笑い声合成における call および吸気音の自動アノテーションの有効性*

☆上田 拡人, 森 大毅 (宇都宮大)

1 はじめに

音声対話システムで笑いを扱う研究は少ない。ロボットが共起笑いを適切に選択し表出することで会話が自然になることを示した研究 [1] などがあるが、これらは実際の人間の笑い声の録音を用いた研究がほとんどである。これは笑い声合成技術が未成熟であることのあらわれである。

笑い声合成が難しい要因の一つは、モデル学習に必要なラベルのついた笑い声のデータを増やすのが難しいことである。統計的音声合成の枠組に基づく笑い声合成 [2] では笑い声の構成要素である call や吸気音の時間情報付きラベルが必要であるが、アノテーションに大きな労力を要することが問題であった。

この問題に対処するため、大規模事前学習モデルを用いて call および吸気音（以下、phone と呼ぶ）レベルのアノテーションを自動作成する方法 [3] が提案されている。

本研究では、笑い声合成に必要な phone レベルのアノテーションを自動作成することの有効性を検証する。自動作成したラベルを用いて笑い声合成を行い、その自然性を人手によるアノテーションを用いたものと比較する。

2 Call および吸気音の認識

本研究の遂行のため、オンラインゲーム音声チャットコーパス（OGVC [4]）の 4 話者およびアクションゲーム音声コミュニケーションコーパス（AGSC [5]）の 7 話者の笑い声に対して人手で call および吸気音のアノテーションを行いデータを整備した。対象とした話者の ID およびデータ数を表 1 に示す。OGVC の評価セットには、先行研究で用いられた 04_MSY と 06_FWA の 2 話者を使用し、他の話者を訓練セットとした。AGSC の話者は、コーパスに収録された笑い声の数が多いことを基準に訓練セットに使用した。

構築した phone 認識モデルの精度を表 2 に示す。

Table 1 phone 認識モデルの訓練セットおよび評価セットとその話者の笑い episode 数

		訓練セット		評価セット	
	sex	speaker	#	speaker	#
AGSC	男性	G001R	188	G010R	136
	女性	G003R	189	G004R	159
		G004L	215		
		G008L	180		
		G008R	182		
OGVC	男性	04_MNN	164	04_MSY	256
	女性	06_FTY	354	06_FWA	226

Table 2 構築した phone 認識モデルの精度 (PBE: phone 境界誤差 [ms]、sub./del./ins.: 置換／脱落／挿入誤り率、auxerr.: 補助記号誤り率)

	PBE	sub.	del.	ins.	auxerr.
既知話者	23.7	0.30	0.12	0.14	0.12
未知話者	23.0	0.30	0.14	0.07	0.15

3 笑い声合成

3.1 笑い声合成モデル

笑い声合成モデルには 3 層の LSTM および線形層からなる統計的パラメトリック DNN 音声合成モデル [2] を用いた。話者情報は one-hot 表現を phone 埋め込みに結合することで表現している。モデル訓練および評価は、表 1 の評価セットに含まれる 4 話者を対象とした。

笑い声合成を行うモデルとして、以下の 4 モデルを作成した。

NoLabel4 phone ラベルを用いず、笑い声全体を直接モデル化。合成対象の 4 話者の笑い声で訓練。

AutoLabel4 合成対象の 4 話者の笑い声で訓練。自動作成したラベルを使用。

AutoLabel33 合成対象の 4 話者の笑い声に、AGSC および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（UUDB [6]）の

*Effectiveness of automatic transcription of calls and inhalation sounds for laughter synthesis. by UEDA, Hiroto, MORI, Hiroki (Utsunomiya Univ.)

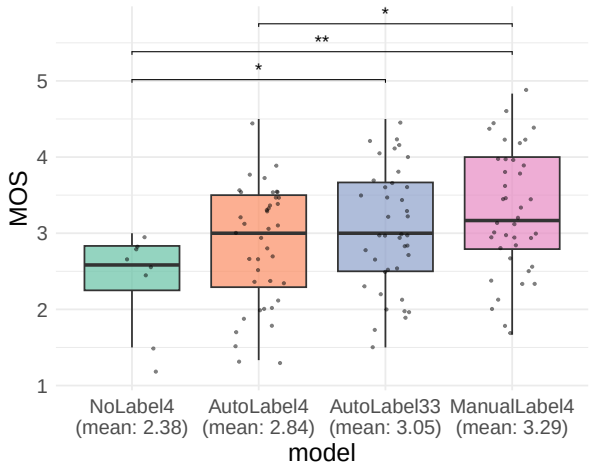

Fig. 1　MOS 評価の結果

29 話者の笑い声を追加したデータセットで訓練。自動作成したラベルを使用。

ManualLabel4　合成対象の 4 話者で訓練。人手で作成したラベルを使用。

AutoLabel4 モデル、AutoLabel33 モデル、ManualLabel4 モデルの 3 モデルへの入力には、phone 言語モデルからのサンプリング [2] を用いず、合成対象話者の笑い声から特殊な phone を含まない各 10 episodes を無作為抽出し、訓練した合成器に入力することで呈示刺激を作成した。

3.2　自然性評価実験

自然性評価実験への参加者は音声の研究室に所属する学生 6 人である。呈示刺激の数は、NoLabel4 モデルについては 4 話者× 2(同一刺激の繰り返し)、AutoLabel4 モデル、AutoLabel33 モデル、ManualLabel4 モデルについてはそれぞれ 4 話者× 10 episodes で、合計 128 episodes である。参加者は音声を聴き、その音声が笑い声として自然であるかを 5. 非常に良い、4. 良い、3. どちらでもない、2. 悪い、1. 非常に悪いの 5 段階で評価した。

3.3　結果

実験結果を図 1 に示す。AutoLabel4 モデルよりも ManualLabel4 モデルの方が自然性が高い ($p = 0.013$)。一方、AutoLabel4 モデルと AutoLabel33 モデルの差は有意ではなく、AutoLabel33 モデルと ManualLabel4 モデルの差は有意ではなかった。

3.4　考察

自動アノテーションにより、多くの笑い声データを笑い声合成モデルの訓練に使用することが可能となる。そこで多数話者笑い声合成モデルの有効性を検討した。しかし AutoLabel4 モデルと AutoLabel33 モデルとの間に有意な差は無く、自然な笑い声の合成に効果があるかは確認できなかった。

しかしその一方で、AutoLabel33 モデルと ManualLabel4 モデルとの間には有意な差は無く、自動アノテーションに基づく笑い声合成が、人手アノテーションに基づくものに迫る自然性を示したと言える。

人手によるラベルの作成は自動ラベルによるデータの作成と比較し多くの人的、時間的リソースが必要である。自動ラベルであれば笑い声合成モデルに多種多様な会話コーパスの笑い声データを追加することも可能である。これらのことから、笑い声合成に自動ラベルを用いるのは合成笑い声の自然性とラベリングのコストを考慮すると有効な選択肢であると言える。

4　おわりに

大規模事前学習モデルを利用した phone の書き起こしモデルで自動ラベルを作成し、これを笑い声合成に用いた時の自然性を評価した。その結果、自動アノテーションに基づく笑い声合成は人手によるものに匹敵する自然性を示した。

本研究では笑い声の自然性のみに焦点を当てたが、感情次元を入力として制御を行う際の有効性など、未検討の課題が残っている。

参考文献

[1] Inoue *et al.*, Front. Robot. AI, 9:933261, 2022.
[2] Mori and Kimura, Proc. Interspeech 2023, 3372–3376, 2023.
[3] H. Mori, Proc. APSIPA ASC 2025, 1134–1139, 2025.
[4] Arimoto *et al.*, Acoust. Sci. Tech., 33:6, 359–369, 2012.
[5] Mori and Kikuchi, Proc. Interspeech 2020, 3132–3135, 2020.
[6] Mori *et al.*, Speech Communication, 53:1, 36–50, 2011.