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

信学技報
IEICE Technical Report
WIT2007-36 (2007-12)

日本人 ALS 患者のための 日英バイリンガル音声合成システムの構築

加島 慎平[*1]　飯田 朱美[*2]　安 啓一[*1]　荒井 隆行[*1]

Development of a Bilingual Speech Synthesis System for a Japanese ALS Patient

Shimpei Kajima[*1], Akemi Iida[*2] Keiichi Yasu[*1] and Takayuki Arai[*1]

Abstract - This study aimed to build a bilingual speech synthesis system for a communication aid for a Japanese amyotrophic lateral sclerosis (ALS) patient. A corpus-based speech synthesis method is ideal for a communication aid for people anticipating the loss of their voice since the synthetic speech made by such method can reflect the speaker's voice quality. However, this system needs a recording of a large amount of speech that is a burden for them. This paper first describes about our work on building an English speech synthesis system with the patient's voice using a corpus-based speech synthesis method with a voice conversion technique that requires smaller amount of speaker's recordings. After that, we report on our ongoing project of developing a HMM-based Japanese speech synthesis system using HTS, a modified version of HTK. The first method synthesizes speech with HTS and the acoustic model made from the recordings of the patient's voice. The second method synthesizes speech with HTS and the acoustic model with the voice quality which was converted to the patient's voice using voice conversion technique which requires fewer burdens to the patients. The result of the perceptual experiment showed that the voice synthesized with latter method was perceived to have a closer voice quality to the patient's natural speech.

Keywords : a bilingual communication aid, speech synthesis, corpus-based, HMM, voice conversion

1. はじめに

規則合成方式を用いたTTS (Text-to-Speech)音声合成システムは、発声器官の障害等により発話能力の低下した患者のコミュニケーション補助システムとして非常に有効である[1]。また、録音した音声をデータベースとするコーパスベース型音声合成システムは、合成音声に使用者の話者性を持たせることが出来る。

本研究の協力者、山口進一氏は 1996 年、58 歳の時に筋萎縮性側索硬化症（ALS）を発症、2006 年 5 月に 67 歳で逝去された。本稿では山口氏の生前からの要望により実名と個人情報を記す。山口氏は、前向きな、頑強な精神の持ち主で、コンピュータについても詳しく、発症後もパソコンの利便性について精力的に講演活動を行っていた。そして、1999 年冬、声を失った後も自分の声で話がしたいと望まれていた。それ以降、我々の協力者として長く研究に参加して下さった。本稿ではこれより山口氏を協力者と記す。

ALSは、思考、知覚、感覚能力は正常でありながら、徐々に筋肉が動かなくなり、やがては全身麻痺となる難病である。発症部位や症状の進行の過程は人それぞれであるが、やがては呼吸筋も動かなくなり、発声が難しくなるケースがほとんどである。医学の進歩により、人工呼吸器を装着した呼吸療養が可能になった現在、意思伝達の術が著しく限定された中で、患者の「生活の質(QOL)」を向上できるかが大きな課題である[2]。患者からも、コミュニケーションがとれないことが一番つらいと言う声が多く寄せられている[3]。本研究の協力者は声を失ったあとも、機械的な合成音声ではなく、自らの声で話したいという強い要望も持っていた[4]。そこで飯田らは、協力者の音声を録音し、協力者の声質を持った日本語コーパスベース型音声合成システム構築を行った[1]。

その後、協力者の要望から、英語音声合成ソフトウェアFestival[5]を用いた英語音声合成システムに着手した。しかし、協力者の症状が悪化するにつれ、多量の音声コーパスの録音が非常に困難となり、システム構築の大きな障害となった[6]。そこで、Festival用データベース録音支援ソフトフェアであるFestvoxに内蔵されている声質変換機能[7]を用いることで、協力者の少ない発話から声質を抽出し、別の英語母語話者のデータベースを協力者の声質に変えることで英語音声合成システムの構築を試みた。また、合成音声の話者性、了解度について評価を行い、良好な結果を得た[8][11]。そして、先行研究[1]と同じコーパスベース型音声合成方式を採用した（株）アニモのFineSpeechを用いて、同社の協力により、同製品用の協力者の日本語音声データベースを構築し、協力者の声で話す日本語音声合成システムを作成、日英バイリンガルシステムの試作を行った[9]。

その後、オープンソースプログラムとして公開されて

*1: 上智大学　理工学部 電気・電子工学科
*2: 東京工科大学　メディア学部
*1: Sophia University
*2: Tokyo University of Technology

いるHMM (hidden Markov model)音声合成ツールキットHTS[10]を用いた日本語音声合成システムの構築を試みた。また、HTSと声質変換を併用し、より患者の負担を軽減した音声合成システムを構築し、評価を行った。本報告の 2 章では、前述の英語音声合成システムについて紹介する。3 章ではHTSを用いた協力者の日本語音声合成システムについて報告し、また、声質変換技術を利用した患者への負担がより少ないシステムの構築についても報告する。4 章では今回構築した協力者の日本語合成音声と協力者の発話の類似度評価するために行った聴取実験について述べる。また、必要な録音量についても検討を行ったので考察で述べる。

2. 英語音声合成システムの構築

2.1 声質変換（Voice Conversion: VC）

声質変換は、元話者（ソース）の声質（主にf0 およびスペクトル包絡）を別の話者（ターゲット）の声質に変換する技術である。本報告で用いた声質変換機能はFestvoxに内蔵されているものである。これは、Todaら[7]によって作成されたもので、一定量（30 文程度）の文章を学習させることで、正規混合分布（GMM）に基づく声質変換を行う。

2.2 英語音声合成システム[8][11]

図 1 にシステムの概略を示す。本研究では、Festival に内蔵されている英語話者のダイフォンデータベース（voice_kal_diphones）をソースとし、このデータベースの声質を協力者（日本人男性）の声質へ変換した。これにより本来 Festival のデータベース構築にかかる録音量（1369 個のダイフォンパターン）よりも少ない録音量で協力者の声質をもったデータベースを構築できる。学習に用いた協力者の音声は英語音声合成用に 2000 年に録音した TIMIT の発話を用いた。発声者の負担を軽くするため、全 460 文の中から、全てのバイフォンが最低一回は出現するような 246 文を用いた。録音時、話者は補助呼吸装置を鼻に装着しており、音声には雑音がたびたび混入していた。すべての音声はサンプリング周波数 16 kHz、16 bit で保存された。また、声質の学習は 24 次メルケプストラム係数で分析、GMM のクラス数 32 で学習を行った。

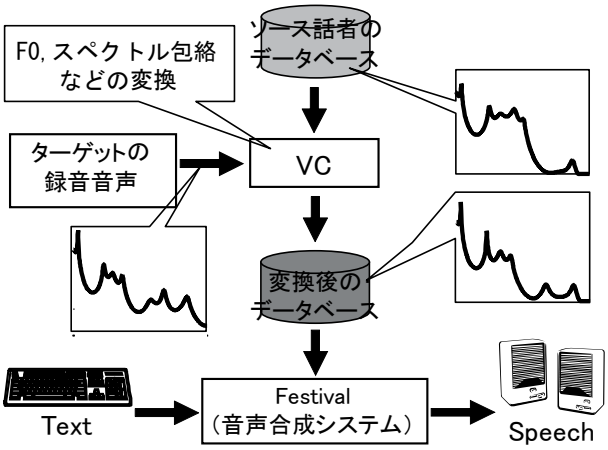

図 1. 英語音声合成システム
Fig.1. English Speech Synthesis System Using Voice Conversion

2.3 客観評価実験[11]

声質変換後の音声の話者性を客観的に評価するために、ターゲット(協力者の TIMIT 読み上げ音声、以降 ys)とソース(kal 合成音声)、および、ys と変換合成音声(ys_conv 合成音声)の間のメルケプストラム歪み（Mel CD）を計算し、比較評価を行った。5 母音/i/, /ɛ/, /ɑ/, /ɔ/, /u/について定常部を ys、kal、ys_conv それぞれについて切り出し（計 22 箇所）、Mel CD を計算した。客観評価では、フレーム長 16 ms, フレームシフト長 4 ms, 24 次メルケプストラム係数で分析を行った。図 2 に客観評価の結果を示す。変換前に比べ、変換後の Mel CD がすべての母音で下回り、歪みの改善が見られることから、声質変換によってソースの声質が協力者の声質に近くなっていることが確認できた。

2.4 主観評価実験[8][11]

声質変換後の合成音声の主観的な評価を聴取実験により行なった。

2.4.1 話者性の評価[11]

20 代の男女 23 名を実験参加者とした。刺激には協力者の日本語の自然発話文(ys_jpn)と声質変換した合成音声（ys_conv）、ソースの合成音声（kal）、その他の合成音声 2 種（ked、rab）で合成した TIMIT 読み上げ音声の 5 刺激で行った。実験参加者には ys_jpn を提示し、続けて同じ文章で合成した ys_conv、kal、ked、rab をランダムに提示した。その後、はじめに提示した ys_jpn と最も声が近いと判断したものをコンピュータに打ち込んでもらった。この過程を 1 セットとし、セットごとに提示文を変えて 40 セット行った。

図 3 に各刺激の選択率と標準偏差を示す。他の刺激と比較して、ys_convの選択率が顕著に高いことがわかる。結果に対し、χ^2検定を行い、有意確率$p < 0.01$ で有意差を得た。以上の結果から、本研究で行った声質変換によって、協力者の話者性を持った音声を合成することがで

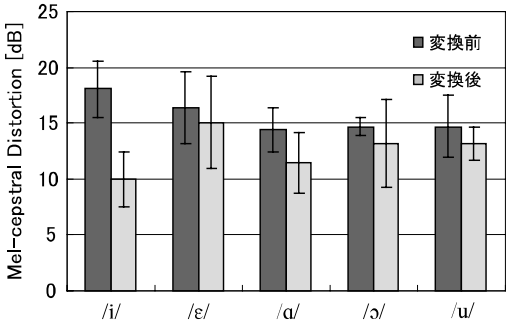

図 2. 客観評価実験の結果[11]

Fig.2. Result of Objective Evaluation

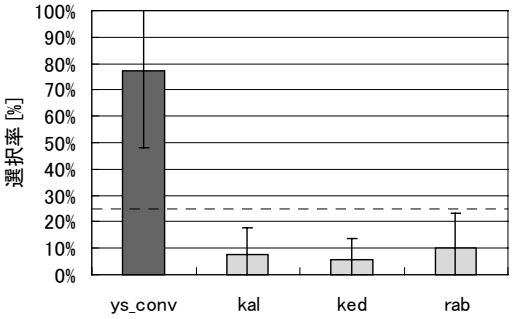

図 3. 話者性評価実験の結果[11]

Fig.3. Result of Speaker Individuality Evaluation

きることがわかった。

2.4.2 了解度の評価[8]

ys_conv の了解度について聴取実験で評価した。以下の 3 種類の刺激文を ys_conv、kal でそれぞれ合成し、7 名の英語話者に提示した。

- 日常語: 日常生活において使用頻度の高いと考えられる 4 単語
- TIMIT 文: 学習に用いた TIMIT 文から長さの適当な 4 文
- 日常会話文: 日常生活において使用頻度の高いと考えられる 4 文

実験参加者には、刺激の内容を聴き取り、聴こえたとおりにコンピュータにタイプ入力してもらった。了解度を（正解の単語数）/（全体の単語数）と定義し、計算した。

結果を図 4 に示す。χ^2検定の結果、日常単語と日常会話文において、ys_convとkalの間に有意差が無かった。本研究で構築した音声合成システムの利用目的は、「日常会話の支援」にある。このことから考えると、日常会話文の評価で変換後の音声について変換前の音声と同等の了解度を得られたことは意義が大きく、利用目的の範囲では実用化が可能であることが示唆された。

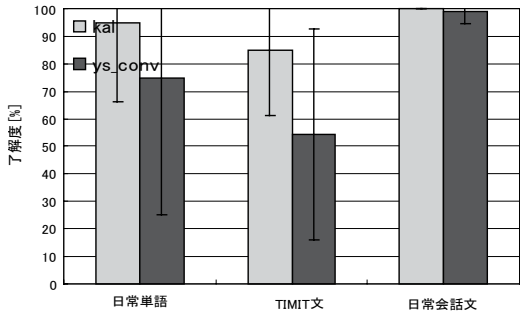

図 4. 了解度評価実験の結果[8]

Fig.4. Result of Speech Intelligibility Evaluation

2.5 バイリンガルシステムの試作[9]

英語音声合成システムと同じくコーパスベース型音声合成方式を採用した（株）アニモの FineSpeech を用いて、日英バイリンガル音声合成システムを試作した。日本語音声合成システムには同社の協力を得て、同製品用の協力者の日本語音声データベースを作成し、本人の声質で音声を合成した。

図 5 に作成したバイリンガルシステムの GUI を示す。ウィンドウ上部にテキスト入力のためのテキストボックス、右下に波形辞書選択のためのプルダウンメニューを配置した。文章入力を終えたら、右下の再生ボタンをクリックすることで音声合成が実行され、音声が出力される。日本語音声合成システムと英語音声合成システムをひとつの GUI で操作できることで、言語の切り替えを簡単に行うことができるようになった。

図 5. バイリンガルシステムの GUI

Fig.5. GUI of the bilingual system

3. 日本語音声合成システムの構築

3.1 HTS を用いた音声合成システム

2.5 章で試作したバイリンガル実装では、日本語合成は合成品質およびサポートの両面で実績の高い（株）アニモの市販ソフトウェアFineSpeechでの実装を同社の協力で行った。しかし、品質やサポートの面での補償がある一方で、録音量やコストの面で、個人が構築するには負担が大きいのも事実である。そこで、次のステップと

してオープンソースでかつ合成音声に話者性を反映することのできる音声合成ソフトウェアであるHMM音声合成を用いて日本語音声合成システムの構築に取り掛かった。HMM音声合成用音響モデルの構築にはHTS[10]を用いた。これは、HMM音声認識用ツールキットHTK (hidden Markov model toolkit)[12]を音声合成システムツールとして拡張するパッチプログラムであり、一定量の文章の読み上げ音声とそのラベルファイルから発話者に固有のHMMモデルを学習し、音声合成ソフトウェアGalatea Talk[13]用の音響モデルを作成する。

HTSによる音響モデルの学習量としては現在のところATR503 文[14]の全文を録音することが妥当であるとされている。しかし、我々が目指す音声合成システムの使用者は発話の困難な障害者であり、そういった使用者にとって、503 文章の読み上げは大きな負担となる。本報告の協力者についても 2000 年にATR503 文の収録を先行研究[1]のために行ったが、本人の負担を考え、すべてのバイフォンが出現するようにATR503 文の中から 129 文を選定し、最終的に 108 文を録音することができた。このように、音声合成システムの構築に 503 文の収録は対象の使用者の負担が大きく、この収録量では断念せざるを得ない対象者も少なくないはずである。また、129 文に減らしたとしても依然として負担は大きい。そのような背景から、最初は協力者の発話した 25 文を用いてHTSによる音響モデル（yam25）の構築を試みた。25 文の根拠は、音声認識・音声対話技術講習会（京都大学学術情報メディアセンター主催、2007）で最少収録量が 20 文だったため、それよりも少し多い量とした。そして、協力者の発話すべて（108 文）を用いて学習した音響モデル（yamfull）も作成した。その結果、声質としては協力者の話者性を感じるものの、yam25 で合成した音声は自然性を大きく欠き、yamfullによる合成音声も自然性において満足の行く結果には至らなかった。原因の一つとして、今回の実験では、ラベルデータの修正を自分たちで行うには至らず、標準のラベルデータを用いたことが考えられる。よって、ラベルデータの修正ができれば、高品質な音声を合成できると考えられ、HTSのより一層の理解も含めて、今後の検討課題である。

3.2 声質変換技術の利用

2 章で述べた英語音声合成システムの評価[8][11]により、声質変換技術を用いることで、患者の話者性を反映したままより少ない録音量で音声合成システムを構築できることがわかった。そこで上記のHTSを用いた日本語音声合成システムにも声質変換の利用を試みた。声質変換利用前のシステムと利用後のシステムの概略図を図 6 に示す。HMMによって構築される音響モデルの話者性、自然性は学習文章量に大きく影響を受ける。そこで、HTSにあらかじめ用意された男性話者の発話によるATR503 文全文とそのラベルデータを用意し、声質変換によって 503 文を協力者の声質に変換した。声質変換に必要な声質の学習には協力者の発話した 108 文から 25 文を用いた。変換後の 503 文でHTSを用いて音響モデルを学習させることで、協力者の発話 25 文から 503 文で学習した音響モデルを得ることができる。この方法により協力者の話し方などのf0 以外の韻律情報はモデル化されないが、高い自然性を持った音声合成が期待できる。また、ラベルデータは標準のものを利用できるという利点もある。

4. 日本語音声合成システムの評価

3 章で構築した録音音声の HTS および声質変換後音声の HTS の 2 種類の方法による協力者の日本語合成音声について、協力者の録音データの比較評価実験を行った。

4.1 刺激

刺激に用いた音声は以下の 5 種類とした。

- yamorg: 協力者の録音音声
- yam25: ATR503 文から協力者の発話した 25 文を用いて HTS で音響モデルを学習した合成音声
- yamfull: ATR503 文から協力者の発話した 108 文を用いて HTS で音響モデルを学習した合成音声
- yamVC: ATR503 文から協力者の発話した 25 文を用いて声質を学習し、声質変換を行った 503 文で HTS による音響モデルを学習した合成音声
- yam25vc108: ATR503 文から協力者の発話した 25 文を用いて声質を学習し、声質変換を行った 108 文で HTS による音響モデルを学習した合成音声

刺激文には協力者が日常的に使用する頻度が多い文章から 6 文選定し、GalateaTalk で合成した。使用した文章を以下に示す。

- エアコンをつけて下さい。
- 体の向きを変えてください。
- よくなりました。
- おはようございます。
- 元気にしています。
- お世話になりました。
- 御苦労さまでした。それでは失礼します。

すべての音声は 16 kHz でサンプリング、16 bit で量子化された。また、HTS による音響モデルの学習は、フレーム長 25 ms、フレームシフト長 5 ms、18 次メルケプストラム係数、HMM 状態数 5 で学習を行った。

4.2 実験方法

20 代から 40 代の男女を実験参加者とした。実験は Web 上で行い、実験環境はヘッドフォン着用を条件とした以外は実験参加者の任意とした。実験参加者には yamorg を聞いてもらい、残りの 4 刺激を聞き、yamorg との音声の類似度を 5 段階で評価（1: まったく似ていない、5: 非常に似ている）してもらった。音声は何度でも聞いてよいものとした。この過程を刺激文ごと（計 6 セット）に行なってもらった。

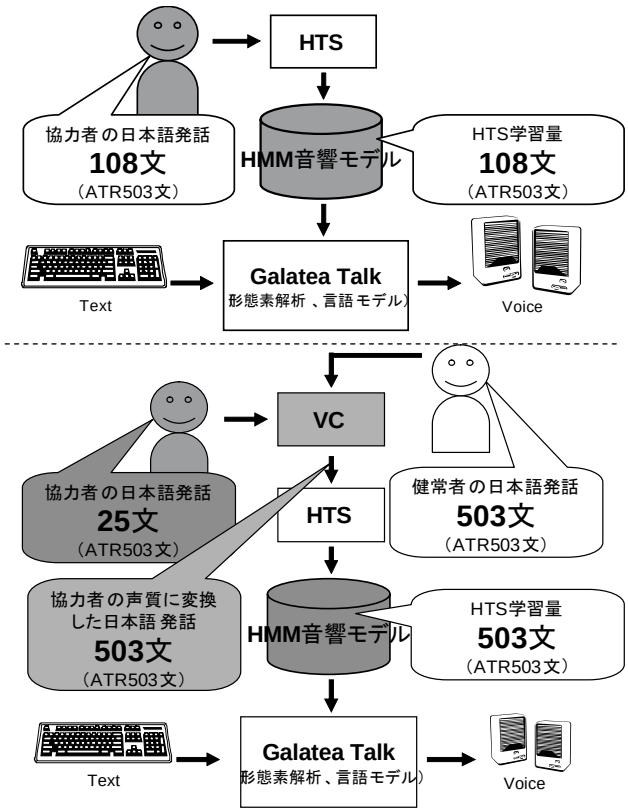

図 6. 日本語音声合成システム
(上：HTS で構築、下：声質変換後、HTS で構築)
Fig.6. Japanese Speech Synthesis Systems
(Upper: Built by HTS,
Lower: Built by HTS after Voice Conversion)

4.3 結果・考察

類似度の平均値と標準偏差を図 7 に示す。χ^2検定の結果、全体として有意に差があるという結果が得られた。声質変換を用いたyamVCとyam25vc108 の 2 刺激がHTSのみの刺激の類似度を上回り、声質変換を利用したほうが協力者本人の発話に近いことが確認できた。

システムを使用する患者の負担となる、学習に必要な録音量に注目すると、yam25 と yamVC は共に必要とする患者の録音量が 25 文である。2 つの刺激の類似度を比較すると、yamVC のほうが高い値を示している。また、協力者の長女の高橋雅子さんにも聴取実験にご協力いただき、その結果も、実験結果の平均と同様に yamVC のスコアが yam25 よりも高かった。このことから、同じ少量の録音量で音声合成システムを構築するのであれば、声質変換を用いる方が、患者の元の声に近い音声を合成できることがわかった。

HTS から音響モデルを構築するのに必要な学習量に注目すると、yamfull と yam25vc108 は HTS での学習量がともに 108 文である。二つの類似度に大きな差は無く、また、HTS の学習量が等しいため、合成音声の自然性の点では大きな差が無いと考えられる。よって、声質変換での学習量が 25 文でも、録音音声で構築した音響モデルと同様の声質を持った音声が合成できることがわかった。しかし、今回は協力者の 108 文の発話にラベルデータは存在しないため、yam25、yamfull では適当な音響モデルが学習できているとは一概には言えない。しかし、ラベルデータの付与は手作業によるところが大きいため、音声合成システムの構築の負担にもなる。そういった意味では、声質変換を利用した方法も本研究の対象者のような場合には十分選択肢となりえると考えられる。

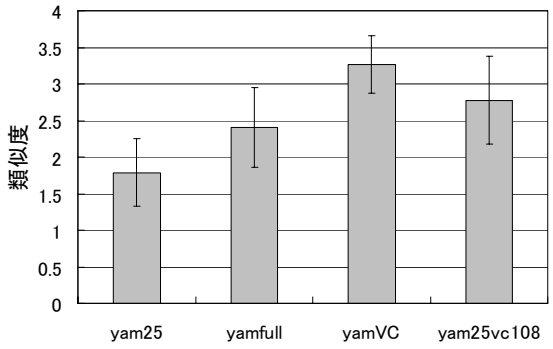

図 7. 各刺激と yamorg との類似度
Fig.7. Similarity Between yamorg and each Stimulus

5. まとめ

日本人 ALS 患者の録音音声を元に、コーパスベース型音声合成手法を用いて患者の声質を持った英語音声合成システムの構築を行った。システムの構築に必要な多量の音声録音が、発話の困難な患者の大きな負担となることから、声質変換技術を用いて、患者の少量の発話から患者の声質を持った英語音声データベースを構築した。英語合成音声の話者性について評価実験を行った結果、合成音声が患者の話者性を持つことを確認した。また、了解度についても評価実験を行い、日常的な使用には問題の無い程度の了解度を得た。英語音声合成システムと、同じくコーパス音声合成手法を用いて、患者の声で出力する日本語音声合成システムを構築し日英バイリンガル音声合成システムを試作した。

次に、患者の HMM 音声合成システムの構築を試みた。患者の発話した 108 文章から HMM 音声合成ツールキット HTS を用いて音響モデルを構築した。また、英語音声合成システムで用いた声質変換モジュールを用い、患者の声に声質変換した健常者の十分な量の発話とラベルデータから音響モデルを学習した。患者の発話から学習した音響モデルと声質変換後の発話から学習した音響モデルの 2 種類のモデルについて聴取実験による評価を行った。その結果、患者の発話量が少ない場合、声質変換を用いたほうが患者の自声に近い音声を合成できることが確認できた。ただし、今回は患者の発話にラベルデータの付与を行っていないため、今後はラベルデータの付与

を行い、再度声質変換を用いない方法でも構築を試みる予定である

謝辞

本研究は科学研究費補助金（A-2, 16203041）の助成を受けて行った。録音に協力してくださった故・山口進一氏とそのご家族の方々、および、山口氏の TIMIT 文読み上げの収録にご協力頂いた ATR のニック・キャンベル氏、実験の考察に関して協力してくださった慶應義塾大学の樋口文人先生、Festival 利用面でご協力して下さった Carnegie Mellon 大学の John Kominek 氏、HTS についてご教示頂いた、名古屋工業大学の酒向慎司先生、全炳河先生に感謝申し上げます。この研究の一部は、文部科学省私立大学学術研究化推進事業上智大学オープン・リサーチ・センター「人間情報科学研究プロジェクト」の支援を受けて行った。

参考文献

[1] A. Iida and N. Campbell: Speech database Design for a Concatenative Text-to-Speech Synthesis System for Individuals with Communicative Disorders; International Journal of Speech Technology 6, pp. 379-392 (2003).

[2] 日本 ALS 協会ホームページ：ALS の症状と生活の変化； Retrieved from http://www.alsjapan.org/contents/whatis/02.html

[3] 豊浦保子：生命のコミュニケーション-筋萎縮性側索硬化症（ALS）患者の記録； 東方出版 (1996).

[4] 山口進一：パソコンを使いこなそう； 日本 ALS 協会福岡支部第四回総会記念講演； Retrieved from http://www.ne.jp/asahi/laconicmako/ikiru/toukou/pasokon_tukaikonasou.pdf

[5] Univ. of Edinburgh: Festival Homepage; Retrieved from http://www.cstr.ed.ac.uk/projects/festival

[6] A. Iida, J. Ito, et al.: Building an English Speech Synthesis System from a Japanese ALS Patient's Voice; IEICE Technical Report SP2005-170, pp. 43-48 (2006).

[7] T. Toda: High-Quality and Flexible Speech Synthesis with Segment Selection and Voice Conversion; Ph.D. Thesis, Nara Institute of Science and Technology (2003).

[8] 加島，飯田，他：声質変換機能を用いた日本語話者のための英語合成音声の了解度評価；日本音響学会春季研究発表会講演論文集 (2007).

[9] 飯田，加島，他：ALS 患者のためのバイリンガル音声合成システムの構築と評価；日本音響学会春季研究発表会講演論文集 (2007).

[10] K. Tokuda, et al.: The HMM-based speech synthesis system (HTS); Retrieved from http://hts.ics.nitech.ac.jp/.

[11] 加島，飯田，他：声質変換技術を用いた日本語話者のための英語音声合成システムの構築；日本音響学会秋季研究発表会講演論文集 (2006).

[12] Univ. of Cambridge: HTK Homepage; Retrieved from http://htk.eng.cam.ac.uk/

[13] 嵯峨山，川本，他：擬人化音声対話エージェントツールキット Galatea； 情報処理学会研究報告, 2002-SLP-45-10, pp.57-64, Feb. (2003)

[14] 阿部，匂坂，他：研究用日本語音声データベース利用解説書（連続音声データ編）; ATR 自動翻訳電話研究所 (1990).