# Defining laughter context for laughter synthesis with spontaneous speech corpus

Tomohiro Nagata and Hiroki Mori

**Abstract**—In this paper, conversational laughter was synthesized by a statistical model-based speech synthesis framework using spontaneous speech corpora. The phonetic transcriptions of natural laughter in these corpora were annotated, and the context required to synthesize the laughter that accompanies speech sounds was defined from the perspective of the (1) phonetic properties of the current segment, (2) phonetic properties of previous and succeeding segments, and (3) positional factors of the current segment or laughter bout. Laughter was synthesized using the defined context and the framework of HMM-based speech synthesis.
To confirm the influence of the contextual factors on the naturalness of speech, a subjective evaluation was performed. As the result of the evaluation, the naturalness of the entire utterance was improved by using the contextual factors defined in this study. This result confirmed the importance of defining the appropriate context to synthesize natural conversational laughter.

**Index Terms**—Laughter, Spontaneous speech corpus, HMM-based speech synthesis

## 1 Introduction

In recent years, human-human and human-machine interactions have attracted increasing interest. To ensure that human-machine interactions resemble those between humans, machine speech must express both linguistic information and paralinguistic information such as the speaker's attitude, intentions, and emotions. To express such information, a speech synthesizer must produce non-lexical sounds in addition to speech sounds.

Laughing is a typical method of conveying non-verbal information and has attracted research interest. Laughter research has a long history and has been studied from various perspectives, including generational, sociological, and physiological ones [1]. In recent years, incorporating laughter-based behavior into human-machine interaction models has become a hot topic. For example, the ILHAIRE project [2] researched the fundamentals and applications of laughter, which include laughter detection, recognition and synthesis [3]–[10].

Laughter is highly diverse. It has many forms, and appears in various situations [11], [12]. Also, it has been suggested that there is a gradual change from smiling to "pure" laughter, i.e., speech-smile → speech-laugh → laughter, each of which has different acoustic characteristics and communicative functions [13]–[16]. All of these phenomena must be taken into consideration in order to make a machine laugh, which cannot be realized merely by replaying recorded laughter sound. To cope with such diversity, a flexible framework of laughter synthesis is essential.

However, laughter synthesis is still a developing field. In [6] and [7], methods for synthesizing laughter vowels using linear prediction were proposed, where the laughter energy envelope was modeled via a mass-spring system. A method of laughter vowel synthesis using formant synthesis to control the expressivity of laughter was developed in [8]. In [9] and [10], a diphone concatenation-based laughter synthesis method was proposed. Additionally, a method of laughter synthesis was developed using a 3D model of the vocal tract in [10].

In more recent studies, a method of laughter synthesis via a hidden Markov model (HMM)-based approach was proposed in [17] and [18]. In the method, the excitation and spectral parameters of laughter are modeled via HMMs; the differences in terms of the acoustic features are represented by the context label. In [19], the acoustic features of laughter were controlled via an arousal-driven method. In relation to these "pure" laughter synthesis, it has been proposed that speech-laughs be synthesized by replacing vowels of the speech-smile model with vowels of the laughter model [20], [21].

However, most corpus-based studies use induced laughter rather than that in conversational scenes, even though the synthesized laughter can be used in human-machine communication. Laughter in conversation occasionally accompanies speech sounds. It has been reported that even if the naturalness of the laughter itself is high, the overall naturalness of an utterance is reduced when it is connected with speech sounds [9]. In [10], laughter synthesized by diphone concatenative synthesis was perceived as unnatural. To properly synthesize laughter in conversational scenes, the laughter model needs to consider the context in which laughter is placed.

In this study, we focus on the statistical parametric speech synthesis framework. In this framework, the prosodic and segmental features of phonemes that vary due to various factors are expressed as the context [23]. Unlike simple concatenative methods, this method can robustly model the acoustic properties of segments, which can be explained by a combination of contextual factors. Thus, laughter that is suitable for a situation can be synthesized using the statistical parametric framework if we can define a proper laughter context.

• *T. Nagata and H. Mori are with Graduate School of Engineering, Utsunomiya University, Japan.*
*E-mail: ken1@speech-lab.org*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2018.2813381, IEEE Transactions on Affective Computing



In the HMM-based laughter synthesis mentioned above [17], [18], the datasets used were not laughs from communication scenes but ones collected by emotion induction using joke videos. The authors believe that no previous work has attempted to synthesize laughter in conversation using a spontaneous dialog speech corpus as a dataset.

The goal of this study is to synthesize natural laughter in conversational scenes. To achieve this, we need to collect natural laughter in spontaneous conversations rather than artificially induced laughter.

To properly model laughter in conversation, we need to identify effective contextual factors. Urbain et al. investigated the effect of contextual information on the naturalness of synthesized laughter [18]. They tried to add "an extended context including the information available thanks to the syllabic annotation (e.g., position of the phone in the syllable, position of the syllable in the word, etc.)" [18] to simple phone information, but failed to prove the effectiveness of the extended context information. In the following papers to [18], contextual factors are basically same [19], [22]. So far, no previous work has further explored the contextual factors for HMM-based laughter synthesis. Hence, the main contribution of this study is to clarify the contextual factors required for laughter synthesis with spontaneous dialog speech. We define a relatively simple laughter context as the first step.

In this paper, the natural laughter that accompanies speech is synthesized with spontaneous speech corpora using the HMM-based speech synthesis framework and this newly defined context. The corpora used in this study is described in Section 2. Section 3 shows the laughter annotation method, and some statistics of laughter annotated for the corpora. In section 4, the context for laughter in a conversation is defined; the laughs are synthesized based on the context using the HMM-based speech synthesis framework. In this study, the position of laughter in an utterance that includes speech sounds (a relative position with speech sounds) was defined, as was the basic context (e.g. phonetic transcription, position of the syllable, etc. [17], [18]). In section 5, the effectiveness of the newly defined context is investigated via a subjective evaluation of naturalness.

## 2 Dialog speech corpus

Laughter in dialog may play a social role; thus, it is necessary to consider what factors are different from laughter alone. For example, laughter overlaps more frequently in conversation than with speech sounds [24]. In fact, overlapping laughter has a different acoustic form than non-overlapping laughter does [25]. Thus, because there may be a difference in acoustic features between laughter sounds in conversation and induced laughter, it may not be appropriate to exclusively use induced laughter to synthesize laughter in conversation. Therefore, we focus on laughter included in spontaneous dialog speech corpora for laughter synthesis. In this study, the Utsunomiya University Spoken Dialogue Database for Paralinguistic Information Studies (UU Database) [26] and the Online Gaming Voice chat Corpus (OGVC) [27] were used.

TABLE 1
The number of laughs in the UU Database.

| Speaker | Num | Speaker | Num | Speaker | Num |
|---|---|---|---|---|---|
| FJK | 8 | FKC | 22 | FMS | 16 |
| FMT | 34 | FNN | 14 | FSA | 26 |
| FSH | 17 | FTH | 5 | FTS | 40 |
| FTY | 19 | FUE | 14 | FYH | 23 |
| MKK | 22 | MKO | 20 | | |

TABLE 2
The number of laughs in the OGVC.

| Speaker | Num | Speaker | Num | Speaker | Num |
|---|---|---|---|---|---|
| 01_MMK | 26 | 03_FMA | 74 | 05_MYH | 52 |
| 01_MAD | 118 | 03_FTY | 41 | 05_MKK | 87 |
| 02_MFM | 142 | 04_MNN | 154 | 06_FTY | 251 |
| 02_MEM | 145 | 04_MSY | 246 | 06_FWA | 175 |

### 2.1 UU Database

The UU Database is a speech corpus for studying linguistic and phonetic phenomena in expressive spoken dialog. The database consists of natural dialog spoken by seven pairs of college students (12 female speakers and 2 male speakers). The task of the dialogs is "four-frame cartoon sorting." Thanks to the amusing nature of the task, the database is characterized by a wide variety of recorded expressive dialog speech. The total number of utterances is 4840.

The speech recorded in the UU Database is annotated with the labels of non-verbal sounds such as laughter, inhalation, and coughing. The total number of laughs is 280. The number of laughs by each speaker is shown in Tab. 1.

### 2.2 OGVC

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 speakers (4 female speakers and 9 male speakers).

In the OGVC, the location of the laughter is annotated in the transcription. The total number of laughs is 1593. The number of laughs by each speaker is shown in Tab. 2.

## 3 Segmentation of laughter

Laughter is difficult to identify. In particular, the distinction between pure "laughter" and "speech-laugh" [14] is important for corpus annotation. We define "laughters" in this paper as laughing sounds that do not overlap speech, as opposed to "speech-laughs", which overlap speech. We shall exclude speech-laughs from the scope of the current study, though the distinction is not straightforward both in theory and practice. We shall avoid the difficulty by relying on the laughter labels provided by the target corpora, the UU Database and the OGVC. In both corpora, the annotated laughters (shown in Tables 1 and 2) do not include speech-laughs, according to the user manuals.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2018.2813381, IEEE Transactions on Affective Computing



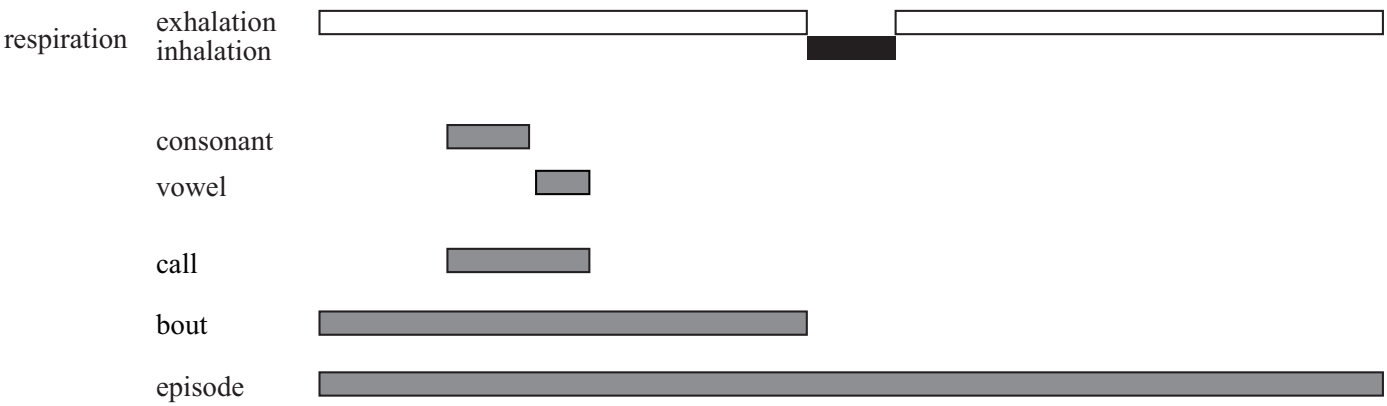



Fig. 1. The hierarchical structure of laughter (adapted from [28]).

The structure of laughter is hierarchical, and its hierarchy is roughly divided into the segmental level, the syllabic level and the phrasal level [28]. Fig. 1 shows the structure of laughter. At the segmental level, laughter is split into consonants and vowels. Here, "vowel" and "consonant" denote sound-pause alternations; therefore, they have a different meanings from a phonetic perspective [28]. At the syllabic level, laughter is segmented into syllable (*calls*), which consist of the initial and final consonants and a vowel. At the phrasal level, laughter is segmented into phrase (*bouts*) consisting of some *calls*. Additionally, multiple phrases establish the sentence level (i.e., the laughter episode).

In this study, each laughter episode is segmented into *bouts* and *calls*.

### 3.1 Description of laughter in this study

#### 3.1.1 Phrasal and sentence level

Laughter episodes can be composed of several *bouts* separated by inhalation. In this study, a laughter episode is represented by a "{laugh}." First, the laughing part of an utterance is segmented by the laughter episode. Next, the laughter episode is segmented by *bouts* and inhalation, which are denoted as "b" and "h", respectively.

#### 3.1.2 Syllabic level

In the AVLaughterCycle Database [29], which was built for automatic laughter processing (i.e., HMM-based laughter synthesis [17], [18]), a detailed description of laughter is provided, e.g., a transcription using the International Phonetic Alphabet (IPA). Although such descriptions are considered useful for high-quality laughter synthesis, professional knowledge and long observation times are required. This is a major barrier when transcribing large-scale speech corpus or voice resources. Therefore, as a simplified method, *calls* are described using *kana* (Japanese syllabic characters) in this study. The description is transcribed once via *kana*, and then it is converted to phoneme sequences. For example, laughter "ハハハ" is described as [hahaha].

It is impossible to express differences between sounds due to phenomena such as unvoicedness, nasalization and prolongation by using only the phoneme symbols. To enable such descriptions, a set of auxiliary symbols, as shown in Fig. 2, was defined.

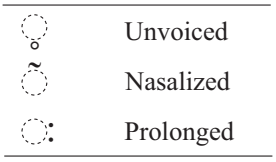



Fig. 2. Auxiliary symbols.

### 3.2 Annotation of laughter transcription

Laughter was annotated using the description system defined in Sect. 3.1. The annotator was the first author and Praat [30] was used for the annotation. For this annotation, the target speakers were limited to females because the HMM synthesis requires that the training data have as similar voice quality as possible in order to train stable models.

The annotation of laughter of speaker FTS in the UU Database and speakers 03_FTY, 06_FTY and 06_FWA in the OGVC was completed. Since the purpose of this study is to synthesize laughter accompanied by speech sounds, isolated laughter (i.e. there are no speech sounds just before or just after the laugh) was excluded from the scope of the annotation.

### 3.3 Annotation results

An actual example of annotated laughter is shown in Fig. 3. Laughter episodes and the transcription of speech sounds are described as layers ①. This example is an utterance containing a laughter episode. In layer ②, the structure of a laughter episode is described. The laughter episode consists of a *bout* and an inhalation. In layer ③, the structure of a *bout* is described using *calls*. The phonetic notation is extended by auxiliary symbols used to describe *calls*.

A histogram of the vowels in *calls* is shown in Fig. 4. For vowels in *call*, /a/, /u/ and /e/ were dominant. The vowels /i/ and /o/ were rare, and some speakers did not produce them at all.

The most frequent vowel was /u/, and there are few *calls* consisting only of vowels. This is shown in Fig. 5. Black and gray bars indicate the number of *calls* containing only vowels and the first *calls* containing only vowels, respectively. In contrast, the vowels /a/ and /e/ had relatively

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2018.2813381, IEEE Transactions on Affective Computing



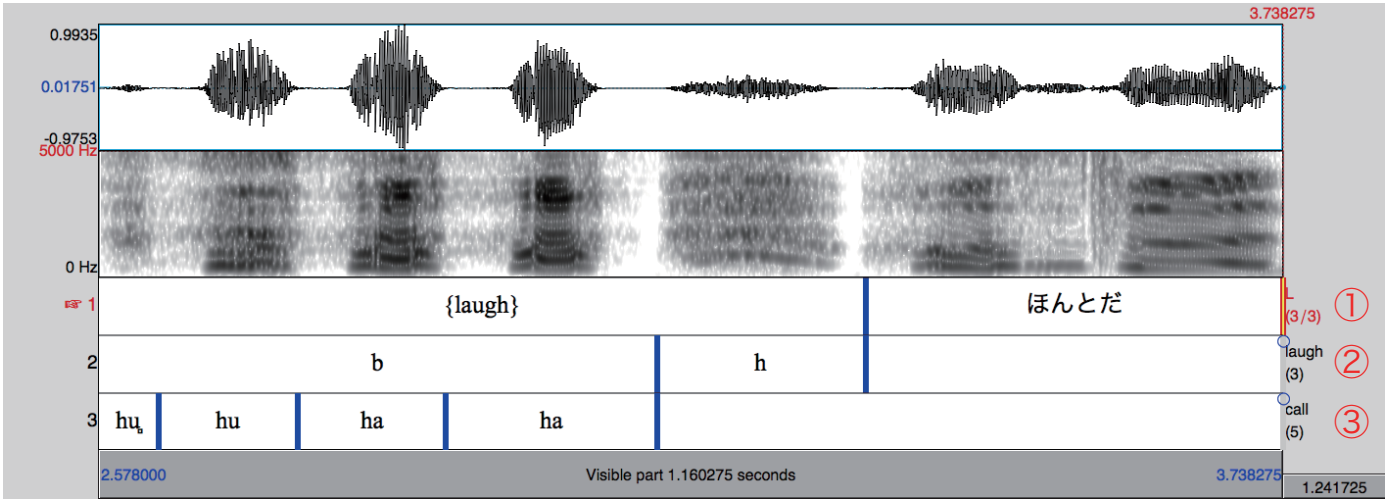



Fig. 3. An example of laughter annotation.

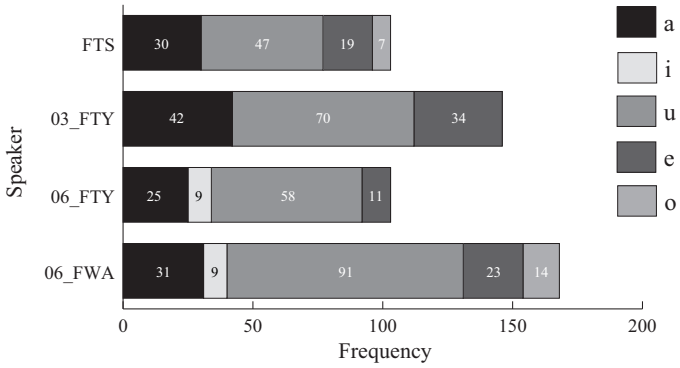



Fig. 4. The histogram of vowels in *calls*.

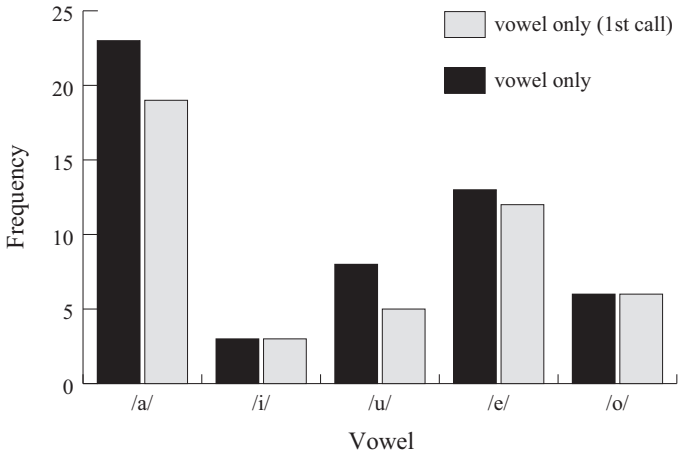



Fig. 5. The number of *calls* in terms of vowels only.

more *calls* consisting only of vowels. Moreover, most of the *calls* tended to appear at the initial position of the *bout*.

Fig. 6 illustrates the breakdown of the *bout*'s structure. *Bouts* consisting of only one *call*, i.e., *single-call bouts*, accounted for 17% of the total, and unvoiced *bouts* accounted for 35%. In contrast, there were few *multi-call bouts* where all of the *calls* were unvoiced. This is the same trend as in [31]. Additionally, the distribution of the voicedness of *calls*

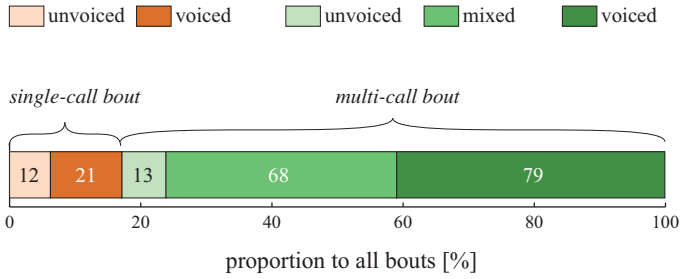



Fig. 6. The breakdown of *bout*'s voiced/unvoiced structure.

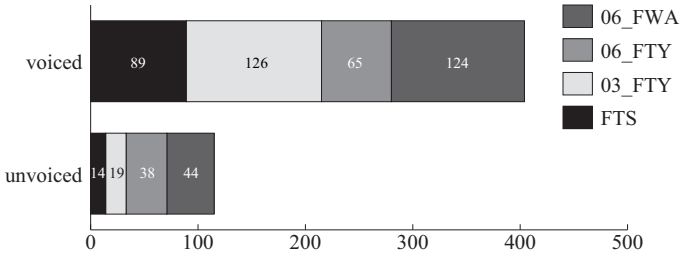



Fig. 7. The number of voiced/unvoiced *call*.

is illustrated in Fig. 7. These results confirm that the number of unvoiced *calls* is small.

## 4 Laughter Synthesis Model

### 4.1 Definition of context

Various kinds of contexts (dialog acts [32], functions as a response token [33], interpersonal stance [34], etc.) may affect the acoustical properties of laughter sounds. In this study, we defined similar contexts that are used in TTS systems as a first step.

The context defined in this study is shown in Tab. 3. To express the difference in sounds characterized by the vocal tract configuration and phonation, the transcription of current *calls* was included. Here, the syllabic characters described in Sect. 3.1 were used. Voiced/unvoiced, oral/nasal and prolonged sounds were clearly distinguished via auxiliary symbols.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2018.2813381, IEEE Transactions on Affective Computing



TABLE 3
Defined context for laughter

| | |
|---|---|
| $c_c$ : The transcription of current *calls* | |
| $c_l$ : The transcription of preceding segment | A |
| $c_r$ : The transcription of succeeding segment | A |
| $p_l$ : The laughter position for an utterance | B |
| $p_c$ : The *call* position for a *bout* | B |
| $n_c$ : The number of *call* for a *bout* | B |

TABLE 4
The training condition for the laughter model

| | |
|---|---|
| Model | 5 states left-to-right HSMM |
| Feature vector | 0-39th mel-cepstral coefficient, logarithmic fundamental frequency, $\Delta$, $\Delta\Delta$ |
| Analysis | Speech signals were sampled at a rate of 16 kHz and windowed by a 25 ms Hamming window with a 5 ms shift |

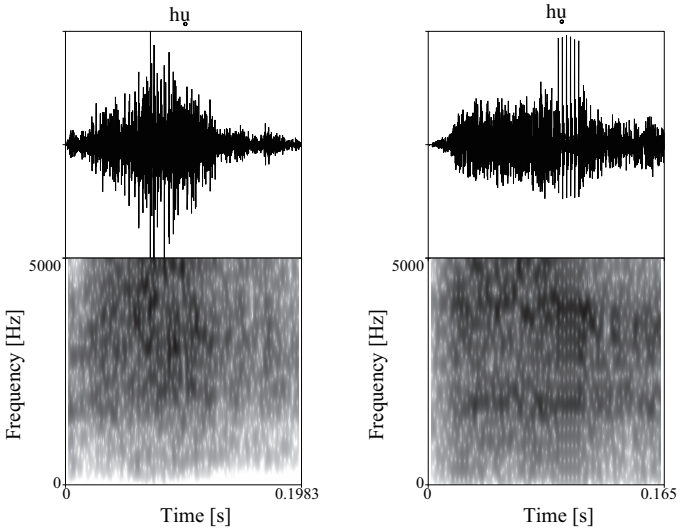
Two contexts were added as well. Context A includes the transcription of the preceding and succeeding segments to consider the influence of the difference between preceding and succeeding sounds. Here, "segments" are either phonemes (for speech sound) or *calls* (for laughter).

Context B includes contextual factors that have more global information. The laughter location in an utterance was added to consider whether the laughter is placed at the beginning, middle, or end of an utterance. The distribution of *call* duration is dependent on the *call* position [11]. Therefore, the *call* position was also added to the context. The number of *calls* in a *bout* was included, similar to the number of morae in an accentual phrase, which is often used in HMM-based speech synthesis.

### 4.2 HMM-based laughter synthesis

The laughter of speakers FTS, 06_FTY and 06_FWA was used for model training. The laughs of speaker 03_FTY were discarded because the amplitude levels of some recorded sounds were too small. The total amount of training data was 109 *bouts*. The HMM-based speech synthesis system (HTS, version 2.2) [35] was used for the model training and laughter synthesis. The speech analysis conditions and training model are shown in Tab. 4. Here, the number of states for HSMMs was determined by seeking from the range of 5–10 by assessing the quality of synthesized laughter; we used 5. The model was trained via the Speaker Adaptive Training technique using a shared decision tree [36]. When constructing the tree, questions about contextual factors in Tab. 3 were applied, and the nodes were split with the maximum likelihood criterion.

The test laughter was synthesized with the model that was trained using all of the training data except the test laughter itself (leave-one-out method). Therefore, the number of synthesized laughs is 109 bouts. Each laughter was synthesized adaptively to fit the model of the original speaker and of her laughter.




(a) natural *single-call bout* (b) synthesized *single-call bout*

Fig. 8. Example of *single-call bout*.

### 4.3 Results of synthesized laughter

As an example of a *single-call bout*, the waveform and spectrum of laughter [hu̥] are shown in Fig. 8. Laughter [hu̥] is a fricative sound without vocal fold vibration, as shown in Fig. 8 (a), and it is a typically a *single-call bout* in a daily conversation scene such as giggles or self-derision. Fig. 8 (b) shows the waveform and spectrum of the synthesized *single-call bout*, from which it can be confirmed that the synthesized [hu̥] is similar to the turbulence of natural laughter.

Similarly, the waveform and spectrum of laughter [hu̥huhu] are shown in Fig. 9 as an example of a *multi-call bout*. From Fig. 9 (b), it can be confirmed that the natural *multi-call bout*'s tendency in which the *call* amplitude gradually decreases is reflected in the synthesized laughter.

## 5 Naturalness Evaluation

In this section, the overall naturalness of laughter accompanying the speech sounds is evaluated.

To evaluate the overall naturalness, synthesized laughter was connected to the speech sounds. Furthermore, to confirm the effectiveness of considering the laughter context, synthesized laughter considering the context was compared to the laughter synthesized without context.

### 5.1 Stimulus

The conditions for synthesizing laughter were as follows:

- baseline (BL)
  The laughter was synthesized using the context label considering only current *calls*.
- baseline plus context A (BL+A)
  The laughter was synthesized using the context label considering the preceding and following segments.
- baseline plus context A and B (BL+AB)
  The laughter was synthesized using the context that considers more global contextual factors in addition to context A.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2018.2813381, IEEE Transactions on Affective Computing



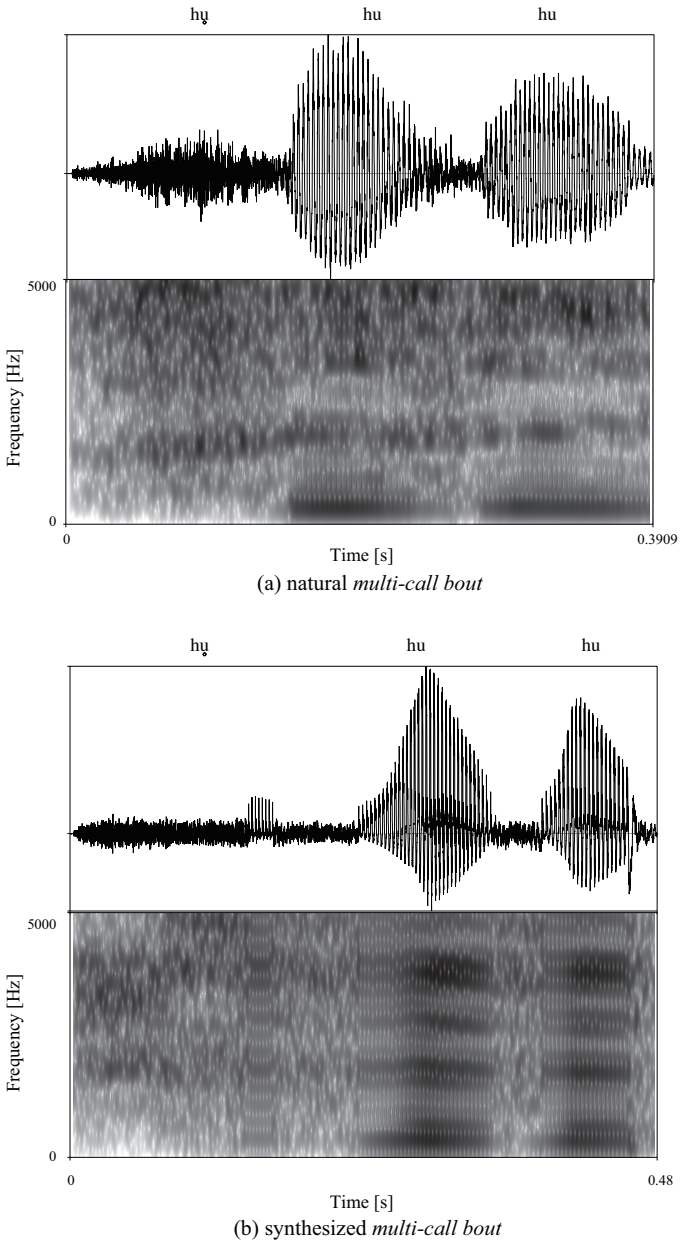



Fig. 9. Example of *multi-call bout*.

- analysis-synthesis (RESYN)
  The laughter was re-synthesized from natural laughter in the test utterance via the analysis-synthesis.

The speech sounds were connected to the laughter synthesized under these conditions. The speech sounds were re-synthesized via analysis-synthesis, to avoid the influence of the quality of the speech part on the naturalness evaluation of the laughter part. The number of stimuli was 45 for each condition.

### 5.2 Experimental conditions

The stimuli for the experiments were created under four conditions. Hence, the number of stimuli was 4 (conditions) $\times$ 45 (test utterances) $=$ 180. The number of *single-call bout* and *multi-call bout* are 9 and 36, respectively. The number of *call* of *multi-call bouts* is distributed within a range of 2 to 5.

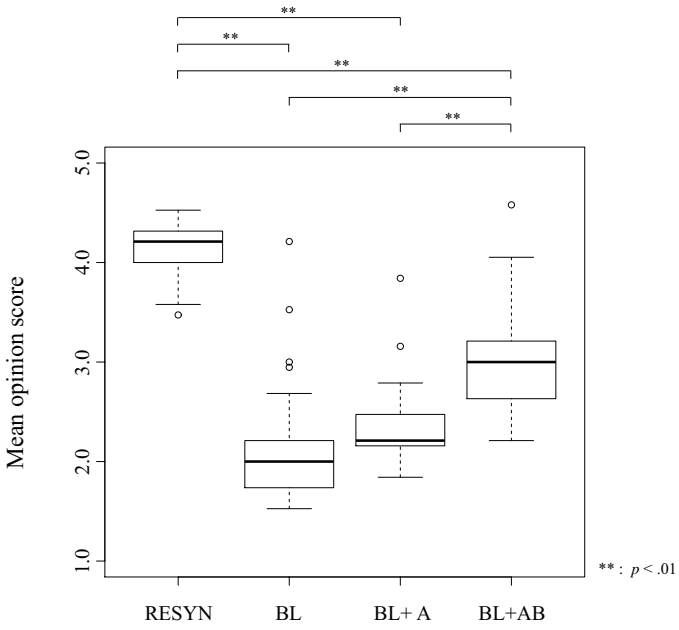



Fig. 10. The distribution of the naturalness score.

Seventeen undergraduate male students and two undergraduate female students from Utsunomiya University, all of whom were Japanese native speakers and had no special knowledge about speech science, voluntarily participated in the experiment. The evaluation was performed via a web-based interface. The birthplace of the subjects was distributed over a wide area of Japan. The subjects evaluated the naturalness of each stimulus on a 5-point scale: "1: unnatural," "2: somewhat unnatural," "3: neither ("どちらともいえない")," "4: somewhat natural," and "5: natural." Naturalness was defined as "the degree of matching of segmental/prosodic features of speech and laughter parts." This also considers the degree of discordancy of paralinguistic information perceived from the speech or laughter parts.

Subjects were asked to listen to the stimuli using headphones in a quiet room. They were instructed to listen to each stimulus only once.

### 5.3 Results

As a result of the naturalness evaluation, the distribution of mean opinion score (MOS) is shown in Fig. 10. The boxplot represents the mean opinion score averaged over all test utterances. The averaged mean opinion score of BL, BL+A, BL+AB and RESYN is 2.10, 2.32, 3.01 and 4.13, respectively. A one-way ANOVA test revealed a significant main effect of the synthesized conditions ($F(3, 176) = 220.8, p < .01$). The result of multiple comparison by Tukey's HSD test revealed a significant difference between RESYN and other conditions ($p < .01$) and between BL+AB and other conditions ($p < .01$). These results confirmed that the overall naturalness was improved by defining the context that has more global information, such as the laughter and *call* position.

Distributions of averaged MOS for each laughter structure are shown in Fig. 11. The averaged MOS of *single-call bouts* and *multi-call bouts* was 3.15 and 2.82, respectively. A

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2018.2813381, IEEE Transactions on Affective Computing



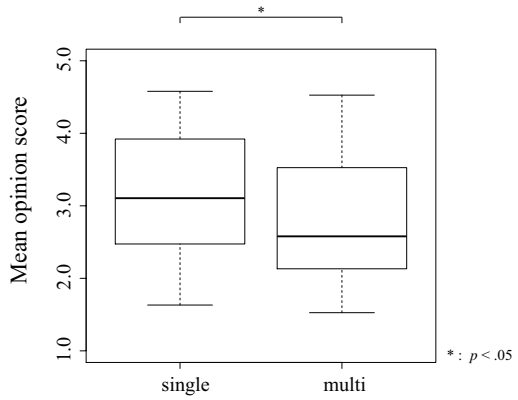



Fig. 11. The distribution of the mean opinion score for each laughter structure.

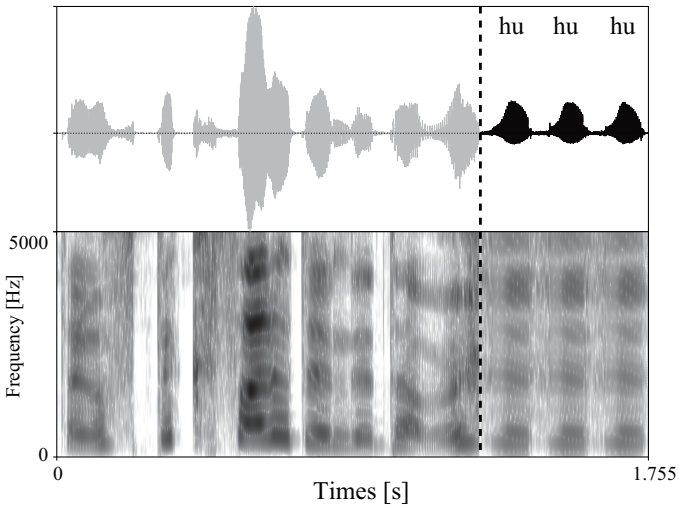
$t$-test revealed a significant effect of the laughter structure ($t(58.4) = -2.14, p < .05$). This result implies the possibility that the structure of laughter affected the naturalness. However, there is also a possibility that the different speech sounds that were connected to the synthesized laughter affected the naturalness. In order to assess the structural effect on the naturalness in more detail, it will be necessary to carefully control the sentences to be connected.

An example of naturalness improvement is shown in Fig. 12. The figure shows the waveform and spectrum of an utterance with laughter. The gray part of the waveform represents the speech sounds, and the black part represents the laughter. Fig. 12 (a) is the synthesized laughter without considering the context. Synthesized laughter is perceived as very unnatural because all *calls* have the same acoustic features. Fig. 12 (b) is synthesized laughter that considers the preceding and succeeding segments. The acoustic feature of the first *call* is different from the others, and the naturalness is somewhat improved. Fig. 12 (c) is the synthesized laughter that considers more global contextual factors. Different acoustic features are generated depending on the *call* position and sound like human laughter. From this, we determined that the *call* position contributes greatly to improving the naturalness.

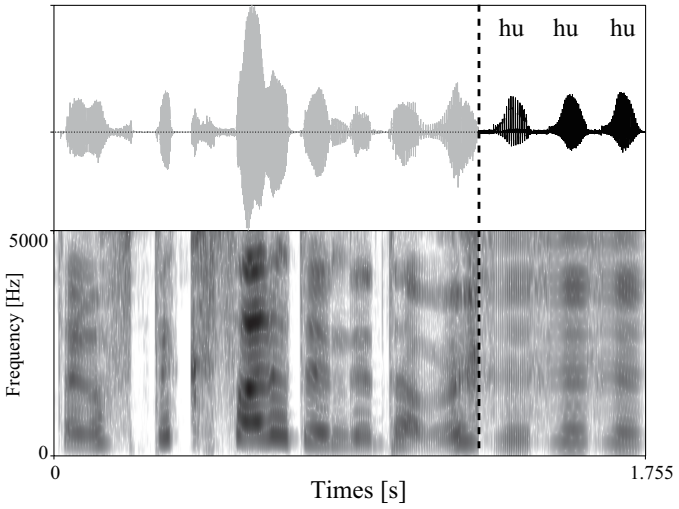
There were also a few utterances for which the naturalness was not improved, even by using the extended context. These are utterances where the naturalness of the laughter itself is low. An example of poorly synthesized laughter is shown in Fig. 13. In this example, [hu] has the same acoustic features that are repeatedly being synthesized. Such monotonous repetition of the same sound causes the degradation of naturalness, known as the "machine gun effect." A possible cause of such a phenomenon is the sparseness of the training data. We observed that some nodes stopped growing at an early stage of the decision tree-based context clustering during the model construction. A potential solution is to add more fine-grained context that describes the position of the *calls* or to increase the training data.

## 6 Conclusions

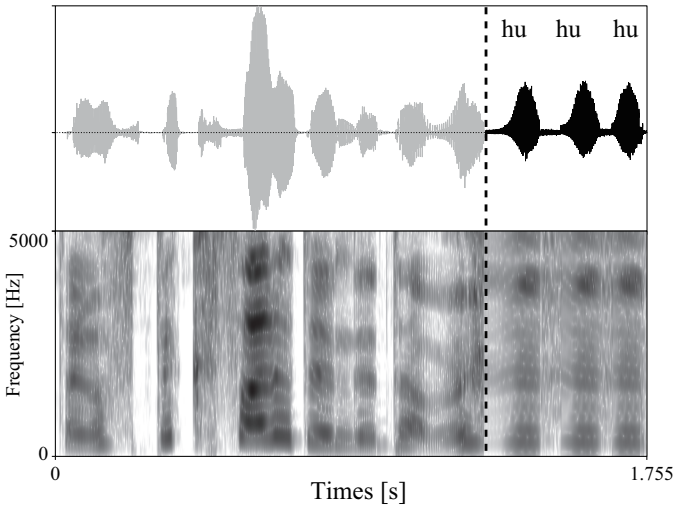
In this paper, conversational laughter, which typically accompanies speech sounds, was synthesized via a statistical model-based speech synthesis framework using spon-




(a) Synthesized with BL




(b) Synthesized with BL+A




(c) Synthesized with BL+AB

Fig. 12. Example that the naturalness of entire speech was increased.

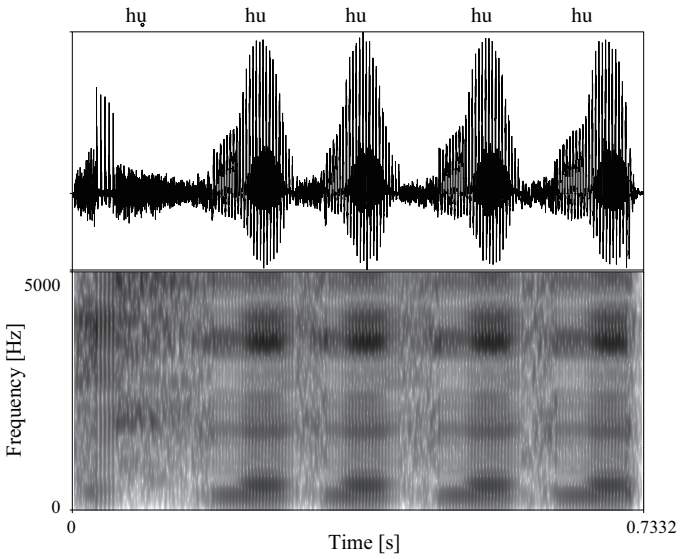



Fig. 13. Example of low naturalness.

taneous speech corpora. To prepare the training data, the transcription of *calls* was annotated for laughter with speech sounds in two spontaneous speech corpora: the UU Database and OGVC. In terms of contextual factors, the transcription of the preceding and succeeding segments, the *call* position, the number of *calls* and the *bout* position were defined. Laughter was synthesized using the defined context and the framework of HMM-based speech synthesis. The synthesized laughter exhibited acoustic features resembling natural laughter.

To confirm the influence of the contextual factors on naturalness, a subjective evaluation was performed. The naturalness was evaluated for utterances that include both speech sound and laughter. The naturalness of the entire utterance was improved by using the defined contextual factors in this study. From this, we concluded that it is necessary to use the appropriate contextual factors to synthesize natural conversational laughter.

This study was conducted as a first step in laughter synthesis that considers the context of natural conversation. Its effectiveness was confirmed under very limited conditions. As future tasks, it is necessary to consider the gender difference of laughter, to improve the naturalness by increasing the annotated training data, and to define further contextual factors. Additionally, paralinguistic information perceived from laughter was not evaluated in this study. Because laughing is a medium of conveying the speaker's emotional and mental states, paralinguistic information perceived from laughter may play an important role in communication. Therefore, future tasks should include examining the influence of laughter on perceived paralinguistic information and the context required to control paralinguistic information.

## ACKNOWLEDGMENTS

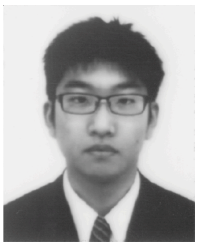
The authors thank Dr. Yoshiko Arimoto for her enthusiastic discussion on corpus use and laughter annotation. The authors also thank Dr. Takashi Nose for his technical advice on HMM-based synthesis.

This work was supported by JSPS KAKENHI Grant Number 26280100.

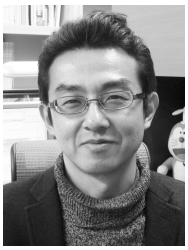
## REFERENCES

[1] R.R. Provine, "Laughter: a scientific investigation," Penguin Books, 2001.
[2] S. Dupont, H. Çakmak, W. Curran, T. Dutoit, J. Hofmann, G. MkKeown, O. Pietquin, T. Platt, W. Ruch and J. Urbain, "Laughter research: a review of the ILHAIRE project," Toward Robotic Socially Believable Behaving Systems - Volume I, 2016, pp. 172–181.
[3] K. P. Truong and D.A. van Leeuwen, "Automatic detection of laughter," Proc. Interspeech, 2005, pp. 485–488.
[4] M. T. Knox and N. Mirghafori, "Automatic laughter detection using neural networks," Proc. Interspeech2007, 2007, pp. 2973–2976.
[5] T. Neuberger, A. Beke and M. Gósy, "Acoustic analysis and automatic detection of laughter in Hungarian spontaneous speech," Proc. ISSP, 2014, pp. 281–284.
[6] S. Sundaram and S. Narayanan, "Automatic acoustic synthesis of human-like laughter," *J. Acoust. Soc. America*, **121**, 2007, pp. 527–535.
[7] A. T. Sathya, K.K. Sudheer and B. Yegnanarayana, "Synthesis of laughter by modifying excitation characteristics," *J. Acoust. Soc. America*, **133**, 2013, pp. 3072–3082.
[8] J. Oh and G. Wang, "Lolol: laugh out loud on laptop," Proc. International Conference on New Musical Instruments, 2013.
[9] J. Trouvain and M. Schroder, "How not to add laughter to synthetic speech," Proc. Workshop Affective Dialog Systems, 2004, pp. 229–232.
[10] E. Lasarcyk and J. Trouvain, "Imitating conversational laughter with an articulatory speech synthesis," Proc. *Interdisciplinary Workshop Phonetics of Laughter*, 2007, pp. 43–48.
[11] J. A. Bachorowski, M. J. Smoski and M. J. Owren, "The acoustic features of human laughter," *J. Acoust. Soc. America*, **110**, 2001, pp. 1581–1597.
[12] M. Schröder, "Experimental study of affect bursts," Proc. *ISCA Workshop "Speech and Emotion"*, 2000, pp. 132–137.
[13] W. Ruch and P. Ekman, "The expressive pattern of laughter," *Emotion, qualia and consciousness*, 2001, pp. 426–443.
[14] J. Trouvain, "Phonetic aspects of "speech-laughs"," Proc. *the 2nd Conference on Orality & Gestuality (ORANGE)*, 2001, pp. 634–639.
[15] K.J. Kohler, "'Speech-laugh', 'speech-laugh', 'laughter' and their sequencing in dialogic interaction," Phonetica, **65**, 2008, pp. 1–18.
[16] K. El Haddad, I. Torre, E. Gilmartin, H. Çakmak, S. Dupont, T. Dutoit and N, Campbell, "Introducing AmuS: The amused speech database," Proc. *International Conference on Statistical Language and Speech Processing*, 2017, pp. 229–240.
[17] J. Urbain, H. Cakmak and T. Dutoit, "Development of HMM-based acoutic laughter synthesis," *Interdisciplinary Workshop Laughter and Other Non-Verbal Vocalisations in Speech*, 2012.
[18] J. Urbain, H. Cakmak and T. Dutoit, "Evaluation of HMM-based laughter synthesis," Proc. ICASSP 2013, 2013, pp. 7835–7839.
[19] J. Urbain, H. Çakmak, A. Charlier, M. Denti and T. Dutoit, "Arousal-driven synthesis of laughter," IEEE Journal of Selected Topics in Signal Processing, **8**, 2014, pp. 273–284.
[20] K. El Haddad, S. Dupont, J. Urbain and T. Dutoit, "Speech-laugh: an HMM-based approach for amused speech synthesis," Proc. ICASP, 2015, pp. 4939–4943.
[21] K. El Haddad, H. Çakmak, S. Dupont and T. Dutoit, "Breath and repeat: An attempt at enhancing speech-laugh synthesis quality." Proc. *Signal Processing Conference*, 2015, pp. 355–358.
[22] H. Çakmak, K. El Haddad and T. Dutoit, "Audio-visual laughter synthesis system," Proc. the 4th Interdisciplinary Workshop on Laughter and Other Non-verbal Vocalisations in Speech, 2015, pp. 11–13.
[23] T. Yoshimura, K. Tokuda, T. Kobayashi and T. Kitamura, "Simultaneous modeling of spectrum pitch and duration in HMM-based speech synthesis," *IEICE Trans. Information & Systems*, **83**, 1999, pp. 2347–2350.
[24] K. Laskowski and S. Burger, "Analysis of the occurrence of laughter in meetings," Proc. Interspeech, 2007, pp. 1258–1261.
[25] K. P. Truong and J. Trouvain, "On the acoustics of overlapping laughter in conversational speech," Proc. Interspeech, 2012, pp. 851–854.

[26] H. Mori, T. Satake, M. Nakamura and H. Kasuya, "Constructing a spoken dialog corpus for studying paralinguistic information in expressive conversation and analyzing its statistical/acoustic characteristics," Speech Communication. **53**, 2011, pp. 36–50.
[27] 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*, **33**, 2012, pp. 359–369.
[28] J. Trouvain, "Segmenting phonetic units in laughter," Proc. *15th Int'l Congress of Phonetic Sciences (ICPhS)*, 2003, pp. 2793–2796.
[29] J. Urbain, T. Dutoit, A. Moinet, R. Niewiadomski, C. Pelachaud, B. Picart, J. Tilmanne and J. Wagner, "The AVLaughterCycle database,", Proc. *7th Int'l Conf. Language Resources and Evaluation (LREC 2010)*, 2010, pp. 2996–3001.
[30] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [Computer program]," `http://www.praat.org/`.
[31] H. Mori, "Toward morphological classification of affect bursts," Proc. *Acoust. Soc. Japan 2015 Spling Meeting*, 2015, pp.405–406 (in Japanese).
[32] H. Bunt, J. Alexandersson, J.-W. Choe, A. Fang, K. Hasida, V. Petukhova, A Popescu-Belis and D. Traum, "ISO 24617-2: A semantically-based standard for dialog annotation," Proc. *8th Int'l Conf. Language Resources and Evaluation (LREC 2012)*, 2012, pp. 430–437.
[33] R. Gardner, "When listeners talk: Response tokens and listener stance," Amsterdam: John Benjamins, 2001.
[34] M. Mehu, "Smiling and laughter in naturally occurring dyadic interactions: Relationship to conversation, body contacts, and displacement activities," *Human Ethology Bulletin*, **26**, 2011, pp. 10–28.
[35] K. Tokuda, H. Zen, J. Yamagishi, T. Masuko, S. Sako, A.W. Black and T. Nose, "The HMM-based speech synthesis system (HTS)," `http://hts.sp.nitech.ac.jp/`.
[36] J. Yamagishi, M. Tamura, T. Masuko, K. Tokuda and T. Kobayashi, "A context clustering technique for average voice models," *IEICE Trans. Information and Systems*, **E86-D**, 2003, pp. 534–542.



**Tomohiro Nagata** received B.E. and M.E. degrees from Utsunomiya University in 2011 and 2013, respectively. He is currently working toward the Doctor's degree. His research interests include spontaneous speech synthesis, affective computing, human-machine interaction, and social signals processing.



**Hiroki Mori** received B.E., M.E. and Ph.D. degrees from Tohoku University, in 1993, 1995 and 1998, respectively. He was with the Graduate School of Engineering, Tohoku University in 1998. He is an associate professor of Utsunomiya University. His research interests include verbal and nonverbal communication, corpus design and development for speech science and conversation analyses, analysis and modeling of prosody and paralanguage, and expressive speech synthesis. He is a member of the IEICE, the Acoustical Society of Japan, IPSJ, and ISCA.