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*

◎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]。

先行研究として飯田ら[3]は、ある日本人筋萎縮性側索硬化症（ALS）患者の英語発話音声を用いて英語音声合成システムの構築を試みたが、患者の使用していた人工呼吸器の駆動音が録音音声に混入してしまったことと、非母語話者の英語の発音の問題で、音声合成に適したデータベースを構築するのは困難であった。また、膨大な録音量が患者の大きな負担となった。そこで、本報告では Festvox に内蔵されている声質変換機能を用いて、別の英語母語話者のダイフォンデータベースを患者の声質に変換することによって、英語音声合成システムの構築を試みた。合成音声の話者性は、客観評価と聴取実験による主観評価で考察した。

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。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[4]によって作成されたもので、一定量の文章を学習させることで、正規混合分布に基づく声質変換（F_0、スペクトル包絡の変換）を行う。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）の日本人男性ys氏の声質へ変換を行った。

学習に用いた ys 氏の音声は英語音声合成用に5年前に録音した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

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）、合成音声の話者性について評価を行った。先行研究では、ys 氏の音声と kal の音声を混合したダイフォンデータベースを用いて音声合成を行ったが[3]、録音音声に混入した人工呼吸器音のためと、非母語話者の英語の発音の難しさから、了解度に欠ける合成結果が多かった。そのため､本手法との比較には及ばなかった。

3.1 客観評価実験

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(ys 氏の 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]。ys と kal、ys_conv との間には時間のアラインメントが大きくずれていたため、本研究では ys と kal、ys_conv の時間的な整合を取るために、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､窓長 16 ms, 1/4 シフト，メルケプストラム係数24次で分析を行った。

Fig. 1 に客観評価の結果を示す。すべての母音について歪みの改善が見られた。特に /i/、での歪みの改善が大きかった。母音の話者性においては､F2 の高い母音がより話者性を反

* Development of an English speech synthesis system using voice conversion from a Japanese male voice, by S. Kajima (Sophia University), A. Iida (Tokyo University of Technology), K. Yasu, T. Arai and T. Sugawara (Sophia University)

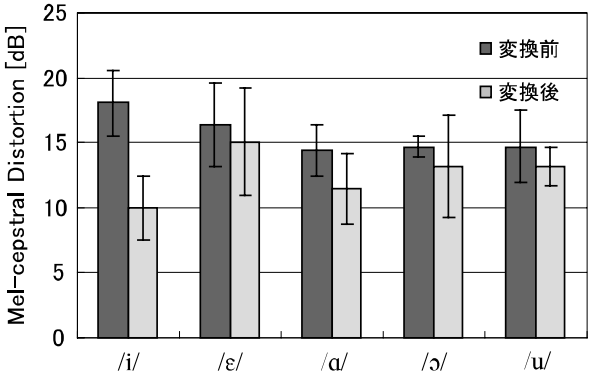

Fig. 1. 客観評価実験の結果

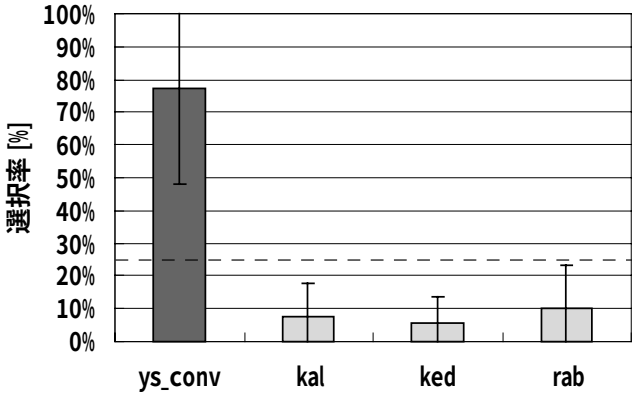

Fig. 2. 主観評価実験における各刺激の選択率　(破線は偶然確率 25%)

映するという報告があり[6]、結果から合成音声に ys 氏の話者性が大きく反映されたことがわかった。また、同様に F2 の高い/ɛ/でも歪みが改善されれば、より話者性が反映されたということが言えるが、本実験では評価の対象として切り出した音声の数が非常に少なく、実験結果からそれらの考察を立証するのは難しかった。

3.2 主観評価実験

話者性の主観的評価は聴取実験を行い、評価した。

・実験方法

20 代の男女 23 名を被験者とした。刺激には ys 氏の日本語の自然発話文(ys_jap)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。被験者には ys_jap 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_jap 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。

・結果と考察

Fig. 2 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。また、実験で得たデータの半分から各刺激の選択率を求め、残りの半分のデータを用いて χ^2 検定を行い、実験値の妥当性を調べた。その結果、有意差は得られず、実験結果が妥当であることが証明された。以上の結果から、本研究で行った声質変換によって、ys 氏の話者性を持った音声を合成することができたことがわかった。

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。合成音声の話者性について、母音における Mel CD を用いた客観評価と、聴取実験による主観評価を行った。客観評価では、すべての母音で歪みの改善が見られた。また、主観評価でも、合成音声が元の話者の話者性を有していることを証明できた。

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、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] A. Iida *et al.*, IEICE Technical Report SP2005-170, pp. 43-48, 2006.
[4] T. Toda, Ph.D. Thesis, Nara Institute of Science and Technology, 2003.
[5] M. Mashimo *et al.*, Proc. Eurospeech, pp. 361-364, Aalborg, Denmark, 2001.
[6] S. Furui *et al.*, 聴覚研資, H85-18, 1985.