On the Correlation between Perceptual and Contextual Aspects of Laughter in Meetings

Kornel Laskowski and Susanne Burger

interACT, Carnegie Mellon University, Pittsburgh PA, USA
kornel|sburger@cs.cmu.edu

ABSTRACT

We have analyzed over 13000 bouts of laughter, in over 65 hours of unscripted, naturally occurring multiparty meetings, to identify discriminative contexts of voiced and unvoiced laughter. Our results show that, in meetings, laughter is quite frequent, accounting for almost 10% of all vocal activity effort by time. Approximately a third of all laughter is unvoiced, but meeting participants vary extensively in how often they employ voicing during laughter. In spite of this variability, laughter appears to exhibit robust temporal characteristics. Voiced laughs are on average longer than unvoiced laughs, and appear to correlate with temporally adjacent voiced laughter from other participants, as well as with speech from the laugher. Unvoiced laughter appears to occur independently of vocal activity from other participants.

1. INTRODUCTION

In recent years, the availability of large multiparty corpora of naturally occurring meetings [2] [7] [3] has focused attention on previously little-explored, natural human-human interaction behaviors [17]. A non-verbal phenomenon belonging to this class is laughter, which has been hypothesized as a means of affecting interlocutors, as well as a signal of various human emotions [14].

In our previous work, we produced an annotation of perceived emotional valence in speakers in the ISL Meeting Corpus [10]. We showed that instances of isolated laughter were strongly predictive of positive valence, as perceived in participants by external observers who had not participated in the meetings. In a subsequent multi-site evaluation of automatic emotional valence classification within the CHIL project [21], we found that transcribed laughter is in general much more indicative of perceived positive valence than any other grouping of spectral, prosodic, contextual, or lexical features. Three-way classification of speaker contributions into negative, neutral and positive valence classes (with neutral valence accounting for 80% of the contributions), using the presence of transcribed laughter as the only feature, resulted in an accuracy of 91.2%. The combination of other features led to an accuracy of only 87% (similar results were produced on this data by [12]). A combination of all features, including the presence of transcribed laughter, produced an accuracy of 91.4%, only marginally better than transcribed laughter alone.

Although these results show that the presence of laughter, as detected by human annotators, was the single most useful feature for automatic valence classification, laughter and positive valence are not completely correlated in the ISL Meeting Corpus. We are ultimately interested in the ability to determine, automatically, whether a particular laugh conveys information about the laughter's valence to an outside observer. The current work is a preliminary step in that effort, in which we characterize laughter along two separate dimensions. First, we determine whether each laugh is voiced or unvoiced. Previous work with this distinction in other domains has shown that voiced laughter may be used strategically in conversation [14].

Second, we attempt to characterize the temporal context of voiced and unvoiced laughter within the multiparticipant vocal activity on-off pattern of a conversation. In the current work, we are interested exclusively in *text-independent* context, which allows us to ignore specific lexical and/or syntactic phenomena having bearing on the occurrence of laughter. The study of laughter in sequence with spontaneous speech has been treated by conversation analysis [8]; the latter has offered solutions for both transcribing and investigating multiparticipant laughter [9] [4], but it has not produced quantitative descriptions or means of obtaining them. Laughter has also been shown to evoke laughing in listeners [15], in this way differing from speech. In particular, laughers do not take turns laughing in the same way that speakers take turns speaking. Vocal activity context therefore appears to provide important cues as to whether ongoing vocal activity is laughter or speech [11]. In the current work, we attempt to determine whether context also disambiguates between voiced and unvoiced laughter.

2. DATA

To study the pragmatics of laughter, we use the relatively large ICSI Meeting Corpus [7]. This corpus consists of 75 unscripted, naturally occurring meetings, amounting to over 71 hours of recording time. Each meeting contains between 3 and 9 participants wearing individual head-mounted microphones, drawn from a pool of 53 unique speakers (13 female, 40 male); several meetings also contain additional participants without microphones.

In this section, we describe the process we followed to produce, for each meeting and for each participant: (1) a *talk spurt* segmentation; (2) a voiced *laugh bout* segmentation; and (3) an unvoiced laugh bout segmentation. A talk spurt is defined [18] as a contiguous interval of speech delineated by non-speech of at least 500 ms in duration; laugh bouts, as used here, were defined in [1].

We note that each meeting recording contains a ritualized interval of read speech, a subtask referred to as Digits, which we have analyzed but excluded from the final segmentations. The temporal distribution of vocal activity in these intervals is markedly different from that in natural conversation. Excluding them limits the total meeting time to 66.3 hours.

2.1. Talk Spurt Segmentation

Talk spurt segmentation for the meetings in the ICSI corpus was produced using word-level forced alignment information, available in a corpus of auxiliary annotations known as in the ICSI Dialog Act Corpus [19]. While 500 ms was used as the minimum inter-spurt duration in [18], we use a 300 ms threshold. This value has recently been adopted for the purposes of building speech activity detection references in the NIST Rich Transcription Meeting Recognition evaluations.

2.2. Selection of Transcribed Laughter Instances

Laughter is transcribed in the ICSI Meeting Corpus orthographic transcriptions in two ways. First, discrete events are annotated as `VocalSound` instances, and appear interspersed among lexical items. Their location among such items is indicative of their temporal extent. We show a small subset of `VocalSound` types in Table 1. As can be seen, the `VocalSound` type `laugh` is the most frequently annotated non-verbal vocal production. The second type of laughter-relevant annotation found in the corpus, `Comment`, describes events of extended duration which were not localized between specific lexical items. In particular, this annotation covers the phenomenon of "laughed speech" [13] We list the top five most frequently occurring `Comment` descriptions pertaining to laughter in Table 2. As with `VocalSound` descriptions, there is a large number of very rich laughter annotations each of which occurs only once or twice.

Table 1: Top 5 most frequently occurring `VocalSound` types in the ICSI Meeting Corpus, and the next 5 most frequently occurring types relevant to laughter.

Freq Rank	Token Count	`VocalSound` Description	Used Here
1	11515	`laugh`	√
2	7091	`breath`	
3	4589	`inbreath`	
4	2223	`mouth`	
5	970	`breath-laugh`	√
11	97	`laugh-breath`	√
46	6	`cough-laugh`	√
63	3	`laugh, "hmmph"`	√
69	3	`breath while smiling`	
75	2	`very long laugh`	√

The description attributes of both the `VocalSound` and `Comment` tags, as produced by the ICSI transcribers, appear to be largely ad hoc, and reflect practical considerations during an annotation pass whose primary aim is to produce an orthographic transcription. In the current work, we used the descriptions only to select and possibly segment laughter, and afterward ignored them.

We identified 12635 transcribed `VocalSound` laughter instances, of which 65 were ascribed to farfield channels. These were excluded from our subsequent analysis, because the ICSI MRDA Corpus includes forced alignment information for nearfield channels only. We also identified 1108 transcribed `Comment` laughter instances, for a total of 13678 transcribed laughter instances in the original ICSI transcriptions.

2.3. Laugh Bout Segmentation

Our strategy for producing accurate endpoints for the laughter instances identified in Subsection 2.2. consisted of a mix of automatic and manual methods. Of the 12570 non-farfield `VocalSound` instances, 11845 were adjacent on both the left and the right to either a time-stamped utterance boundary, or a lexical item. We were thus able to automatically deduce start and end times for 87% of the laughter instances treated in this work. Each automatically segmented instance was inspected by at least one of our annotators; disagreement as to the presence of laughter was investigated by both authors together, and in a small handful of cases ($<$3%), when there appeared to be ample counter-evidence,

Table 2: Top 5 most frequently occurring `Comment` descriptions containing the substring "`laugh`" or "`smil`". We listened to all utterances whose transcription contained these descriptions, but portions were included in our final laugh bout segmentation only if the utterances contained laughter (in particular, intervals annotated with "`while smiling`" were not automatically included.)

Freq Rank	Token Count	`Comment` Description
2	980	`while laughing`
16	59	`while smiling`
44	13	`last two words while laughing`
125	4	`last word while laughing`
145	3	`vocal gesture, a mock laugh`

we discarded the instance.

The remaining 725 non-farfield `VocalSound` instances were not adjacent to an available timestamp on either or both of the left and the right. These instances were segmented manually, by listening to the entire utterance containing them[1]; since the absence of a timestamp was due mostly to a transcribed, non-lexical item before and/or after the laughter instance, segmentation consisted of determining a boundary between laughter and, for example, throat-clearing. We did not attempt to segment one bout of laughter from another.

All of the 1108 `Comment` instances were segmented manually. This task was more demanding than manual segmentation of `VocalSound` laughter. We were guided by the content of the `Comment` description, which sometimes provided cues as to the location and extent of the laugh (ie. `last two words while laughing`). We placed laughter start points where the speaker's respiratory function was perceived to deviate from that during speech; in determining the end of laughter, we included the audible final recovery inhalation which often accompanies laughter [6].

A quarter of the manually segmented `Comment` instances were checked by the second author. The final laugh bout segmentation was formed by combining the automatically segmented `VocalSound` laughter, the manually segmented `VocalSound` laughter, and the manually segmented `Comment` laughter; due to overlap, a small number of laugh segments were merged, to yield 13259 distinct segments.

We note that the resulting laugh bout segmentation differs from that recently produced for the same corpus in [20] at least in the number of bouts. The authors of [20] report using only 3574 laughter segments; it is unclear how these were selected, except that the authors state that they excluded speech and inaudible laughter after listening to all the ICSI-transcribed instances.

2.4. Laugh Bout Voicing Classification

In the last preprocessing task, we classified each laughter instance as either voiced or unvoiced. Our distinction of voiced versus unvoiced was made according to [14]. Voiced laughter, like voiced speech, occurs when the energy source is quasi-periodic vocal-fold vibration. This class includes melodic, "song-like" bouts, as well as most chuckles and giggles. Unvoiced laughter results from fricative excitation, and is analogous to whispered speech. It includes open-mouth, pant-like sounds, as well as closed-mouth grunts and nasal snorts. Additionally, we decided that bouts consisting of both voiced and unvoiced calls should receive the voiced label when taken together. Instances of "laughed speech" were automatically assigned the voiced label.

Voicing classification was performed by two annotators, who were shown all the close-talk channels per meeting in parallel, for all segmented instances of laughter from Subsection 2.3. with their original ICSI `VocalSound` or `Comment` annotation. For each instance, they were able to select and listen to the foreground channel, the same time interval on any of the remaining channels, and the temporal context on the foreground and remaining channels[2]. Annotators were encouraged to insert ad-hoc comments in addition to their voiced/unvoiced label.

58 meetings were labeled by one of two annotators, 14 were labeled by one annotator and were then checked by the other, and 3 were independently labeled by both annotators. Finally, all laughter instances which received a comment during classification were subsequently listened to by both authors.

Interlabeler agreement on the classification of voicing was computed using the three meetings which were labeled independently by both annotators, `Bmr016`, `Bmr018` and `Bmr019`. Agreement was between 88% and 91%, and chance-corrected κ-values [5] for the three meetings fell in the range 0.76-0.79. This is lower than we expected, having had assumed that assessment of voicing is not a very subjective task. Inspection of the disagreements revealed that they occurred for `VocalSound` instances whose endpoints had been inferred from inaccurate forced alignment timestamps of the adjacent words. In many cases the annotators had labeled the presence of laugher speech inside laugh bouts; since commented cases were revisited by both authors, a portion of the disagreement cases were resolved. In the remainder, we kept the voicing label

of that of our two annotators who had worked on the larger number of meetings.

In a final verification effort (following the publication of [11]), the second author checked the voicing label and boundaries of every instance, which led to a change of voicing label in 942 instances. Endpoints were modified in 306 instances, and 50 instances were removed. 11961 laughter segments (90% of the total) were not modified.

3. ANALYSIS

In this section, we describe the results of our investigations into the differences between voiced and unvoiced bouts of laughter, in terms of total time spent in laughter, bout duration, and multiparticipant vocal activity context.

3.1. Quantity

Of the 13209 bouts identified in the previous section, 33.5% were labeled as unvoiced while 66.5% were labeled as voiced.

We were also interested in the total proportion of time spent laughing. For each participant, and for each of voiced and unvoiced laughter categories, we summed the time spent laughing, and normalized this quantity by the total time of those meetings which were attended by that participant. Since a given participant may not have been present for the entirety of each meeting, the results we show represent ceiling numbers.

We found that the average participant spends 0.98% of their total meeting time in voiced laughter, and 0.35% of their total meeting time in unvoiced laughter. For contrast, in [11], we showed that the average participant spends 14.8% of their total meeting time on speaking. It can be seen in Figure 1, that the time spent laughing and the proportion of voiced to unvoiced laughter vary considerably from participant to participant.

Visually, there appears to be only a very weak correlation between the amount of individual participants' voiced laughter and their amount of unvoiced laughter. The majority of participants appears capable of both modes of laughter production.

3.2. Duration

Next, we analyze the durations of bouts to determine whether there is a difference for voiced and unvoiced laughter. The results are shown in Figure 2. Although bout durations vary much less than talkspurt durations, the modes for all three of voiced laughter bouts, unvoiced laughter bouts, and talkspurts fall between approximately 1 second and 1.5 seconds. On average, voiced bouts appear to be slightly longer than unvoiced bouts.

Figure 1: Proportion of total recorded time per participant spent in voiced and in unvoiced laughter. Participants are shown in order of ascending proportion of voiced laughter.

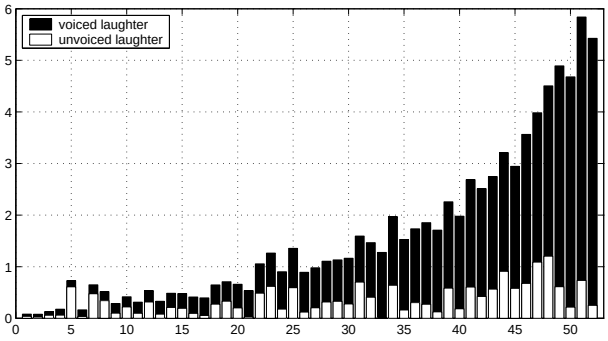
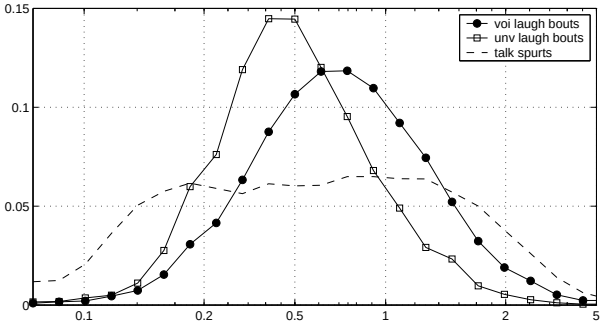


Figure 2: Normalized distributions of duration in seconds for voiced laughter bouts, unvoiced laughter bouts, and talk spurts.

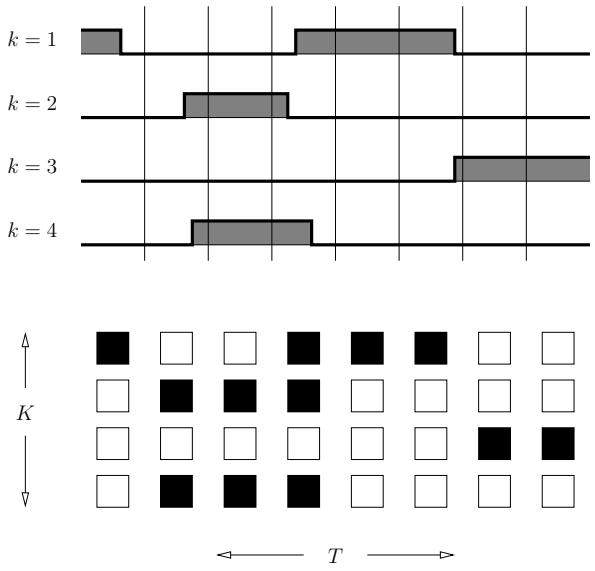


3.3. Interaction

Finally, we attempt to analyze local (short-time) differences in conversational context for voiced and unvoiced laughter. We are interested in whether a choice of voicing during laughter has a significant impact on the kinds of vocal interaction which immediately follow, or whether preceding interaction has a significant impact on whether a laugher will employ voicing. For each bout, we study only the *vocal interaction* context; in particular, we ignore the specific words spoken and focus only on whether each participant is silent, laughing (in either voiced or unvoiced mode), speaking, or both.

We accomplish this analysis in a time-synchronous fashion as follows, accumulating statistics over all meetings in the ICSI corpus. For every meeting, we begin with the reference on-off patterns corresponding to speech (Subsection 2.1.), for each of K participants. We discretize these patterns using 1-second non-overlapping windows, as shown in Figure 3. We do the same with the

Figure 3: Discretization of multichannel speech (or voiced or unvoiced laughter) segmentation references using a non-overlapping window size of 1 second. When participant k is vocalizing for more that 10% of the duration of frame t, the cell (t, k) is assigned the value 1 (black); otherwise it is assigned 0 (white).

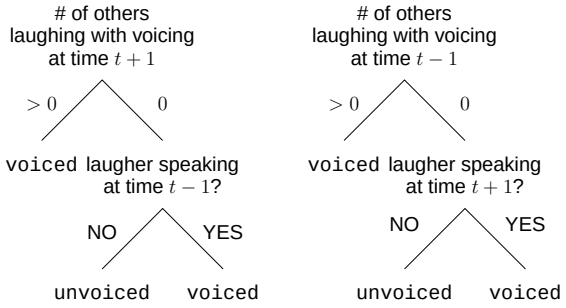


on-off voiced laughter segmentation and the on-off unvoiced laughter segmentation, producing for each meeting 3 binary-value matrices of size $K \times T$, where T is the number of 1-second non-overlapping frames.

For each meeting in the corpus, we inspect the reference matrices described above to determine whether participant k is laughing at time t. If so, we collect 11 features describing the conversational context of cell (t, k): binary-valued features whether participant k, the laugher, is speaking at times $t - 1$ and $t + 1$; the number of *other* participants speaking at times $t - 1$, t, and $t + 1$; the number of *other* participants laughing with voicing at times $t - 1$, t, and $t + 1$; and the number of *other* participants laughing without voicing at times $t - 1$, t, and $t + 1$.

We wish to analyze interactional aspects during laughter initiation, laughter termination, and laughter continuation, separately. To determine whether voiced and unvoiced bouts of laughter differ in terms of their short-time conversational context during initiation, we take all laughter frames which are preceded immediately by not-laughter, from all meetings and all participants, and measure the statistical significance of association between the 11 features we collect and the binary voiced or unvoiced attribute of the laughter frame. Although a standard χ^2-test is possible, we choose instead to determine whether *the voicing attribute during laughter initiation is predictable from context.* We do this by inferring the parameters of a decision tree [16], followed by pruning. Tree nodes which survive pruning are statistically significant; structurally, the decision tree can be thought of as a nested χ^2-test.

Figure 4: Automatically identified decision trees for detecting voiced versus unvoiced laughter based on multiparticipant vocal activity context; laughter initiation context on left, termination context on right.



We repeat the same procedure for both laughter termination and for laughter continuation. Our experiment identifies no significant distinction between the conversational context of voiced and unvoiced laughter continuation. That is, there appears to be no significant difference in the kinds of interactions that occur during voiced and unvoiced laughter, nor does voicing during laughter appear to have a significant impact on the interactions that occur during it.

For initiation and termination frames, we show the inferred classification trees in Figure 4. It is surprising that the two trees are symmetric. In attempting to predict the voicing of a frame which initiates a bout, the most useful contextual feature, of those studied here, is whether others will be laughing at $t + 1$; in other words, voicing during laughter is significantly more likely to cause at least one other participant to subsequently laugh with voicing. In attempting to predict the voicing of a frame which terminates a bout, the most useful feature is whether others were laughing with voicing at $t - 1$. Again, this suggests that voicing during laughter is much more likely if others were previously laughing with voicing. The next most useful feature is whether the laugher is speaking before or after laughing. For bout-initiating frames, if no others subsequently laugh with voicing and the laugher was not previously speaking, they are much more likely to be lau-

ghing without voicing, and symmetrically for bout-terminating frames.

4. CONCLUSIONS

We have produced a complete voiced and unvoiced laughter segmentation for the entire ICSI Meeting Corpus, including isolated instances as well as instances of laughter co-occurring with the laugher's speech. We have shown that on average, voiced laughter accounts for 66.5% of all observed laughter in this corpus, but that participants vary widely in their use of voicing while laughing. Most importantly, we have shown that in spite of inter-participant differences, voiced and unvoiced laughs are correlated with different vocal interaction contexts. Voiced laughter seems to differ from unvoiced laughter in that voiced laughter from other participants follows its initiation and precedes its termination. Voiced laughter also seems more interdependent with the laugher's speech; in cases where laughter follows speech or precedes laugher's speech, it is more likely to be voiced than unvoiced.

5. ACKNOWLEDGMENTS

This work was funded in part by the European Union under the integrated project CHIL (IST-506909), Computers in the Human Interaction Loop (`http://chil.server.de`) [21].

6. REFERENCES

[1] Bachorowski, J.-A., Smoski, M., Owren, M. 2001. The acoustic features of human laughter. *J. Acoustical Society of America* 110(3), 1581–1597.

[2] Burger, S., MacLaren, V., Yu, H. 2002. The ISL Meeting Corpus: The impact of meeting type on speech style. *Proc. ICSLP* Denver CO, USA. 301–304.

[3] Carletta, J. e. a. 2005. The AMI Meeting Corpus: A pre-announcement. *Proc. MLMI (Springer Lecture Notes in Computer Science 3869)* Edinburgh, UK. 28–39.

[4] Chafe, W. 2003. *Linguistics, Language, and the Real World: Discourse and Beyond* chapter Laughing while Talking, 36–49. Georgetown University Press.

[5] Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20, 37–46.

[6] Filippelli, M. e. a. 2001. Respiratory dynamics during laughter. *J. Applied Physiology* 90(4), 1441–1446.

[7] Janin, A. e. a. 2003. The ICSI Meeting Corpus. *Proc. ICASSP* Hong Kong, China. 364–367.

[8] Jefferson, G. 1979. *Everyday Language: Studies in Ethnomethodology* chapter A Technique for Inviting Laughter and its Subsequent Acceptance Declination, 79–96. Irvington Publishers.

[9] Jefferson, G. 1985. *Handbook of discourse analysis* volume 3 chapter An exercise in the transcription and analysis of laughter, 25–34. Academic Press.

[10] Laskowski, K., Burger, S. 2006. Annotation and analysis of emotionally relevant behavior in the ISL Meeting Corpus. *Proc. LREC* Genoa, Italy.

[11] Laskowski, K., Burger, S. 2007. Analysis of the occurrence of laughter in meetings. *Proc. INTERSPEECH (to appear)* Antwerpen, Belgium.

[12] Neiberg, D., Elenius, K., Laskowski, K. 2006. Emotion recognition in spontaneous speech using GMMs. *Proc. INTERSPEECH* Pittsburgh PA, USA. 809–812.

[13] Nwokah, E., Hsu, H.-C., Davies, P., Fogel, A. 1999. The integration of laughter and speech in vocal communication: a dynamic systems perspective. *J. Speech, Language & Hearing Research* 42, 880–894.

[14] Owren, M., Bachorowski, J.-A. 2003. Reconsidering the evolution of nonlinguistic communication: The case of laughter. *J. Nonverbal Behavior* 27(3), 183–199.

[15] Provine, R. 1992. Contagious laughter: Laughter is a sufficient stimulus for laughs and smiles. *Bull. Psychonomic Society* (30), 1–4.

[16] Quinlan, J. 1986. Induction of decision trees. *Machine Learning* 1(1), 81–1006.

[17] Shriberg, E. 2005. Spontaneous speech: How people really talk, and why engineers should care. *Proc. INTERSPEECH* Lisbon, Portugal. 1781–1784.

[18] Shriberg, E., Stolcke, A., Baron, D. 2001. Observations on overlap: Findings and implications for automatic processing of multi-party conversation. *Proc. EUROSPEECH* Aalborg, Denmark. 1359–1362.

[19] Shriberg, E. e. a. 2004. The ICSI Meeting Recorder Dialog Act (MRDA) Corpus. *Proc. SIGdial* Cambridge MA, USA. 97–100.

[20] Truong, K., van Leeuwen, D. February 2007. Automatic discrimination between laughter and speech. *Speech Communication* 49(2), 144–158.

[21] Waibel, A., Steusloff, H., Stiefelhagen, R. 2004. CHIL: Computers in the Human Interaction Loop. *Proc. ICASSP2004 Meeting Recognition Workshop* Montreal, Canada.

[1] We used the freely available Audacity© for this task. Only the foreground channel for each laughter instance was inspected.

[2] We used our in-house multichannel annotation tool TransEdit for this task