HMM 音声合成における複数の自然対話コーパス (UUDB,OGVC) 併用の有効性 *

☆嘉屋和樹, 森大毅 (宇都宮大)

1 はじめに

話者の感情などを表現できるような、実際の人間同士の対話に近い音声の合成を対話音声合成と読んでいる。これまで宇都宮大学パラ言語情報研究向け音声対話データベース (UUDB)[1] を利用した HMM 音声合成方式に基づく対話音声合成の研究を進めてきた [2][3]。しかし、現状では十分な品質の音声合成が実現できていない。その原因の 1 つにモデルの学習に使用する音声データが少ないことが挙げられる。

本研究はモデルの学習に使用する話者数の増加により音声合成の自然性を向上させることを目的とする。特に複数の自然対話コーパスを併用して話者数を増やす検討を行う。

本研究では UUDB と感情評定値付きオンラインゲーム音声チャットコーパス (OGVC)[4] を研究対象とする。UUDB と OGVC は共に仲の良い友人同士の自然対話を収録したコーパスであるため、併用による効果が期待できる。

2 自然対話コーパスの分析

分析の対象とした話者は UUDB に収録されている女性話者 1 名 (11.6 分) と、OGVC に収録されている女性話者 1 名 (4.57 分) である。

自然対話コーパスでは「うん」や「あ」と言ったような同内容の発話が多く収録されていることが多いため、音声合成ではカバレッジが問題となる可能性がある。各自然対話コーパスにおける単語数、出現コンテキスト数を Fig. 1 に示す。

Fig. 1 を見るとコーパスによって出現する単語、出現コンテキストが大きく違うことがわかる。よって、複数の自然対話コーパスを併用することでカバレッジを向上できると考えられる。

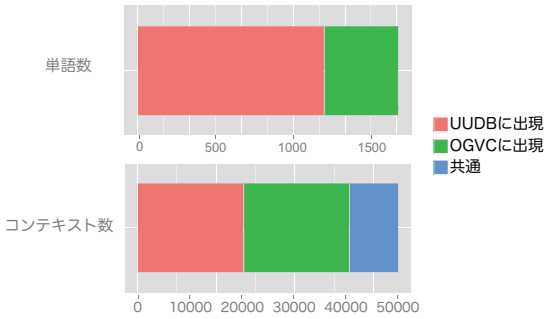

Fig. 1 単語数と出現コンテキスト数

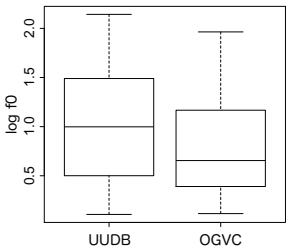

Fig. 2 F0 レンジ

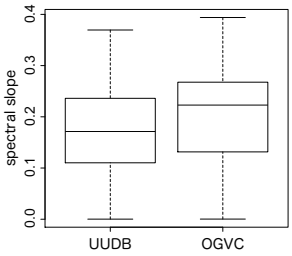

Fig. 3 スペクトル傾斜

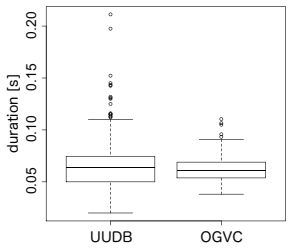

Fig. 4 音素継続時間

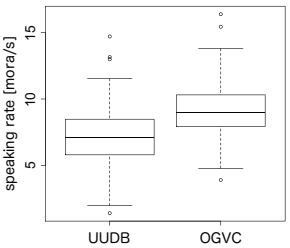

Fig. 5 話速

自然対話コーパスでは対話を収録するための環境や状況が異なる。例えば UUDB は防音室で四コマ漫画の正しい順序を推理させるタスク指向対話を収録したものだが、OGVC は研究室内の会議スペースでオンラインゲームをプレイ中の自由対話を収録したものである。そのため、コーパス間で様々な音響的性質のに違いが生じると考えられる。それぞれのコーパスの F0 レンジ、スペクトル傾斜、継続時間、話速の分析結果をそれぞれ Fig. 2〜5 に示す。

分析の結果を見ると、F0 レンジ、スペクトル傾斜、話速に違いがあることがわかる。この原因は自然対話コーパスの対話を収録する環境や対話状況の違いによるものと考えられる。

3 複数の自然対話コーパスを併用した HMM 音声合成

OGVC 単体で話者数の異なるモデルを 2 種類、UUDB 単体で話者数の異なるモデルを 4 種類、そして UUDB と OGVC を併用して学習したモデルを 1 種類作成し、これら 7 種類のモデルより合成した音

---

*Effectiveness of spontaneous speech corpus combination in HMM-based speech synthesis . by KAYA, Kazuki, MORI, Hiroki (Utsunomiya University)

Table 1　モデルの学習条件

モデル	5 状態 left-to-right HSMM
特徴ベクトル	0-34 次のメルケプストラム係数, 対数基本周波数, Δ, ΔΔ
分析条件	サンプリング周波数 16 kHz の音声に対し窓長 25 ms, 分析周期 5 ms のハミング窓

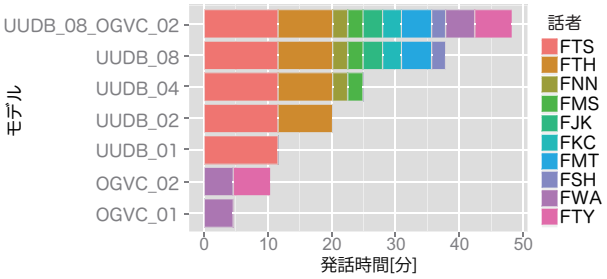

Fig. 6　モデルの概要

声を比較する。各モデルの学習に使用した発話の継続時間を Fig. 6 に示す。モデルの学習には話者正規化学習法 [5] を適用した。その他の条件は Table. 1 に示す。

合成した音声を用いて自然性に関する主観評価実験を行った。モデルの学習に使用した話者は以下の通りである。

- OGVC 単体 (1 名,2 名)
- UUDB 単体 (1 名,2 名,4 名,8 名)
- OGVC 2 名+UUDB 8 名

評価対象合成音声は学習に使用していない 45 発話×モデル数で、被験者は大学生 5 名である。評価は合成音声の自然性を 5 段階評価 (1:非常に悪い,5:非常に良い) で評価するように指示した。主観評価実験の結果を Fig. 7 に示す。

Fig. 7 を見ると、まず OGVC で学習した場合には、UUDB で学習した場合と比べ自然性が低いことがわかる。この原因として、Fig. 5 を見てもわかるように、OGVC は話速が速い発話が多く、音韻のスペクトル変化が小さいため、HMM でうまくモデルが学習できなかったことが考えられる。

また、UUDB で学習した場合には、話者数の増加による自然性の向上が確認できた。Fig. 6 を見てもわかるように学習に使用する音声データが大幅に増加したことによるものと考えられる。しかし、OGVC で学習した場合には話者数を増やしても自然性の向上は見られなかった。また、UUDB の 8 名に OGVC の 2 名を追加したデータで学習した場合には、かえって自然性が低下する傾向が見られた。この原因として、OGVC は音韻のスペクトル変化が小さいため、

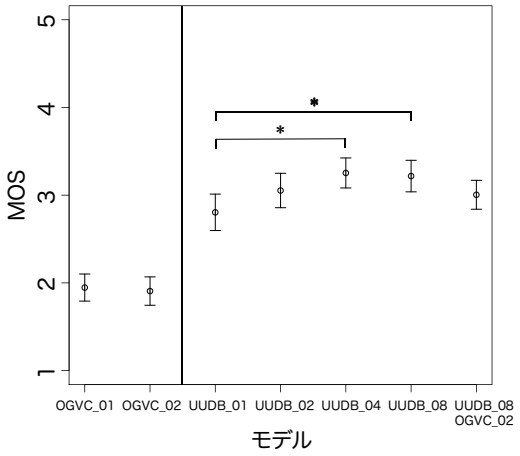

Fig. 7　主観評価実験結果

UUDB に追加した際に HMM でうまくモデルが学習できなくなったこと、また、前節で述べた自然対話コーパスの対話を収録する環境の違いが、スペクトルの推定などに悪影響を与えたのではないかと考えられる。

4　おわりに

本研究では、自然対話コーパスに基づく HMM 音声合成による対話音声合成の品質向上を目標に、2 種類のコーパスの分析と、複数の自然対話コーパスを利用して学習したモデルから合成した音声の主観評価実験を行った。その結果、UUDB で学習した場合には自然性の向上が確認できたが、OGVC と UUDB の単純な併用では合成音声の自然性が向上しないことが確認できた。

今後は、OGVC のような非タスク指向型対話音声から合成した音声の自然性が低下する原因を詳細に分析し、モデルの構造および学習法の工夫などにより品質を改善する必要がある。

参考文献

[1] Mori et al., Speech Communication, No. 53, pp. 36–50, 2011.

[2] Mori et al., Proc. Oriental COCOSDA 2012, pp. 135–140, 2012.

[3] Nagata et al., Proc. Interspeech 2013, pp. 1549–1553, 2013.

[4] Arimoto et al., Acoustical Science and Technology, No. 33, pp. 359–369, 2012.

[5] Yamagishi et al., Proc. ICASSP 2005, pp. 365–368, 2005.