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*

◎加島慎平（上智大・理工），飯田朱美（東京工科大・メディア），安啓一，
荒井隆行（上智大・理工），菅原勉（上智大・外）

1　はじめに

規則合成方式を用いたTTS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]。特にFestival[2]のようなコーパスベース型音声合成システムは、録音した音声をデータベースとすることで合成音声に話者性を持たせることが出来る。しかし、システム使用者の病状によっては多量の音声コーパスの録音が非常に困難であるケースも多く存在し、合成音声が話者性を保つ最低限の量でのコーパス設計が必要とされる[1]。

前回の報告[3]では、Festvoxに内蔵された声質変換機能を用いて、英語話者のダイフォンデータベースをある日本人筋萎縮性側索硬化症（ALS）患者の声質に変換することで、その話者の英語音声合成システムの構築を行なった。合成音声の話者性を聴取実験とメルケプストラム歪による客観評価実験で評価した結果、変換後の合成音声がALS患者本人の話者性を十分持つことが確認された[3]。本報告では変換後の合成音声の了解度の評価実験について報告する。評価はWebベースの聴取実験システム[4]を作成して行なった。実験システムの構築にはプログラム言語にPHP、リレーショナルデータベースにMySQLを用いた。

2　声質変換

声質変換は、元話者（ソース）の声質（主にF_0、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。前回および本報告で用いた声質変換機能は、Todaら[5]によって作成されたFestvoxに内蔵されているものである。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、日本人男性ALS患者ys氏の声質へ変換した。学習には、ys氏がTIMIT文を発話した音声246文（全460文から発話者の負担を減らすためにすべてのバイフォンが最低一回出現するように選んだもの）を使用した。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、GMMのクラス数32で学習を行った。

3　評価実験

声質変換を用いて作成したys氏のデータベースから音声を合成し（ys_conv）、合成音声の了解度について聴取実験を行い評価した。

3.1 実験方法

評価実験はMySQLとPHPを用いた聴取実験システム[4]を用い、Web上で行なった。実験環境は、ヘッドフォン着用を条件とした以外は被験者の任意とした。20代から60代の男女48名（うち、英語話者7名）を被験者とした。Table 1に使用した刺激を示す。刺激は以下の基準で選定した。

- 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

すべての刺激文は声質変換後の音声ys_convとFestivalに内蔵された英語話者の音声kalで合成した。被験者はA、Bの2グループに分け、Aグループには Table 1の左半分の文をys_convで合成した6刺激、および、右半分の文をkalで合成した6刺激の計12刺激を提示し、Bグループには逆に左半分をkalで、右半分をys_convで合成した計12刺激を提示した。被験者には提示した刺激が何と言っているかをWebブラウザ上のフォームにそのままタイプ入力してもらい、（正解の単語数）/（全体の単語数）を計算して了解度とした。

3.2 結果と考察

結果は以下の条件で被験者を分けて集計し、考察を行なった。

- STANDARD: 平均的な英語習熟度（TOEIC 700点未満）を持つ27人の日本人
- HIGH: 高い英語習熟度（TOEIC 700点以上）を持つ14人の日本人
- NATIVE: 7人の非常に高い英語習熟度をもつ（うち5人は英語母語話者）外国人

*Evaluating intelligibility of English speech synthesized from a Japanese individual's voice using voice conversion technique, by S. Kajima (Sophia University), A. Iida (Tokyo University of Technology), K. Yasu, T. Arai and T. Sugawara (Sophia University)

Table 1. 用いた刺激語と刺激文

日常単語	keyboard	glasses	television	water
TIMIT 文	Is she going with you?		They enjoy it when I audition.	
	Did you eat lunch yesterday?		I took her word for it.	
日常会話文	Hello, my name is Yamaguchi. I am sixty eight years old.		My favorite food is potato salad. It is delicious.	
	Please turn off the air conditioner.		Please open the window and close the door.	

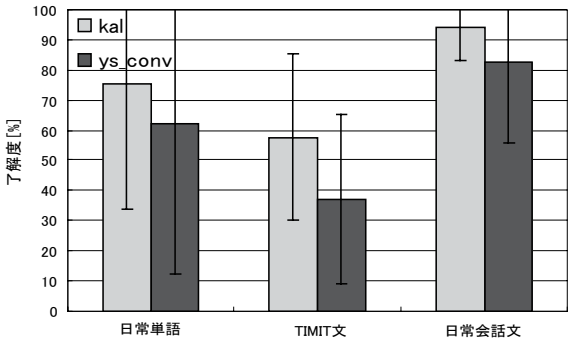

Fig. 1. 了解度（STANDARD）

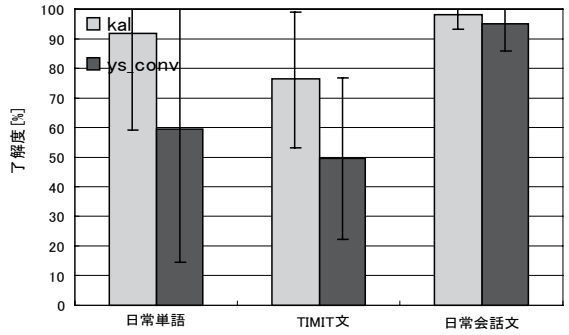

Fig. 2. 了解度（HIGH）

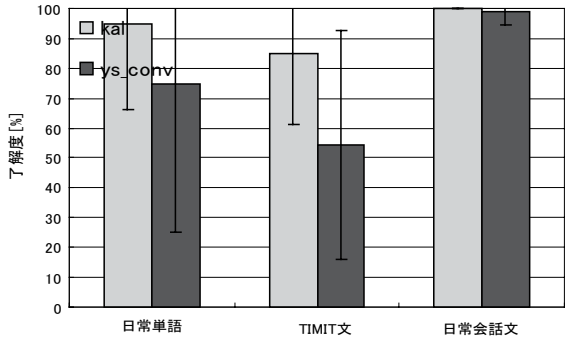

Fig. 3. 了解度（NATIVE）

Fig. 1 にSTANDARD、Fig. 2 にHIGH、Fig. 3 にNATIVEの各グループの了解度と標準偏差を示す。全体的にはすべてのグループ、刺激について変換後の了解度のほうが変換前を下回ったが、有意水準 0.05 でχ^2検定を行った結果、STANDARDグループでは日常単語、HIGHグループでは日常会話文、NATIVEグループでは日常単語と日常会話文において有意差が無かった。

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、今回、STANDARD グループより英語聴取能力の高いと判断できる HIGH, NATIVE グループの日常会話文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る。

4　まとめ

本研究では、Festvox に内蔵された声質変換機能を用いて、日本人 ALS 患者の英語の録音音声から英語音声を合成した。合成音声の了解度について、MySQL と PHP を用いて Web 上ベースの聴取実験システムを作成し、評価実験を行なった。その結果、声質変換後の合成音声では変換前よりも了解度が下がる傾向が見られたものの、コミュニケーション支援として最も重要な日常会話の了解度については、変換前と同等の了解度が変換後も得られることがわかった。

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・山口進一さんとそのご家族の方々、および、被験者の TIMIT 文読み上げの収録にご協力頂いた ATR のニック・キャンベル博士、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に感謝申し上げます。

参考文献

[1] A. Iida *et al.*, International Journal of Speech Technology 6, pp. 379-392, 2003.

[2] Festival Homepage, Retrieved from http://www.cstr.ed.ac.uk/projects/festival/

[3] 加島他, 日本音響学会秋季研究発表会講演論文集, pp. 251-252, 2006.

[4] 飯田他, “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.

[5] T. Toda, Ph.D. Thesis, Nara Institute of Science and Technology, 2003.