第 6 回 東北地区音響学研究会（2023 年 11 月 25 日）資料番号 6−6

音声認識による笑い声表現と音声合成の検討

◎須藤智貴，近藤和弘（山形大）

1　はじめに

人間の発話には非言語的な情報が含まれている。笑い声や叫び声など、これらは典型的な情動表現であり、特に笑い声は他者への伝染やストレス反応の抑制など社会的、生理的に重要な役割を備えている。これを自然に合成することは、合成音声による感情表現をより豊かにし、自動対話システムなどによる人間と機械のコミュニケーションをより円滑なものにすることができると考えられる。

通常の発話を対象にした機械学習による Text-to-Speech では一般的に合成対象音声の言語情報を文字や音素などで表現し、それを用いて合成モデルの学習を行う。笑い声でも同様の試みが行われているがその自然性は低い。原因に関する議論では、笑い声の生成過程と通常音声の生成過程の不一致による音声分析合成時点のエラーや、ラベリング済みデータが不足していること、笑い声の適切な表現方法がはっきりしていないことなどが指摘されている。

笑い声の表現方法に関して、音声パワーの情報を使用したモデル[1]や感情ラベルを使用したモデル[2]、非笑い声の MFCC を入力とした笑い声生成[3]、自己教師あり学習モデルで抽出した特徴量をクラスタリングによって擬似的な音素トークンとして使用する方法[4]など様々な手法が提案されている。

本研究では、音声認識を使用して得られる事後確率を笑い声の表現として音声合成を行うことを検討する。これにより、コストが高く不安定な、人間による文字起こしを行わずとも、言語的な認識とより近い表現から笑い声合成可能となることを期待する。

2　使用した音声データ

日本語母語話者による大規模笑い声コーパス[4]を使用した。これは Youtbe 動画から収集した単独話者による笑い声データで、584 名の話者による 11413 発話が含まれている。

公開されているデータは人力で単一話者であることを確認され、深層学習を用いた音源分離モデルを使用してデノイジングが施されている。サンプリング周波数は 24 kHz で公開されているが、本研究では 16 kHz にダウンサンプリングして使用した。

3　Wav2Vec2.0 による音声認識

音声認識には wav2vec2.0[5]の事前学習済みモデル[1]を使用した。これは英語の音声認識モデルで、Convolutional Encoder 部分と Transformer による Context Representation 部分で構成されている。音声波形を入力して、最終層出力は区切り文字などのトークンとアルファベットを含む 32 種の確率を出力する。Fig.1 は音声認識結果の例である。

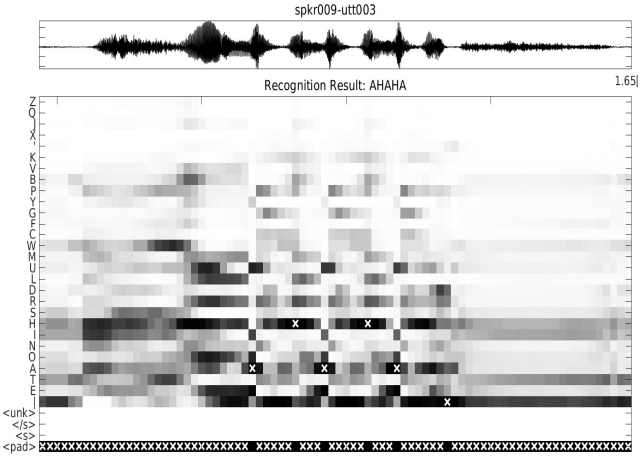

Fig.1 笑い声波形と音声認識の結果
バツ印が最高確率のトークンを示す

図のように各フレームに対し全トークンの確率を与えることで音声認識による事後確率表現を得る。

4　事後確率を入力とした笑い声合成

前章で例示した事後確率表現を入力とする音声合成モデルを構築した。モデルの構造は Fig.2 に示す。音声データのうち、テスト用に特定の話者の音声を除いた後、残りの 70%に

---

* Toward Laughter Expression and Speech Synthesis Using Speech Recognition, by SUTO, Noritake and KONDO, Kazuhiro (Yamagata University).

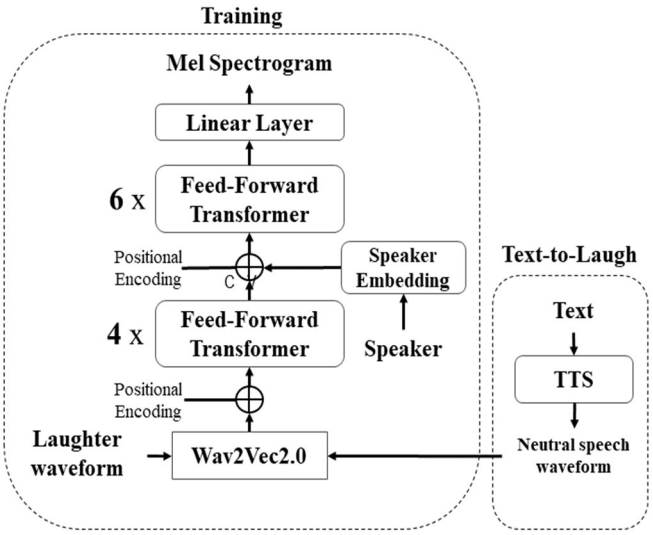

Fig.2 モデルの構造

あたる 7972 個を train データとしてランダムに選択し、Mel-Spectrogram は音声認識モデルに合わせてホップ長を 320 で算出した。

学習はバッチサイズ 8 で 500 エポック行われた。音声波形の生成には Xin ら[4]によって公開[2]されている日本語音声で学習された HiFi-GAN[6]を使用する。

5　客観評価

Mel-cepstral distortion(MCD)によって合成された笑い声の客観的な評価を行う。MCD は原音と合成音声のスペクトル特徴（低次ケプストラム係数）の距離をもとに算出され、値が小さいほど原音をよく再現していると言える。ここでは比較対象として HiFi-GAN、疑似音素トークン笑い声合成(PPT)、原音にホワイトノイズを加算した音声(SNR n dB)を用いる。ホワイトノイズを加算した音声の MCD は、必ずしも合成による音質の劣化を再現するものではないが、参考として比較する。

テストには学習に使用されていない音声で、コーパス全体の 10%である 1163 個を使用した。結果を Table 1 に示す。

Table 1 MCD による品質評価. PPT は算出ではなく引用である。

Model	MCD
SNR 10 dB	9.92
SNR 35 dB	2.26
HiFi-GAN	2.32
PPT [4]	(11.41)
Proposed	**9.77**

結果から、事後確率表現による合成笑い声は既存の手法に比べて品質の良い音声であるといえる。ただしニューラルボコーダーによる再合成に比べ原音との差は大きく、やや劣化した音声と同程度の大きさであることから、その再現性や音質を改めて評価する必要がある。

6　まとめと今後

本研究では音声認識結果を笑い声の表現として音声合成モデルを構築した。客観評価により比較した結果、既存手法と比べ優れた性能を示した。

今後、検討・調査すべきポイントは、

①　笑い声合成モデルの性能評価

②　多様な笑い声への対応

③　テキスト等から笑い声を生成する手法

の 3 つが考えられる。

①について、人間による聴取テストを実施する必要がある。ただし以前の研究[3]で、「自然性」と「笑い声らしさ」は異なるカテゴリで評価されていることが考察されている。合成された笑い声の評価では、その音声の利用が想定される状況で笑い声として違和感なく許容できるかを評価する必要があると考えられる。

②について、今回使用した音声コーパスはシンプルな笑い声が大量に含まれている。しかし笑い声/笑い方は多様で、特に対話中は発話と共起する speech-laugh が存在し、言語的・音響的特徴が典型的な笑い声と異なることが知られている。自然対話コーパスを用いるなどでデータを増やすことは可能だと考えられるが、品質と量で音声合成モデルの学習に耐えうるか疑問がある。

③について、音声合成ソフトウェアとして考えた場合、事後確率表現では制御が困難であるため、直感的に操作し易いインターフェースを考える必要がある。例えば、Fig.2 右部のように既存の TTS モデルと組み合わせる方法が考えられる。

この手法によって合成した笑い声のスペクトログラムを Fig.3 に示す。Neutral speech の TTS には VOICEVOX[3]を使用した。テキストは典型的に笑い声らしいと考えられる文字列を使用した。

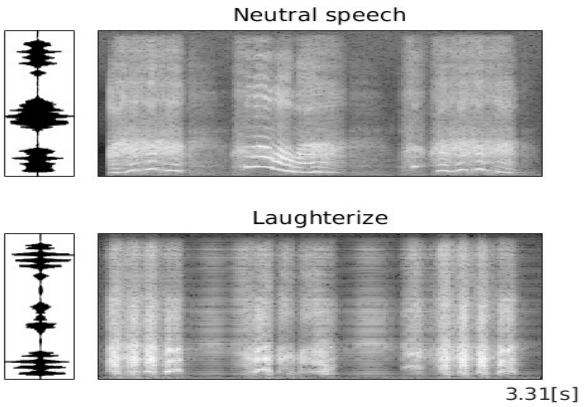

Fig.3 通常の TTS と笑い声合成のスペクトログラム

笑い声の形態らしく call 一つ一つが明確になっていることが確認できる。

この例のように、文字など直感的に操作可能なインターフェースによる、笑い声合成手法を検討する必要がある。

参考文献

[1] Mori *et al.*, Proc. INTERSPEECH 2019, pp.520-523, 2019.
[2] Matsumoto *et al.*, Proc. INTERSPEECH 2020, pp.3412-3425, 2020.
[3] Suto *et al.,* IEEE GCCE 2022, pp. 1-2, 2022.
[4] Xin *et al.*, Proc. INTERSPECH 2023, pp.117-21, 2023.
[5] Baevski *et. al.*, arXiv:2006.11477v3, 2020.
[6] Kong *et al.*, Proc. NeurIPS, vol. 33, pp. 17022–17033, 2020.

[1]https://huggingface.co/facebook/wav2vec2-base-960h
[2]https://github.com/Aria-K-Alethia/laughter-synthesis
[3]https://voicevox.hiroshiba.jp