INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

# A Generative Framework for Conversational Laughter: Its 'Language Model' and Laughter Sound Synthesis

*Hiroki Mori*[1], *Shunya Kimura*[1]

[1]Utsunomiya University, Japan
hiroki@speech-lab.org

## Abstract

As the phonetic and acoustic manifestations of laughter in conversation are highly diverse, laughter synthesis should be capable of accommodating such diversity while maintaining high controllability. This paper proposes a generative model of laughter in conversation that can produce a wide variety of laughter by utilizing the emotion dimension as a conversational context. The model comprises two parts: the laughter "phones generator," which generates various, but realistic, combinations of laughter components for a given speaker ID and emotional state, and the laughter "sound synthesizer," which receives the laughter phone sequence and produces acoustic features that reflect the speaker's individuality and emotional state. The results of a listening experiment indicated that conditioning both the phones generator and the sound synthesizer on emotion dimensions resulted in the most effective control of the perceived emotion in synthesized laughter.

**Index Terms**: laughter synthesis, generative model, language model of laughter, emotional conditioning

## 1. Introduction

Laughing is a basic and essential emotional behavior for humans. Nevertheless, almost all of the conversational agents that interact with humans do not laugh. Part of the reason for this is attributed to the fact that we ourselves do not well understand why, when, and how we laugh. A recent study on conversational robots by a Kyoto-U team aimed at the positive effect of the robot's laughter on empathy [1]. By focusing on "shared laughter," they cleared the *when* problem. For the *how* problem, however, they avoided laughter synthesis and randomly picked one from the pools of "mirthful" or "social" laughs.

The current laughter synthesis study focuses on *how* conversational agents should laugh. Laughter synthesis is an emerging technology and is gaining importance as the human-agent interaction becomes more advanced and popular in our daily lives [2, 3, 4, 5, 6, 7]. A large part of previous work has employed a similar framework to text-to-speech systems. An open problem here is how to construct input sequences for the synthesizer. As the phonetic structure and its functional aspects of laughter in conversation has not been fully understood, most previous work simply used exemplars of natural laughter for input, which limits flexibility. Recent research in non-speech vocalization synthesis [8] also points to the need for some kind of "language model".

This paper proposes a generative model of laughter in conversation that can produce a wide variety of laughter. A highlighted feature is the "language model" of laughter, which serves as a laughter sequence generator. This model generates various but realistic combinations of laughter components for a given speaker ID and emotional state. The generated sequence is then fed into the laughter "sound synthesizer," which produces acoustic features that reflect the speaker's individuality and emotional state.

In this paper, we will be using specific laughter-related terminology, following to [9]. A "laughter episode" will refer to a series of acoustic events that correspond to exhalation or inhalation. A "bout" will refer to an event that corresponds to an exhalation and is composed of one or more laughter calls. A "call" will be used to describe an individual unit of laughter, analogous to a syllable. Therefore, a typical bout "hahaha" is a 3-call bout.

## 2. Morphology of laughter sounds

A typical method for collecting laughter data has been induction by funny movies [10, 11]. Provine criticized past studies for focusing solely on audience-oriented, passive laughter [12]. He argued that laughter is social and that speakers actually laugh more than listeners. Since we are interested in laughter in agent-human interaction, we need to collect laughter that occurs naturally in conversation. In this study, we used the Online Gaming Voice chat Corpus (OGVC) [13], a speech corpus containing spontaneous dialogue during massively multiplayer online role-playing games (MMORPGs), which has a larger number of laughs than other Japanese conversational corpora used in emotion studies.

Bout- and call-level annotation was performed for the top three speakers with the highest frequency of laughter in OGVC. An example of the annotation is shown in Fig. 1. The annotation has a hierarchical structure: Bouts and inhalation sounds that comprise each laughter episode were annotated, as well as calls that comprise each bout.

The consonant and vowel of each call were transcribed as a romanization of Japanese syllable, rather than in a phonetic way. Therefore, laughter vowels are classified into one of a, e, i, u, or o. The proportions of vowels are shown in Fig. 2. The most common vowel was /u/, followed by /a/. However, these are not contrastive, and most laughter sounds are realized around the mid central vowel [ə].

In addition to consonants and vowels, phonetic variants, including unvoiced (e.g. hu̥), nasal (e.g. hũ), and consonant prolongation (e.g. h:u), were also transcribed. Among them, the voicelessness of laughter sound has received much attention due to its functional importance. For example, voiced laughter induces significantly more positive emotional responses in listeners than unvoiced laughter does [14].

The proportion of bout length (number of calls) is shown in Fig. 3. It is worth noting that the proportion of single-call bouts is surprisingly large. The proportion of unvoiced calls in single-

10.21437/Interspeech.2023-2453

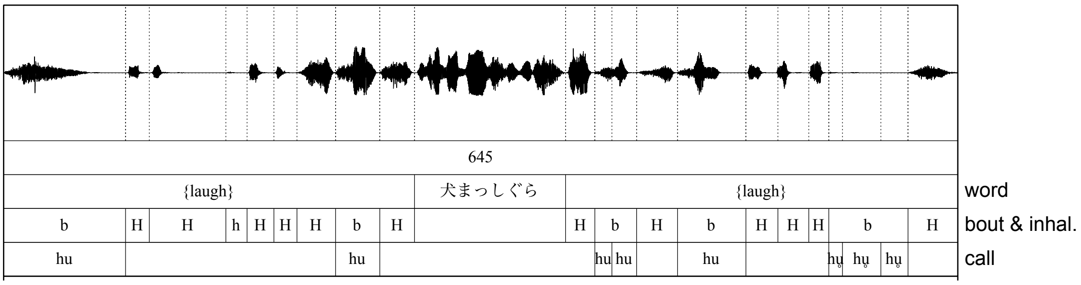


Figure 1: *Bout- and call-level annotation of laughter.*

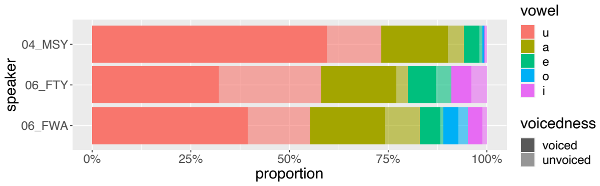


Figure 2: *Vowel proportions of calls. Each darker color stands for voiced, and lighter color for unvoiced.*

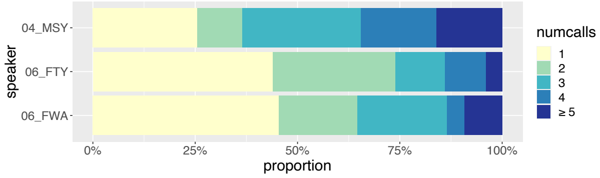


Figure 3: *Proportions of the number of calls per bout.*

call bouts (55.4 %) is significantly larger than that of multi-call bouts (18.7 %). This suggests that single-call bouts tend to be accompanied by negative emotions [14].

Individual inhalation sounds were identified as h (unvoiced) or H (voiced), and annotated at the same tier as bouts. Inhalation sounds often accompany vocal fold vibration (voiced), some of which constitute a main part of a laughter sound. This voiced/unvoiced distinction is crucial because of its relation to perceived emotion. Arimoto et al. [7] showed that laughs containing voiced inhalation sounds tend to be perceived as more pleasant and aroused. Voiced inhalation sounds are also important in characterizing the individuality of laughing speakers. For the top seven OGVC speakers with the highest frequency of laughter, the proportion of episodes with mid-laugh voiced inhalations is less than 1 % for two speakers, around 10 % for three speakers, and 21 % and 27 % for the remaining two speakers. This implies that there are speakers who almost exclusively use egressive laughter, as well as those who frequently produce ingressive laughter.

## 3. Emotion perception from laughter

The morphological variation of laughter depends on its discourse and social context. However, it is difficult to encode such contexts in a comprehensive and adequate way. As a first-order approximation, this study attempts to use the speaker's emotion perceived from laughter as an explanatory variable for modeling laughter forms [7].

This requires an evaluation of the perceived emotion for the laughs in the corpus. For this purpose, emotion categories such as "big six" emotions [15] seem virtually useless. In this study, we annotated the emotion perceived from laughter with two emotion dimensions, pleasantness and arousal. Dimensional descriptions of emotions have a long history and are well established in psychology. A number of studies have stated that two or three dimensions are sufficient to account for a good portion of emotional variation. Among all, the pleasantness (also known as valence) and arousal (also known as activation) dimensions have been regarded as fundamental [16].

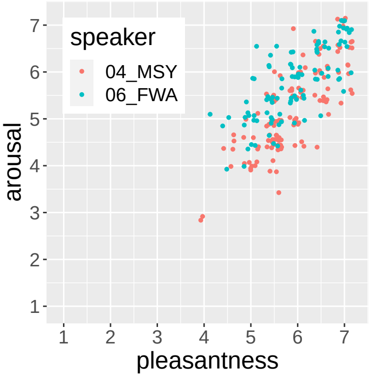


Figure 4: *Distribution of the ground-truth pleasantness and arousal dimensions evaluated for laughter sounds. Points are jittered to avoid overplotting of laughters with identical values.*

Prior to the emotion annotation, the first author checked the laughter sounds of a male speaker 04_MSY and a female speaker 06_FWA, then filtered out subtle or less audible ones, which yielded 125 and 100 laughter episodes for the two speakers as our laughter dataset.

The two authors individually annotated the perceived pleasantness (1: extremely unpleasant, 7: extremely pleasant) and arousal (1: extremely sleepy, 7: extremely aroused). The ground-truth values were obtained by averaging them. Figure 4 shows the distribution of the emotion dimensions for the two speakers. Most laughter sounds were evaluated as more pleasant and more aroused than neutral (4). Mean pleasantness and arousal were 5.86 and 5.19 for the male speaker 04_MSY, and 5.95 and 5.78 for the female speaker 06_FWA.

## 4. Phones generator: The "language model" of laughter

Contrary to the notion that laughter sounds have a homogenious structure such as "hahaha," "hehehe," or "huhuhu," there are so many variations that a closed lexicon of laughter cannot

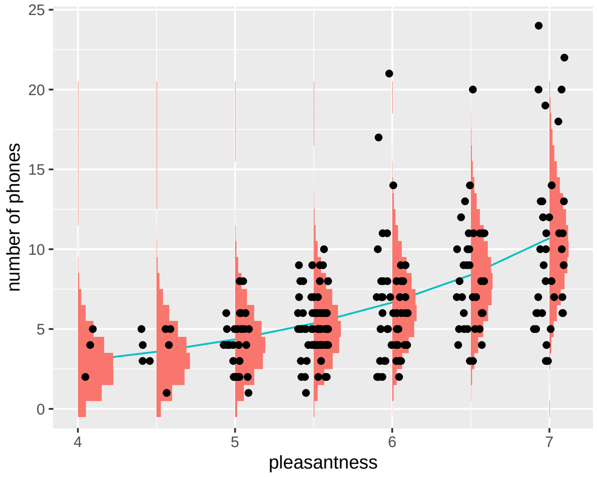


Figure 5: *Laugh length distribution (points) and its probabilistic model with Poisson regression. Points are horizontally jittered to avoid overplotting.*

be defined. At the same time, we barely hear laughter sounds such like "hahohaho," which implies that there are some constraints that prescribe possible combinations of laughter calls. Provine [12] suggested biological constraints against producing such mixed-call laughs, but he also pointed out that one can easily switch call types in mid-laugh, as in "hahahoho." His observation implied the existence of some laughter *grammar*, but he did not discuss a computational model of laughter calls that could be applied to laughter synthesis.

A desired laughter language model should not only regulate such possible combinations (as opposed to the random arrangement [7]), but also account for morphological preferences related to discourse and social context. As described in Sect. 2, the length of laughter is related to its emotion. Therefore, we modeled the length first, then the components. Hereafter, we regard either a call or a single inhalation as a component and refer to each component as a "phone." For example, the phone sequence corresponding to the second laughter episode in Fig. 1 is "H hu hu H hu H H H hu̥ hu̥ hu̥ H."

Figure 5 shows the distribution of the laugh length (number of phones) versus pleasantness by black points. As these could be modeled by a Poisson regression, the fitted mean parameter $\lambda$ (green line) and probability mass function (red bars) are overlaid (here the arousal value was set equal to the pleasantness value for simplicity). A generalized linear model with Poisson distribution was obtained through variable selection using AIC (Akaike Information Criterion) [17]. The fitted laugh length model was as below:

$$\log(\lambda_i) = b + 0.527 x_i^{\text{ple}} + 0.750 x_i^{\text{ple}} x_i^{\text{aro}}, \tag{1}$$

$$y_i \sim \text{Pois}(\lambda_i), \tag{2}$$

where $y_i$ is the length of $i$-th laughter, $x_i^{\text{ple}}$ and $x_i^{\text{aro}}$ are the pleasantness and arousal dimensions, whose range is linearly transformed from $[1, 7]$ to $[-1, 1]$, and $b$ is the speaker specific baseline (1.433 for 04_MSY, 0.936 for 06_FWA).

In the generation phase, the decision to stop generating is determined dynamically and randomly. Here we define $P_{\text{end}}(n)$ as the probability that the $n$-th generated phone is the last one:

$$P_{\text{end}}(n) = \frac{f(n; \lambda)}{1 - F(n-1; \lambda)}, \tag{3}$$

| Speaker: 04_MSY | Speaker: 06_FWA |
|---|---|
| Pleasantness: 4 Arousal: 4 | Pleasantness: 4 Arousal: 4 |
| hu hu hu h | hu |
| ha ha ha h | hu |
| hu hu hu hu h | hu |
| hũ | hu |
| hu hu hu | hu |
| hu hu | u hu hu h |
| hu hu hu h | hu |
| hu hũ hũ̥ h | hu |
| hu | hu |
| hũ hũ hu | hu |
| Pleasantness: 5 Arousal: 5 | Pleasantness: 5 Arousal: 5 |
| ha ha ha ha ha h | ha ha |
| hu hu hu hu h | hu h |
| hu hu hu hu | ha |
| hu hu hu hu h | hu hi |
| hu hu | hu |
| hu hu | hu h |
| ha ha ha ha ha h | h |
| hu hu hu hu h | h |
| ha ha ha | ho ho |
| hi h hu hu hu | hu h |
| Pleasantness: 6 Arousal: 6 | Pleasantness: 6 Arousal: 6 |
| hu hu hu ha H | H hu hu |
| hu hu hu ha H | H hu̥ ha ha ʔu̥ |
| hu hu hu ha H | u hu hu hu h |
| hu hu hu ha H | a ha ha h |
| ha ha H | H hu hu |
| hu hu ha ha H | u hu hu hu h |
| hu hu hu ha | u hu hu hu h |
| hu hu hu ha H | ha H hu H |
| a ha | H |
| hu hu ha ha H | hi H |
| Pleasantness: 7 Arousal: 7 | Pleasantness: 7 Arousal: 7 |
| hu ha ha hḁ hḁ hḁ hḁ hḁ hḁ H hḁ h | a ha ha H |
| hu ha ha hḁ hḁ hḁ hḁ hḁ hḁ H hḁ h | ʔa ʔa H |
| h a ha ha ha ha H H H hu̥ H | u hu hu |
| a ha ha ha hḁ h h h h h h h | hi H H H hu h |
| he ʔu̥ ʔu̥ h hu ha ha H ha ha hu hu̥ ʔu̥ h | a ha ha H |
| a ha ha ha hḁ h h h h h h h h h h H H H | a ha ha H |
| a ha ha ha hḁ h h h h h h h | he he he ʔe H H H he |
| hu ha ha hḁ hḁ hḁ hḁ hḁ hḁ H hḁ h | e ha ha |
| hu ha ha hḁ hḁ hḁ hḁ hḁ hḁ H hḁ h | H ha h hu H |
| a ha ha ha hḁ h h h h h h h h h h | H hu H h hu H |

Figure 6: *An excerpt from generated laughter phones (10 draws per condition). See the multimedia file for the complete list.*

where $f(k; \lambda)$ and $F(k; \lambda)$ are the probability mass function and cumulative distribution function of Pois($\lambda$), respectively. For each generated phone, an "end-of-laughter" is drawn according to $P_{\text{end}}(n)$. This ensures that the length distribution of generated laughs follows the Poisson distribution, whose mean is determined by Eq. (1).

Thirty-two different phones appeared in the laughter dataset described in Sect. 3. By replacing phones that appeared only once (e.g. hːa, hi̥, na) with similar ones, we obtained a phone list comprising 22 different calls and inhalations. In the modeling, phone sequences that constitute each laughter episode were converted into a sequence of 64-dimensional embedding vectors.

Similar to neural language models [18], the call sequence of laughter was modeled with a recurrent neural network. We used an architecture with an LSTM layer with 128 hidden dimensions, a linear layer, and a softmax layer. The dimensionality of the input was 64 (phone embedding) + 1 ($P_{\text{end}}(n)$) + 1 (speaker) + 2 (emotion dimensions) = 68.

Generated phones resulting from 10 draws for several combinations of emotion dimensions are shown in Fig. 6. Note that these are random draws without any cherry-picking, so many duplicates exist in the lists. From the figure, it is apparent that emotion and speaker individuality are reflected not only in the length of laughter but also in the pattern of laughter phones.

For example, more pleasant and aroused laughter contains more /a/'s and voiced inhalations.

## 5. Laughter sound synthesizer

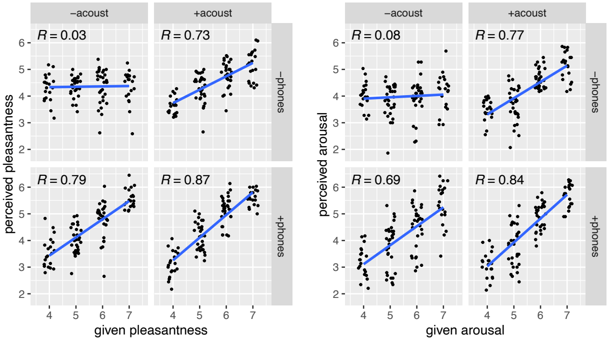
The current waveform synthesizer is basically a vocoder-based parametric speech synthesis [19], which can model human vocalization better than end-to-end models for limited data sizes, such as the one used in our case. The input feature set for duration modeling consisted of the identity of current consonant-vowel (19), 2 phonetic variations (voicedness, nasality) and their left and right context (×3), phone position (1), laughter length (1), and 2 emotion dimensions (67 in total). For acoustic modeling, phone duration and 3 numerical features for coarse-coded frame position in the current phone [20] were added to the input, and 59th-order Mel-cepstrum, $\log f_o$, aperiodicity, their $\Delta$, $\Delta\Delta$, and the voicedness were inferred as the output. The network was composed of a three-layer stacked bidirectional LSTM with 128 hidden dimensions and a linear layer. For the subsequent experiment, the model was trained with the 04_MSY dataset whose waveform was downsampled to 16 kHz.

## 6. Experiment

To investigate emotion controllability in the proposed laughter synthesis, we conducted an ablation study on both the phones generator and the laughter sound synthesizer. Hereafter, we denote the absence or presence of emotion inputs to the phones generator as −/+phones, and similarly, the absence or presence of emotion inputs to the laughter sound synthesizer as −/+acoust. The emotion inputs were masked at the training and inference stages in the −phones and −acoust conditions. For each of 10 pleasantness and arousal combinations (4, 4), (4, 5), (5, 4), (5, 5), (5, 6), (6, 5), (6, 6), (6, 7), (7, 6), and (7, 7), twenty sequences were generated using the phones generator. Then, the corresponding laughter waveform was synthesized from the acoustic features generated for each sequences using WORLD [21]. The generated phones and synthesized waveforms are provided as the multimedia files for this paper.

For each condition, the first 10 phone sequences were used in the listening test (see Fig. 6). The number of stimuli was 10 (target emotion dimensions) $\times$ 10 (phone sequences) $\times$ 2 (−/+phones) $\times$ 2 (−/+acoust) plus two reference real laughter sounds for subject screening $\times$ 4 (repetitions) $=$ 408. Thirty-one undergraduate and graduate students who were not involved in speech research participated in the listening test. First, they watched a video that described the objectives of the experiment and an introduction to the theory of emotion dimensions. The subjects then used a web interface to listen to the stimulus sounds in a random order and evaluated perceived pleasantness and arousal on a 7-point scale, as in Sect. 3. From the results of the screening test, two subjects were found not to meet our criteria (distinguishing between obviously pleasant/unpleasant laughter and responding consistently to identical stimuli), so their responses were excluded from later analysis.

The perceived pleasantness and arousal for the 400 synthesized laughter sounds were averaged over the subjects. Figure 7 shows the distribution of perceived pleasantness and arousal. For both dimensions, the +phones+acoust model showed the best controllability, as the correlation coefficient is as high as 0.87 (pleasantness) and 0.84 (arousal). This means that the emotion input to the phones generator and the emotion input to the laughter sound synthesizer are individually effective, but the emotion input to the both modules is even more effective. A statistical test for the difference between two paired correlations revealed that the correlation coefficient for the +phones+acoust model is significantly higher than that for the +phones−acoust model for both dimensions ($p < 0.01$).



(a) Pleasantness

(b) Arousal

Figure 7: *Relationship between target and perceived emotion from synthesized laughter for (a) pleasantness, and (b) arousal. Points are horizontally jittered to avoid overplotting.*

Best linear models to predict responses from the target dimension were obtained through variable selection using AIC:

$$\begin{aligned}\hat{y}^{\text{ple}} =& 0.112 - 0.301\delta_{\text{phones}} - 0.199\delta_{\text{acoust}} + 0.141\delta_{\text{phones}}\delta_{\text{acoust}} \\ &+ (0.669\delta_{\text{phones}} + 0.489\delta_{\text{acoust}} - 0.321\delta_{\text{phones}}\delta_{\text{acoust}})x^{\text{ple}},\end{aligned} \tag{4}$$

$$\begin{aligned}\hat{y}^{\text{aro}} =& -0.265\delta_{\text{phones}} - 0.198\delta_{\text{acoust}} + 0.171\delta_{\text{phones}}\delta_{\text{acoust}} \\ &+ (0.661\delta_{\text{phones}} + 0.561\delta_{\text{acoust}} - 0.375\delta_{\text{phones}}\delta_{\text{acoust}})x^{\text{aro}},\end{aligned} \tag{5}$$

where $\delta_{\text{phones}}$ and $\delta_{\text{acoust}}$ are the dummy (0/1) variables corresponding to the −/+phones and −/+acoust conditions. The coefficients of $x^{\text{ple}}$ and $x^{\text{aro}}$ in Eqs. (4) and (5) clearly demonstrate the synergistic effect gained by controlling both the phones generator and the sound synthesizer.

## 7. Conclusions

In this paper, we proposed a generative model for laughter in conversation, which allows for the production of a wide variety of laughter that can be controlled by emotion dimensions. Our results indicate that conditioning both the phones generator and the laughter sound synthesizer on emotion dimensions is most effective in controlling perceived pleasantness ($R = 0.87$) and arousal ($R = 0.84$).

One limitation of the current study is the lack of scalability, as call-level annotation for new datasets could become a bottleneck. Although state-of-the-art speech recognition systems such as Whisper can transcribe laughter calls to some extent, they cannot distinguish the phonetic variants necessary for laughter synthesis. One potential solution is to fine-tune the model using richly annotated laughter data such as the one built in this study.

## 8. Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers 22K12107 and 22K18477.

# 9. References

[1] K. Inoue, D. Lala, and T. Kawahara, “Can a robot laugh with you?: Shared laughter generation for empathetic spoken dialogue,” *Frontiers in Robotics and AI*, vol. 9, no. 933261, pp. 1–11, 2022.

[2] J. Trouvain and M. Schröder, “How (not) to add laughter to synthetic speech laughter in human interactions,” in *Proc. Workshop on Affective Dialogue Systems*, 2004, pp. 229–232.

[3] J. Urbain, H. Çakmak, A. Charlier, M. Denti, T. Dutoit, and S. Dupont, “Arousal-driven synthesis of laughter,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 2, pp. 273–284, 2014.

[4] H. Mori, T. Nagata, and Y. Arimoto, “Conversational and social laughter synthesis with WaveNet,” in *Proc. Interspeech 2019*, 2019, pp. 520–523.

[5] T. Nagata and H. Mori, “Defining laughter context for laughter synthesis with spontaneous speech corpus,” *IEEE Transactions on Affective Computing*, vol. 11, no. 3, pp. 553–559, 2020.

[6] N. Tits, K. El Haddad, and T. Dutoit, “Laughter synthesis: Combining seq2seq modeling with transfer learning,” in *Proc. Interspeech 2020*, 2020, pp. 3401–3405.

[7] 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.

[8] C.-C. Hsu, “Synthesizing personalized non-speech vocalization from discrete speech representations,” in *Proc. ICML Expressive Vocalizations Workshop & Competition 2022*, 2022.

[9] J. Trouvain, “Segmenting phonetic units in laughter,” in *Proc. ICPhS ’03*, 2003, pp. 2793–2796.

[10] 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.

[11] J. Urbain, E. Bevacqua, T. Dutoit, A. Moinet, R. Niewiadomski, C. Pelachaud, B. Picart, J. Tilmanne, and J. Wagner, “The AVLaughterCycle database,” in *Proc. LREC 2010*, 2010, pp. 2996–3001.

[12] R. R. Provine, *Laughter: A Scientific Investigation*. New York: Viking, 2000.

[13] 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.

[14] 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.

[15] P. Ekman, “Basic emotions,” in *Handbook of Cognition and Emotion*, T. Dalgleish and M. J. Power, Eds. Chichester, UK: John Wiley & Sons, Ltd, 1999, ch. 3.

[16] J. A. Russell, “How shall an emotion be called?” in *Circumplex Models of Personality and Emotions*, R. Plutchik and H. R. Conte, Eds. Washington, DC: American Psychological Association, 1997, pp. 205–220.

[17] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning with Applications in R*. New York: Springer, 2013, ch. 6.

[18] S. Merity, N. S. Keskar, and R. Socher, “Regularizing and optimizing LSTM language models,” in *Proc. International Conference on Learning Representations*, 2018.

[19] H. Zen, A. Senior, and M. Schuster, “Statistical parametric speech synthesis using deep neural networks,” in *Proc. ICASSP 2013*, 2013, pp. 7962–7966.

[20] H. Zen and H. Sak, “Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis,” in *Proc. ICASSP 2015*, 2015, pp. 4470–4474.

[21] M. Morise, F. Yokomori, and K. Ozawa, “WORLD: A vocoder-based high-quality speech synthesis system for real-time applications,” *IEICE Transactions on Information and Systems*, vol. E99-D, no. 7, pp. 1877–1884, 2016.