複数の自発音声コーパスの併用による end-to-end 対話音声合成の高品質化*

☆西野広直, 森大毅 (宇都宮大)

1 はじめに

これまで、感情次元の制御が可能な対話音声合成は統計的パラメトリック音声合成の枠組みで行われてきた [1]. その一方で, end-to-end 音声合成の一つである Tacotron 2 [2] はこの枠組みに比べて高品質であり, 対話音声合成への応用も期待できる. しかし, 感情次元ラベルが付与された自然対話音声コーパスは小規模であり, データ量に品質が大きく依存する end-to-end 音声合成への応用は難しい.

そこで本稿では, 自動的に付与された感情次元ラベル付きの大規模自発音声コーパスによるモデルの事前学習を提案する. 感情次元を考慮した事前学習により, 感情次元の制御性能と対話合成音声の品質が向上することを期待する.

2 感情次元の制御が可能な end-to-end 音声合成

感情次元の制御と複数話者に対応させるために, Tacotron 2 の入力に話者埋め込みと感情次元ラベルを追加した. 話者埋め込みと感情次元ラベルはそれぞれ 2 層の線形層を経て, エンコーダの出力に追加される.

メルスペクトログラム算出のための短時間フーリエ変換の条件はフレーム長 50ms, フレームシフト 12.5ms で, ハン窓を使用した.

3 感情次元を考慮した事前学習

Tacotron 2 の学習には, 2 種類のコーパスを使用する. 事前学習には大規模自発音声コーパスを使用し, fine tuning の際には自然対話音声コーパスを用いる.

自然対話音声コーパスとして UUDB (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)[3] を使用する. モデルの学習には女性話者 12 名, 約 1 時間 12 分を用いる. UUDB に含まれる発話には, ラベラ 3 名による快-不快, 覚醒-睡眠など 6 次元の評価値が 7 段階 (4:中立) で付与されている. 本研究では快-不快および覚醒-睡眠の 2 次元を制御の対象とし，3 名の平均評価値を学習時に入力する感情次元として使用した.

大規模自発音声コーパスには CSJ (日本語話し言葉コーパス) を使用する. モデルの学習には模擬講演音声 (話者 368 名, 約 58 時間 40 分) を用いる. CSJ には感情次元ラベルが付与されておらず, 人手でラベリングをすることは困難であるため, 感情次元認識器による自動ラベリングを行う. 感情次元認識器は, 2 つのたたみこみ層 (3 × 3, チャンネル数 64, 2 × 2 max プーリング), 線形層, 双方向 LSTM (次元数 128), 感情次元ごとの local attention [4], および線形層で構成される. 入力は 80 次元のメルスペクトログラム, 出力は感情次元の推定値である. 感情次元認識器の学習には UUDB の全話者を用いた. 評価セットに対する正解と推定値の相関係数は, 快-不快で 0.739, 覚醒-睡眠で 0.919 であり, どちらの感情次元もおおよそ正しく推定できていることがわかる.

Table 1　入力した感情次元と対数平均 F0 の相関係数

	快-不快	覚醒-睡眠
CSJ→UUDB-wEmo	0.24	0.42
CSJ-wEmo→UUDB-wEmo	**0.39**	**0.56**

音声合成モデルの事前学習において自動ラベリングされた感情次元を用いることの有効性を調査するために, 以下の 2 つのモデルを比較する.

CSJ→UUDB-wEmo
CSJ で事前学習を行い, 感情次元付き UUDB で追加学習したモデル. 感情次元入力のノードは事前学習の後に追加.

CSJ-wEmo→UUDB-wEmo
感情次元ラベル付き CSJ で事前学習を行い, 感情次元付き UUDB で追加学習したモデル.

4 感情次元制御実験

評価対象音声は, UUDB の評価セットから選んだ 18 発話に対し, 感情次元 (快-不快, 覚醒-睡眠) の組み合わせを (3,3), (4,4), (4,5), (4,6), (5,5), (6,6) の 7 通りとした計 126 発話を, 2 つのモデルで合成したものである. また, ニューラルボコーダは MelGAN [5] を使用し, 標本化周波数 16 kHz で合成音声の波形を生成した.

4.1 客観評価

入力した感情次元と合成音声の対数平均 F0 の相関係数を Table 1 に示す. CSJ-wEmo→UUDB-wEmo は, CSJ→UUDB-wEmo に比べて各次元とも対数平均 F0 との相関係数が大きくなることがわかった. このことから, 事前学習において感情次元を考慮した方が, 出力される音声の F0 の変化が感情次元の変化により敏感になったことがわかる.

*Improving the quality of end-to-end dialogue speech synthesis by semi-supervised pre-training with a large-scale spontaneous speech corpus. by NISHINO, Hironao, MORI, Hiroki (Utsunomiya University)

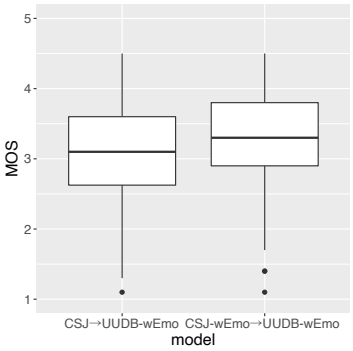

Fig. 1　明瞭性評価実験結果

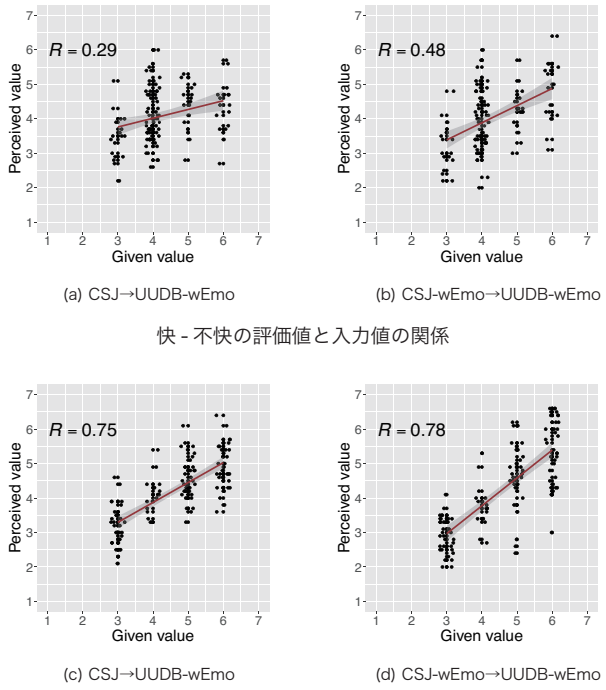

Fig. 2　感情次元評価実験結果

4.2　主観評価

各モデルで合成した音声を, 明瞭性と感情次元の点から知覚実験によって比較検討する. 被験者は音声の研究に従事していない大学院生 10 名である. 被験者には各モデルで合成した音声計 252 発話をヘッドホンにより聞かせ, 合成音声の明瞭性を 5 段階, 各感情次元を 7 段階で評価させた.

明瞭性の結果を Fig. 1 に示す. 各手法の平均は, CSJ→UUDB-wEmo で 3.13, CSJ-wEmo→UUDB-wEmo で 3.25 であり, これらの差は統計的に有意ではなかった.

感情次元知覚の結果を Fig. 2 に示す. 入力した感情次元の値と知覚された値の相関を調べた結果, 快-不快では CSJ-wEmo→UUDB-wEmo の方が相関係数が大きいことがわかった ($p < 0.05$). また, 覚醒-睡眠の相関係数の差は統計的に有意ではなかった.

Table 2　AB 法に基づく主観評価実験結果 (%)

CSJ-wEmo	CSJ-wEmo→UUDB-wEmo
28.5	**71.5**

5　独話ベース vs 対話ベース

事前学習に使用した CSJ は独話音声であった. 事前学習の段階で合成音声の品質は既に高いので, これが音声対話システムの声として適していれば, 対話音声による追加学習の必要はない. そこで, 対話音声による追加学習の効果を検証するため, 事前学習だけのモデル CSJ-wEmo と, UUDB で追加学習したモデル CSJ-wEmo→UUDB-wEmo の合成音声に対し, AB テストに基づく主観評価実験を行った. 被験者は感情次元制御実験と同じである. 被験者は 2 種類のモデルで合成された A, B のペアのうち,「あなたと 1 対 1 でおしゃべりをするならば, 話し相手は A と B のどちらの話し方が良いですか?」の質問に対する答えを選択する. 評価用の発話内容は UUDB の評価セットから 48 発話選んだ.

主観評価実験の結果を Table 2 に示す. CSJ-wEmo→UUDB-wEmo の選好率が CSJ-wEmo を上回る結果となった ($p < 0.05$). 以上より, 対話システムの音声合成では, 独話音声にはない対話音声独自のスタイルを学習させることが必要だとわかる.

6　おわりに

本論文では, 感情次元が制御可能な対話音声合成において, 感情次元ラベルが自動的に付与された大規模自発音声コーパスによる事前学習の有効性を検証した. その結果, 事前学習において感情次元を考慮した方が, 入力する感情次元の変化に対して出力される音声の変化がより敏感な傾向であり, 合成音声から知覚される快-不快の制御性能が高いことがわかった. また, 独話ベースと対話ベースで学習したモデルを比較した結果, 対話ベースの方がより対話音声合成に適していることが主観評価実験の結果から明らかとなった.

本研究より, 感情次元を考慮した事前学習が有効であることが示された. 今後は, 快-不快の制御性能を向上させるのに有効な特徴量に関して分析する.

謝辞　本研究は JSPS 科研費 19H01252 の助成を受けている.

参考文献

[1] Yokoyama *et al.*, Proc. Interspeech 2018, pp. 3053–3056, 2018.
[2] Shen *et al.*, ICASSP, pp. 4779–4783, 2018.
[3] Mori *et al.*, Speech Communication, No. 53, pp. 36–50, 2011.
[4] Mirsamadi *et al.*, ICASSP, pp. 2227–2231, 2017.
[5] Kumar *et al.*, NIPS, pp. 14881–14892, 2019.