Time-Aligned Laughter Sound Event Recognition for Conversational Laughter Analysis and Synthesis

Hiroki Mori*
* Utsunomiya University, Japan
E-mail: hiroki-public@speech-lab.org

***Abstract*—Although time-aligned laughter data is essential for various studies including phonetics and conversational laughter synthesis, existing corpora lack annotations for laughter sound events, and manual annotation is highly costly. Accurately recognizing laughter components remains challenging with existing speech recognition models. In particular, the conventional CTC-based approach fails to effectively utilize temporal information, resulting in poor boundary estimation accuracy. To address this issue, we propose a laughter sound event recognition model that minimizes frame-wise loss instead of CTC loss while fully leveraging the boundary information provided in the training dataset. We also introduce a new evaluation metric to assess recognition accuracy. Experimental results demonstrate that the proposed method achieves phone boundary errors (PBE) of 16.8 ms for seen speakers, significantly outperforming the conventional CTC-based approach by reducing PBE to approximately one-fifth.**

I. Introduction

Laughter exhibits diverse forms. Studies on laughter have investigated its morphological features, conversational context, and emotional effects [1], [2], [3], [4], [5], [6], [7], [8], [9], [10]. For example, voiced laughter induces significantly more positive emotional responses in listeners than unvoiced laughter does [4]. In terms of its structure, laughter consists of calls, each corresponding to a single syllable, which combine to form bouts—events that correspond to single exhalations [5]. Notably, the proportion of unvoiced calls in a single-call bout is significantly larger than that in a multi-call bout [11]. If we aim to equip conversational agents with the ability to produce a variety of laughter similar to that of humans, it is necessary to annotate the laughter events that constitute laughter in various conversational contexts, and build a laughter generation model based on this data.

However, conversational corpora rarely provide annotations for the phonetic components of laughter, namely, calls and inhalations. Particularly in laughter synthesis based on statistical speech synthesis frameworks [11], [12], temporal information on these components (hereafter referred to as "phones") is required, yet no existing corpus provides such information. Since annotation requires expertise in laughter structure and involves an immense workload, it is impractical to annotate large-scale conversational corpora, making this a major bottleneck in laughter synthesis research. Thus, automating the annotation of laughter phones is highly desirable.

On the other hand, precisely recognizing laughter phones remains a challenge with current technology. Although speech data used for training speech recognition models has become more diverse, and it is common to include conversational speech containing laughter, in most cases, laughter is merely transcribed as {laugh} without any breakdown into individual phones. Therefore, even state-of-the-art Transformer-based speech recognition systems, such as Whisper, tend to either output "LoL" or internet slang such as "wwww" as a whole, or produce a sequence of laughter syllables that does not correspond to the actual number of calls or inhalations, as illustrated in Fig. 1.

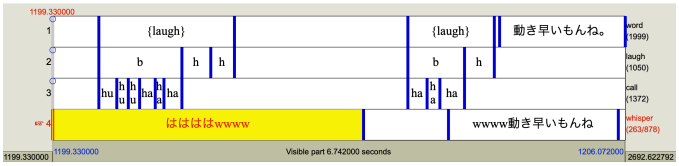

Fig. 1. Whisper does not transcribe laughter phones in a phone-by-phone manner. (Tier 1: Ground truth (episode level), Tier 2: Ground truth (bout/inhalation level), Tier 3: Ground truth (call level), Tier 4: Output of Whisper.)

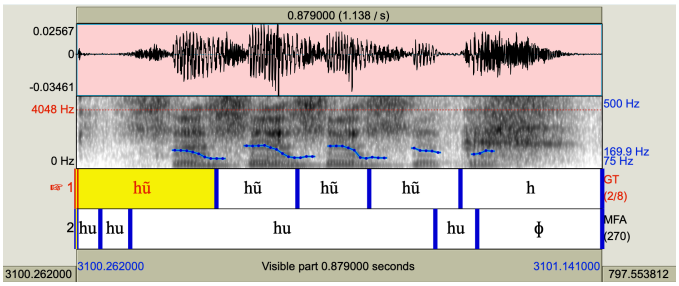

Fig. 2. Montreal Forced Aligner does not correctly segment laughter phones, even if a ground-truth phonetic transcription is provided.

Laughter syllables do not form a phonological system with contrastive units as in speech. The variation in vowel quality and consonant-like sounds is largely continuous and graded rather than categorical, reflecting physiological and expressive factors rather than linguistic contrasts [13], [14], complicating consistent phonetic labeling [15].

In addition to phonetic identity, annotators reportedly exhibit substantial disagreement in the temporal segmentation of spontaneous laughter [16]. From our experience, segmenting laughter into phones is also extremely difficult for automatic alignment software such as Montreal Forced Aligner [17], for which a failure example is shown in Fig. 2.

In this study, we present an investigation of laughter sound

979-8-3315-7206-8/25/$31.00

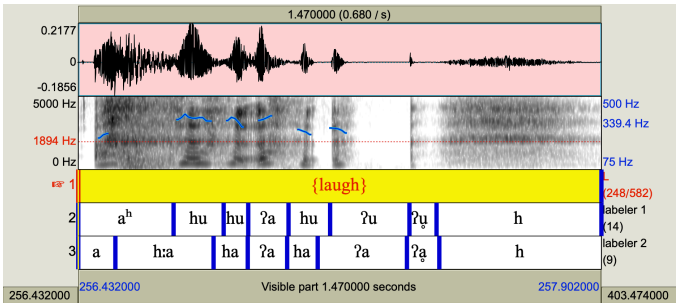

Fig. 3. Manual annotation of laughter sounds by two labelers.

event recognition aimed at expanding available laughter data for corpus linguistics and conversational laughter synthesis research. Taking the phonetic ambiguity of laughter sounds described above into account, we place the highest priority on alignment accuracy—specifically, the temporal accuracy of laughter phone boundaries. Still, while segmental contrast in laughter is indeed ambiguous, discourse and social context influence certain morphological features of laughter syllables, particularly in terms of vowel characteristics, the presence or absence of voicing and nasalization, and the distinction between exhalation and inhalation [4], [11], [18]. Therefore, we also investigate the accuracy of consonant-vowel identification and morphological feature identifiication.

In Section III, we explain why conventional speech recognition methods based on CTC are inadequate for laughter phone recognition and propose a novel framewise phone recognition method. Furthermore, in Section V, we discuss why phonemic identity-based evaluation metrics are unsuitable for laughter phone recognition and introduce a method for assessing accuracy by aligning with ground truth without reliance on segmental phonemes.

II. Laughter Sound Event Annotation

Laughter consists of acoustic events corresponding to exhalation and inhalation. A single laughter "phrase," which corresponds to one exhalation, comprises one or more laughter "syllables" (calls). The consonant and vowel of each call were transcribed using the romanization of Japanese syllables rather than a purely phonetic representation. Therefore, laughter vowels were classified as one of a, e, i, u, or o. Individual inhalation sounds were also identified as h (unvoiced) or H (voiced). This voiced/unvoiced distinction is crucial because of its relation to perceived emotion [19]. A phone refers to a single call or inhalation sound, which may serve as a synthesis unit.

The identification of these units is highly ambiguous because laughter sounds are generally unarticulated and non-contrastive [3]. Fig. 3 illustrates this ambiguity: The annotated vowels fluctuated between "a" and "u," depending on the labelers.

For 482 laughter episodes of two speakers from the Online Gaming Voice chat Corpus (OGVC) [20], we have reference data that includes consonants and vowels in calls, variations such as unvoiced (e.g., hụ), nasal (e.g., hũ), consonant prolongation (e.g., h:u), unvoiced/voiced inhalation sounds (h, H), and temporal information for each phone, which serve as the ground truth. In this study, the entire dataset was divided into a training set (377 episodes) and a test set (105 episodes). The test set was further expanded by a small number of additionally annotated laughter episodes picked from other conversational corpora, namely the Utsunomiya University Spoken Dialogue Database for Paralinguistic Information Studies (UUDB [21]; 74 episodes from six speakers), and the Action Gameplay Social Communication corpus (AGSC [22]; 39 episodes from one speaker).

III. Laughter Sound Event Recognition

Large-scale pre-trained models are known to be effective for speech recognition in low-resource languages [23]. Since laughter can be considered a kind of low-resource language, our laughter phone recognition models are trained based on the large-scale pre-trained model wav2vec 2.0 [24].

Although CTC (connectionist temporal classification) loss is widely used in training speech recognition models, it may not be effective for laughter phone recognition for several reasons. First, laughter phone sequences contain very little information, unlike speech recognition. Laughter consists mainly of a small set of calls (such as "ha" and "hu") and inhalation sounds, and the distinction between these is highly ambiguous, as mentioned earlier. Consequently, the contrast in loss becomes relatively small compared to speech recognition. Second, models trained with CTC loss provide poor temporal information about the recognized units. It is known that the output posterior distribution of CTC models tends to be temporally peaky [25], [26]. Fig. 4 shows an example of automatic annotation using **CTC** under the experimental conditions described in Section VI. Tier (1) in the figure presents the phone with the highest output probability at each frame using a model trained with CTC loss. From the figure, it is evident that the blanks "_" dominate the predicted sequence, leading to inaccurate estimation of the start and end times of phones.

In the current problem setting, we can use time-aligned ground-truth labels of phones for training. A disadvantage of CTC is its inability to effectively utilize this temporal information. Instead of minimizing loss at the episode level using CTC loss, we can minimize the cross-entropy loss directly at the frame level. Tier (2) in Fig. 4 shows the result obtained using this **framewise** model. While blank outputs are eliminated, multiple phones are merged into a single unit. This occurs because consecutive frames with the same output phone are merged into a single phone.

To improve the framewise approach, we propose a method in which the reference label for the final frame in each phone is replaced with a blank symbol "_". This allows boundaries to be marked even when the same phone appears consecutively. Tier (3) in Fig. 4 shows the result based on this method referred to as **framewise+**. It can be seen that the positions of each phone are accurately estimated. Although blank outputs remain in this method as in CTC, they are generally short (1–2 frames) and

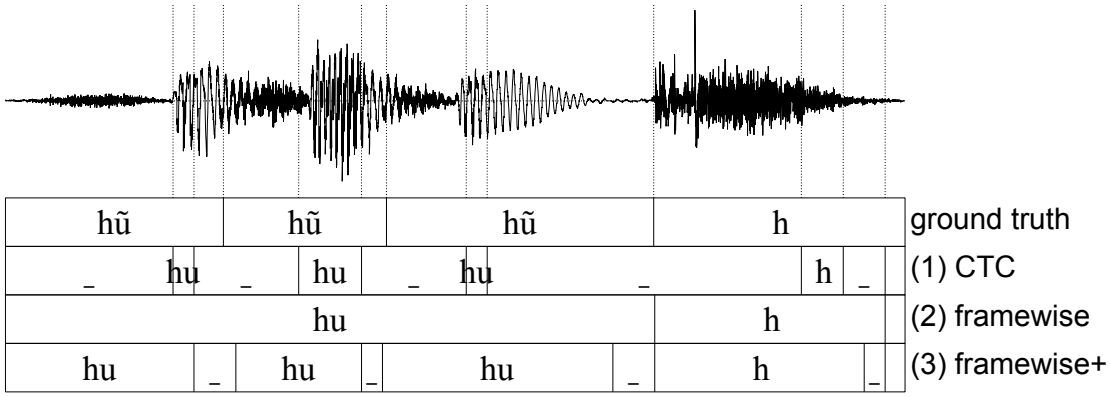

Fig. 4. Example of model outputs (without postprocessing).

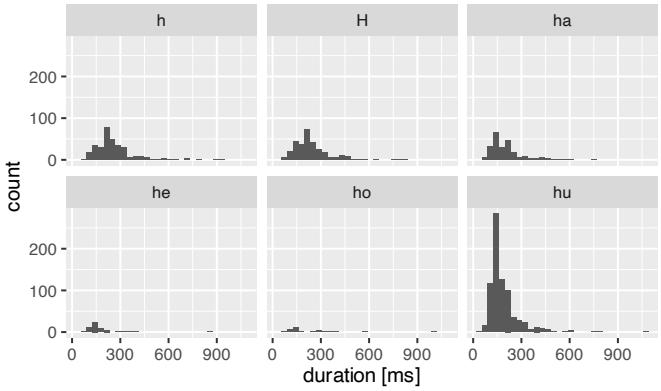

Fig. 5. Duration distribution for the six most frequent laughter phones.

do not introduce ambiguity in boundary timing as in the case of CTC.

IV. Postprocessing

As shown in Fig. 4, the framewise+ method can generally estimate phone boundaries accurately in most cases. However, it occasionally produces extremely short phone segments, as will be discussed in Section VI.

To mitigate this issue, we introduce a postprocessing step for framewise+ using a phone duration model. Fig. 5 shows the duration histograms for most frequent laughter phones. The durations of most phones are typically in the range of several hundred milliseconds, depending on the phone identity. For example, ingressive phones (h and H) tend to have longer durations.

Here, we introduce a phone duration model $P(d|w_k)$, which gives the probability that the k-th phone w_k has a duration of d frames. Our objective is to find the optimal alignment path π^* among all possible alignment paths $\pi \in \mathcal{V}^T$, given a feature sequence $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_1, \dots \mathbf{x}_T)$ for an episode, where $\mathcal{V}$ is the phone vocabulary.

Let $w_k \in \mathcal{V}$ denote the k-th phone, and let d_k be its duration, where $k \in [1..K]$ and $d_1 + \cdots d_K = T$. The alignment probability is then modeled as:

$$P(\pi = w_1^{d_1} w_2^{d_2} \cdots w_K^{d_K} | \mathbf{X}) = \prod_{k=1}^{K} P(d_k|w_k)^\gamma P(w_k|\mathbf{X}_k), \tag{1}$$

where $\mathbf{X}_k = (\mathbf{x}_{p_k}, \dots, \mathbf{x}_{q_k})$ is the feature sequence corresponding to the k-th phone, and $p_k = 1 + \sum_{i=1}^{k-1} d_i$, $q_k = d_k + \sum_{i=1}^{k-1} d_i$ represent the start and end times of the k-th phone, respectively. Here, γ is a weighting hyperparameter for the duration model; setting $\gamma = 0$ reduces the model to the original framewise+ method.

Given the posterior probabilities $y^t_{\pi_t}$ of each label $\pi_t \in \mathcal{V} \cup \{\_\}$ at time t, predicted by the model for input $\mathbf{X}$, $P(w_k|\mathbf{X}_k)$ is defined according to the framewise+ probability formulation from Section III:

$$P(w_k|\mathbf{X}_k) = \left(\prod_{t=p_k}^{q_k-1} y^t_{w_k} \right) y^{q_k}_{\_}, \tag{2}$$

that is, the posterior probability of the final frame is replaced by that of the blank symbol. The optimal alignment is then obtained as:

$$\max_{(w_1,w_2,\dots w_K)\in\mathcal{V}^*} \max_{(d_1,d_2,\dots,d_K)\in\mathbb{N}^K} \prod_{k=1}^{K} P(d_k|w_k)^\gamma P(w_k|\mathbf{X}_k). \tag{3}$$

As the duration model, we use a generalized linear model with a Gamma distribution, where the predictor variable is a one-hot vector representing the six most frequent phones plus a category for all others. This configuration was selected based on model selection using AIC (Akaike's Information Criterion). The weighting parameter for the duration model was set to $\gamma = 0.2$. Hereafter, the framewise+ method with the duration model is referred to as **framewise+d**.

V. Evaluation of Laughter Sound Event Recognition and Alignment Performance

To evaluate the automatic annotation of laughter, it is necessary to determine how each recognized laughter phone corresponds to the ground truth phone. Unlike speech recognition, however, aligning the ground truth with the recognition

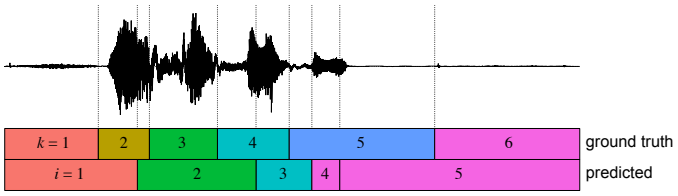

Fig. 6. Illustrated example of alignment calculation.

results based on phone identity is not very meaningful, because of the its lower phonetic contrast discussed in Section II.

From the perspective of laughter synthesis, the temporal accuracy of each recognized phone takes precedence over phone identity. Therefore, an alignment method and a performance evaluation method that do not rely on phone identity are required. In this paper, we determine the alignment $\pi \in \{\pi_1 \cdots \pi_I | \pi_i \in \{1, 2, \ldots, K\}, \forall i < j \ \pi_i \leq \pi_j\}$ that minimizes the phone boundary error (PBE) [17] with deletion penalty, which is defined in Equation (4).

$$\mathrm{PBE_d} = \mathrm{PBE} + C_{\mathrm{del}} |\{\{1, 2, \ldots, K\} - \{\pi_i\}|, \tag{4}$$

$$\mathrm{PBE} = \sum_{k=1}^{K} \frac{1}{2} (|p_k^{\mathrm{ref}} - p_{b(k)}^{\mathrm{pred}}| + |q_k^{\mathrm{ref}} - q_{e(k)}^{\mathrm{pred}}|), \tag{5}$$

$$b(k) = \min_{\{i|\pi_i=k\}} i, \quad e(k) = \max_{\{i|\pi_i=k\}} i, \tag{6}$$

where K and I are the length of reference (ground truth) and predicted phone sequence, respectively. p_k and q_k denote the start and end times of the k-th phone in the reference/predicted sequence. $b(k)$ and $e(k)$ represent the first and last indices of the predicted phone sequence corresponding to the k-th phone of the reference. C_{del} is a penalty weight for deletions. Fig. 6 illustrates an example of alignment. In this case, phones $k = 2$ and $k = 5$ do not have counterparts in the predicted sequence and are thus interpreted as deletions. On the other hand, phone $k = 6$ corresponds to multiple predicted phones $i = 4, 5$, resulting in an insertion.

VI. Experiment

The network structure is consistent across all training settings. The model accepts a laughter waveform as input and generates the posterior distribution of phone categories and the presence of auxiliary features for each output frame. The phone categories consist of 14 consonant-vowel pairs of laughter calls (e.g. ha), inhalation sounds (h, H), and the blank symbol "_". The auxiliary features include nasal, consonant prolongation, and unvoiced.

The wav2vec 2.0 model generates 1024-dimensional feature representations for each frame, which are then converted into posterior distributions by a classification head using a linear layer and softmax/logistic functions. As the pre-trained model, we used the multilingual speech recognition model XLSR-53 large. As a finalizing step, blank regions were filled with the preceding phones.

Table I presents the evaluation results for different training methods. Among the four training/postprocessing conditions,

TABLE I
Overall results of phone alignment and recognition. The unit of PBE is ms. Phone recognition errors are shown separately for substitution, deletion, and insertion. Error rates for auxiliary feature recognition are also displayed.

(a) seen speakers

		phone			aux. feat.
	PBE	subst.	del.	ins.	error
CTC	80.7	0.224	0.307	0.032	0.388
framewise	60.5	0.186	0.356	0.084	0.046
framewise+	20.7	0.256	0.092	0.186	0.073
framewise+d	16.8	0.245	0.094	0.078	0.075

(b) unseen speakers

		phone			aux. feat.
	PBE	subst.	del.	ins.	error
CTC	135.7	0.266	0.490	0.053	0.297
framewise	77.1	0.365	0.334	0.260	0.221
framewise+	36.5	0.494	0.121	0.461	0.285
framewise+d	34.9	0.463	0.129	0.154	0.277

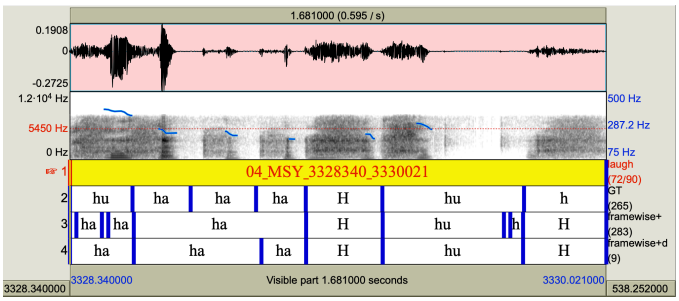

Fig. 7. Example of phone alignment and recognition for a seen speaker 04_MSY.

framewise+d achieved the best performance; a phone boundary error (PBE) of 16.8 ms for seen speakers (i.e., the OGVC test set) is satisfactory, considering that the mean phone duration is 227.1 ms (SD = 140.0 ms). With framewise+, PBE was reduced to approximately one-fourth of that obrained by CTC-based alignment. Introducing the duration-based postprocessing improved the precision further.

The most apparent benefit of framewise+d is the reduction of insertion errors. An example of phone recognition and alignment using framewise+ and framewise+d for a seen speaker is shown in Fig. 7. In this example, the first and sixth calls were erroneously split into short segments when using framewise+. These errors were successfully corrected with framewise+d.

The alignment performance deteriorated for unseen speakers; however, a PBE of 34.9 ms may still be acceptable for certain applications.

Table II presents the evaluation results with the framewise+d model for different phone types that appeared at least five times. For seen speakers. the accuracy for frequently occurring phones such as hu, ha, h and H, is relatively high. The recognition accuracy for infrequent phones tends to be lower. For unseen speakers, on the other hand, the accuracy is not satisfactory except for inhalation sounds.

Table III presents the evaluation results for each speaker

TABLE II
PHONE ALIGNMENT AND RECOGNITION ACCURACY FOR DIFFERENT PHONE TYPES. THE UNIT OF PBE IS MS.

(a) seen speakers

		phone			aux. feat.	
	PBE	subst.	del.	ins.	error	*N*
hu	16.78	0.17	0.09	0.10	0.05	135
ha	25.54	0.32	0.15	0.03	0.11	65
he	18.94	0.90	0.10	0.10	0.20	10
ʔa	22.54	0.86	0.14	0.00	0.14	7
a	29.88	0.60	0.40	0.00	0.00	5
u	49.07	0.60	0.40	0.00	0.20	5
h	12.41	0.26	0.04	0.08	0.12	77
H	11.53	0.12	0.07	0.08	0.02	59

(b) unseen speakers

		phone			aux. feat.	
	PBE	subst.	del.	ins.	error	*N*
hu	38.41	0.55	0.12	0.19	0.32	136
ha	52.08	0.69	0.12	0.14	0.46	118
a	29.98	0.62	0.38	0.00	0.04	26
ʔu	35.62	0.68	0.24	0.16	0.32	25
e	39.96	0.68	0.32	0.16	0.00	19
he	49.97	0.62	0.23	0.31	0.31	13
ho	24.09	0.75	0.12	0.00	0.50	8
u	18.64	0.71	0.29	0.14	0.14	7
hi	72.84	0.83	0.00	0.17	0.83	6
ʔa	32.78	0.83	0.17	0.17	0.33	6
h	12.50	0.30	0.01	0.13	0.10	67
H	13.08	0.19	0.02	0.21	0.05	42

TABLE III
PHONE ALIGNMENT AND RECOGNITION ACCURACY FOR EACH SPEAKER. THE UNIT OF PBE IS MS.

		phone			aux. feat.
	PBE	subst.	del.	ins.	error
04_MSY	17.3	0.225	0.101	0.041	0.073
06_FWA	16.1	0.307	0.085	0.131	0.078
FTS	36.3	0.433	0.167	0.000	0.267
FTY	20.9	0.583	0.083	0.083	0.333
FMT	40.4	0.482	0.173	0.100	0.191
FYH	28.3	0.487	0.154	0.154	0.128
MKK	32.8	0.553	0.128	0.106	0.277
MKO	37.4	0.475	0.300	0.075	0.275
G004L	35.0	0.652	0.066	0.242	0.348

using the framewise+d model. The top two rows correspond to seen speakers, while the remaining rows represent unseen speakers. The table shows that accuracy varies across speakers. This suggests that laughing style is highly speaker-dependent. We have repeatedly observed that some speakers prefer specific inventories, such like prolonged consonants (e.g. hːa), glottal sounds (e.g. ʔu̯), and voiced inhalations (H). In short, not all laughs are alike among speakers. Deletion errors for speaker MKO is particularly frequent, likely because the consonant portions of MKO's laughter calls frequently becomes voiced, making the boundaries ambiguous (Fig. 8). The result also suggests that merely achieving accurate recognition of phone sequences does not always lead to precise alignment, as seen in the result for speaker FTS, for whom phone recognition is the most accurate among unseen speakers. This observation may justify the advantage of the proposed time-aligned training of laughter phones over training with CTC.

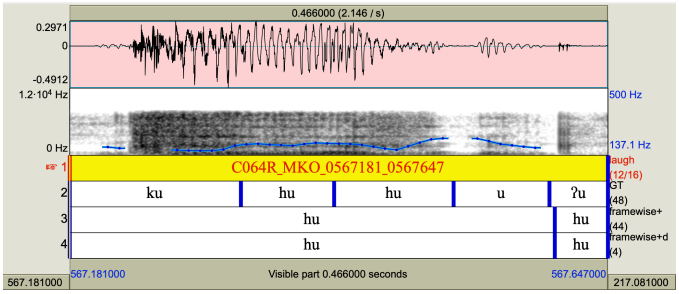

Fig. 8. Another example of phone recognition and alignment for an unseen speaker MKO.

VII. CONCLUSIONS

We have presented an automatic recognition and alignment method for laughter sound events using a large-scale pre-trained model. With the enhanced framewise recognition model and duration-based postprocessing, we achieved a phone boundary error (PBE) of 16.8 ms for seen speakers and 34.9 ms for unseen speakers, reducing the error to approximately one-fifth and one-fourth of that obtained with CTC, respectively.

In the current study, we were only able to use laughter data from two speakers for model training. Experimental results indicate that generalizing the laughing styles of just two speakers to others is challenging. To enable large-scale laughter annotation across diverse conversational corpora, it is necessary to improve recognition accuracy for unseen speakers to a level suitable for semi-supervised learning. Moving forward, we aim to expand the dataset through manual labeling and develop laughter sound recognition that accommodates the diverse laughing styles observed across different speakers.

ACKNOWLEDGMENT

The author thanks Mr. Ryosuke Fujitsuka and Mr. Hiroto Ueda for their assistance. This work was supported by JSPS KAKENHI Grant Numbers 22K12107 and 22K18477.

REFERENCES

[1] G. Jefferson, "A technique for inviting laughter and its subsequent acceptance declination," in *Everyday Language: Studies in Ethnomethodology*, G. Psathas, Ed., New York: Irvington, 1979, pp. 79–96.

[2] R. R. Provine and Y. L. Yong, "Laughter: A stereotyped human vocalization," *Ethology*, vol. 89, no. 2, pp. 115–124, 1991.

[3] J.-A. Bachorowski, M. J. Smoski, and M. J. Owren, "The acoustic features of human laughter," *J. Acoust. Soc. Am.*, vol. 110, no. 3, pp. 1581–1597, Sep. 2001.

[4] J.-A. Bachorowski and M. J. Owren, "Not all laughs are alike: Voiced but not unvoiced laughter readily elicits positive affect," *Psychological Science*, vol. 12, no. 3, pp. 252–257, 2001.

[5] J. Trouvain, "Segmenting phonetic units in laughter," in *Proc. ICPhS '03*, 2003, pp. 2793–2796.

[6] M. Gervais and D. S. Wilson, "The evolution and functions of laughter and humor: A synthetic approach," *Q. Rev. Biol.*, vol. 80, no. 4, pp. 395–430, 2005.
[7] J. Urbain et al., "The AVLaughterCycle database," in *Proc. LREC 2010*, 2010, pp. 2996–3001.
[8] P. Glenn and E. Holt, *Studies of Laughter in Interaction*. London: Bloomsbury, 2013.
[9] H. Tanaka and N. Campbell, "Classification of social laughter in natural conversational speech," *Comput. Speech Lang.*, vol. 28, no. 1, pp. 314–325, 2014.
[10] S. Scott, N. Lavan, S. Chen, and C. McGettigan, "The social life of laughter," *Trends Cogn. Sci.*, vol. 18, no. 12, pp. 618–620, 2014.
[11] H. Mori and S. Kimura, "A generative framework for conversational laughter: Its 'language model' and laughter sound synthesis," in *Proc. Interspeech 2023*, 2023.
[12] T. Nagata, H. Mori, and T. Nose, "Dimensional paralinguistic information control based on multiple-regression HSMM for spontaneous dialogue speech synthesis with robust parameter estimation," *Speech Communication*, vol. 88, pp. 137–148, 2017.
[13] D. P. Szameitat, C. J. Darwin, A. J. Szameitat, D. Wildgruber, and K. Alter, "Formant characteristics of human laughter," *Journal of Voice*, vol. 25, no. 1, pp. 32–37, 2011.
[14] J. Trouvain and N. Campbell, "Phonetics as a laughing matter," in *Proc. 19th International Congress of Phonetic Sciences (ICPhS)*, 2019, pp. 62–66.
[15] J. Urbain and T. Dutoit, "A phonetic analysis of natural laughter, for use in automatic laughter processing systems," in *Affective Computing and Intelligent Interaction*, S. D'Mello, A. Graesser, B. Schuller, and J.-C. Martin, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 397–406.
[16] K. P. Truong, J. Trouvain, and M.-P. Jansen, "Towards an annotation scheme for complex laughter in speech corpora," in *Proc. Interspeech 2019*, 2019.
[17] McAuliffe, "Montreal Forced Aligner: Trainable text-speech alignment using Kaldi," in *Proc. Interspeech 2017*, 2017, pp. 498–502.
[18] M. E. Kret, D. Venneker, B. Evans, I. Samara, and D. Sauter, "The ontogeny of human laughter," *Biol. Lett.*, vol. 17, no. 9, pp. 1–25, 2021.
[19] Y. Arimoto, R. Imanishi, and H. Mori, "Laughter components estimation using emotional information towards natural and expressive laughter synthesis," *Transactions of Information Processing Society of Japan*, vol. 63, no. 4, pp. 1159–1169, 2022.
[20] Y. Arimoto, H. Kawatsu, S. Ohno, and H. Iida, "Naturalistic emotional speech collection paradigm with online game and its psychological and acoustical assessment," *Acoustical Science and Technology*, vol. 33, no. 6, pp. 359–369, 2012.
[21] H. Mori, T. Satake, M. Nakamura, and H. Kasuya, "Constructing a spoken dialogue corpus for studying paralinguistic information in expressive conversation and analyzing its statistical/acoustic characteristics," *Speech Communication*, vol. 53, no. 1, pp. 36–50, 2011.
[22] H. Mori and Y. Kikuchi, "Gaming corpus for studying social screams," in *Proc. Interspeech 2020*, 2020, pp. 3132–3135.
[23] K. Soky, S. Li, C. Chu, and T. Kawahara, "Domain and language adaptation using heterogeneous datasets for wav2vec2.0-based speech recognition of low-resource language," in *Proc. ICASSP 2023*, 2023, pp. 1–5.
[24] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. NIPS 2020*, 2020.
[25] A. Zeyer, R. Schlüter, and H. Ney, "Why does CTC result in peaky behavior?" arXiv:2105.14849, 2021.
[26] R. Huang et al., "Less peaky and more accurate CTC forced alignment by label priors," in *Proc. ICASSP 2024*, 2024, pp. 11 831–11 835.