会話コーパスに含まれる笑い声を構成する call および吸気音の大規模事前学習モデルによる書き起こし *

○森　大毅 (宇都宮大)

1　はじめに

笑い声は多様な形態を持っており、その形態的特徴と会話場面、文脈、知覚される感情などとの関連が研究されている [1]。会話エージェントに人間と同様の多様な笑い声を表出する機能を持たせようとするならば、会話コーパスに含まれるさまざまな会話場面と文脈における笑い声のアノテーション [2] を実施し、それに基づいて笑い声合成のモデルを構築する必要がある。

しかし、会話コーパスにおいて音素に対応する笑い声の構成要素 (call および吸気音) のアノテーションが提供されていることは稀である。特に、統計的音声合成の枠組に基づく笑い声合成 [3] では構成要素の時間情報が必要だが、これを提供しているコーパスは皆無である。アノテーションは笑い声の構造に精通した研究者でなければ行えず、また極めて大きな労力を必要とするため、大規模な会話コーパスに対して実施することは現実的ではなく、これが笑い声合成研究のボトルネックになっている。

本論文では、会話における笑い声データの拡充を目的に、音声認識で広く用いられるようになった大規模事前学習モデルを利用した笑い声の自動アノテーションを試みる。

2　Call および吸気音の書き起こし

笑い声は呼気／吸気に対応する音響イベントからなり、1 回の呼気に対応する笑い “句” は 1 個以上の笑い “音節” (call) からなる。ここでは [3] にならい、1 つの call または 1 つの吸気音を合成単位とみなして “phone” と呼称する。

本論文では、笑い声のアノテーションを (1) phone 列の書き起こしと (2) 波形へのアラインメントの 2 段階で実施する。会話コーパス OGVC に含まれる 2 話者の笑い声 482 episodes については、call の子音・母音に加え、無声 (e.g. hu̥), 鼻音 (e.g. hũ), 子音伸長 (e.g. h:u) などの変異、無声／有声吸気音 (h, H)、および各 phone の時間情報が記述されたデータがあり、これを正解とする。

wav2vec 2.0 に基づき笑い声からの phone 列認識モデルを構築した。Wav2Vec2-XLS-R-1B を事前学習モデルとし、CTC (Connectionist Temporal Classification) をヘッドとして、笑い声の訓練セット (385 episodes) によりモデルを学習した。

OGVC の残り 97 episodes に加え、会話コーパス UUDB に含まれる 6 話者の笑い声 74 episodes について新たに正解ラベルを作成し、合計 171 episodes をテストセットとした。テストセットに対する phone 列認識モデルの精度は PER (phone エラー率) で 0.505 となった。訓練セットに含まれる既知話者 2 名に対する性能は PER 0.368 であり、笑い声の構成要素およびその認識性能には話者依存性があることが示唆される。

3　アラインメントおよび総合性能評価

モデルの出力には、phone に対応するクラスのほか、CTC で利用されるブランクという特殊トークンのクラスが含まれる。各フレームの確率最大のクラスを拾うことによって phone 列認識結果とアラインメントの両方が得られるが、ほとんどのフレームでブランクが確率最大となるため、phone の境界時刻を同定するのが困難であることが知られている [4]。本論文では、(A1) ブランク以外に認識された phone の開始時刻を直前の phone の終了時刻まで延長する方法、(A2) 認識された phone 列を Montreal Forced Aligner (MFA) [5] に与えて強制アラインメントを行う方法、の 2 手法で認識された phone の時間情報を推定する。(A2) ではアラインメント用の GMM-HMM を笑い声の訓練セットにより学習した。

笑い声の自動アノテーションを評価するためには、認識結果の各 phone が正解 phone とどのように対応するのかを同定する必要がある。しかし、音声認識の場合と異なり、正解と認識結果とのア

---

* Transcribing calls and inhalation sounds that constitute laughter in conversational corpora using a large-scale pretrained model.
by MORI, Hiroki (Utsunomiya Univ.)

ラインメントを phone の同一性に基づいて行うことはあまり意味をなさない。これは、笑い声の構成要素が言語音に比べて音響的なコントラストが乏しく、識別が困難[1]なためである。笑い声合成の立場からは、認識された phone の種類よりも各 phone の時間情報が正確であることの方が重要だと予想されることからも、phone 同一性に依存しないアラインメント方法、および性能評価方法が求められる。そこで、本論文では式 (1) で定義される脱落ペナルティ付き phone 境界誤差 (PBE)[5] を最小にする対応 $\pi \in \{\pi_1 \cdots \pi_I | \pi_i \in R, \forall i < j \ \ \pi_i \leq \pi_j\}$ を求める。

$$\mathrm{PBE_d} = \mathrm{PBE} + C_{\mathrm{del}}|R - \{\pi_i|\pi_i \in \pi\}| \quad (1)$$

$$\mathrm{PBE} = \sum_{k=1}^{K} \frac{1}{2}(|p_k^{\mathrm{ref}} - p_{b(k)}^{\mathrm{pred}}| + |q_k^{\mathrm{ref}} - q_{e(k)}^{\mathrm{pred}}|) \quad (2)$$

$$b(k) = \min_{\{i|\pi_i=k\}} i, \quad e(k) = \max_{\{i|\pi_i=k\}} i \quad (3)$$

ただし、R は正解 phone 列のインデックス集合、K は認識結果 phone 列長、p_k および q_k は k 番目の phone の開始時刻および終了時刻、C_{del} は phone 脱落のペナルティ重みである。

PBE、phone 誤り率 (PER)、カバー率 (正解 phone 区間に占める対応する認識結果 phone 区間との重なり率) の平均は、アラインメント方法 A1 ではそれぞれ 50.48 ms, 0.791, 0.371, アラインメント方法 A2 ではそれぞれ 50.48 ms, 0.793, 0.329 であった。以下では A1 の結果のみ示す。

既知話者の平均はそれぞれ 53.29 ms, 0.703, 0.450, 未知話者の平均はそれぞれ 47.38 ms, 0.915, 0.259 であり、未知話者では phone 誤り率と平均カバー率で大きく性能が下がることがわかる。

10 回以上出現した phone の種類別の評価結果を Table 1 に示す。例数が比較的多い hu 系 call の境界時刻推定および phone 認識精度は比較的高いが、全体として精度は低い。また、置換・脱落・挿入のいずれも多い。

自動アノテーションの例を Fig. 1 に示す。この例は比較的正確であるが、3 番目の call の継続長が非常に短く推定されてしまっている。人手によるアノテーションでは呼気音と吸気音の間に存在する短い無音区間が境界位置の手がかりとなるが、自動アノテーションではこの例のように境界が不可解な位置に推定されることが多い。ブランクを抑制するような CTC モデル学習 [4] の有効性検証が必要である。

[1] Phone 列認識の結果が正解 phone 列と一致部分を持たないことも珍しくない。

Table 1 笑い声の構成要素アノテーションの評価結果。PBE の単位は ms。PER は phone 誤り率、(S), (D), (I) はうち置換・脱落・挿入分。

	PBE	PER	(S)	(D)	(I)	cov.	N
hu	45.21	0.667	0.25	0.08	0.33	0.53	132
ha	52.52	0.892	0.43	0.16	0.30	0.26	83
hu̥	58.71	0.902	0.39	0.24	0.27	0.29	41
hũ	28.86	0.500	0.21	0.03	0.26	0.64	38
hḁ	49.07	1.048	0.76	0.14	0.14	0.08	21
a	25.70	1.000	0.73	0.13	0.13	0.12	15
e	13.05	1.067	0.80	0.20	0.07	0.00	15
he	47.11	1.214	0.86	0.00	0.36	0.10	14
ʔu̥	66.04	0.643	0.14	0.50	0.00	0.24	14
hːa	45.91	1.364	1.00	0.00	0.36	0.00	11
hːu	64.54	1.600	1.00	0.00	0.60	0.00	10
h	62.69	0.525	0.22	0.18	0.13	0.54	118
H	67.48	0.680	0.17	0.24	0.27	0.49	75

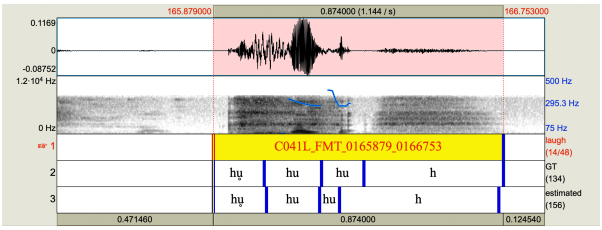

Fig. 1 笑い声の自動アノテーション例

4 おわりに

大規模事前学習モデルを利用した笑い声の構成要素の自動書き起こし、およびその笑い声波形とのアラインメント手法について述べた。現在は特に未知話者に対する精度が低い。訓練セットに含まれる話者数を拡大し、未知話者に対して頑健なモデルとする必要がある。また、提案手法により拡充した笑い声データセットに基づく笑い声合成による評価が必要である。

謝辞 研究協力者の藤塚亮佑、上田拡人の両氏に感謝します。本研究は JSPS 科研費 22K18477 の助成を受けている。

参考文献

[1] 森, 音講論 (秋), 1535–1538 (2023).
[2] 森, 有本, 永田, 音講論 (秋), 217–218 (2017).
[3] H. Mori and S. Kimura, Proc. Interspeech 2023, 3372–3376 (2023).
[4] Huang et al., arXiv:2406.02560 (2024).
[5] McAuliffe et al., Proc. Interspeech 2017, 498–502 (2017).