社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

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

西本　卓也† 　酒向　慎司† 　嵯峨山茂樹† 　小田　浩一†† 　渡辺　隆行††

† 東京大学大学院情報理工学系研究科
†† 東京女子大学現代文化学部
E-mail: {nishi,sako,sagayama}@hil.t.u-tokyo.ac.jp, k-oda@twcu.ac.jp, nabe@lab.twcu.ac.jp

あらまし　視覚障害者はスクリーンリーダなどの利用において，速い話速でありながら聞き取りやすい合成音声を求めている．我々は，テキスト音声合成エンジン GalateaTalk と早口音声コーパスを用いて，聞き取りやすい早口の合成音声を実現した．また，4 桁の乱数の読み上げ音声を聴取する実験を行い，早口コーパスから学習した統計量を用いることで早口でかつ十分な音質の合成音声を得られること，早口音声の聴取タスクにおいては短時間の訓練によって大きな学習効果が得られ，その効果は時間が経過しても持続すること，などの知見を得た．

キーワード　音声合成，話速，早口，学習効果，GalateaTalk

Evaluation of text-to-speech synthesizers at fast speaking rates

Takuya NISHIMOTO†, Shinji SAKO†, Shigeki SAGAYAMA†, Koichi ODA††, and Takayuki WATANABE††

† The University of Tokyo
†† Tokyo Woman's Christian University
E-mail: {nishi,sako,sagayama}@hil.t.u-tokyo.ac.jp, k-oda@twcu.ac.jp, nabe@lab.twcu.ac.jp

Abstract　The text-to-speech synthesizer that makes easy-to-listening voices at very fast speaking rates is expected to help visually impaired persons to acquire information effectively with screen reading softwares. We realized a fast and intelligible text-to-speech engine, using GalateaTalk synthesizer with the corpus with fast speaking rate. Our preliminary experiments of listening random 4 digits numbers showed that the statistical models trained with the fast speaking corpus was effective, and prominent learning effect was observed and the effect sustains for at least several days.

Key words　Text-to-speech, Speaking rate, Fast speaking rate, Learning effect, GalateaTalk

1. はじめに

視覚障害者がコンピュータを利用できることの社会的意義は大きい．電子メールやウェブを使うことで，即時性のある情報の送受信やコミュニケーションの手段を得ることができる．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．

音声出力による情報提示においては，短い時間で効率よく情報を得ることが重要である．画面表示，特に GUI(Graphical User Interface) を用いたコンピュータ操作では，表示する情報を構造化しデザインやレイアウトに配慮することで，ユーザは，重要な情報と読み飛ばすべき情報を区別し,「ななめ読み」を行うことが可能である．音声においては「ななめ聞き」は一般に困難であるが，統語情報を用いて品詞単位で音声合成エンジンの話速を制御する手法 [1] などが提案されている．

視覚障害者が短時間で多くの情報を確実に得るためには，音声合成エンジンそのものが，話速が速い場合においても明瞭性が高く，内容も聴取可能である，といった性能を有することが重要である．

スクリーンリーダを利用する視覚障害者の実態調査 [2] によると，多くのユーザは合成音声において話速の設定を最高値にしている．これは一般的な朗読音声の約 2 倍の速度であるが，さらに高速での読み上げを望む声もある．

また，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 [3] 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／分である，という知見が得られている．これに対して，既存の音声合成エ

ンジンの最高速度は 878 モーラ／分であり，より高速な読み上げに対応した音声合成エンジンが必要であるとされる．

従来，音声合成エンジン（日本語テキスト音声合成システム）は通常の話速での自然性や表現力などを高めることを主な目標として開発されてきたため，視覚障害者が望む早口での合成を前提として開発されていない．たとえ人間が発話できる程度の速さであっても，通常話速で合成される音声波形を伸縮して出力する場合には，自然な発話とは異なる早口合成音声が生成される可能性が高い．また，そもそも人間に発話できないほどの早口の音声については，どのような音声が聞き取りやすいのか，明確な指針はない．

早口音声の聴取については，訓練の効果も考慮する必要がある．例えば，多くの視覚障害者は，訓練によって，自然性や明瞭性の乏しい早口合成音声を聞き取る能力を得ていると考えられる．一方で，スクリーンリーダの利用法の学習を容易にしたり，利用者の疲労を防ぐためには，できるだけ訓練をしなくても聴取できる合成音声を用いることが重要である．

本稿では，Galatea（ガラティア）プロジェクト [4], [5] において開発中のオープンソースのテキスト音声合成エンジン GalateaTalk（ガラティアトーク）を用いて，聞き取りやすい早口での音声合成を実現するための検討を行う．また，実現した早口音声の性能を評価するために，4 桁の乱数を聴取する実験を行う．さらに，訓練による慣れの効果についても検討を行う．

本報告の第 2 章では，使用する音声合成エンジン GalateaTalk の概要について述べる．第 3 章では，話速の異なるコーパスの収集と構築できる話者モデルに関する検討を行う．第 4 章では，本研究で行う実験で比較する音響モデルについて述べる．第 5 章では，乱数の聴取実験の手順について述べる．第 6 章では実験の結果について，第 7 章では考察について述べる．第 8 章ではまとめと今後の課題について述べる．

2. GalateaTalk の概要

GalateaTalk は HMM (Hidden Markov Model, 隠れマルコフモデル) 方式によるテキスト音声合成エンジンである．2000 年〜2002 年度に情報処理技術振興協会 (IPA) の支援で行われた Galatea プロジェクトの成果物として 2003 年 8 月より無償配布を行っている擬人化音声対話エージェントツールキット (Galatea Toolkit) の一部である．東工大・名工大・奈良先端大・東大などに所属する研究者が中心となって開発を進めており，音声対話技術コンソーシアム (ISTC) [6] の活動として改良が行われている．ISTC が毎年開催する講習会では音声認識や顔画像合成などの技術と合わせて GalateaTalk の技術に関する講義や演習も行っている．

GalateaTalk は，テキスト解析部，音声合成部，コマンド解析部，音声出力部の 4 つから構成される．各部分の概要は次の通りである．

（1） テキスト解析部：読みとアクセントの辞書に基づいてテキスト（文字列）を読みやアクセントなどの情報が付与された音素ラベル系列に変換する．テキスト解析部は，日本語形態素解析システム「茶筅」，日本語電子化辞書 UniDic,「茶筅」の出力の後処理を行う XML 処理系 ChaOne から構成される．

（2） 音声合成部：あらかじめ学習された話者モデルと，テキスト解析部で得られたラベル系列を用いて，音声波形を合成する．話者モデルは，音声コーパスから学習されたピッチ，スペクトル情報（メルケプストラム），継続長モデルから構成され，これらは木構造クラスタリングされた統計量によって表現される．

（3） コマンド解析部：GalateaTalk の通信仕様に従ってコマンドを処理する．例えば，以下のコマンドで音声を出力することができる．

set Speaker = male01
set Text = こんにちは.
set Speak = NOW

（4） 音声出力部：音声出力デバイスの制御を行う．

現在，GalateaTalk には Linux(UNIX) 版および Windows 版が存在する．本稿では音声対話技術コンソーシアム (ISTC) が配布中の Windows 版 GalateaTalk (注1) を使用する．

GalateaTalk では日本語テキスト音声合成用記号 (JEIDA-62-2000) に準拠したマークアップを入力テキストに含めることができ，文字列の一部の話速や韻律などを変更する，といったことが可能である．

3. コーパスからの話者モデル構築

GalateaTalk で利用可能な話者モデルは，HTS(HMM-Based Speech Synthesis System) [7] と呼ばれるツールを用いて音声コーパス（大規模な音声データベース）から独自に構築することができる．経験的には，ATR 音素バランス 503 文の読み上げ音声と，その音声に関するラベル（品詞やアクセント位置などの情報）があれば，高品質な話者モデルを自動学習によって得ることができる．

我々は，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た．しかし，多数の文章を早口で読み上げることは容易ではない．そこで，本研究では ATR 音素バランス 503 文のうち，短くて発話が容易な 100 文を選んだ．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 2 倍速音声(注2) を聞かせて同じように発話させた音声（早口コーパス）を収集した．

２種類のコーパスで学習した話者モデルを用いて，文字列「番号はイチニサンヨンです.」（14 モーラ）を読み上げた場合の話速の比較を表 1 に示す．また，生成される音声の，各音素の継続時間長（GalateaTalk が出力する内部情報）を以下に示す．ただし数値の単位はミリ秒である．GalateaTalk はフレームシフト（音声を分析・合成する時間区間の進み方の単位）を 5 ミリ秒として処理を行うため，音素の継続時間は 5 ミリ秒の倍数となっている．

(注1)：独自の修正を施しているが，ベースとなるモジュールのバージョンは次の通りである．gtalk-1.2beta, chasen-2.3.3p1, chaone-1.1.0, unidic-sjis-1.1.2

(注2)：標準コーパスの音声を変換ツール sox の stretch 機能で 2 倍速にした音声.

表 1　文字列「番号はイチニサンヨンです」の話速の比較

コーパス	文継続長 (ms)	速度（モーラ／分）
標準コーパス	2090	402
早口コーパス	1325	634

表 2　早口音声の話者モデル

モデル名	音素継続長	ピッチ	スペクトル
fast1	早口	早口	早口
fast2	早口	標準	標準

- 標準コーパス：

b[70] a[95] N[100] g[65] o[100] o[80] w[95] a[100] i[95] ch[105] i[90] n[55] i[95] s[100] a[90] N[100] y[85] o[85] N[100] d[45] e[95] s[165] U[80]

- 早口コーパス：

b[45] a[70] N[60] g[45] o[45] o[70] w[45] a[50] i[60] ch[65] i[40] n[50] i[60] s[75] a[65] N[65] y[70] o[55] N[65] d[30] e[55] s[80] U[60]

早口コーパスを用いることで，標準コーパスに対して約 1.6 倍の早口音声を生成できることが確認された．

4. 評価する話者モデル

評価実験に用いる早口の話者モデルとしては表 2 に示す 2 種類を用いる．fast1 モデルはすべての情報を早口コーパスの統計量を用いたものである．fast2 モデルは，音素継続長のみ早口コーパスの統計量を用い，ピッチおよびスペクトルの情報は標準コーパスの統計量を用いている(注3)(注4)．

fast2 モデルでは，性質の異なるコーパスの統計量を併用するため，モデル間のミスマッチによる品質の劣化が懸念される．しかし，ピッチとスペクトルの情報に関しては，学習に用いる音声の総時間が長いほど，多くの学習データに基づく統計量が得られるため，音質が向上する可能性がある．

また，もし fast1 と fast2 の音質に明確な差がないとすれば，多様な声質の音声コーパスを標準話速のみで収集し，別の一人の話者の早口コーパスから得た音素継続長と併用することで，多様な声質の早口合成音声を容易に実現できる可能性がある．

5. 聴取実験の手順

前述の 2 種類の早口話者モデルを，さまざまな話速で使用し，GalateaTalk における音声の聞き取りやすさを評価する聴取実験を行った．

被験者は晴眼者の大学生（女性）3 名である．

聴取する課題は，文脈による内容の予想を回避し，内容に対する被験者の親密度の差が生じることを防ぐために，4 桁の乱数の聴取とした．提示する音声は「番号は XXXX です」という定型文の中に埋め込み，位取りをせず「イチニサンヨン」のように 1 桁ずつ読み上げることとした．

表 3　聴取実験の構成

実験日	課題	課題数	所要時間
1 日目	試行 t1	200	約 15-20 分
	訓練	50	約 7-10 分
	試行 t2	200	約 15-20 分
2 日目	試行 t3	200	約 15-20 分
	訓練	50	約 7-10 分
	試行 t4	200	約 15-20 分

評価する話速は，GalateaTalk の話速制御タグ (例：<RATE SPEED=”0.5”> こんにちは </RATE>) によって，0.3, 0.4, 0.5, 0.6, 0.7 の 5 段階とした．5 種類の速度と 2 種類の話者モデルの 10 種類の組み合わせが各 20 回ずつ（計 200 回）ランダムな順序で出現するように課題を設定した．読み上げる 4 桁の数字はランダムな数値とした．

実験構成を表 3 に示す．1 日目は試行 t1（200 課題）に続いて，訓練（50 課題）を行い，試行 t2（200 課題）を行う．1 日目の実験の 3 日後に 2 日目の実験として，試行 t3（200 課題）に続いて，訓練（50 課題）を行い，試行 t4（200 課題）を行う．試行 t1-t4 は同一の内容の課題であるが，出題順序を入れ換えている．

訓練は，GalateaTalk の合成音声に慣れることを目的とする．ただし，試行 t1-t4 で用いる話者モデルとは異なる話者（男性 3 話者，女性 5 話者の計 8 話者，ATR 503 文で学習したモデル）をランダムに用いる．また，fast2 と同様に，音素継続長に関しては早口コーパスの統計量を用いる．

訓練で提示する音声は試行 t1-t4 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．

本実験のために Eclipse SWT(Standard Widget Toolkit) で実装した GUI を有する Java アプリケーションを実装した．画面を図 1 に示す．このアプリケーションから Windows 版 GalateaTalk を外部プロセスとして呼び出し，課題提示時に逐次，音声合成を行っている．なお，実験に用いた PC は Panasonic Let’s Note CF-T1 (Windows XP Pro) である．音声の聴取にはヘッドフォン (SONY MDR-CD480) を使用した．操作はマウスで行うこととした．

被験者は数字ボタンを 4 回押し，回答ボタンを押して次の課題に進む．わからない桁がある場合は数字の代わりに「？」を選んでもよいこととした．

被験者に行った教示は次のようなものである．

（1） ヘッドフォンから「番号は XXXX です」という音声が聞こえます．

（2） 音声の速度は基本的にすべて速く，いろいろな種類の速度で読み上げられます．

（3） 聞き取った 4 つの数字をマウスで入力し，[Answer] ボ

(注3)：現在のバージョンの GalateaTalk では，set Speaker コマンドで早口コーパスのモデルを設定し，<VOICE OPTIONAL=”xxx”> タグにより標準コーパスの話者モデルを指定すると，このような設定を簡単に実現できる．

(注4)：本来は異なる話者モデル間で音素の各状態が正しく対応するとは限らない．したがって，継続長モデルを取り換えることと音素継続長を流用することは意味が異なる．本実験ではこのミスマッチは無視できると仮定している．

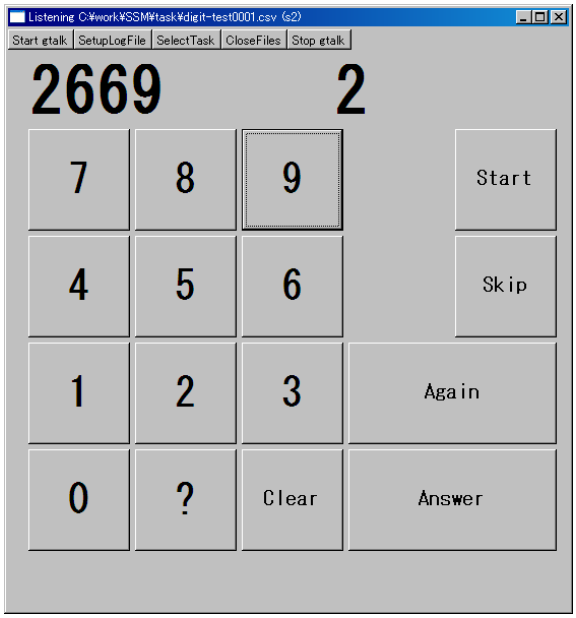

図 1　聴取実験用アプリケーションの画面

タンを押します．自動的に次の問題が始まります．

（4） うまく聞き取れず，数字が分からない場合は，[?] ボタンを押します．(例： [1] [?] [3] [0])

（5） 入力を間違えた場合は，[Clear] ボタンを押すと 4 つすべての数字が消去されます．

（6） 音声を聞き直したいときは [Again] ボタンを押すと，再度聞くことができます．ただし，なるべく 2 度聞きはせず，1 回で回答すること．

（7） 数字を 4 つ入力したことを確認してから [Answer] ボタンを押してください．

（8） 回答時間などに焦る必要はありません．

正解数の計算については，桁の脱落や挿入などの誤りは考慮せず，各桁について単純に正解と回答を比較し，一致した数を合計した．最高値は 4 である．

6. 実験結果

本課題で生成された音声の文継続長（発話文全体の長さ）を，RATE タグで指定した SPEED 属性値（以後「RATE 値」と表記する）ごとに集計した値を表 4 および図 2 に示す．ただし，RATE 値の 0.8-1.2 は聴取実験では使用しなかった値である．なお，fast1 と fast2 は同じ継続長モデルを用いているため，同一の文においては生成される音声の継続長は同じとなる．

聴取実験の正解数（1 つの課題で 4 桁すべて正解していれば 4 とする）被験者 s1-s3 および RATE 値 (0.3-0.7) についてそれぞれ平均を求めたものを表 5 および図 3 に示す．

聴取実験の平均正解数における学習効果を表 6 および図 4 に示す．

7. 考　　察

表 4 および図 2 から，GalateaTalk の RATE タグで指定した値は生成される文継続長にほぼ比例しており，GalateaTalk がタグの指定に従って適切に話速制御を行ったことが確認できる．ただし，GalateaTalk においては，生成される音声が均等に時間伸縮するのではなく，5 状態で構成される音素の中で，

表 4　文字列「番号は XXXX です」の平均文継続長

RATE 値	平均文継続長 (ms)
0.3	429
0.4	543
0.5	679
0.6	819
0.7	952
0.8	1092
0.9	1210
1.0	1362
1.1	1504
1.2	1628

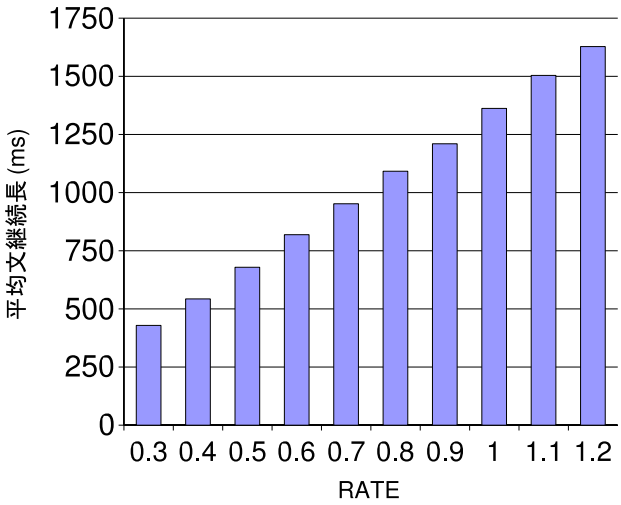

図 2　文字列「番号は XXXX です」の平均文継続長

表 5　聴取実験の平均正解数 (最高値=4)

RATE 値	平均正解数 (fast1)	平均正解数 (fast2)
0.3	1.033	0.954
0.4	1.946	1.946
0.5	2.667	2.717
0.6	3.292	3.408
0.7	3.587	3.663

表 6　聴取実験の平均正解数における学習効果 (s1-s3 は被験者)

試行	全員	s1	s2	s3
t1	2.250	2.065	2.145	2.540
t2	2.590	2.320	2.595	2.855
t3	2.582	2.440	2.520	2.785
t4	2.663	2.460	2.625	2.905

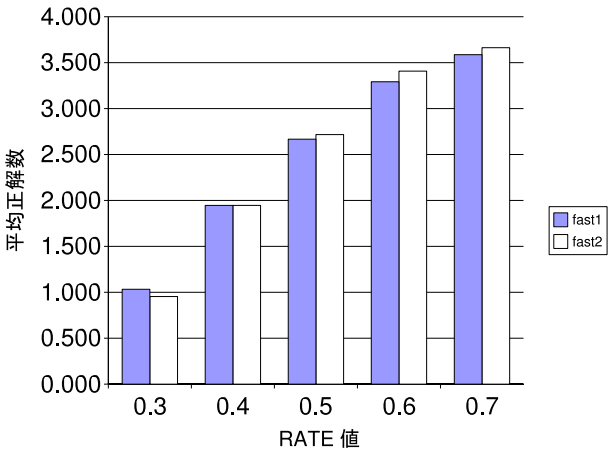

図 3　聴取実験の平均正解数 (最高値=4)

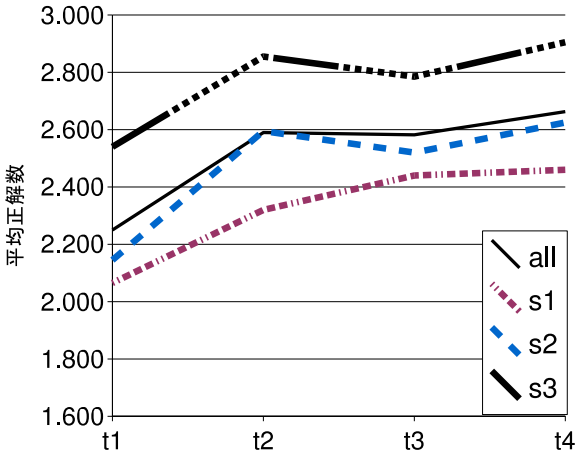

図 4　聴取実験の平均正解数における学習効果 (s1-s3 は被験者，all は全員の平均)

各状態の継続時間長の統計量（分散）に応じて伸縮させている．したがって，学習データに忠実で自然性の高い時間伸縮が実現できる．

次に，表 5 および図 3 からは，RATE 値が大きくなる（話速が遅くなる）ほど，平均正解数が増えていること，モデルの違いによる平均正解数の差はみられないこと，などが確認できた．特に，RATE 値 0.4 という条件で正解率がほぼ 50%になっており，先行研究 [3] が提案する「最高速度」の定義とほぼ一致する条件となる．このときの話速は約 1500-1600 モーラ／分である．タスクが異なるため単純な比較はできないが，上級者が要求する最高速度に匹敵する．これらのことから，本研究で検討した早口音声は実用に耐えるとことが期待される．

また，RATE 値 0.4 では約 2100 モーラ／分となるが，この条件での平均正解率は約 1/4 となり，現時点では実用的とはいえない．これについては，さらに音響モデルやエンジンの実装を改良することで，より聴取されやすくすることが可能であるか，今後の検討が必要である．

表 6 および図 4 で示される学習効果については，被験者ごとの個人差が見られるものの，試行 t1 から t2 にかけて特に大きな成績の向上が見られることから，短時間の訓練によって大きな学習効果が得られたと考えられる．今回の実験で行った訓練では，試行と異なる話者モデルのみを用いたが，音素継続長に関しては共通の統計量を用いたため,「ある話速において特定の数字の特定の音素がどう聞こえにくくなるか」といった現象が学習されたと考えられる．

また，試行 t2 から t3 にかけての成績の低下が大きくないことから，学習効果は 3 日という時間が経過しても持続する，と考えられる．

8. ま と め

テキスト音声合成エンジン GalateaTalk と早口音声コーパスを用いて，聞き取りやすい早口の合成音声を実現した．また，4 桁の乱数の読み上げ音声を聴取する実験を行い，その有効性を検討した．早口コーパスから学習した統計量のうち，音素継続長のモデルのみを用いることで早口でかつ十分な音質の合成音声を得られること，早口音声の聴取タスクにおいては短時間の訓練によって大きな学習効果が得られ，その効果は時間が経過しても持続すること，などの知見が得られた．

今後の課題を以下に挙げる．

（1） 視覚障害者を被験者とした実験を行う．既存の合成音声エンジンへの慣れの程度や，個人による好みなどを考慮し，幅広いユーザを対象とする．

（2） GalateaTalk を他の手法で早口化した場合（話者固有の音素継続長モデルを用いつつ RATE 値をさらに小さくして早口化した場合など），および，他の音声合成エンジンで同等の早口音声を生成した場合と，本稿で提案した手法の性能比較を行う．特に，早口に適したエンジンとそうでないエンジン，習熟しやすいエンジンとそうでないエンジン，といった観点での比較を行う．

（3） RATE の値を非常に小さくした場合に，継続長が 0 になる音素が生じる．このような音素が音響的に完全に欠落しないための考慮を行う．

（4） 数字聞き取り以外のタスクに関する検討を行う．

（5） 本手法で生成した音声が，関連研究 [3] で行われている肉声の伸縮変換と比較して，どのような音響的特性を持っているのか，また，どのような特性を実現していれば早口での聴取が容易になるのか，などの観点から，早口音声合成エンジンに求められる特性を明確にする．

（6） GalateaTalk はソース・フィルタ型（パラメトリック方式）の音声合成システムの一種であり，ポストフィルタを適切に設定しないと切れが悪い音やこもった音が生成される．これらの問題が早口音声の品質に与える影響を検討する．

（7） 本実験で構築した話者モデルは，自動ラベリングからの学習で自動構築したものであり，音質としては下限に近いものである．話者モデルの構築に用いるコーパスに，コーパスの読み誤りやアクセントの相違などの精緻なラベリングを施すことで，音質を改善する余地がある．また，スペクトルおよび

ピッチ情報に関しては，学習データを増やしてモデルを精緻化することも検討する．

（8） GalateaTalk を本研究のような早口音声合成に用いるにあたって，音素モデルの状態数，フレームシフト，コンテクストクラスタリングにおけるツリーのノード数，音声分析のパラメータ次元数など，デフォルトの値が適切であるとは限らない．これらの動作条件について検討する．

（9） 音素継続長のみを早口モデルで置き換えることの問題点について詳細に検討する．音素継続長に関する統計モデルは，各状態の自己遷移回数をガウス分布でモデル化したもの（状態数を次元数とするベクトル値）であるが，スペクトルの変化の大きさなどを考慮した継続長の学習を行うことで，話速変化を考慮した継続長モデルが構築できる可能性がある．

（10） 本研究の成果を生かし，ウェブおよび情報家電を効率的に利用できる視覚障害者向け音声対話システムの試作および評価を行う [8].

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（視覚障害者の聴覚認知の解明と音声対話への利用）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．安村通晃・慶應義塾大学教授，新城 直・View-Net 神奈川代表，各プロジェクトのメンバーの方々に感謝する．

文　　献

[1] 鳥原信一, “「ななめ聞き」システム［統語情報を用いた，視覚障害者のための「速聞き」システム］,” 電子情報通信学会技術報告, 第 5 回福祉情報工学研究会 WIT00-28, Nov 2000.

[2] 渡辺哲也, 視覚障害者のコンピュータ利用を考慮した速い話速音声に関する音響学的研究, 科学研究費補助金研究成果報告書, 国立特殊教育総合研究所, 特殊研 F-117, 2004.

[3] 浅川智恵子, 高木啓伸, 井野秀一, 伊福部達：“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.

[4] 嵯峨山茂樹, 伊藤克亘, 宇津呂武仁, 甲斐充彦, 小林隆夫, 下平博, 伝康晴, 徳田恵一, 中村哲, 西本卓也, 新田恒雄, 広瀬啓吉, 峯松信明, 森島繁生, 山下洋一, 山田篤, 李晃伸, “擬人化音声対話エージェント基本ソフトウェアの開発プロジェクト報告,” 情報処理学会研究報告, 2003-SLP-49, pp.319-324, Dec. 2003.

[5] http://hil.t.u-tokyo.ac.jp/∼galatea/

[6] http://www.astem.or.jp/istc/

[7] http://hts.ics.nitech.ac.jp/

[8] 渡辺隆行, 安村通晃, 小田浩一, 西本卓也: “視覚障害者の聴覚認知の解明と音声対話への利用に向けて,” 電子情報通信学会技術報告, WIT2004-74, pp.7-12, Mar 2005.