WaveNet による笑い声の合成*

○森　大毅，永田　智洋 (宇都宮大)，有本 泰子 (帝京大)

1　はじめに

人と機械のコミュニケーションにおける笑い声の合成では、合成単位の環境や感情状態などの文脈を適切に反映させることが求められる。

これまで我々は、コミュニケーション中の笑い声の分析に基づき定義したコンテキストがHMM音声合成の枠組による合成笑い声の自然性を改善することを示してきた [1] が、未だ十分に自然な笑い声は得られていない。呼気の噴出による声帯振動と強い乱流雑音を既存の音声生成のモデルで表現するのが難しいこと、呼気音の減衰的な断続を単純なコンテキスト依存 HMM で表現するのが難しいことなど、音声学の知見に基づいたモデルベースでの笑い声合成は現状では課題が多い。

本研究では、DNN に基づく波形合成方式である WaveNet [2] に着目し、自然な笑い声合成への適用可能性を検討する。WaveNet は波形生成の過程にほとんど何らの仮定も置かないモデルであり、音声に比べて複雑な生成過程を持つ笑い声の波形を自然にモデル化できる可能性がある。

2　コーパス

笑い声 (laughter episode) は，1 回以上の呼気／吸気に対応する音響イベントからなり，これらが知覚的なひとかたまりを形成する。1 回の呼気に対応する笑い “句” (bout) は 1 個以上の笑い “音節” (call) からなる [3]。

本研究では、表情豊かな自然対話コーパスである OGVC [4] を笑い声合成のためのデータとして用いる。これまでの研究で、OGVC に含まれる笑い episode に対し、bout および有声／無声吸気音のアノテーションを行った [5]。これらには非常に微弱なものも含まれており、それらを全て笑い声として同列に扱うのは初期検討の段階では適当ではない。そこで、今回は第 1 著者が主観的に「主要な笑い声」の認定を行った。この認定は、顕著な有声口音 call からなる multi-call bout を含むことを基準として行った。今回は、男性話者 04_MSY の 127 episodes, 女性話者 06_FWA の 102 episodes が実験データとして選ばれた。

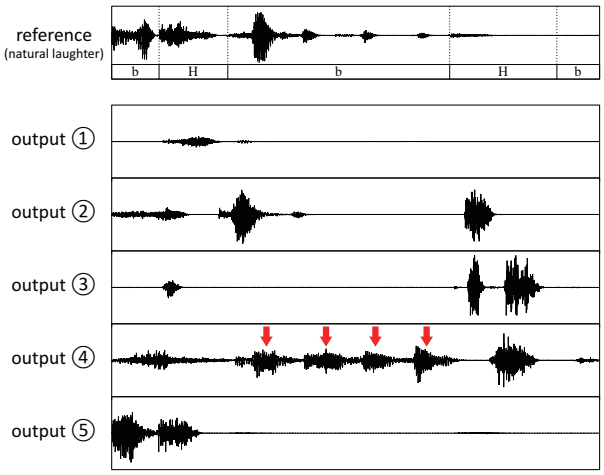

Fig. 1　Examples of generated laughter waveforms. (b: bout, H: voiced inhalation)

3　WaveNet の構成

WaveNet は過去のサンプル点を入力として現時点のサンプル点を予測するモデルであるが、音韻情報などを補助的な入力として条件付けすることもできる。今回は、笑い声の構造をごく緩く指定するため、当該位置が bout であるか吸気音であるかの情報だけで条件付けした。また、吸気音の場合、その有声／無声の別も与えた。

1 つの bout は、時には 1 秒以上続く長いイベントである。減衰的に Call が連続する典型的な bout のパターンを表現するためには、少なくとも 2 calls 以上の十分な長さの履歴を参照する必要があると考えられる。そこで、10 層 $(1, 2, \ldots, 512)$ の膨張畳み込みを 3 組積層し、受容野を 6139 サンプル (383 ms) と大きく取った。出力は 8 ビット μ-law (256 ユニット softmax) とした。

4　出力波形の例

4.1　緩い条件付け

図 1 に 04_MSY のデータから学習した WaveNet から生成した笑い声波形の例を示す。緩い条件付けのもとでは、出力波形は偶然に左

*Laughter synthesis with WaveNet.
by MORI, Hiroki, NAGATA, Tomohiro (Utsunomiya Univ.), and ARIMOTO, Yoshiko (Teikyo Univ.)

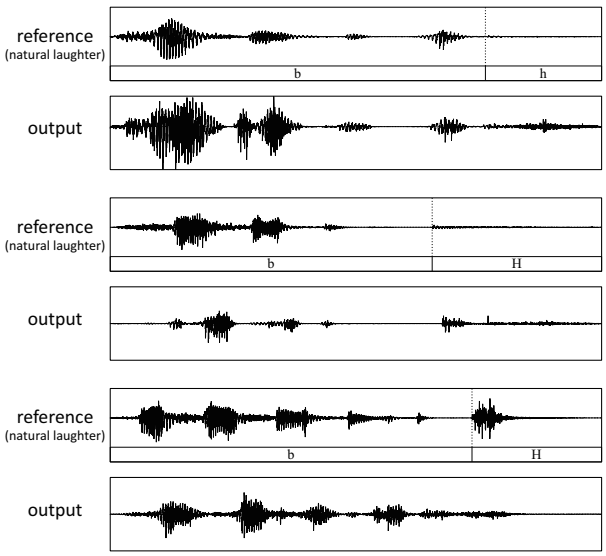

Fig. 2 Generated waveforms conditioned by power contours.

右される。この図では、同じ自然笑い声の bout および吸気音の継続時間を参照して 5 回合成を行っている。毎回、波形がかなり異なるだけでなく、ほぼ全体が無音になる場合も少なくない。

注目したいのは、④の波形である。今回使った学習データには bout および吸気音のラベルは付与されているが、call のラベルはない。にも関わらず、④では 4 つの call によって bout が表現されているように見える (矢印で示した部分)。これは、非常に長い受容野によって、WaveNet が複数の call からなる bout のパターンを獲得できていることを示唆する。

4.2 パワーによる条件付け

上で示したように、WaveNet は比較的自然な笑い声を合成できる潜在的能力を持っている。しかし、応用面からは偶然性は不便であり、自然な笑い声波形を安定して合成できることが望まれる。

ここでは補助入力を 1 次元増やし、パワー情報により合成笑い声の振幅を制御することを検討する。図 2 に、04_MSY のデータから学習した WaveNet の補助入力に参照音声のパワー軌道を与えて合成した波形の例を示す。参照波形に類似した笑い声が安定して合成できていることがわかる。

5 自然性評価実験

WaveNet による合成笑い声の自然性を評価するため主観評価実験を行った。従来法である HMM 合成による笑い声 [1] および原音声 (Natural) も同時に評価した。提案法と異なり、HMM 合成はコーパスに call レベルのコンテキストラベルを必要とする。このラベルは現在 04_MSY には付与されていないので、HMM 合成の評価は 06_FWA のデータのみを対象とした。評価対象は、04_MSY の 31 episodes × {WaveNet, Natural} と 06_FWA の 31 episodes × {HMM, WaveNet, Natural} の計 155 episodes である。

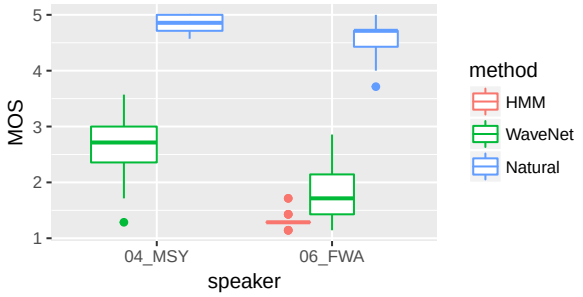

Fig. 3 Naturalness evaluation.

WaveNet 合成ではパワーによる条件付けを行った。合成に必要な継続時間長とパワーは、04_MSY の場合は合成対象笑い声のものをそのまま用い、06_FWA の場合は HMM から予測されたものを用いた。

評価者は大学生・大学院生 7 名であり、呈示された笑い声の肉声らしさを 5 段階で評価した。

結果を図 3 に示す。WaveNet 合成の MOS は 04_MSY では 2.64 と中程度の自然性であった。06_FWA では耳ざわりな雑音がしばしば出力され、MOS は 1.76 と低いが、HMM 合成の MOS 1.31 に比べ自然性は有意に高かった。

6 おわりに

WaveNet による笑い声合成の可能性を検討した。コーパスとして OGVC を用いた初期検討の結果、男性話者については比較的自然な笑い声を合成することができた。音韻環境による変形への対応や、笑い声が伝達するパラ言語情報の制御については、今後の課題とする。

参考文献

[1] Nagata and Mori, IEEE Trans. Affect. Comput., 2018. doi:10.1109/TAFFC.2018.2813381
[2] van den Oord et al., arXiv:1609.03499 [cs.SD]
[3] Bachorowski et al., J. Acoust. Soc. Am. **110**, 1581–1597, 2001.
[4] Arimoto et al., Acoust. Sci. Tech., **33**, 359–369, 2012.
[5] 森, 有本, 永田, 音講論 (秋), 217–218, 2017.