早口音声の聴取における学習効果と加齢の影響

西本　卓也[*1]　渡辺　隆行[*2]

The Learning Effects and Age-Related Effects on Listening to Ultra-Fast Speech

Takuya Nishimoto[*1]　Takayuki Watanabe[*2]

Abstract – **We investigated the intelligibility of synthesized voices at fast speaking rates. Four-digit random numbers are used as the vocabulary of the recall test. As the result, elderly persons can recall fast speech at some levels. However, their average recall rates are lower than the young university students and the individual differences are significant. The results of this tasks we consider are affected by the difficulty of auditory perception itself and the difficulty of recall the numbers in the correct order. Leaving out the effect of latter factor, it turned out that the task performances of elderly persons and young students are almost same. The learning effects of the elderly persons are not significant in either case, though those of the young students are significant for several weeks.**

Keywords : **Listening, Ultra-Fast Speech, Learning Effects, Age-Related Effects**

1.　はじめに

パーソナルコンピュータ (PC) およびインターネットを音声で使用することにより，視覚障害を持つ人の社会的活動への参加が容易になる．このようなシステムを使いやすくするために，早口であっても正確に聞き取れるテキスト音声合成システム (Text-to-Speech System, TTS) が求められている．

渡辺 [1] は日本国内で視覚障害を持つ PC 利用者がスクリーンリーダの音声をどのように設定しているか実状を調査した．多くの利用者はソフトウェアにおいて設定可能な最高の読み上げ速度を選択しており，これは一般的な読み上げ速度の約 2 倍であった．

視覚障害を持つ人々がどの程度の早口音声を聞き取ることができるか，といった検討を浅川らが行っている [2], [3]．浅川らは音声波形編集ソフトウェア CoolEdit の時間伸縮機能を用いて，日本語の文章を読み上げた音声を早口化した．視覚に障害を持ちスクリーンリーダの熟練者である複数の被験者がこれを聴取した．文章に含まれる単語の約 90%を聞き取ることができる話速を「最適速度」，約 50%の単語を聞き取ることができる話速を「最高速度」と定義し，合成音声の利用に熟練した被験者によって評価した結果，最適速度は約 18 モーラ／秒，最高速度は約 23 モーラ／秒であった．これに対して一般的な TTS エンジンの最高速度は 900 モーラ／分（15 モーラ／秒）以下であった．

我々は，独自に早口の合成音声を作成し，晴眼の若年者および 65 歳以上の高齢者がどのような話速の合成音声をどの程度聞き取れるのか，聞き取り能力は学習によってどのように変化するのか，といった検討を行ってきた [4]~[9]．4 桁の数字を対象とした聴取実験を通じて，早口音声は最初は聞き取りにくいが，しばらく聞いていればある程度は聞き取れるようになる，という学習効果が確認できた．

また，実験結果に対して記憶の影響を相殺した分析を行った結果，若年者と高齢者は実験開始時においては 4 桁の数字を同じ程度聞き取れているが，特に高齢者は数字の順序を含めて正しく記憶することが困難であること，訓練による了解度の向上は若年者には顕著だが高齢者においては顕著ではないこと，などが明らかになった．

2.　研究の目的

2.1　音声聴取における学習の効果

聞き取りやすく疲れにくい早口合成音声を実現するためには，まず既存技術による早口合成音声がどのくらい聴取可能であるかを，話速ごとに詳細に調査する必要がある．スクリーンリーダ利用者は非常に高速の音声を聞き取ることができるが，これが学習によって可能であるならば，その学習の過程を明らかにする必要がある．また，若年者と高齢者でどのような差があるかを知ることも重要である．

本研究における実験のモデルを図 1 に示す．本研究では，人間による音声の聴取は「知覚レベル」「意味レベル」の 2 段階で行われている，と考える．これらは，聴覚からの入力と記憶している音響的イベントや語彙を照合して単語を同定する処理であると考える．

*1: 東京大学
*2: 東京女子大学
*1: The University of Tokyo
*2: Tokyo Woman's Christian University

また，音声を提示して書き取らせる実験を行う場合には，聞き取った語句は入力されるまで短期記憶に格納される．

本研究では，語彙に関する知識の個人差をなくすために，誰もが知ってる小規模の語彙として「ゼロ，イチ，ニー，サン，……」といった数字を 4 桁ずつ提示する．また，提示する音声の話速は標準の 2〜5 倍の合成音声とし，なんとか聞き取れる早口音声から，ほとんど聞き取れない「超早口音声」までを含めることとする．話速が速くなるほど単語理解度が偶然のレベル（数字の場合は 0.1）に近づくことが予想される．

2.2 加齢の影響に関する仮説

高齢者の知覚や認知に関する特徴は文献[10]によると以下が挙げられる．

- 視覚：焦点の合わせにくさ，光の強弱における順応の遅さ，視野の狭まり，色覚の弱さ
- 聴覚：特に高音が聞こえにくくなる，周囲の雑音と聴き取るべき音の区別が難しくなる，音源の方向を察知しにくい，聞こえてはいるが意味がわからない
- 反応速度：主に脳内での処理作業能力が低下
- 記憶：記憶すべき項目数が短期記憶容量を超えたり，情報操作が必要であったり，注意を分割する必要があると，学習が困難
- 対応力：新しい場面に適応する際に働く知能は大きく低下

本研究では聴覚に障害のない若年者と高齢者を被験者とする．このような実験の結果を比較することで，以下の仮説について検証を行う．

- **仮説 1:** 早口音声聴取の訓練が行われていない状態では，健聴の若年者と高齢者の聴取能力は「知覚」「意味」の両レベルで同等である．ただし短期記憶は加齢によって衰えやすいため，高齢者は順序を含めて正しく回答することが困難になる．
- **仮説 2:** 早口音声が音として聞き取りにくい場合には，既知の単語を思い出してその聴取した音声に当てはめて聞こうとするような適応学習が行なわれる．この学習の効果は聴覚だけでなく認知や記憶にも依存しており，長期記憶と同じように持続する．また，この学習効果は加齢によって衰えやすい．

なお，語彙の統制方法としては単語親密度を用いる方法が考えられる．これに対して数字を用いた今回の実験は「非常に親密度の高い語彙」に相当する．これに対して親密度の低い語彙を用いた聴取実験は，意味レベルのトップダウン情報を用いることができないため，知覚レベルの聴取能力の影響を受けやすい評価になる．単語親密度を用いた実験については別の発表[13]で報告する予定である．

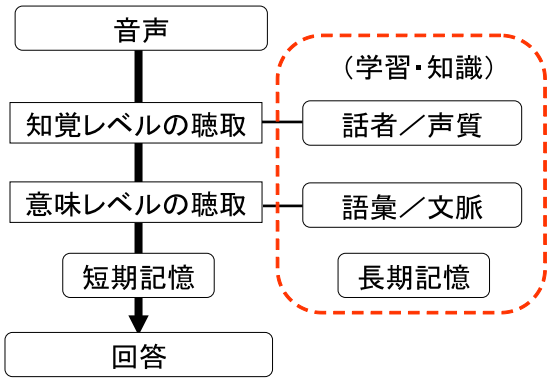

図 1　音声聴取実験のモデル化

3. 実験方法

3.1 評価する合成音声

我々はオープンソース・ソフトウェアとして開発されている HTS (HMM-Based Text-to-Speech Synthesis System) を使用し，HMM 方式によって早口合成音声を作成した．これはオープンソースの日本語テキスト音声合成エンジン GalateaTalk のベースになっているツールである．フレームシフトは標準の 5ms から 2ms に変更した[11],[12]．男性アナウンサー 1 名によって，日本語の音素バランス文の音声コーパスを収録する．合計 503 文のうち，203 文を標準的な話速で，150 文を早口で，150 文をゆっくり発話するように教示する．収録した音声の話速平均値は，早口，標準，ゆっくりのそれぞれについて 12.00，7.29, 5.38（モーラ／秒）であった．このコーパスから以下の統計的話者モデルを構築した．

A モデル： スペクトルとピッチの情報は全ての音声（503 文）で学習する．継続時間長は早口音声（150 文）で学習する．

B モデル： スペクトルとピッチの情報は全ての音声（503 文）で学習する．継続時間長は標準音声（203 文のうち 150 文）で学習する．

早口音声から得られた継続時間長の統計モデルが超早口音声の合成に有用ではないかという期待から，上記 2 つのモデルを作成し，了解度を比較した．過去の実験[6],[7]からは早口音声コーパスから学習したスペクトルやピッチの統計量は有効性が明らかでなかったため，本実験では時間継続長のみに着目した．しかし現段階では後述の実験結果において有意差は見られないため，次章以降では A/B 各モデルの有効性についての議論は省略する．

3.2 語彙と話速

聴取させる音声は 4 桁の数字を含む文「番号は NNNN です」である．聞き取りやすさを考慮して，数

表 1　話速の実現値（グループごとの平均）

	話速（モーラ／秒）	
グループ	モデル A	モデル B
R1	17.60	17.57
R2	20.12	20.17
R3	23.07	23.32
R4	27.27	27.01
R5	31.10	29.67

字の読みは「ゼロ，イチ，ニー，サン，ヨン，ゴー，ロク，ナナ，ハチ，キュー」のように，すべての数字が 2 モーラとなるようにする．

2 種類の話者モデル（A/B）について，全ての数字（0〜9）が全ての桁位置（1〜4 桁目）に全ての目標話速（17〜30 モーラ／秒，5 グループ）で同じ回数だけ出現するように考慮した 100 種類の音声を 2 セット作成する（「課題セット 1」および「課題セット 2」）．目標話速は最低で約 17 モーラ／秒，最高で約 30 モーラ／秒とする．話速の実現値の平均を表 1 に示す．

3.3　順序の誤りを考慮した集計

前述の高齢者の特性を考慮すると，聴取した 4 桁の数字を回答するまで正しく記憶しておくことの負荷は無視できない．特に回答に PC を用いる場合には，キーボード入力のメンタルワークロードによって短期記憶が妨げられる可能性もある．

本報告では strict score および loose score の 2 種類の集計を行い，特に後者において順序に関する記憶の影響の相殺を試みる．

- **strict score:**　各桁ごとに数字の一致を数える．
  例：正解 1234 ／ 入力 1324 ：単語了解度 50%
  これは知覚と記憶の両方が影響する評価となる．
- **loose score:**　順序の入れ替わりを許して一致を数える．
  例：正解 1234 ／ 入力 1324 ：単語了解度 100%
  これは strict score に比べて記憶誤りの影響を減らした評価となる．

なお，本報告の 4 章および 5 章では strict score の結果について述べ，6 章において strict score と loose score の比較を行う．

4.　若年者を対象とした実験

4.1　実験の手順

被験者は課題の音声を聴取したことがない大学生（女性）で日本語を母国語とする晴眼者かつ健聴者である．21 人の被験者のうち途中で欠席した 1 人を除いて 20 人を分析対象とした．

1 人 1 台のノート PC(Microsoft Windows XP) とヘッドフォン (Panasonic RP-H750-S) を使用した．実験に先立ってサンプル音声を提示し，被験者に音量を調節してもらった．被験者自身にウェブブラウザを操作させて，実験用に作成したウェブページに含まれる「番号は XXXX です」という音声を 1 つずつ聴取させ，聞こえた数字をフォームに入力させた．被験者の操作した画面の例を図 2 に示す．

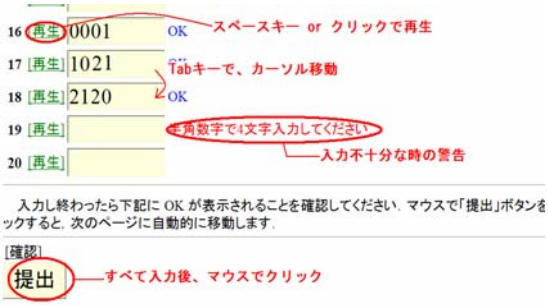

図 2　音声聴取実験の画面

約 90 分の聴取実験を 1 週間に 1 回行った．提示する音声は 3 週間にわたって同一である．課題セット 1 および 2 をそれぞれランダムに並べ替えた後，前半 50 課題と後半 50 課題に分割し 8 つのページを作成した．各ページには 4 桁の数字の聴取が 50 課題ずつ含まれる．のべ 1600 個（4 桁× 50 課題× 8 ページ）の数字聴取を 3 週間繰り返した．

1 ページ聴取させるごとにグループの全員が終了するのを待ち，次のページに進むように指示を出した．また，各ページの所要時間が 7〜8 分になるように適宜休憩を取った．これは疲労の影響を回避するためである．

4.2　結果

学習効果による単語了解度の変化を図 3 に示す．各週 (w1-w3) ごとの試行（p1-p4，課題セット 2 種類を 2 回ずつ反復する各課題群）についての被験者 20 人の単語了解度の平均と標準偏差である．この結果から，聴取課題を繰り返すことのみで学習効果が得られている．第 1 週 (w1) における試行 p1-p4 の効果は統計的に有意であり（$F = 4.515, p = 0.0057$），有意水準 5%で Post Hoc テスト（Fisher PLSD）を行った結果，p1 と他 3 群の間で平均値の有意差があり，実験開始直後の学習効果が顕著である．

被験者 20 人の第 1 週から第 3 週にかけての単語了解度の変化を図 4 に示す．試行の効果は統計的に有意であり（$F = 10.791, p < 0.0001$），有意水準 5%で Post Hoc テスト（Fisher PLSD）を行った結果，w1-w2 および w1-w3 の組み合わせで平均値の有意差があった．このことから，単語了解度の向上は第 1 週から第 3 週まで続き，まったく学習を行わないで 1 週間が経過しても学習効果は持続すること，第 3 週の実験からは顕著な学習効果がみられなくなること，などが示唆された．

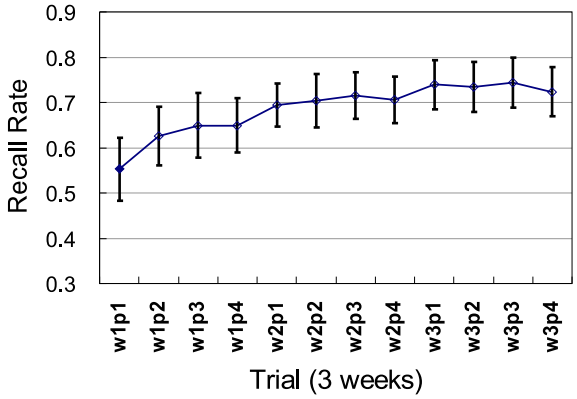

図 3　若年者の実験における単語了解度の変化（バーは標準偏差）

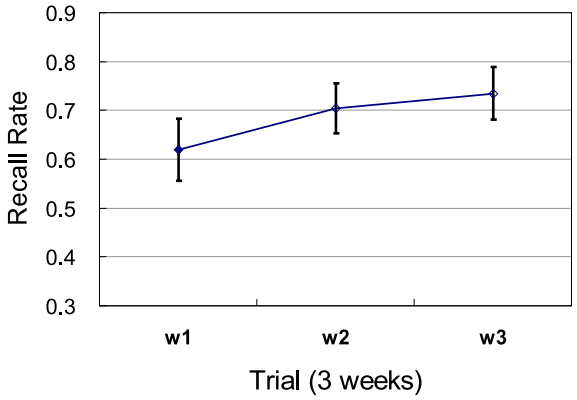

図 4　若年者の第 1 週から第 3 週にかけての単語了解度の変化（バーは標準偏差）

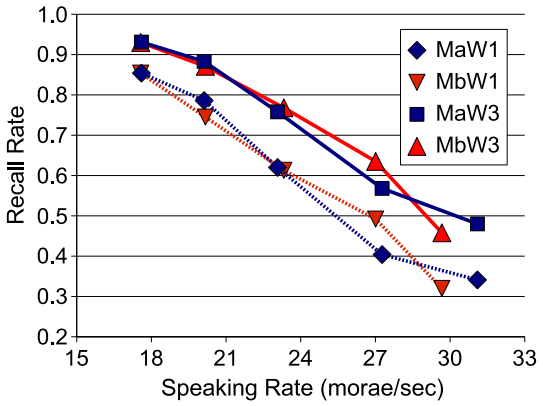

図 5　若年者の第 1 週と第 3 週における話者モデルごとの話速と単語了解度.

第 1 週 (W1) および第 3 週 (W3) における話速と単語了解度を図 5 に示す．話者モデルによる差はみられないが，いずれの話者モデルにおいても，第 3 週では話速 30 モーラ／秒で単語了解度が約 50%になっており，学習によって高速での聞き取りが可能になることが示された.

5.　高齢者を対象とした実験

5.1　実験の手順

若年者の実験と同様に，1 人 1 台 PC とヘッドフォンを使用し，ウェブページ上で課題音声を聴取させ，聞こえた数字をフォームに入力させた．高齢者は若年者よりも疲労の影響が大きいと考え，課題音声は 1 回につき 200 問に減らし，1 ページの課題数を 20 問とし，ページごとに休憩を取った．また若年者の実験よりも実験期間を増やして 4 週間とした.

5.2　被験者

被験者の年齢は 65 歳から 78 歳で，男性 8 人，女性 5 人である．簡単なキーボード操作ができることを条件として募集した.

被験者全員についてオージオメーター（リオン AA-30W）を用いて，第 1 週から第 4 週のいずれかの実験後に聴力の測定を行った．左右それぞれ 500Hz, 1000Hz, 2000Hz, 4000Hz の測定を行った．その結果，特に聴力に障害のある被験者は含まれていなかった（4000Hz 付近の音声が聞き取りにくい被験者が存在したが，後述する実験の結果からは，その影響を確認できなかった）.

5.3　結果

高齢者 13 人の聴取実験における 4 週間の単語了解度の推移を図 6 に示す．試行の効果は有意傾向であった（$F = 2.221, p = 0.0990$）．有意水準 5%で Post Hoc テスト（Fisher PLSD）を行った結果，w1-w3 および w1-w4 の組み合わせで平均値の有意差があった．話速と単語了解度の関係を図 7 に示す.

若年者と高齢者の単語了解度の比較を行った結果を図 8 に示す．両者は 1 週間ごとの課題数が異なるため，200 課題の聴取ごとに単語了解度を比較する．つまり若年者実験における第一週の前半および後半の各 200 課題が，高齢者実験の第 1 週の 200 課題，第 2 週の 200 課題に対応すると見なす．また，最も遅い話速 (R1) についての結果を図 9 に，最も早い話速 (R5) についての結果を図 10 に示す.

高齢者の単語了解度は若年者と比較すると低い．高齢者の実験においてもある程度の学習効果が確認されたが，若年者と比較して有意な差が生じるまでに多くの試行を要している．この理由としては高齢者の「聴覚の機能低下」「対応力の低下」などが考えられる．実験条件における最低話速（図 9）において，若年者は短期間に 90%程度の単語了解度を達成したが，高齢者は 4 週間後も約 80%であった.

6.　記憶の影響に関する考察

本章では順序の誤りを許す loose score を用いて，前述した実験における記憶の影響について考察する.

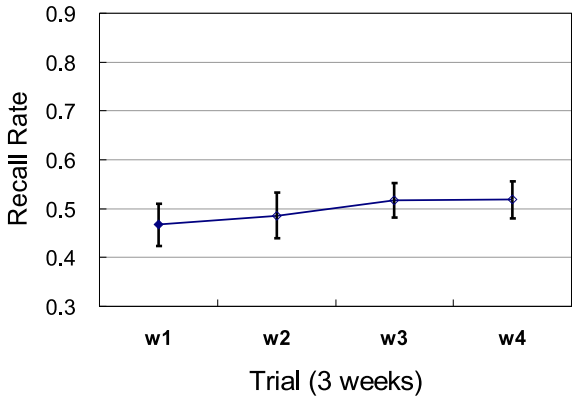

図 6　高齢者における合成音声への慣れ（4 週間，バーは標準偏差）

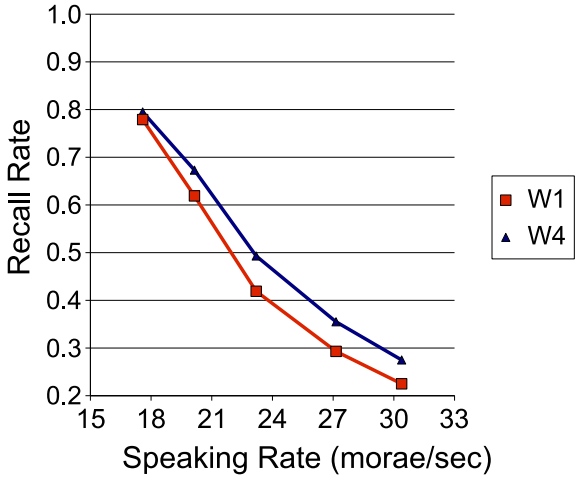

図 7　高齢者における話速と単語了解度の関係（第 1 週および第 4 週）

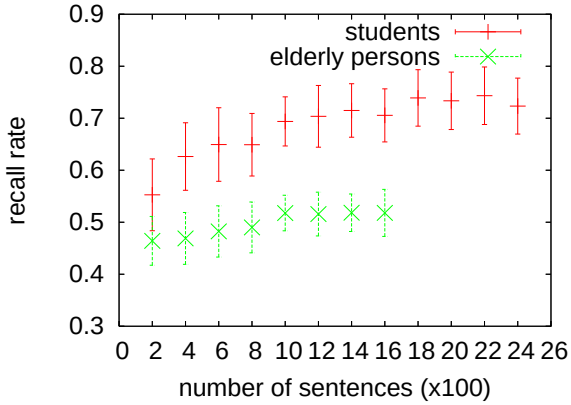

図 8　若年者と高齢者の平均単語了解度の比較（200 問単位，エラーバーは標準偏差）

loose score による 200 課題ごとの単語理解度を，若年者と高齢者について比較したものを図 11 に示す．w1, w2 などは 200 課題ごとの各試行を表す．分散分析の結果，若年者については w1-w6 の群間の効果が有意であり，Post Hoc テスト (Fisher PLSD) の結果，

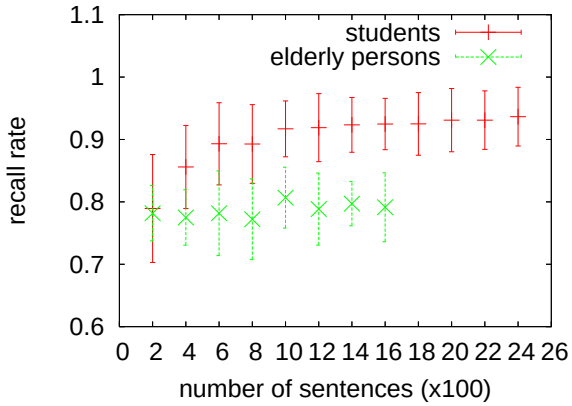

図 9　若年者と高齢者の平均単語了解度の比較 (話速 R1, 200 問単位，エラーバーは標準偏差)

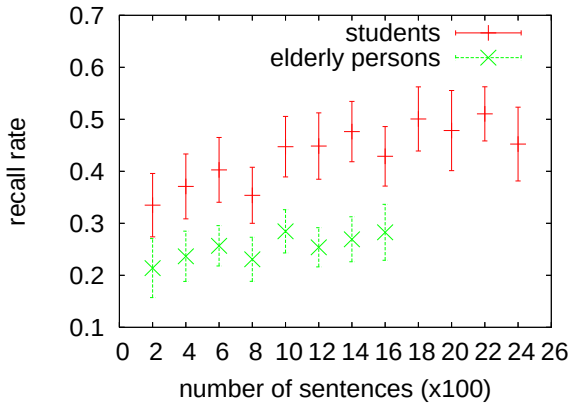

図 10　若年者と高齢者の平均単語了解度の比較 (話速 R5, 200 問単位，エラーバーは標準偏差)

w1 および w2 についてのみ他の各群と 5%水準で有意差が見られた．高齢者については群間の有意差は見られなかった．

また，話速 R5 に限って集計した結果を図 12 に示す．分散分析の結果，若年者については w1-w6 の群間の効果が有意であり，Post Hoc テスト (Fisher PLSD) の結果，w1 および w2 についてのみ他の各群と（w1-w2 間を除いて）5%水準で有意差が見られた．高齢者については群間の有意差は見られなかった．

これらの結果から，順序に関する記憶誤りの影響を除くと，学習がまったく行われていない状態 (w1) では，若年者と高齢者の単語了解度に有意差がないことが分かる．また，学習の効果は若年者にのみ有意であり，その効果は第 1 週の 400 課題の試行で収束すると考えられる．

また strict score と loose score の差が順序に関する記憶負荷に起因すると考えると，若年者ではこの差が 0.10〜0.11 であるのに対して高齢者では 0.21〜0.23 となっている．音声聴取における加齢の影響を検討する際には，記憶負荷の考慮が重要であると考えられる．

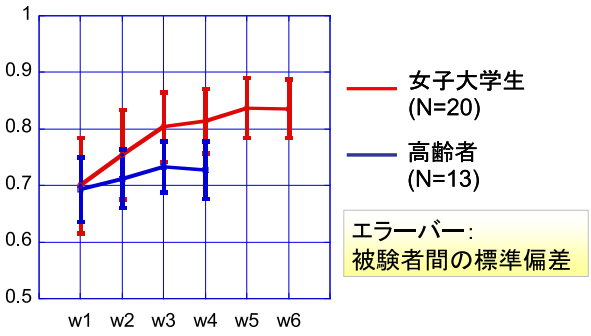

図 11　loose score による 200 課題ごとの単語理解度（全話速．エラーバーは被験者間の標準偏差）

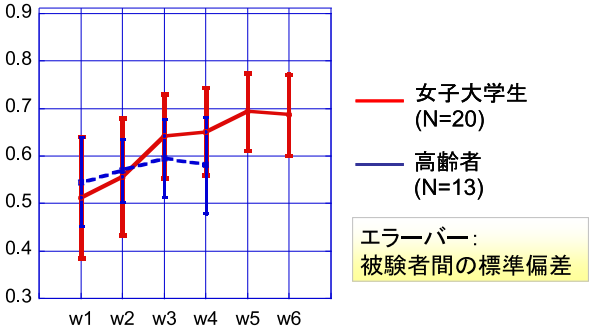

図 12　loose score による 200 課題ごとの単語理解度（話速 R5 のみ．エラーバーは被験者間の標準偏差）

7.　まとめ

晴眼者の若年者および高齢者の早口合成音声への学習効果を検討した．特に，学習効果は聴取量に依存すること，加齢による単語了解度の低下が，記憶力や学習効果の低下に起因することなどが示唆された．

今後は，メンタルワークロードの観点から早口音声を評価するために，NASA-TLX による主観評価実験を行う．また，数字以外の語彙に関する検討として，親密度別単語了解度試験用音声 FW03（NTT・東北大学）を使用した実験を行う予定である．

これらの結果を踏まえて，早口合成音声のためのスペクトル，ピッチ，継続長制御などの効果的な制御方法について検討を行い[12]，視覚障害を持つ人を対象に有効性を評価したい．

謝辞

本研究は科学研究費補助金特定研究「情報福祉の基礎」計画研究「視覚障害者の聴覚認知の解明と音声対話への利用」（課題番号 16091210）の支援を受けた．実験を実施した東京女子大学・大島一恵さん，小野友理子さんに感謝する．また実験に御協力いただいた NPO シニア SOHO 普及サロン・三鷹，三鷹市高齢者社会活動マッチング推進事業事務局の皆様に深く感謝する．また本研究において御議論・御協力をいただいた東京女子大学・小田浩一教授，慶應義塾大学・安村通晃教授，東京大学・酒向慎司氏（現在・名古屋工業大学），嵯峨山茂樹教授，小野順貴講師に感謝する．

参考文献

[1] 渡辺 哲也: “スクリーンリーダの速度・ピッチ・性別の設定状況,” 電子情報通信学会論文誌 D-I, Vol. J88-D-I, No.8, pp.1257-1260, Aug 2005.

[2] C. Asakawa, H. Takagi, S. Ino, T. Ifukube, ”Maximum listening speeds for the blind,” Proceedings Conference of International Community for Auditory Display 2003, pp. 276-279, 2003.

[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] 西本 卓也, 酒向 慎司, 嵯峨山 茂樹, 小田 浩一, 渡辺 隆行: “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,” 電子情報通信学会技術報告, WIT2005-5, pp.23-28, May 2005.

[5] 西本 卓也, 酒向 慎司, 嵯峨山 茂樹, 大島 一恵, 小田 浩一, 渡辺 隆行: “早口合成音声に対する聴取者の慣れの効果の検討,” 日本音響学会 2005 年秋季研究発表会講演論文集, 3-6-14, pp.355-356, Sep 2005.

[6] 大島 一恵, 西本 卓也, 渡辺 隆行: “視覚障害者用早口合成音声による慣れの効果,” 電子情報通信学会技術報告, WIT2005-43/SP2005-81, pp.19-24, Oct 2005.

[7] Takuya Nishimoto, Shinji Sako, Shigeki Sagayama, Kazue Ohshima, Koichi Oda, Takayuki Watanabe: “Effect of Learning on Listening to Ultra-Fast Synthesized Speech,” Proceedings of the 28th IEEE Engineering in Medicine and Biology Society Annual International Conference (EMBC2006), pp.5691-5694, New York, Sep 2006.

[8] 小野 友理子, 渡辺 隆行, 西本 卓也: “早口合成音声に対する高齢者の慣れ,” 電子情報通信学会技術報告, WIT2006-70, pp.115-120, Dec 2006.

[9] 西本 卓也, 酒向 慎司, 嵯峨山 茂樹, 小田 浩一, 渡辺 隆行: ”視覚障害者用早口合成音声に対する慣れの効果,” 日本音響学会 2007 年春季研究発表会講演論文集, 2-8-13, pp.357-360 (in CD-ROM), Mar 2007.

[10] Roy J. Shephard (柴田 博, 青柳 幸利, 新開 省二 訳): シェパード老年学, 大修館書店, 2005.

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

[12] 酒向 慎司, 西本 卓也, 嵯峨山 茂樹: “HMM 音声合成手法による早口音声合成の検討,” 日本音響学会 2005 年秋季研究発表会, 3-6-15, pp.357-358, Sep 2005.

[13] 西本 卓也, 狩谷 幸香, 渡辺 隆行: ”早口音声聴取における単語親密度と心的負荷の検討,” 日本音響学会 2007 年秋季研究発表会後援論文集, Sep 2007 （発表予定）.