, Sagayama Shigeki (The University of Tokyo), Ohshima Kazue, Oda Koichi and Watanabe Takayuki (Tokyo Woman's Christian University)

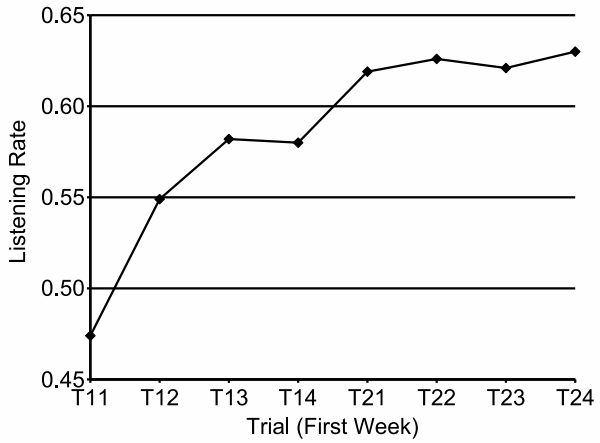

Fig. 1　第 1 週の各課題セットにおける聴取率

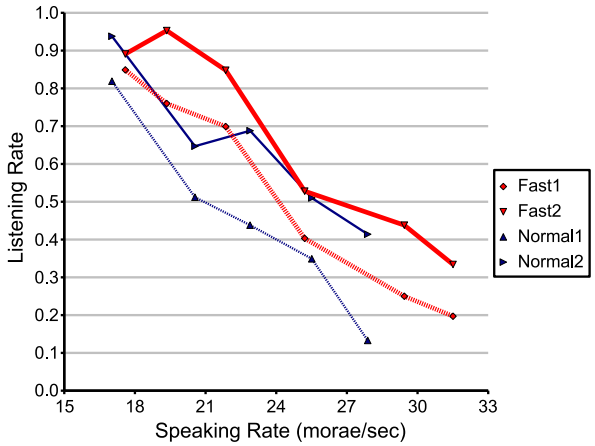

Fig. 2　第 1 週の最初と最後の課題セットにおける話者モデルごとの話速と聴取率

4　実験結果

Fig. 1 に，第 1 週の実験において同一の課題セットを 8 回行った間の聴取率の変化を示す．課題セットは 50 個の 4 桁数字の聴取であり，前半に 4 セット (T11-T14)，前述した訓練を挟んで後半 4 セット (T21-T24) である．聴取率は 4 桁を個別に集計した被験者 16 人の平均値である．特に最初の段階での慣れの効果が大きく，また，T14 と T21 の間で行った訓練にも効果が認められ，訓練後は成績が安定している．

Fig. 2 に，第 1 週における話者モデルごとの話速と聴取率（被験者 16 人の平均値）の関係を示す．Fast1/Fast2 はそれぞれ T11（最初の課題セット）および T24 （最後の課題セット）における早口モデルでの聴取率であり，Normal1/Normal2 は標準モデルでの聴取率である．50 種類の課題を話速に応じて 6 群に分割し，話者モデルごとの各群の平均話速値と平均聴取率を求めた．この結果によると，早口モデルと標準モデルの両者において，慣れの効果により聴取率が上昇している．また，慣れる前 (T11) および慣れた後 (T24) のいずれの場合でも早口モデルは標準モデルと比較して，同じ話速における聴取率が高く，早口モデルの有効性が認められる．

Fig. 3 に，第 1 週 (W1) から第 3 週 (W3) までの各試行 (T1, T2) の聴取率を示す．聴取率は被験者 4 人の平均値である．試行 T1 は前半の 4 課題セット (T11-T14) を，試行 T2 は後半の 4 課題セット (T21-T24) をそれぞれ平均した．この結果によると，聴取率は 3 週間にわたって上昇し続けている．また，W1T2 と W2T1 の間のように，1 週間をおいて実験を行った場合でも，聴取率は低下せず，逆に高まっている．このことから慣れの効果が長期間にわたって持続することが伺える．

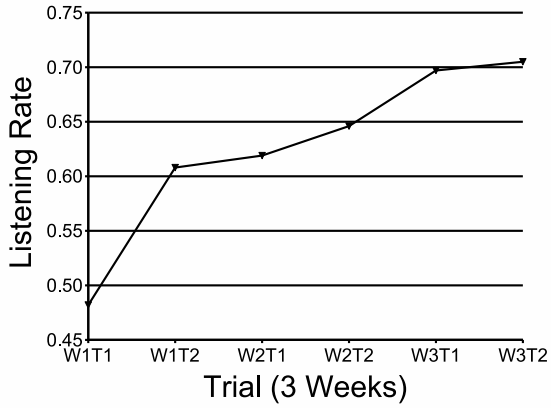

Fig. 3　第 1 週から第 3 週までの聴取率の変化

5　まとめ

HMM 方式による早口音声合成の検討を行い，聴取率による評価を試みた．その結果，聴取者の慣れによる効果の大きさや持続性，早口話者モデルの有効性などが確認された．今後は聴取率に影響を与える要因についてさらに詳細な検討を行いたい．また，数字聴取以外のタスクでの評価を行いたい．

謝辞　本研究の一部は科研費特定研究「情報福祉の基礎」Kiki 班および「リアルスクリーンリーダー」プロジェクトの成果である．

参考文献

[1] 浅川, 高木, 井野, 伊福部：視覚障害者への音声提示における最適・最高速度, ヒューマンインタフェース学会論文誌, Vol.7, No.1, pp.105-111, Feb 2005.
[2] 西本, 酒向, 嵯峨山, 小田, 渡辺: 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, May 2005.
[3] http://hil.t.u-tokyo.ac.jp/~galatea/
[4] http://www.astem.or.jp/istc/
[5] http://hts.ics.nitech.ac.jp/