INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Conversational and Social Laughter Synthesis with WaveNet

Hiroki Mori[1], *Tomohiro Nagata*[1], *Yoshiko Arimoto*[2]

[1]Utsunomiya University, Japan
[2]Chiba Institute of Technology, Japan
hiroki@speech-lab.org

Abstract

The studies of laughter synthesis are relatively few, and they are still in a preliminary stage. We explored the possibility of applying WaveNet to laughter synthesis. WaveNet is potentially more suitable to model laughter waveforms that do not have a well-established theory of production like speech signals. Conversational laughter was modelled with a spontaneous dialogue speech corpus based on WaveNet. To obtain more stable laughter generation, conditioning WaveNet by power contour was proposed. Experimental results showed that the synthesized laughter by WaveNet was perceived as closer to natural laughter than HMM-based synthesized laughter.

Index Terms: Laughter, dialogue, conversation, speech synthesis, spontaneous speech, affect burst, social signals

1. Introduction

Laughter is one of the most fundamental display of emotion. It plays essential roles in human communication. Nevertheless, laughter has long been paid less attention in the studies of speech communication between human and machine, especially in the area of speech synthesis.

One of the main reasons of few attempts to synthesize laughter may be that the way of evaluating synthesized laughter is not established. Unlike speech, laughter does not need intelligibility. On the other hand, natural laughter has many forms depending on its contexts [1, 2]. It is quite convincing that each form of laughter voice is not mutually compatible, and therefore it is bound by some contextual relevance. Thus, laughter synthesis for human-machine communication should appropriately reflect not only local contexts such as surrounding speech sounds but also more global ones such as discourse functions or emotional states, as discussed in our first laughter synthesis paper [3]. This cannot be realized by simply playing back recorded laughter sound.

Previous studies of laughter synthesis include ones based on an oscillatory system [4], formant synthesis [5], diphone synthesis [6], articulatory speech synthesis [7], and hidden Markov model (HMM)-based synthesis [8, 9]. We have been investigating laughter synthesis for conversational agents based on spontaneous spoken dialogue corpora. Our previous study showed that an extended set of contextual factors that are defined to describe the global characteristics improved the overall naturalness of synthesized laughter [3]. Nevertheless, the quality of synthesized laughter is still far from satisfactory. Current laughter synthesis frameworks have a lot of difficulties, which include modelling irregular vocal fold vibration and strong aspiration driven by a sudden burst of the airflow, modelling complicated dynamics of laughter sounds by means of a simple context-dependent HMM, and so on.

In this paper, we focus on the WaveNet [10], a statistical waveform synthesis framework based on the convolutional neural network, and explore the possibility of applying WaveNet to laughter synthesis. Because WaveNet is a model that makes almost no assumption on the production process of waveform, it is potentially more suitable for modeling laughter waveform that lacks a theoretical model of production.

2. Corpus

Laughter is social [11]. Laughter "invites both naturalistic and socially based explanations, as a mood indicator outwardly manifesting an inner emotion or as a signal that can be employed strategically" (Glenn & Holt, 2013, p. 1 [12]). Most previous laughter studies use induced laughter, for example using joke videos, rather than that in conversational scenes. We think this is not appropriate as a corpus for conversational laughter synthesis, because we should also model the social aspect of laughter. Therefore, we focus on laughter included in a spontaneous dialogue speech corpus for laughter synthesis. In this study, the Online Gaming Voice chat Corpus (OGVC) was used [13]. The OGVC is an emotional speech corpus that can compare spontaneous speech and acted speech. In this paper, spontaneous speech is exclusively used. The voice chat during the online game was recorded as spontaneous speech. In total, 9114 spontaneous utterances were recorded by 13 (4 female and 9 male) speakers.

Although the laughter is already indicated as {laugh} in the original transcription of OGVC, the authors established a guideline for laughter annotation and conducted a re-annotation of laughter from scratch, because:

- The original identification of laughter relied solely on the transcribers' intuition without any guideline for laughter annotation,
- We also need a fine-grained annotation for the laughter synthesis, which presuppose reliable identification of laughter, and
- The original annotation often excludes inhalation (breathing-in) sounds from laughter episodes. We have found that inhalation sounds form a perceptually important part of laughter sounds. Therefore, inhalation sound that accompanies a laughter should be identified as a part of the laughter.

The structure of laughter was hierarchically annotated according to Trouvain's schema [14] (Fig. 1). Laughter is composed of one or more acoustic events each of which corresponds to an exhalation or inhalation. These together form a percept, which is called a laughter *episode* and often indicated as {laugh} in the transcription. An event corresponding to an exhalation is called a laughter *bout*, which is analogous to a word or phrase. A bout is composed of one or more laughter *call*s. A call is analogous to a syllable. Therefore, a typical bout "hahaha" is a 3-call bout. An inhalation sound is often

http://dx.doi.org/10.21437/Interspeech.2019-2131

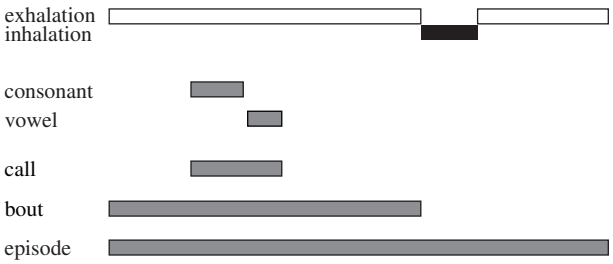

Figure 1: *Temporal structure of laughter (adapted from [14]).*

heard immediately after the main body of laughter produced by exhalation [15]. Some inhalation sounds characterize the laughter together with followed or following bouts.

The annotation consists of two stages. The first stage (L1) is the bout-level annotation, which involves the segmentation of bouts (b) and inhalations. Because inhalation within laughter is often accompanied by audible vocal fold vibration, unvoiced inhalation (h) and voiced inhalation (H) was distinguished in the L1 annotation. The second stage (L2) is the call-level annotation, which involves the segmentation and phonetic transcription of calls. An example of the annotation is shown in Fig. 2.

All the identified laughter sounds are not equally audible; some of them are slight, or even extremely weak. Regarding all of them equally as "laughter" and using the whole data as the training set may be challenging at the stage of initial study. In this trial, the first author objectively sorted out the "major" laughter episodes from the subtle ones. The criteria was that a major laughter episode should include at least one multi-call bout composed of prominent voiced oral calls. The selection was performed for the recordings of a male speaker (04_MSY) and a female speaker (06_FWA). Consequently, 125 episodes of the speaker 04_MSY and 101 episodes of the speaker 06_FWA were selected as the dataset for the laughter synthesis.

3. The WaveNet architecture

WaveNet is a model that predicts the current sample point for given past sample points as the condition. It is also possible to append auxiliary features to the condition. Namely, if linguistic features are provided, WaveNet will become a text-to-speech [10]; if speech parameters such as the excitation and spectral features are provided, it will become a WaveNet vocoder [16].

In the current study, the following two conditioning were considered:

1. Loose Conditioning
2. Conditioning by Power

In the Loose Conditioning, only whether current position is a bout or an inhalation is provided in a one-hot encoding as the condition. Additionally, voiced/unvoiced distinction (1/0) is appended for inhalations. Namely, a frame of a bout (b), an unvoiced inhalation (h), and a voiced inhalation (H) is represented as [1,0,0], [0,1,0], and [0,1,1], respectively. The advantage of the Loose Conditioning is that it only requires a specification of laughter structure (e.g. b, bh, Hbh, etc.) in the synthesis stage.

The Conditioning by Power employs the information of power contour of the laughter signal as well as the bout/inhalation information. The motivation of introducing power contour was to reduce an instability in synthesis, which will be discussed in the later section.

Table 1: *The structure of WaveNet.*

# layers	30
# channels of aux input	3 (Loose) or 4 (Power)
# channels of residual blocks	256
# channels of skip-out	256
# channels of output	256 (softmax)

A laughter bout is a long event that can stretch over a few seconds. To capture the typical pattern of bout, which consists of consecutive calls with gradually decreasing pitch and amplitude, WaveNet should refer to the history of enough time length stretching over at least two calls. We stacked 3 repetitions of 10 layers of dilated causal convolutions whose dilation was $1, 2, 4, \cdots, 512$, which in total corresponds to a receptive field of 6139 samples (383 ms). The output was 8-bit μ-law. A detailed specification of the WaveNet architecture is shown in Table 1.

4. Synthesized waveforms

4.1. Loose Conditioning

Figure 3 shows the examples of laughter waveforms generated by the WaveNet trained with the 04_MSY data. Under the Loose Conditioning, the output depends heavily on chance. In this figure, we synthesized the laughter five times referring to the bout/inhalation information of an identical natural laughter waveform. These five waveforms look very different. We often observed that a large part of some synthesized waveforms became silence.

Among the examples shown, the output waveform ④ is remarkable. The waveform was synthesized without call-level (L2) information. Despite this, the second bout seems to be composed of 4 calls (indicated by arrows). This implies that the WaveNet is successfully capturing the bout pattern composed of multiple calls by the very long receptive field described in the previous section.

4.2. Conditioning by Power

As shown in the above, WaveNet has the potential to synthesize fairly natural laughter. However, the dependence on chance is inconvenient from a practical perspective. It is desirable to be able to stably generate natural laughter waveform.

We tried to append an input feature as the condition for mitigating the instability. Specifically, we examined the effectiveness of introducing power information into the auxiliary input to control the amplitude envelope of synthesized laughter. Figure 4 shows the example waveforms generated by providing the power contour of a reference laughter waveform as the auxiliary input of WaveNet. Hereinafter, the 0th mel-cepstral coefficient, which is a byproduct of the spectral analysis for HMM-based synthesis described later, of a given frame is regarded as its power parameter. For each reference waveform, it can be seen that a laughter waveform resembling its reference waveform is synthesized in a stable manner.

5. Naturalness evaluation

A listening test was conducted for evaluating the WaveNet-based laughter synthesis. The evaluation was done for the laughter synthesized by HMM [3] and by WaveNet, as well as the natural laughter for reference.

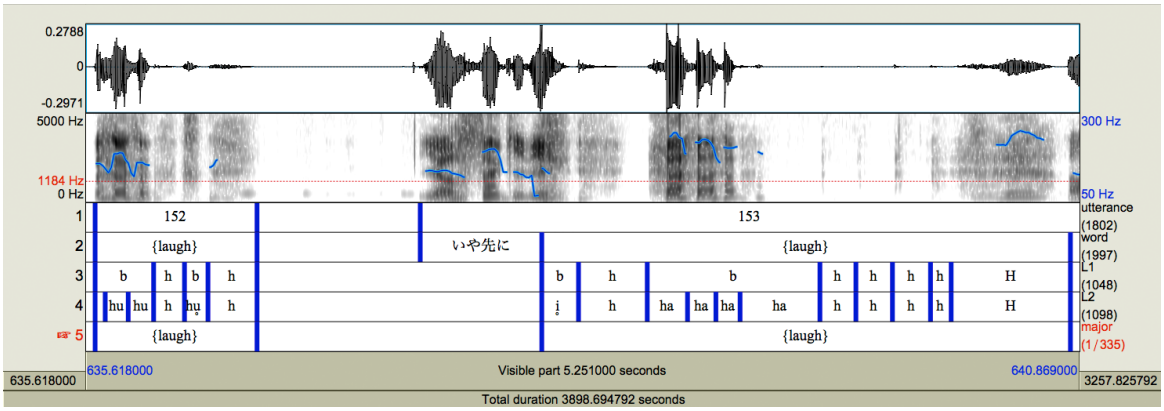

Figure 2: *L1 and L2 annotation.*

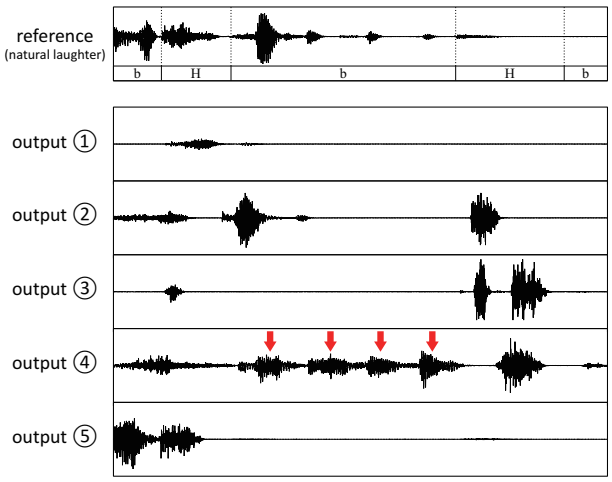

Figure 3: *Examples of generated laughter waveforms under the loose conditioning. (b: bout, H: voiced inhalation)*

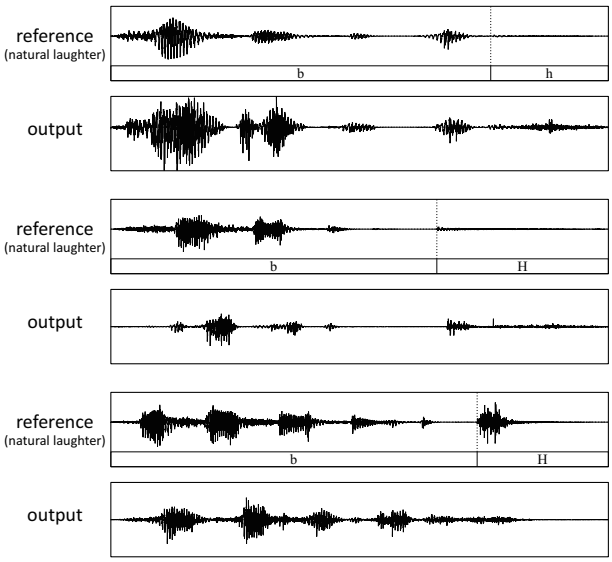

Figure 4: *Examples of generated laughter waveforms conditioned by power contours. (b: bout, h: unvoiced inhalation, H: voiced inhalation)*

The test set was composed of 30 episodes of 04_MSY and 30 episodes of 06_FWA. These were selected at random from the pool of b/h/H sequences appeared in the corpus. The methods to be compared were HMM, WaveNet and Natural. Therefore, the set of stimulus includes $(30+30)\times 3 = 180$ episodes. The sampling rate was 16 kHz.

The laughter synthesis with HMM requires the call-level (L2) annotation. For the HMM synthesis, the employed context included the phonetic identity of current / preceding / succeeding calls, the laughter position within utterance, the call position within bout, and the number of calls within bout [3]. The F0 and spectral information was analysed by WORLD [17] with every 5 ms, then the spectral information was converted to 40th-order mel-cepstral coefficients. The speech parameters consisted of 40 mel-cepstra, log F0, BAP, and their delta and delta-delta, summed up to 126 dimensions. The model structure was a 5-state left-to-right hidden semi-Markov model with a single Gaussian distribution.

The WaveNet synthesis was carried out with the Conditioning by Power. For the duration of bouts and inhalations required for the input to the WaveNet, and for the power contour required for the conditioning, the HMM-predicted parameters described above were used.

The subjects were twelve undergraduate and graduate students. They were instructed to evaluate how close the presented laughter is to real laughter in a 5-point scale.

The results of naturalness evaluation are shown in Fig. 5 as the distribution of the mean opinion score (MOS) for each stimulus. The MOS averaged over all stimuli synthesized from 04_MSY's laughter was 2.45 for HMM and 3.14 for WaveNet. The difference of average MOS was significant ($p < 1.0 \times 10^{-7}$, paired t-test). Although the quality of most synthesized laughter by HMM was acceptable, WaveNet has an apparent superiority in the reality of synthesized laughter to the HMM. In fact, there is much room for improvement, as the average MOS for WaveNet was 1.6 points below the natural laughter. In our impression, the WaveNet laughter is generally real, but it tends to sound a bit noisy and echoic. We think the most serious bottleneck is the shortage of data. However, we achieved a certain breakthrough to laughter synthesis using WaveNet.

The MOS averaged over all stimuli synthesized from 06_FWA's laughter was 1.97 for HMM and 2.16 for WaveNet.

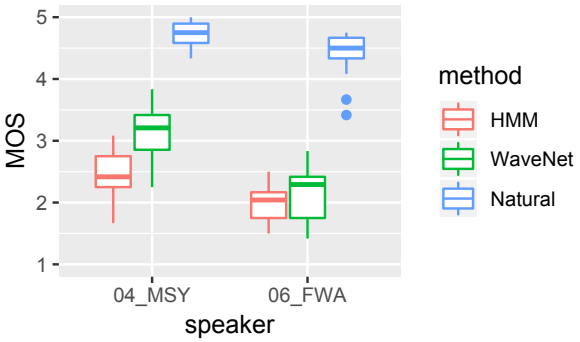

Figure 5: *Naturalness evaluation.*

The difference of average MOS was significant ($p = 0.048$, paired t-test). Generally, laughter synthesis for the speaker 06_FWA is much more difficult than for the speaker 04_MSY, mainly because her laughing style is diverse and not necessarily typical. A considerable part of her laughter consists of giggles and chuckles rather than typical laughter bouts. Some laughter is hardly distinguishable from breathing. Even natural laughter was occasionally evaluated not-so-natural (3.42 or 3.67 in MOS). WaveNet certainly contributed to improve the naturalness compared to HMM. However, the current Conditioning by Power method relies on the power contour generated by HMM. If a poorly estimated power contour misguides the WaveNet, the result also becomes poor. A more reliable method to estimate the power contour of the laughter to synthesize is needed for the improvement.

6. Conclusions

This paper presented a preliminary study of laughter synthesis in conversation using WaveNet. To obtain more stable laughter generation, conditioning WaveNet by power contour was proposed. The result of the subjective evaluation test revealed that artificial laughter generated by WaveNet was perceived as closer to natural laughter than HMM-based synthesized laughter.

WaveNet has a potential to control the acoustic characteristics of waveforms to generate by feeding extra explanatory variables to the auxiliary input. A straight forward extension to the current study is to introduce phonetic context of laughter [3] for seamless transition to/from speech, and the global characteristics to adjust, for example, relative prominence to following / followed speech. A more intriguing but challenging goal is to control the perceptual effect of laughter directly. Several dialogue speech corpora that include laughter provide the speakers' emotional states annotated by human [18, 19], so conditioning WaveNet by emotional states is attractive. However, there is considerable evidence that laughter form depends on emotional states [20, 21]. Controlling emotion in laughter synthesis is therefore involved not only in the generation of waveform but also in the prediction of laughter structure.

7. Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers 19H01252, 18H04128 and 16H03421.

8. References

[1] J.-A. Bachorowski, M. J. Smoski, and M. J. Owren, "The acoustic features of human laughter," *The Journal of the Acoustical Society of America*, vol. 110, no. 3, pp. 1581–1597, sep 2001.

[2] M. Schröder, "Experimental study of affect bursts," *Speech Communication*, vol. 40, no. 1-2, pp. 99–116, 2003.

[3] T. Nagata and H. Mori, "Defining laughter context for laughter synthesis with spontaneous speech corpus," *IEEE Transactions on Affective Computing*, 2018. doi: 10.1109/TAFFC.2018.2813381

[4] S. Sundaram and S. S. Narayanan, "Automatic acoustic synthesis of human-like laughter," *Journal of the Acoustical Society of America*, vol. 121, no. 1, pp. 527–535, 2007.

[5] J. Oh and G. Wang, "LOLOL: Laugh out loud on laptop," in *Proc. NIME 2013*, 2013, pp. 190–195.

[6] 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.

[7] E. Lasarcyk and J. Trouvain, "Imitating conversational laughter with an articulatory speech synthesis," in *Proc. Interdisciplinary Workshop on the Phonetics of Laughter*, 2008, pp. 43–48.

[8] J. Urbain, H. Çakmak, and T. Dutoit, "Evaluation of HMM-based laughter synthesis," in *Proc. ICASSP 2013*, 2013, pp. 7835–7839.

[9] 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.

[10] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," *CoRR*, vol. abs/1609.03499, 2016.

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

[12] P. Glenn and E. Holt, *Studies of Laughter in Interaction*. London: Bloomsbury, 2013.

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

[15] J. Urbain, E. Bevacqua, T. Dutoit, A. Moinet, R. Niewiadomski, C. Pelachaud, B. Picart, J. Tilmanne, and J. Wagner, "The avlaughtercycle database," in *Proc. 7th Int'l Conf. Language Resources and Evaluation (LREC 2010)*, 2010, pp. 2996–3001.

[16] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, "Speaker-dependent WaveNet vocoder," in *Proc. Interspeech 2017*, 2017, pp. 1118–1122.

[17] 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.

[18] 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.

[19] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, 2013, pp. 1–8.

[20] J.-A. Bachorowski and M. J. Owren, "Vocal expression of emotion: Acoustic properties of speech are associated with emotional intensity and context," *Psychological Science*, vol. 6, pp. 219–224, 1995.

[21] ——, "Not all laughs are alike: Voiced but not unvoiced laughter readily elicits positive affect," *Psychological Science*, vol. 12, no. 3, pp. 252–257, 2001.