



# Assessment of the synthetic quality and controllability of laughing onset in speech-laugh synthesis

Ryo Setoguchi<sup>1</sup>, Yoshiko Arimoto<sup>1</sup>

<sup>1</sup>Chiba Institute of Technology, Japan

setoguchi@mac-lab.org, ar@mac-lab.org

## Abstract

This study is the first challenge of building a synthetic speech-laugh model via a deep learning technique. To maintain the phonetic intelligibility of synthesized speech-laugh, the model was trained with nonlaughing read speech material for both phones of speech-laugh (SL) and of speech (SP). To control laughing onset in SL, the model was also trained using SL material only for the phones of SL instances. The listening tests revealed that the naturalness score for synthesized female SL was as high as that for human SL and that the laughter-likeness score for synthesized SL was higher than that for synthesized SP in almost all conditions. The dictation test revealed that the training for phonetic intelligibility in SL synthesis was highly effective for synthesized SL. However, the difference between segmented SL onset and correct onset was greater for synthesized SL with phonetic intelligibility training than for that without training.

**Index Terms:** speech-laugh synthesis, paralinguistic information, laughter onset controllability, naturality, intelligibility

## 1. Introduction

Verbal communication and nonverbal cues, such as laughter, are essential components of human interaction [1, 2]. Advances in Text-to-Speech (TTS) technology, driven by deep learning, have led to synthesized verbal speech that is nearly indistinguishable from human vocalization [3, 4]. To enhance the expressiveness of human communication using synthesized speech, research has also explored laughter synthesis [5, 6, 7, 8, 9, 10, 11, 12], primarily focusing on pure laughter that has no linguistic information [13]. Although laughter synthesis has also significantly improved with TTS advancements, existing laughter synthesis models cannot generate speech-laugh that has two characteristics of laughter and speech because it occurs while we are speaking. Since speech-laugh is acoustically and phonetically different from pure laughter [14, 15], it is crucial to develop a synthetic model specialized for speech-laugh and to establish methods for assessing its synthetic quality.

Speech-laugh is a speech phenomenon accompanied by the strong exhalations of laughter [16], and the phonetic form involved in speech-laugh may be interfered with by the production of laughter. Therefore, successful speech-laugh synthesis should prevent linguistic information from being obscured by laughter. Additionally, laughing information should be perceivable from synthesized speech-laugh. Unless these two requirements are fulfilled, speech-laugh synthesis cannot be achieved.

Speech-laugh can occur at the beginning of speech, but it can also occur in the middle of utterances [17]. Hence, nonlaughing speech and speech-laugh can be sequentially produced during a single utterance. According to a phonological study examining which phones are more likely to occur as initial phones

in speech-laugh [17], speech-laugh was more likely to occur with phones articulated via vocal tract narrowing, such as the vowel /i/ and the alveolar consonants /s/ and /t/. On the basis of the results of [17], it is necessary to control laughter onset on a phone basis and to synthesize speech-laugh starting with a phone that is likely to occur initially. A study comparing the acoustic features of the initial phones of speech-laugh and the phones in speech without laughter [18] revealed that the first formant frequencies extracted from the vowels /a/ and /o/ in speech-laugh are significantly greater than those extracted from the vowels /a/ and /o/ as phones of speech. Since the acoustic features differ between the initial phones of speech-laugh and the phones in speech, the position of laughter onset should be added to an input sentence in TTS-based speech-laugh synthesis to clearly distinguish speech-laugh from speech.

The goal of this study is to build a model for speech-laugh synthesis via the text-to-speech model and to assess the controllability of laughing onset in speech-laugh synthesis. To preserve the phonetic intelligibility of synthesized speech-laugh, pretraining is introduced to train a model with phones of speech as phones of speech-laugh. After pretraining, the model is trained with phones of speech-laugh to represent the laughter-likeness in the synthesized speech-laugh and to control laughing onset. The largest contribution of our study is that our synthetic speech-laugh model provides a foundation for future studies that control laughter onset. A few studies have addressed speech-laugh synthesis by developing the model or changing acoustic features extracted from speech [19, 20, 21], however, those studies did not validate and assess the controllability of laughter onset in speech-laugh. A synthetic model of speech-laugh is needed to control where speech-laugh should start within a single utterance. Establishing the methods of assessing the laughter onset of speech-laugh will contribute to more humanlike laughing speech generation.

## 2. Speech-laugh synthesis

### 2.1. Model training

Variational Inference with adversarial learning for end-to-end Text-to-Speech (VITS) [4] is selected to synthesize speech-laugh in this study since the naturality of the synthesized speech synthesized by VITS (4.43) is close to that of natural speech (4.46) [4]. Figure 1 shows a diagram of the training process. The phone symbols in speech-laugh and those in speech were prepared separately. Speech-laugh synthesis involves two types of training: pretraining for phone intelligibility and fine-tuning for laughter expressiveness. In pretraining, a model is trained using the same nonlaughing speech for both phone symbols of speech-laugh and speech to preserve phonetic intelligibility. To train the model on laughter-likeness characteristics, fine-tuning

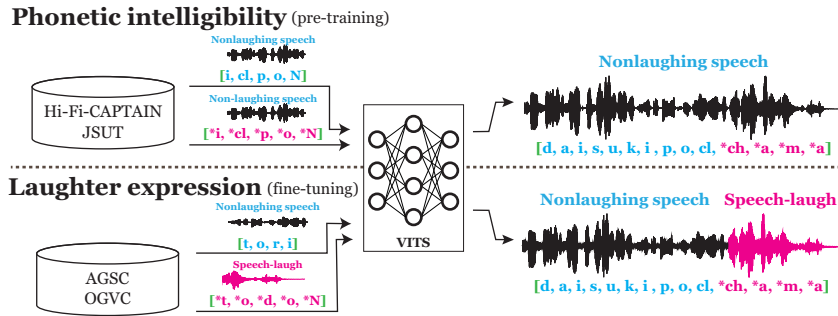


Figure 1: Diagram of the training process for phonetic intelligibility and laughter expression in speech-laugh. (The blue letters indicate phones for speech, and the pink letters indicate phones for speech-laugh.)

is performed to add laughter information to synthesized speech-laugh. In fine-tuning, the model was trained using speech-laugh for the phone symbols of speech-laugh and using nonlaughing speech for the phone symbols of speech. The batch size was set to 2 for pretraining and 1 for fine-tuning because the training wasn't in progress in fine-tuning, and the model was trained for up to 1M steps on an NVIDIA GeForce RTX 3070.

## 2.2. Speech material

Two Japanese read speech materials were adopted for training phones for both speech-laugh and speech in pretraining. One was High-fidelity and high-capacity conversational speech synthesis corpus developed by NICT (Hi-Fi-CAPTAIN) [22], and the other was the Japanese speech corpus of Saruwatari-lab., University of Tokyo (JSUT) [23]. The male speaker's speech (5,000 sentences, approximately 6 hours) from Hi-Fi-CAPTAIN and the female speaker's speech (5,000 sentences, approximately 7 hours) from basic5000 in JSUT were used for training phones for both speech-laugh and speech. To train laughter expression for speech-laugh, 2 Japanese spontaneous speech materials were adopted for fine-tuning. One was the action game spoken communication corpus (AGSC) [24], and the other was the online gaming voice chat corpus with emotional labels (OGVC) [25]. Since spontaneous speech material, including much expressive speech, is important for computational approaches to ensure naturalistic emotion [26], these corpora were adopted for our study because they consist of game-playing dialogs and contain affect bursts [27], which are short emotional nonverbal expressions, such as laughter. This study used one male speaker's speech (G011.L) from the AGSC, which included the most speech-laugh (85 speech-laugh segments and 1,353 utterances in a one-hour recording), and one female speaker's speech (06.FWA) from the OGVC, which enabled stable laughter synthesis [8] (114 speech-laugh segments and 646 utterances in a one-hour recording), for fine-tuning.

Speech-laugh annotation and segmentation for AGSC and OGVC were performed by [17, 18]. The annotated speech-laugh segments were used for fine-tuning. The phones used in the training process were extracted using pyopenjtalk in the Python library. Hi-Fi-CAPTAIN was sampled at 48 kHz and digitized to 24 bits, whereas the other three speech materials were sampled at 48 kHz and digitized to 16 bits. The sampling frequency and digitized bit rate were adjusted to 22 kHz and 16 bits, respectively, when training. Hi-Fi-CAPTAIN and AGSC were used for training a male speaker model, whereas JSUT and OGVC were used for a female speaker model. A total of 4,960 sentences from Hi-Fi-CAPTAIN or JSUT were used for training, and 20 sentences were used for each round of vali-

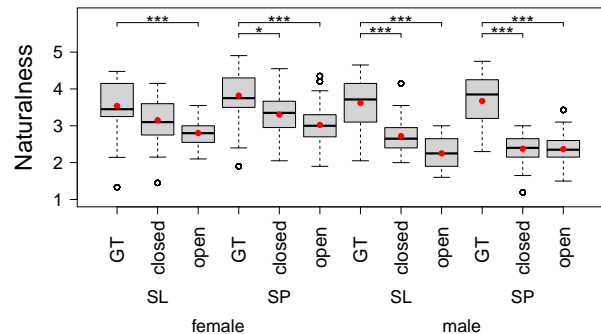


Figure 2: Results of the naturalness evaluation

dation and testing. Ten utterances of speech-laugh and speech from AGSC or OGVC were used for validation, and the other utterances were used for training.

## 3. Experiment

To assess our approach for speech-laugh synthesis, 4 experiments were conducted: naturalness and laughter-likeness listening tests for synthetic quality, a dictation tests for phonetic intelligibility, and speech-laugh segmentation tests for controllability of laughter onset.

The naturalness and laughter-likeness of the synthesized speech-laugh were evaluated by listening tests. Three conditions were set for the test; the closed condition (closed, the texts are identical to the training), the open condition (open, the texts aren't identical) and the human vocalization (GT). Naturalness was evaluated on a 5-point Likert scale of 1 (Very Bad), 2 (Poor), 3 (Fair), 4 (Good), or 5 (Excellent) on the basis of how humanlike each presented speech sample was. Laughter-likeness was measured on a 5-point Likert scale of 1 (not laughter), 2 (not like laughter), 3 (fairly like laughter), 4 (very like laughter), or 5 (laughter) on the basis of how much like laughter the speech sample was. The datasets used in the listening test consisted of a combination of synthesis conditions (GT, closed and open), speech type (speech-laugh (SL) or speech (SP)), and speaker (male or female), for a total of 12 combinations. A total of 360 speeches, 30 for each combination, were used in the naturalness and laughter-likeness listening tests.

The listening test was conducted via the crowdsourcing service Lancers (Lancers, Inc., Tokyo). After providing informed consent and signing an agreement for the test, the evaluators evaluated each speech sample. A total of 120 workers participated in the experiment, and each speech sample was evaluated 20 times. The evaluated ratings were averaged across participants for each speech sample as the evaluation score. With ref-

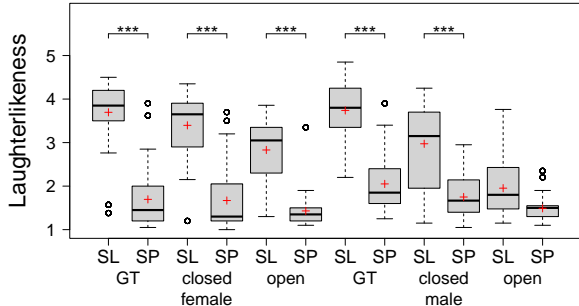


Figure 3: Results of the laughter-likeness evaluation

erence to [28], one evaluator whose average response time was less than 3 seconds and nine evaluators whose ratios of identical evaluations across speech samples were greater than 0.6 were excluded from the average calculation. With the calculated mean values of naturalness and laughter-likeness, three-way ANOVA tests were performed on the factors of synthesis condition, speech type, and speaker. The Tukey HSD test was performed as a post hoc test.

The dictation test for phonetic intelligibility was performed with the synthesized speech-laugh and speech generated with the female speaker model. Ten synthesized speech-laugh or speech instances were generated from the same text across the 4 synthesis conditions, which were the 4 combinations of modeling types, i.e., whether pretraining was performed or not (w or w/o), and vocal expression types (speech-laugh or speech). The text used in the synthesis consisted of shortened sentences originating from the test set in JSUT. Four test sets were prepared. Each test set was composed of 10 synthesized instances of speech-laugh or speech selected from the 4 synthesis conditions. There was no duplication of the 10 texts used for generating the synthesized speech in each of the 4 test sets. Twelve skilled annotators participated in the dictation test. Each participant evaluated one of the 4 test sets and transcribed the synthesized speech-laugh or speech in *kana* characters. The character error rate (CER) was used as a measure for phonetic intelligibility.

The speech-laugh segmentation test was performed on the synthesized speech-laugh instances. Three datasets were prepared for the speech-laugh segmentation test; the human speech-laugh dataset, the synthesized speech-laugh dataset with pretraining, and the synthesized speech-laugh dataset without pretraining. The human speech-laugh dataset included 10 instances of speech-laugh spoken by 06\_FWA from OGVC. Each of the other two synthesized speech-laugh datasets also included 10 speech-laugh instances generated from the texts of the utterances spoken by the other speakers in OGVC. Thus, a total of 30 speech-laugh instances were prepared for the test. Since the occurrence of speech-laugh depends on the specific phones [17], the number and types of phones at laughter onset were selected on the basis of the proportion of each phone and counterbalanced across the three datasets. Five skilled speech-laugh annotators participated in the test and annotated and segmented speech-laugh instances using an annotation tool, Praat. The difference between the position of an initial mora of the speech-laugh segmented by the annotator and the correct position set by the experimenters when generating synthesized speech-laugh was calculated for each speech-laugh instance. To normalize the difference in the number of morae in the texts among the speech-laugh instances, the calculated difference in the initial positions was divided by the number of morae in the text.

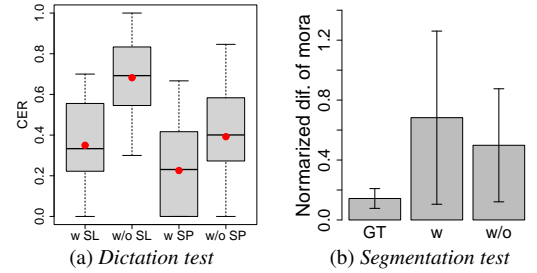


Figure 4: Results of the dictation and segmentation tests (normalized difference)

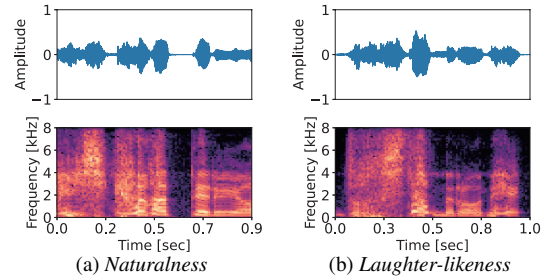


Figure 5: Synthesized speech-laugh with a high naturalness score (a) and high laughter-likeness score (b)

## 4. Results

The three-way ANOVA test for naturalness revealed significant main effects of the speaker ( $F(1, 348) = 53.90, p < .001, \eta_p^2 = 0.082$ ) and synthesis condition ( $F(2, 348) = 109.33, p < .001, \eta_p^2 = 0.333$ ). However, no main effect of the speech type was observed ( $F(1, 348) = 1.61, p = 0.21$ ). Figure 2 shows the results of the naturalness evaluation. GT had the highest naturalness score and open had the lowest. Moreover, the synthesized female speech-laugh and speech (closed and open) had higher naturalness than the male.

The three-way ANOVA test for laughter-likeness revealed significant main effects of the speech type ( $F(1, 348) = 399.99, p < .001, \eta_p^2 = 0.437$ ) and synthesis condition ( $F(2, 348) = 50.51, p < .001, \eta_p^2 = 0.110$ ). However, no main effect of the speaker was observed ( $F(1, 348) = 3.35, p = 0.068$ ). Figure 3 shows the results of the laughter-likeness evaluation for speech-laugh and speech. Figure 3 shows that speech-laugh had greater laughter-likeness than speech, regardless of the synthesis condition, except in the male open condition.

Figure 4(a) shows the results of the dictation experiment. The label w on the x-axis indicates synthesized speech-laugh or speech generated with the pretrained model, and w/o indicates instances generated without the pretrained model. For the model with pretraining, the mean and 95% CI of the CER for speech-laugh were  $0.350 \pm 0.080$ , and those for speech were  $0.226 \pm 0.075$ . For the model without pretraining, the mean and 95% CI of the CER for speech-laugh were  $0.682 \pm 0.070$ , and those for speech were  $0.392 \pm 0.087$ .

Figure 4(b) shows the results of the segmentation test. The mean difference and 95% CI (normalized value) between the segmented position and the correct label for human speech-laugh were  $2.86 \pm 1.70$  ( $0.143 \pm 0.077$ ) morae, and those for synthesized speech-laugh generated by the model with pretraining and for speech-laugh generated without pretraining were  $11.52 \pm 2.73$  ( $0.683 \pm 0.105$ ) morae and  $9.06 \pm 2.55$  ( $0.498 \pm 0.121$ ) morae, respectively.

## 5. Discussion

A significant main effect of the speaker was observed for synthesized speech-laugh and speech in the naturalness evaluation ( $2.83 \pm 0.12$  for the male speaker,  $3.26 \pm 0.10$  for the female speaker). Therefore, the naturalness of the synthesized speech was found to vary depending on the speaker. The limited amount of speech material for model training may have resulted in a lower evaluation score for the male speaker model, which had less data (85 speech-laugh segments) than did the female speaker model (114 speech-laugh segments). As shown by the results of the multiple-comparison test in Fig. 2, there are no differences between the synthesized speech-laugh in the closed condition (closed) and the human speech-laugh (GT) for the female speaker model. Therefore, the synthesized speech-laugh generated by the female speaker model has equivalent naturalness to human speech-laugh when the input text used to generate the speech-laugh model is the same as that used for fine-tuning. Figure 5(a) shows an example of synthesized speech-laugh with a high naturalness evaluation score. From this synthesized speech-laugh, the linguistic information could be perceived distinctly. It was suggested that pretraining contributed not only to preserving the phonetic intelligibility of the synthesized speech-laugh but also to improving its naturalness.

A significant main effect of the speech type was observed ( $3.09 \pm 0.14$  for speech-laugh,  $1.69 \pm 0.09$  for speech) in the laughter-likeness evaluation. According to the results of the multiple-comparison test in Fig. 3, speech-laugh can be perceived as more laughter-like than speech, except in the open condition with the male speaker. Figure 5(b) shows an example of a speech-laugh with a high value in the laughter-likeness evaluation. This synthesized speech-laugh has the acoustic characteristics of voice tremor and breathiness, similar to human speech-laugh [16]. Therefore, our approach can generate synthesized speech-laugh with diverse expressiveness and high laughter-likeness, similar to human vocalized speech-laugh.

In the dictation experiment, the CER for the synthesized speech-laugh generated using the model with pretraining ( $0.350 \pm 0.080$ ) was lower than the CER of that generated without the pretrained model ( $0.682 \pm 0.070$ ). This suggests that the phonetic intelligibility was greater for the synthesized speech-laugh generated using the model with pretraining. Figure 6 shows two synthesized speech-laugh examples generated either with pretraining or without pretraining. These two examples were generated using the same text but show the greatest difference in CER. In the synthesized speech-laugh without pretraining, some of the phonetic information was lost, and its linguistic information was drowned out by laughter. However, in the synthesized speech-laugh with pretraining, the phonetic information was maintained. By pretraining the model using the nonlaughing speech signal for the phonetic symbols of speech-laugh, our model can generate synthesized speech-laugh instances with high phonetic intelligibility.

The difference between the segmented position of laughter onset for the synthesized speech-laugh and the correct position is greater when the instance is generated with pretraining ( $11.52 \pm 2.73$  ( $0.683 \pm 0.105$ )) than without pretraining ( $9.06 \pm 2.55$  ( $0.498 \pm 0.121$ )). This finding suggests that laughter onset is more perceptible for the model without pretraining than for the model with pretraining. Some synthesized speech-laugh instances with pretraining were not annotated and segmented because it was difficult for evaluators to perceive laughter information from them. Our approach has two training phases: pretraining to learn phonetic intelligibility and fine-tuning to

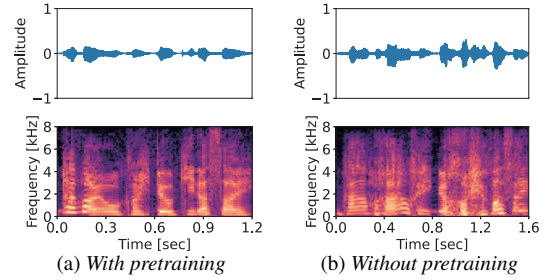


Figure 6: Two synthesized speech-laugh examples with great differences in CER

learn laughter expression. These results suggest that while pretraining allows synthesized speech-laugh to retain its phonetic intelligibility, it does not adequately represent the information of laughter. Consequently, these two training phases may trade off with each other: With pretraining, phonetic intelligibility is improved, but the controllability of laughter onset decreases. Without pretraining, phonetic intelligibility is degraded, but the controllability of laughter onset is improved.

## 6. Conclusion

The goal of this study was to obtain a speech-laugh synthesis model that preserves phonetic intelligibility and controllability of laughter onset by individually preparing phonetic symbols for speech-laugh and speech and by training on speech-laugh phones using nonlaughing speech. The results of the naturalness evaluation revealed that speech-laugh synthesized with the female speaker model was evaluated as equivalent in naturalness to human speech-laugh. The results of the laughter-likeness evaluation revealed that synthesized speech-laugh was evaluated as more laughter-like than nonlaughing speech. The results of the dictation test showed that speech-laugh synthesis with pretraining can more clearly convey linguistic information to listeners, indicating that pretraining contributes to preserving phonetic intelligibility in speech-laugh synthesis. The results of the laughter segmentation test revealed that the model without pretraining achieves more accurate laughter onset in speech-laugh than the model with pretraining does.

This study has three limitations. First, the number of speech-laugh segments and the number of speakers were small, approximately 100 segments from each of only two speakers. Expanding the amount of data via other speech materials and data augmentation could improve the naturalness and laughter-likeness of the synthesized speech. Also, a multi-speaker framework for speech-laugh synthesis needs to be demonstrated. Second, human speech-laugh was not tested for phonetic intelligibility via the dictation test because its context and phones at laughter onset could not be controlled in the test owing to the small number of speech-laugh instances in the corpora. Thus, it was not determined whether we can understand the linguistic information of human vocalized speech-laugh with vocal tremors and breathy characteristics. To fully evaluate the importance of preserving the linguistic information of synthesized speech-laugh, it is necessary to examine whether linguistic information can be perceived in human vocalized speech-laugh. Third, the naturalness of the synthesized speech-laugh was not as high as GT except for the closed condition of a female speaker. The pretraining and fine-tuning method used in this work follows common practices already established in the literature. Therefore, other approaches may improve naturalness.

## 7. Acknowledgements

This study was partially supported by JSPS KAKENHI Grant Numbers JP22K18477 and JP22K12107 and by a Kayamori Foundation for Information Science Advancement Research Grant (K36XXIX No. 662).

## 8. References

- [1] N. Perkins Booker, M. Cohn, and G. Zellou, "Linguistic patterning of laughter in human-socialbot interactions," *Frontiers in Communication*, vol. 9, 2024.
- [2] V. Krepsz, V. Horváth, A. Huszár, T. Neuberger, and D. Gyarmathy, "'should we laugh?' acoustic features of (in)voluntary laughers in spontaneous conversations," *Cognitive Processing*, vol. 25, no. 1, pp. 89–106, 2024.
- [3] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech: Fast, robust and controllable text to speech," *Advances in neural information processing systems*, vol. 32, 2019.
- [4] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *Proceedings of the 38th International Conference on Machine Learning*, p. 5530.
- [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] J. Urbain, H. Çakmak, and T. Dutoit, "Evaluation of hmm-based laughter synthesis," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7835–7839.
- [7] N. Mansouri and Z. Lachiri, "Dnn-based laughter synthesis," in *2019 International Conference on Control, Automation and Diagnosis (ICCAD)*, 2019, pp. 1–6.
- [8] H. Mori, T. Nagata, and Y. Arimoto, "Conversational and social laughter synthesis with wavenet," in *Interspeech 2019*, 2019, pp. 520–523.
- [9] N. Mansouri and Z. Lachiri, "Laughter synthesis: A comparison between variational autoencoder and autoencoder," in *2020 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, 2020, pp. 1–6.
- [10] H.-T. Luong and J. Yamagishi, "Laughnet: synthesizing laughter utterances from waveform silhouettes and a single laughter example," 2022. [Online]. Available: <https://arxiv.org/abs/2110.04946>
- [11] H. Mori and S. Kimura, "A generative framework for conversational laughter: Its 'language model' and laughter sound synthesis," in *Interspeech 2023*, 2023, pp. 3372–3376.
- [12] D. Xin, S. Takamichi, A. Morimatsu, and H. Saruwatari, "Laughter synthesis using pseudo phonetic tokens with a large-scale in-the-wild laughter corpus," in *Interspeech 2023*, 2023, pp. 17–21.
- [13] V. K. Mittal and B. Yegnanarayana, "Analysis of production characteristics of laughter," *Computer Speech and Language*, vol. 30, no. 1, pp. 99–115, 2015.
- [14] S. H. Dumpala, K. V. Sridaran, S. V. Gangashetty, and B. Yegnanarayana, "Analysis of laughter and speech-laugh signals using excitation source information," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 975–979.
- [15] C. Menezes and Y. Igarashi, "The speech laugh spectrum," in *Proc. Speech Production*, 2006.
- [16] J. Trouvain, "Phonetic aspects of "speech-laugh",," in *Oralité et Gestualité: Actes du colloque ORAGE, Aix-en-Provence. Paris: L' Harmattan*, 2001, pp. 634–639.
- [17] Y. Arimoto, "Phonetic analysis on speech-laugh occurrence in spontaneous gaming dialog," *Acoustical Science and Technology*, vol. 44, no. 1, pp. 36–39, 2023.
- [18] R. Setoguchi and Y. Arimoto, "Acoustical analysis of the initial phones in speech-laugh," in *Interspeech 2024*, 2024, pp. 3170–3174.
- [19] J. Oh and G. Wang, "Laughter modulation: from speech to speech-laugh," in *Proc. Interspeech 2013*, 2013, pp. 754–755.
- [20] K. E. Haddad, H. Çakmak, S. Dupont, and T. Dutoit, "Breath and repeat: An attempt at enhancing speech-laugh synthesis quality," in *2015 23rd European Signal Processing Conference (EUSIPCO)*, 2015, pp. 355–358.
- [21] K. E. Haddad, S. Dupont, J. Urbain, and T. Dutoit, "Speech-laugh: An hmm-based approach for amused speech synthesis," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4939–4943.
- [22] "Hi-Fi-CAPTAIN: High-fidelity and high-capacity conversational speech synthesis corpus developed by NICT," <https://astrec.nict.go.jp/en/release/hi-fi-captain/>, 2023.
- [23] R. Sonobe, S. Takamichi, and H. Saruwatari, "Jsut corpus: free large-scale japanese speech corpus for end-to-end speech synthesis," 2017. [Online]. Available: <https://arxiv.org/abs/1711.00354>
- [24] H. Mori and Y. Kikuchi, "Gaming corpus for studying social screams," in *Proc. Interspeech 2020*, 2020, pp. 3132–3135.
- [25] 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, pp. 356–369, 2012.
- [26] A. Nagaoka, H. Mori, and Y. Arimoto, "Utilizing existing labels for automatic cross-corpus emotion labeling," *Acoustical Science and Technology*, vol. 73, no. 11, pp. 682–693, 2017, (in Japanese).
- [27] M. Schröder, "Experimental study of affect bursts," *Speech Communication*, vol. 40, no. 1, pp. 99–116, 2003.
- [28] Y. Arimoto, D. Oishi, and M. Okubo, "A comparison between crowdsourcing and in-person listening tests on emotion rating for spontaneous screams and shouts," *Acoustical Science and Technology*, vol. advpub, p. e24.58, 2024.