感情音声を活用した親しみやすい音声対話システムのための感情合成音の検討

Investigation of Synthesizing Emotional Speech for Familiar Spoken Dialogue System With Emotional Speech

加藤 大地

Daichi Kato

岡山大学 阿部研究室

Abe Laboratory, Okayama University

概要 音声対話システムにおいて感情的な音声を応答に活用することができれば，人間らしい自然な応答が可能になり，対話の質を向上させることができると考えられる．本稿では，近年 TTS の品質を向上させている DNN による TTS を用いた感情合成音を活用した音声対話システムについて検討する．まず，音声対話システムが満たすべき要件について検討し，感情音声の合成を行った．

1 はじめに

音声対話システムとは，音声を使って人と機械が対話する為のシステムである．対話は人の日常的な行為の 1 つであり，音声対話システムはそれと同等な対話を行えることを目指して研究開発がされている．人間同士の対話において行われる表現の 1 つに非言語情報としての感情表現がある．これは，例えば，嬉しい時には声が明るくなったり，悲しい時には声が沈んだりと，言語情報ではなく声質やピッチに現れる非言語情報としての表現である．また，同じ人物であっても，感情表現の程度は一定ではなく，感情豊かになったり，感情表現を抑えたりすることがある．音声対話システムにおいても応答に適切な感情表現を含めることができれば，人間らしく親しみやすい自然な応答が可能になり対話の質が向上すると考えられる．しかし，製品化されている音声対話システムは，中立的な感情の音声のみで応答するものが一般的である．

先行研究では，システムの応答に感情的な表現を含む音声を用いることで対話の質が向上することが確認されている [1, 2]．これらの研究では，隠れマルコフモデル（hidden Markov model; HMM）により感情音声を合成している．また，ユーザ発話あるいはシステム応答に含まれる単語を極性辞書に照らし合わせ，常に一定のルールに基づきシステム応答の感情を推定している．

近年，DNN の発展により，テキスト音声合成（Text-to-Speech; TTS）で高品質な音声を合成することが可能になっている．システムの応答に用いる合成音声の品質も対話の質に貢献すると考えられるため，本研究では，DNN による TTS で感情音声を合成し，音声対話システムに組み込むことで対話の質を向上させる方法について検討する．本稿では，まず感情音声を応答に用いる音声対話システムが満たすべき要件について述べ，感情音声を Transformer-TTS により合成した結果について述べる．

図 1: TTS の流れ

2 感情音声を使う音声対話システムの要件

2.1 感情音声の合成

感情音声を音声対話システムで使う為には，システムの応答文テキストから目的の感情表現を含んだ音声を合成できる必要がある．合成する感情については，中立的，楽しい，悲しいのような感情の音声が音声対話システムにおいて有効であると考えられる．

一般的に TTS は図 1 のような流れで行われる．まず，テキストをテキスト解析器で言語特徴量に変換する．次に，言語特徴量を音響モデルで音響特徴量に変換する．そして，音響特徴量をボコーダで音声に変換する．DNN による TTS では，音響モデルとして Tacotron2[3] や Transformer[4]，FastSpeech[5] のようなモデルが，ボコーダとして WaveNet[6] や Parallel WaveGAN[7] のようなモデルが研究されている．

2.2 応答文の感情推定

悲しい場面で楽しそうに応答するなど，適切でない感情の音声で応答するとユーザが不快に感じてしまい，かえって対話の質が下げてしまう可能性がある．そのため，システムがどのような感情で応答することが適切か推定する必要がある．また，対話全体を通して常に感情豊かに応答するのか，だんだんと感情豊かに応答するようになるのかのような応答戦略についても検討すべきである．

2.3 リアルタイム合成

ユーザ発話が終了した後，システム応答が始まるまでに過度に時間を要する場合，ユーザを待たせてしまい，対話の質の低下につながる．TTS の合成速度が十分に短い場合，合成してから再生，あるいは合成しながら再生することができるが，そうでない場合は，あらかじめ対話で使用する応答文を全て合成しておく必要がある．TTS の合成速度が満たすべき要件については，

音声の長さや合成，再生部分の実装に依存するが，合成する音声の長さに対する合成に要する時間の比（real time factor; RTF）が 1 未満である必要があると考える．高速に合成できる DNN-TTS モデルとしては，音響特徴量の生成は FastSpeech，ニューラルボコーダは Parallel WaveGAN がある．

3 感情音声の合成

3.1 音声コーパス

本稿で使用した音声コーパスについて以下に説明する．

声優統計コーパス [8] プロの女性声優 3 名が 3 パターンの感情（normal，happy，angry）で読み上げた日本語音声が 100 文ずつ含まれている．

JSUT[9] 東京大学猿渡研究室で作成されたコーパスである．1 人の日本語女性話者の 10 時間分の音声が含まれている．内容は，常用漢字の音読み・訓読みを全てカバーしたテキスト 5000 文，助数詞，日本語オノマトペ，声優統計コーパスと同じテキストなどの読み上げである．

JVS[10] 東京大学猿渡研究室で作成されたコーパスである．声優や俳優など，100 人のプロの日本語話者の音声が合計 30 時間分含まれている．内容は，声優統計コーパスと同じテキストの読み上げが全話者共通で含まれており，話者ごとに異なるテキストの読み上げも含まれている．

3.2 合成器の学習

要件の 1 つ目である感情音声の合成について行った．テキスト解析器には Open JTalk[11] を使用し，日本語テキストを音素系列に変換する．音響モデルには Transformer-TTS[4] モデルを使用し，音素系列をメルスペクトログラムに変換する．ボコーダとして Griffin-Lim アルゴリズムを使用し，メルスペクトログラムを音声波形に変換する．

音響モデルは声優統計コーパスのテキストを音素系列に変換したものを入力とし，対応する音声のメルスペクトログラムを出力するよう学習させた．コーパス内の 100 文のうち学習に 90 文，検証と評価に 5 文ずつ使用し，各話者各感情ごとに音響モデルを学習した．ただし，声優統計コーパスだけではデータが少ないため，あらかじめ JSUT で学習済みの Transformer-TTS のパラメータを初期値として再学習させた．学習には ESPnet[12] という End-to-End 音声処理ツールキットを使用した．ESPnet には JVS 用 TTS レシピが含まれているこれは，JSUT で学習済みの音響モデルを JVS 内のいずれかの話者 1 名のデータで再学習させるものである．本稿での音響モデルの学習時のハイパーパラメータはこの JVS 用 TTS レシピに基づく．

3.3 合成結果

合成された音声を聴取して確認したところ，元の話者が元の感情で発話している際の声質を再現できているように感じられた．発話中のアクセントについては再現できていない部分があるものの，語尾上げのような特徴的な表現は再現できていると感じられた．

音声の感情による影響が現れる特徴の 1 つに基本周波数 F0 がある．各話者各感情の元音声と合成音を WORLD[13] により F0 を分析し，常用対数をとりヒストグラムとしてプロットしたものを図 2 に示す．青色が元音声，橙色が合成音を表す．この図より，F0 の分布は概ね再現できているが，図 2c，図 2f，図 2g のように，元の音声にはない低い F0 が多く合成されている部分もあることがわかる．

また，ボコーダとして JSUT で学習済みの Parallel WaveGAN モデルにより合成してみたところ，一部の音声，特に happy の音声が掠れたような音声になってしまい，うまく合成できなかった．これは，JSUT の音声と声優統計コーパスの happy の音声の特徴に大きな差異があるためだと考えられる．図 3 に JSUT と声優統計コーパスに含まれる音声の F0 の分布を示す．この図から，声優統計コーパスの happy の音声の F0 は JSUT の F0 に比べて高いことがわかる．この差異により，happy の音声をうまく合成できなかったと考えられる．

4 まとめ

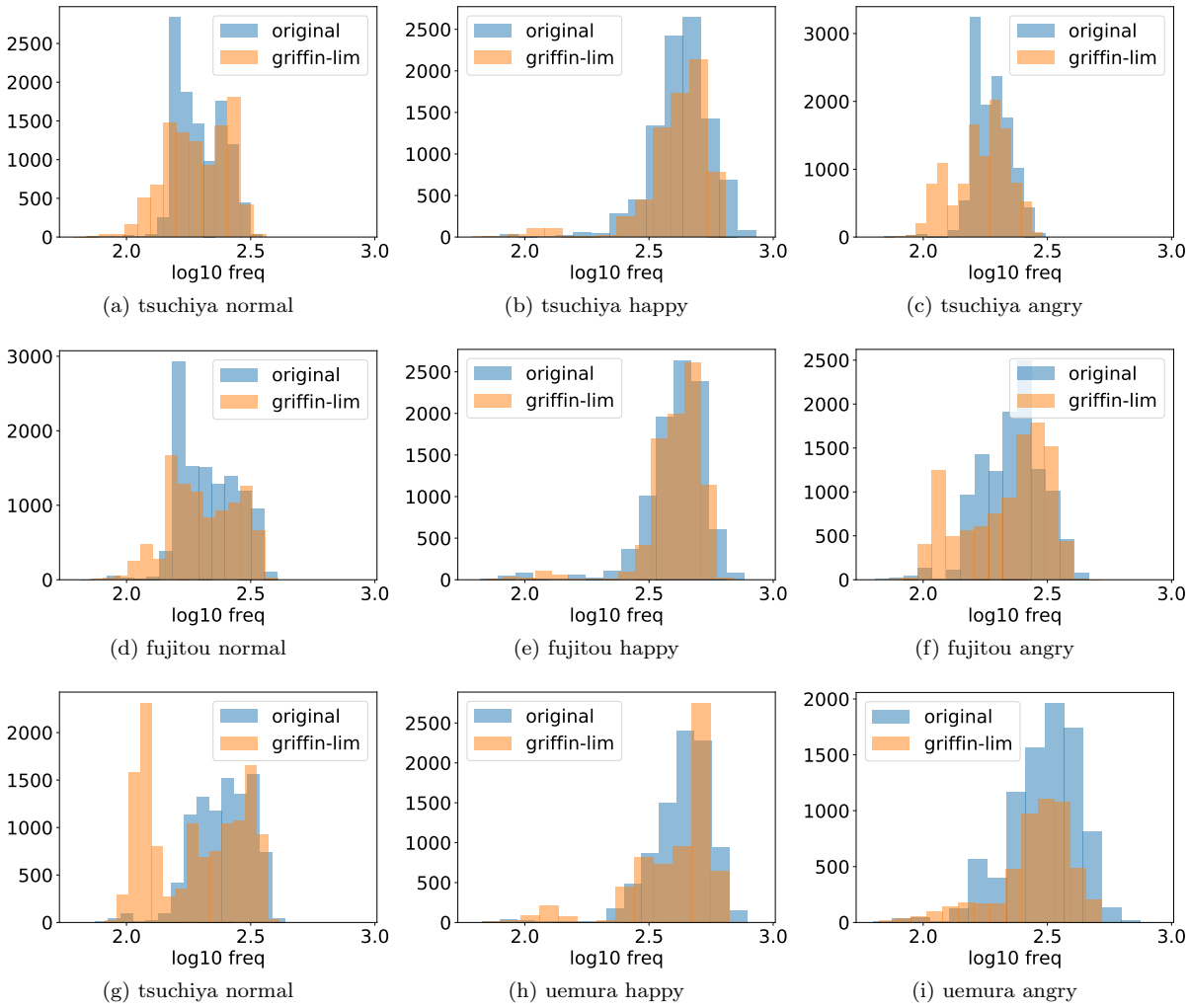
本稿では，応答に感情音声を用いる音声対話システムが満たすべき要件として，感情音声の合成，応答文の感情推定，リアルタイム合成について述べ，1 つ目の要件である感情音声の合成を行った．

今後の課題としては，ニューラルボコーダによる高品質な音声合成，合成音声が本当に目的の感情に感じられるかの主観評価，システムの応答で感情表現を行うか，どのような感情表現を行うかの推定方式の検討，リアルタイムな TTS が可能かの検討などが挙げられる．

参考文献

[1] T. Kase, T. Nose, and A. Ito, “On Appropriateness and Estimation of the Emotion of Synthesized Response Speech in a Spoken Dialogue System,” HCI International 2015 - Posters’ Extended Abstracts, ed. C. Stephanidis, pp. 747–752, Cham, 2015, Springer International Publishing.

[2] Y. Chiba, T. Nose, T. Kase, M. Yamanaka, and A. Ito, “An Analysis of the Effect of Emotional Speech Synthesis on Non-Task-Oriented Dialogue System,” Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dia-



(a) tsuchiya normal (b) tsuchiya happy (c) tsuchiya angry

(d) fujitou normal (e) fujitou happy (f) fujitou angry

(g) tsuchiya normal (h) uemura happy (i) uemura angry

図 2: 各話者各感情ごとの F0 の分布

logue, pp. 371–375, July 2018.

[3] J. Shen, R. Pang, R.J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R.J. Skerry-Ryan, R.A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions," CoRR, vol. abs/1712.05884, 2017.

[4] N. Li, S. Liu, Y. Liu, S. Zhao, M. Liu, and M. Zhou, "Close to Human Quality TTS with Transformer," CoRR, vol. abs/1809.08895, 2018.

[5] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Liu, "FastSpeech: Fast, Robust and Controllable Text to Speech," CoRR, vol. abs/1905.09263, 2019.

[6] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A.W. Senior, and K. Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," CoRR, vol. abs/1609.03499, 2016.

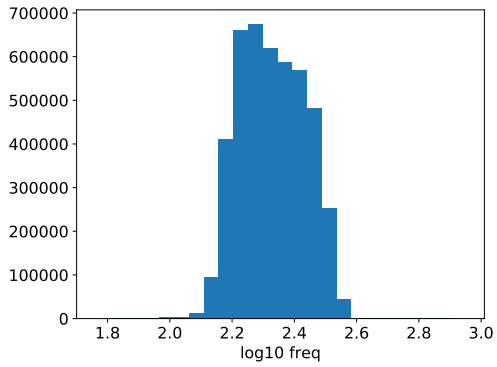
[7] R. Yamamoto, E. Song, and J.M. Kim, "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6199–6203IEEE, 2020.

[8] y_benjo, and MagnesiumRibbon, "Voice-Actress Corpus," `http://voice-statistics.github.io/`, accessed on Jul. 19, 2020.

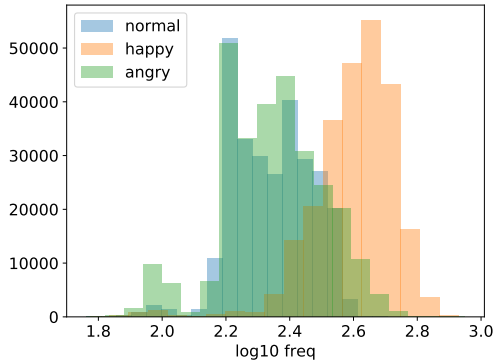
[9] R. Sonobe, S. Takamichi, and H. Saruwatari, "JSUT corpus: free large-scale japanese speech corpus for end-to-end speech synthesis," CoRR, vol. abs/1711.00354, 2017.

[10] S. Takamichi, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji, and H. Saruwatari, "JVS corpus: free Japanese multi-speaker voice corpus," arXiv preprint arXiv:1908.06248, 2019.

[11] "Open JTalk," `http://open-jtalk.sourceforge.net/`accessed on Jan. 28, 2020.



(a) JSUT



(b) 声優統計コーパス

図 3: コーパス中の F0 の分布

[12] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "ESPnet: End-to-End Speech Processing Toolkit," Interspeech, pp. 2207–2211, 2018.

[13] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," IEICE TRANSACTIONS on Information and Systems, vol. 99, no. 7, pp. 1877–1884, 2016.