音声合成用コーパスおよび日常会話コーパスのハイブリッドモデリングによる日常会話音声の合成*

☆古川晃大，森大毅（宇都宮大）

1　はじめに

今現在，会話エージェントに使用されている音声合成システムは，指定した文を声のプロフェッショナルに読み上げさせたコーパスから構築されている。しかし，読み上げ音声はインタラクション性，自発性，非流暢性などの面において会話音声と異なる。そのような読み上げ音声から合成した音声で話す会話エージェントに対して，人間は他人と会話する時のような聞き手反応を示さない [1]。これでは，相槌やフィラーなどの反応から聞き手の理解状態を監視し適応的にふるまう対話システム [2] の利点が活かされない。他方，会話エージェントの声を会話コーパスベースの合成音声にするだけで，ユーザの相槌は増える [1]。

本研究は，これまでの会話音声合成の研究を一歩進め，日常会話音声の合成を目指す。日常会話には，インタラクションの背景となる環境の違いや，アドレス先との社会的関係の違いに起因するパラ言語的特徴の違いを有する特徴がある。会話音声合成研究の目標は，これらの違いをモデル化することで，様々なインタラクション場面に対応できる音声合成を実現することである。本稿ではこの前段階として，日本語日常会話コーパス [3] から韻律情報を，別の高品質なコーパスからスペクトル情報を学習することにより，読み上げ音声の合成と同等な品質を保ちつつ，イントネーションや会話のリズムは会話音声そのものであるような音声を合成することを試みる。

2　コーパス

本研究では，日本語日常会話コーパス (CEJC) [3] を用いる。CEJC には日常場面で自発的に生じた会話が収録されているため，会話音声の持つ特質を適切にモデル化できる可能性がある。

CEJC の会話のうち調査協力者 3 名 (K001, K002, K003) の音声を転記単位で分割し，笑い声とマスク音を含む会話を取り除いた。さらに今回は，発声様式が対面会話と大きく異なる電話音声を除外した。また，CEJC は録音レベルが統制されていないため，発話内強度最大値のセッションごとの平均が一定値になるよう，セッション単位で振幅の正規化を行った。

3　会話音声合成モデル

CEJC は録音品質が悪く，音声合成には向いていない。今回は，FastSpeech 2 [4] を使用した，音声合成用コーパスおよび日常会話コーパスのハイブリッドモデリングを試みた。

まず，CEJC を用いてモデル全体を学習する。次に JSUT [5] でファインチューニングを行う。FastSpeech 2 には，音声の韻律的特徴である duration, pitch, energy の予測器が組み込まれており，訓練用コーパスからそれぞれ個別に学習される。ファインチューニングの時には，duration 予測器と pitch 予測器のモデルパラメータの更新を行わずに，CEJC で学習したモデルパラメータを保存しておく。これにより韻律は日常会話らしさを残しつつ，読み上げ音声合成と同等の品質を実現できることが期待される。

4　合成音声の評価実験

4.1　実験方法

評価対象の合成音声として，以下の 3 条件で音声合成モデルを学習し，合成音声を作成した。

JSUT JSUT のみで学習したモデル

CEJC CEJC のみで学習したモデル

Hybrid CEJC モデルのスペクトル予測部のみを JSUT でファインチューニングしたモデル

訓練データは，JSUT が話者 1 名の読み上げ音声 (5.5 時間)，CEJC が話者 3 名の会話音声 (4.1 時間) である。テスト用発話は，CEJC の話者 K001 の発話のうち訓練データに含まれない連続した 50 発話 (ただし，同一発話は除く) とし，上記の 3 条件で合成した計 150 発話を，順序は

*Synthesis of daily conversation speech using a hybrid model of read and conversational speech corpora. By FURUKAWA, Kota, MORI, Hiroki (Utsunomiya University)

Table 1　評価実験の結果 (5 段階平均)

	JSUT	CEJC	Hybrid
明瞭性	3.87	2.04	2.98
日常会話音声らしさ	2.56	3.40	3.34

変えずに条件だけをランダム化して刺激セットを作成した。

評価実験は，20 代の男性 8 人, 女性 2 人の計 10 人に対して実施した。実験参加者には，評価用ウェブページから刺激を 1 発話ずつ再生し，明瞭性と日常会話音声らしさをそれぞれ 5 段階で評価するよう指示した。日常会話音声らしさは，5 を「家族，親戚，知人などと話しているように聞こえる」，1 を「何かを読み上げているように聞こえる」と設定した。

4.2　実験結果

合成音声を聴き比べたところ，Hybrid モデルは，CEJC モデルと比べて雑音の少ない音声を合成することができた。

Table 1 に，明瞭性および日常会話音声らしさの平均評価値を示す。また，Fig. 1 と Fig. 2 に，各刺激に対する明瞭性と日常会話音声らしさの平均評価値の分布を示す。

明瞭性および日常会話音声らしさの評価に音声合成モデルが与える影響を検討した。各刺激に対する明瞭性および日常会話音声らしさの評価の実験参加者全体にわたる平均値の分布に対し分散分析を実施したところ，音声合成モデル (3 水準) の主効果がいずれも有意であった ($p < 0.001$)。また，多重比較を行ったところ，日常会話音声らしさにおける CEJC と Hybrid の間の差を除き，全ての組み合わせで平均値の差が有意であった ($p < 0.001$)。

まとめると，日常会話音声らしさは CEJC モデルと Hybrid モデルが JSUT モデルに比べ高く評価された。また，Hybrid モデルでは CEJC モデルと同程度の日常会話音声らしさを保ちつつ，CEJC モデルよりも明瞭な音声が合成できることが示された。

4.3　考察

提案法によって，日常会話音声らしさを残しつつ，合成音声の品質を改善できた。しかし，読み上げ音声のモデルに比べて明瞭性は約 0.9 ポイント低い。会話音声には聞き手反応のような短い

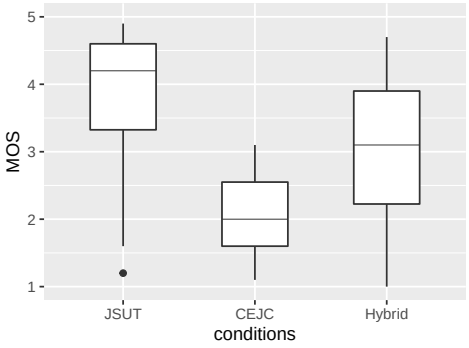

Fig. 1　明瞭性の評価実験結果

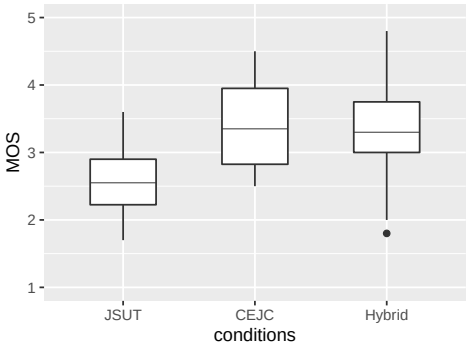

Fig. 2　日常会話音声らしさの評価実験結果

発話が多く含まれるが，このような発話の合成音声の品質は低くなる傾向があった。会話音声はそもそも読み上げ音声ほど明瞭ではないので，明瞭性は高いほど良いという単純なものではなく，より適切な品質評価法が必要である。

5　おわりに

本研究では，日常会話コーパスの韻律情報を用いた日常会話音声合成を提案した。実験の結果，ハイブリッドモデリングによって，日常会話らしさを残しつつ，品質がある程度改善された音声を合成することが可能ということが示された。

参考文献

[1] Iizuka and Mori, IEEE Access, **10**, 111042–111051, 2022.
[2] 森, 森本, 人工知能学会論文誌, **39**, 3, 2024 (accepted).
[3] 小磯他, 国立国語研究所論集, **24**, 153–168, 2023.
[4] Ren et al., ICLR 2021, 2021.
[5] Sonobe, Takamichi and Saruwatari, arXiv 1711.00354, 2017.