発話モード適応による対話音声合成の高品質化*

☆嘉屋和樹, 森大毅 (宇都宮大)

1 はじめに

話者の意図, 態度, 感情状態などのパラ言語情報を表現することができる, 表情豊かな話し言葉の音声合成を対話音声合成と呼んでいる. 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] を用いた 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]. これらの研究ではパラ言語情報の制御の有効性が示されたが, 学習に用いる対話音声のデータが少なく, 合成音声の品質が低いという問題がある.

著者らは, これまで複数の自然対話コーパス併用による対話音声合成の高品質化を図ったが, その有効性は確認できなかった [4].

そこで, 本研究では, 異種ではあるが大量の音声データを有するコーパスによりベースモデルの合成音声の品質を確保し, その後, 対話音声らしいものへと変換するモデル適応の有効性を検討する.

本研究ではベースモデルの学習に日本語話し言葉コーパス (CSJ)[5] の独話音声 (模擬講演音声) を用い, 変換先の音声データに UUDB を用いる. これらは発話のモード (独話 vs 対話) が異なるが, 比較的近い発話スタイルであるため, 本研究の目的に適していると考えられる.

2 発話モード適応

本研究では独話音声から対話音声への発話モード適応を最尤線形回帰 (MLLR) と最大事後確率 (MAP) 推定を組み合わせた手法 [6] によって行う. MAP 推定における事前確率分布の平均ベクトルとしては, MLLR によって変換された平均ベクトルを用いる.

本研究では, 変換元の音声データ量に比べ, 変換先の音声データ量が非常に少なく, ベースモデルの決定木のリーフノードに対応するコンテキストを持つ適応先のインスタンス数がゼロまたは少量となる場合が多い. そこで, 適応データに基づいて回帰クラスを作成し, 同一の回帰クラスに属するガウス分布の変換行列を同一のものとして, 適応先のデータ数を確保した上で MLLR を行う.

Table 1 モデルの条件

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

3 CSJ と UUDB を用いた HMM 音声合成

学習データとして UUDB の女性話者 1 名 (551 発話, 約 11 分) だけを用いたモデル (UUDB), CSJ の模擬講演女性話者 54 名 (13266 発話, 約 5 時間 4 分) だけを用いたモデル (CSJ), そして変換先を UUDB として CSJ から発話モード適応を行ったモデル (ADAPT) の 3 つのモデルを作成し, 合成音声を比較する. モデルの条件は Table 1 に示す.

従来の音声合成は, 肉声らしさや明瞭度を基準として品質が評価されてきた. しかし, 対話音声合成の評価はそれだけでは不十分である. なぜなら, 人と機械の音声コミュニケーションを人同士のそれに近い自然なものにしようとするならば, そこで利用される合成音声は, アナウンサーの読み上げ音声のように単に明瞭であればよいわけではなく, 対話という場面にふさわしい話し方でなければならないからである. また, 日常場面での対話音声は, プロのアナウンサーや声優の音声のように必ずしも明瞭ではなく, またそこまで明瞭である必要もない. そこで本研究では明瞭度に加え, 対話音声らしさの評価を行った. 評価対象合成音声は UUDB の女性話者 1 名の学習に用いていない 50 発話で, 被験者は大学生 12 名である.

明瞭度評価実験では評価対象合成音声の聞き取りやすさ 5 段階 (5:非常に良い, 1:非常に悪い) で評価させた.

対話音声らしさ評価実験では評価対象合成音声が対話音声と独話音声のどちらに聞こえるかを評価するように指示した. 評価実験に先立ち, まず「対話」, 「独話」という用語を説明した. 次に対話と独話の自然音声の例をそれぞれ 10 発話ずつ聞かせた. 次に対話と独話の自然音声をそれぞれ 5 発話ずつ用いて評価の練習をさせた. その後, 評価対象合成音声の対話音声らしさを 5 段階 (5:対話音声に聞こえる, 1:独話音声に聞こえる) で評価させた.

*Improving dialogue speech synthesis based on speech mode adaptation. by KAYA, Kazuki, and MORI, Hiroki (Utsunomiya University)

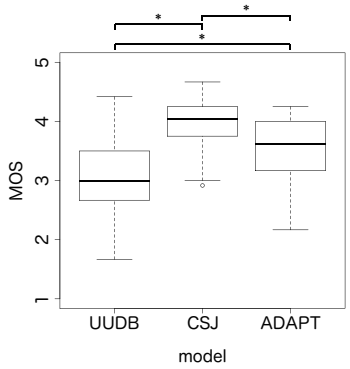

Fig. 1　明瞭度評価実験結果 ($*: p < .05$)

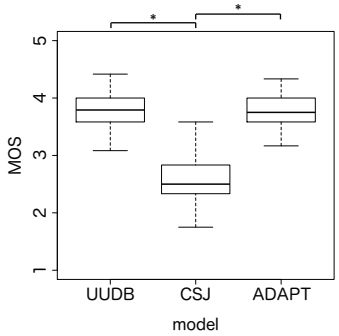

Fig. 2　対話音声らしさ評価実験結果 ($*: p < .05$)

明瞭度評価実験の結果を Fig. 1 に示す. この結果から, UUDB だけから学習したモデルよりも ADAPT の評価値が高いことがわかる. よって, 発話モード適応により合成音声の明瞭度を向上させることができたと言える. しかし, この評価値は CSJ だけから学習したモデルの合成音声の評価値よりも低く, モデル適応前に比べ若干の明瞭度の低下が認められた.

対話音声らしさ評価実験の結果を Fig. 2 に示す. この結果から, ADAPT は UUDB だけから学習したモデルと同程度の対話音声らしさであることがわかる. また, その評価値は CSJ だけで学習したモデルよりも明らかに高く, 5 段階評価で 1.2 ポイント改善できている.

以上の結果から, CSJ からの発話モード適応により, 合成音声の対話音声らしさを損なうことなく明瞭性を改善することができたと言える.

CSJ だけから学習したモデル, 及び UUDB だけから学習したモデルの F0 の決定木のリーフノード数は, それぞれ 1951, 352 であった. CSJ から学習したモデルは UUDB だけから学習したモデルと比べ詳細なモデルが学習できていることがわかる. ADAPT も同じ詳細さを持っており, このために UUDB よりも高い明瞭度となったと考えられる. また, 回帰クラス木のノード数は 111 であった. これは平均 3 個程度のガウス分布に対して 1 つの変換行列が得られていることを意味し, MLLR が比較的詳細に実行されていると言える.

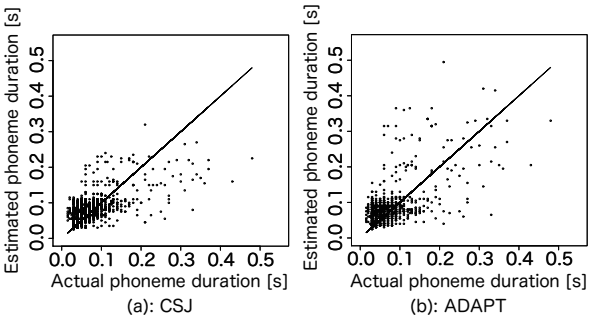

Fig. 3　音素継続時間 (横軸:自然音声, 縦軸:合成音声)

UUDB の自然音声と CSJ から学習したモデル, 及び ADAPT の合成音声の音素継続時間の散布図を Fig. 3 にそれぞれ示す. 横軸はテストデータ 50 発話の実際の継続時間である. 自然音声には 200ms を超える長い継続時間を持つ音素が含まれるが, CSJ から学習したモデルでは, これらの音素が実際よりも短く推定されていることがわかる. これに対し, ADAPT ではその一部が自然音声と同程度に長く推定されている. これは対話音声の特徴である発話末の音素継続時間の伸びなどが, 適応により再現できるようになっていることを示している.

4　おわりに

本研究では, 対話音声合成の高品質化を目的に, 独話音声を用いたモデルの学習を行い, その後, 対話音声を用いた MLLR と MAP の組み合わせによる発話モード適応によって対話音声らしいものへと変換する手法を提案した. 合成音声の主観評価実験の結果, 本手法は対話音声合成の品質向上に有効であることがわかった.

今後は, 独話音声と対話音声について分析を行い, その違いに着目した変換の手法を検討する.

謝辞 本研究は JSPS 科研費 26280100 の助成を受けている。

参考文献

[1] Mori et al., Speech Communication, Vol. 53, pp. 36–50, 2011.
[2] Mori et al., Proc. Oriental COCOSDA 2012, pp. 135–140, 2012.
[3] Nagata et al., Speech Communication, Vol. 88, pp. 137–148, 2017.
[4] 嘉屋他, 音講論 (秋), pp. 161–162, 2016.
[5] 前川, 日本語科学, 15, pp. 111–133, 2004.
[6] Digalakis et al., IEEE Trans. Speech Audio Process., Vol. 3, pp. 357–366, 1995.