Analysis of the Occurrence of Laughter in Meetings

Kornel Laskowski[1,2] *and Susanne Burger*[2]

[1]interACT, Universität Karlsruhe, Karlsruhe, Germany
[2]interACT, Carnegie Mellon University, Pittsburgh PA, USA
kornel@ira.uka.de, sburger@cs.cmu.edu

Abstract

Automatic speech understanding in natural multiparty conversation settings stands to gain from parsing not only verbal but also non-verbal vocal communicative behaviors. In this work, we study the most frequently annotated non-verbal behavior, laughter, whose detection has clear implications for speech understanding tasks, and for the automatic recognition of affect in particular. To complement existing acoustic descriptions of the phenomenon, we explore the temporal patterning of laughter over the course of conversation, with a view towards its automatic segmentation and detection. We demonstrate that participants vary extensively in their use of laughter, and that laughter differs from speech in its duration and in the regularity of its occurrence. We also show that laughter and speech are quite dissimilar in terms of the degree of simultaneous vocalization by multiple participants, and in terms of the probability of transitioning into and out of vocalization overlap states.
Index Terms: laughter, vocal interaction, multiparty meetings, Markov models.

1. Introduction

In recent years, the availability of large multiparty corpora of naturally occurring meetings has shifted attention to previously little-explored human-human interaction behaviors [1]. A non-verbal phenomenon belonging to this class is laughter, which has been hypothesized as a strategic means of affecting interlocutors, as well as a signal of various human emotions [2]. In meetings, laughter has been shown to correlate significantly with speakers' emotional valence, as perceived by outside observers [3]. Laughter detection is therefore positioned to become a key component in the automatic characterization of affect in multi-party converation [4].

To date, laughter detection has remained a challenging task. Farfield detection of group laughter in meetings was attempted by [5], whose system appears to have relied on approximately concurrent vocalization by a majority of participants. Since there is ample evidence to suggest that simultaneous speaking is dispreferred (speakers take turns in speaking) [6] but that laughter is contagious (laughers do not take turns in laughing) [7], it is not clear from the results in [5] that the automatic features they experimented with would differentiate between speech and laughter when applied to single-participant vocalization.

Discrimination between the two types of vocal activity in the meeting domain was most recently explored in [4], leveraging the reported acoustic dissimilarities between speech and laughter [8]. Equal error rates of 3% were achieved on manually presegmented snippets of audio. The authors did not treat the problem of first finding the segments automatically.

In the present work, we analyze the *occurrence* of laughter in meetings, for the purposes of designing models to find it. Although work in conversation analysis touched on laughter [9], it did not lead to the quantitative treatment we present. Using the same meeting corpus as was used in [4] and [5], we first describe our procedure for producing a ground truth laughter segmentation. Then, for the first time in the context of meetings, we attempt to answer the following three questions:

1. What is the quantity of laughter, relative to the quantity of speech?
2. How does the durational distribution of episodes of laughter differ from that of episodes of speech?
3. How do meeting participants affect each other in their use of laughter, relative to their use of speech?

We intend for the answers to inform the future construction of hidden Markov model topologies to locate laughter in interaction, in conjunction with acoustic models such as those proposed in [4] and [5].

2. Analysis Framework

To describe laughter, we adopt the terminology in [8]. Laughter occurs in *bouts*, which consist of one or more *calls*. Same-bout calls are typically seperated by pauses in the expulsion of air. For the purposes of the current work, laughter is treated as a binary per-participant quantity, "on" from the start of the first call of each bout to the end of the last call of the same bout. We describe this segmentation in Section 4.3.

In analyzing the occurrence of laughter, we constrast it with that of speech. Analogously to laugh bouts, we use talk *spurts* to characterize the duration and relative location of speech; talk spurts have been defined as "speech regions uninterrupted by pauses longer than 500 ms" in [10]. As in the case of laughter, we treat speech as a binary per-participant quantity, "on" from the start of the first word of each spurt to the end of the last word of the same spurt.

While we do not claim that the occurrence of laughter is independent of that of speech, we move away from the assumption that the two are mutually exclusive, as has often been done in past work (with some exceptions, i.e. [11]). Our observations suggest that laughter deserves an autosegmental description vis-à-vis speech. Here, we allow laughter and speech to be independent for convenience of analysis, and measure the amount of time that a given participant spends simultaneously talking and laughing. We refer to this phenomenon as *"laughed speech"*.

3. Data

To study the pragmatics of laughter, we use the relatively large ICSI Meeting Corpus [12]. 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).

4. Data Pre-processing

In this section, we describe the process we followed to produce, for each meeting and each participant with an individual head-mounted microphone: (1) a talk spurt segmentation, $\mathcal{S}$; and (2) a laugh bout segmentation, $\mathcal{L}$.

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.

4.1. Talk Spurt Segmentation

Talk spurt segmentation was produced using the word-level forced alignments in the ICSI Dialog Act (MRDA) Corpus [13]. While 500 ms was used as the minimum inter-spurt duration in [10], 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.

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`	√

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.

4.2. Selection of Annotated Laughter Instances

Laughter is annotated in the ICSI Meeting Corpus orthographic transcriptions (`.stm`) 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 often cannot be uniquely localized between specific lexical items. In particular, this annotation covers the phenomenon of "laughed speech". 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.

We identified 12635 annotated `VocalSound` laughter instances, of which 65 were ascribed to farfield channels and which we excluded. We also identified 1108 annotated `Comment` laughter instances, for a total of 13678 annotated laughter instances in the original ICSI transcriptions.

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`

Table 2: Top 5 most frequently occurring `Comment` descriptions containing the substring "`laugh`" or "`smil`".

4.3. Laugh Bout Segmentation

We employed a mix of automatic and manual methods to produce accurate endpoints for the identified laughter instances.

Of the 12570 non-farfield `VocalSound` instances, 11845 were adjacent on both the left and the right to either a time-stamped `.stm` utterance boundary, or a lexical item. This allowed us to automatically deduce start and end times for 87% of the laughter instances treated here.

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 `.stm` utterance containing them[1]. Each of the 12570 segmented `VocalSound` descriptions was checked by at least one annotator, as part of another task, in its complete multichannel context[2]. The 1108 `Comment` instances were also segmented manually. A quarter of these was checked by one of the authors. Manual segmentation of these 1823 instances took a total of 18 hours.

Merging immediately adjacent instances and discarding a small proportion of annotated laughs for which we could find no supporting evidence resulted in 13259 distinct bouts of laughter.

5. Analysis

5.1. Quantity of laughter

Our first variable of interest was the quantity of laughter by elapsed time rather than by number of bouts. Using the $\mathcal{L}$ and $\mathcal{S}$ segmentations produced in the previous section, we found that the average participant vocalizes for 14.8% of the time that they spend in meetings. Of this effort, 8.6% is spent on laughing and an additional 0.8% is spent on laughing while talking.

We also wished to know to what extent the amount of laughter varies from participant to participant. As Figure 1 shows, participants differ in both how much time they spend vocalizing, and what proportion of that is laughter. Importantly, laughing time and speaking time do not appear to be correlated across participants.

5.2. Laughter duration and separation

In a second suite of analyses, we were interested in the duration of laugh bouts, as well as the temporal separation between two bouts produced by the same participant. We show these distributions, normalized such that the area under each curve sums to one, in the top two panels of Figure 2. Alongside them we show the same distributions for talk spurts (in dashed gray).

[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 annotation tool TransEdit for this task.

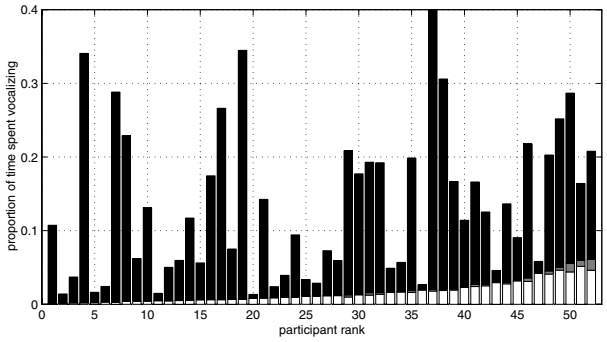

Figure 1: Proportion of time each participant spends on producing only laughter (white), only speech (black), or both speech and laughter simultaneously (gray), normalized by the total recorded duration of all meetings attended by that participant (participant 37, clipped in the diagram, vocalizes 65% of the time). Participants are ordered by increasing proportion of laughter.

We also computed the duration and separation of "islands" of laughter, produced by merging overlapping bouts from all laughing participants; the distribution over these variables is shown in the bottom two panels of Figure 2. The separation between such "islands" is the duration between *any* two participants laughing. The same was done for spurts of talk.

As the plots show, the bout and bout "island" durations follow a lognormal distribution, while spurt and spurt "island" durations appear to be the sum of two lognormal distributions; we suspect the shorter one corresponds to backchannels. Bout durations and bout "island" durations have an apparently identical distribution, suggesting that bouts are committed either in isolation or in synchrony, since bout "island" construction does not lead to longer phenomena. In contrast, construction of speech "islands" does appear to affect the distribution, as expected.

The distribution of bout and bout "island" separations appears to be the sum of two lognormal distributions. The most likely separation between two bouts from the same participant is approximately 46 seconds. The most likely separation between any two laughs, from possibly different laughers, is 4.6 seconds. As expected, both talk spurts and spurt "islands" recur much more frequently; the location of the peak reflects our choice of minimum gap duration employed in talk spurt construction (Section 4.1).

5.3. Interactive aspects

In a final set of experiments, we explore emergent multiparticipant behaviors. We first compute the distribution over different degrees of overlap. Since laughter is reported to be contagious [7], we expect significantly higher proportions of laughter overlap than are reported for speech [10]. We show the results for segmentations involving various logical combinations of speech ($\mathcal{S}$) and laughter ($\mathcal{L}$) in Table 3.

Two main observations can be made based on these numbers. First, laughter does in fact incur significantly more overlap than speech; in relative terms, the ratio is 8.1% of meeting speech time versus 39.7% of meeting laughter time. Comparing $\mathcal{S} \cup \mathcal{L}$ to $\mathcal{S}$, the amount of time in which 3 participants are vocalizing simultaneously is more than 3 times higher when laughter is considered with speech (0.88 hrs) than when speech is

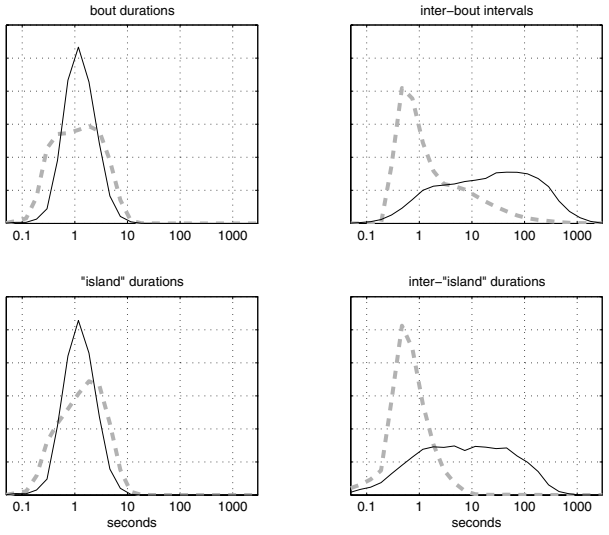

Figure 2: Normalized distributions of the durations of: (*top left*) individual laugh bouts; (*top right*) intervals between laugh bouts talk spurts produced by the same participant; (*bottom left*) multiparticipant laugh bout "islands" (see text); and (*bottom right*) intervals between any two consecutive laugh bouts. Dashed gray lines represent similar distributions for talk spurts and intervals between talk, for comparison. The x-axis represents time in seconds.

considered with "laughed speech" (0.27 hrs) or without it (0.25 hrs). Similarly, the amount of time spent in which 4 or more participants are simultaneously vocalizing is 25 times higher when laugher is considered. This partly explains the success of group laughter detection in the farfield [5], namely that the overwhelming majority of time when more than half of all participants is simultaneously vocalizing appears to be due to laughter.

Second, "laughed speech" ($\mathcal{S} \cap \mathcal{L}$) represents only about 6% of all meeting time spent in laughter. Although the estimates

Vocal Activity	Vocalizing Time, hrs					
	per part	per meet	number of simultaneously vocalizing participants			
			1	2	3	≥ 4
$\mathcal{S}$	55.2	50.8	46.7	3.8	0.27	0.02
$\mathcal{S} - \mathcal{S} \cap \mathcal{L}$	52.1	48.0	44.3	3.5	0.25	0.02
$\mathcal{L}$	5.6	3.3	2.0	0.7	0.31	0.27
$\mathcal{L} - \mathcal{S} \cap \mathcal{L}$	5.2	3.1	1.9	0.7	0.29	0.23
$\mathcal{S} \cap \mathcal{L}$	0.2	0.2	0.2	0.0	0.0	0
$\mathcal{S} \cup \mathcal{L}$	60.3	52.0	45.7	4.8	0.88	0.49

Table 3: Overlap for segmentations under different combinations of speech ($\mathcal{S}$) and laughter ($\mathcal{L}$), for all meetings and all participants. Total recorded meeting time is 66.3 hours. Column 2 shows the total vocalization (Voc. Time) time in hours, for all participants (Part.) over all meetings; column 3 shows the total meeting time (Meet.) for which at least one participant is vocalizing. Columns 4 to 7 show the absolute numbers of hours spent while one, two, three and four or more participants, respectively, vocalize simultaneously.

EDO Transition			100 ms frames		500 ms frames	
n_i	o_{ij}	n_j	$\mathcal{S}$	$\mathcal{L}$	$\mathcal{S}$	$\mathcal{L}$
1	1	1	93.83	87.61	82.94	57.96
1	1	2	1.57	3.68	6.21	8.43
1	1	≥ 3	0.02	0.24	0.39	2.39
2	1	1	19.69	10.58	45.49	26.37
2	2	2	77.05	82.78	40.88	46.93
2	2	≥ 3	1.98	5.48	4.46	13.65
≥ 3	1	1	3.56	0.63	19.24	6.69
≥ 3	2	2	26.30	6.84	40.94	17.45
≥ 3	2	≥ 3	1.06	0.32	5.99	2.83
≥ 3	≥ 3	≥ 3	68.70	92.17	29.44	71.04

Table 4: Selected transition probabilities for EDO models trained on $\mathcal{S}$ and $\mathcal{L}$, at two different frame sizes.

may be unreliable for this reason, Table 3 shows that when considering "laughed speech" alone, very little overlap is observed. Additionally, the exclusion of "laughed speech" from speech ($\mathcal{S} - \mathcal{S} \cap \mathcal{L}$) has negligible impact on the distribution of overlap by degree. This suggests that "laughed" speech is unlikely to be overlapped with other speech, whether the latter is "laughed" or not. However, the exclusion of "laughed" speech from laughter does appear to affect the proportion of time when 4 or more participants are vocalizing. This suggests that "laughed" speech is likely to be overlapped with laughter from other participants.

We also performed a dynamic analysis of interaction, by training participant-independent Extended Degree-of-Overlap (EDO) transition models on $\mathcal{S}$ and $\mathcal{L}$, separately. A detailed description of this model and its training is given in [14]. Briefly, the model yields probabilities of transition between various degrees of overlap, based on training material in the form of a discretized segmentation. We show a select number of transition types for a frame size of 100 ms, as well as a longer frame size of 500 ms, in Table 4. Each transition type is characterized by a triplet (n_i, o_{ij}, n_j), where n_i represents the number of simultaneously vocalizing participants in the "from" state, n_j represents the number of simultaneously vocalizing participants in the "to" state, and o_{ij} represents the number of same participants vocalizing in both states. Here, we consider only states of 0, 1, 2, and ≥ 3 simultaneously vocalizing participants. Table 4 shows that, especially for long frame sizes, when 3 or more participants are laughing the most likely next transition is to the same state. In contrast, when 3 or more participants are speaking, the most likely next state is one in which one of the participants stops vocalizing (a (3,2,2) transition). A similar effect can be seen when 2 participants are laughing, versus when 2 participants are speaking. In general, the table shows that transitions to higher degrees of overlap are more likely with laughter than with speech.

6. Conclusions

We have analyzed a large, publicly available corpus of meetings for the occurrence of laughter. We have shown that laughter accounts for approximately 9.5% of all vocalizing time, which varies extensively from participant to participant and appears not to be correlated with speaking time. Laugh bout durations have a smaller variance than talk spurt durations. We have also shown that laughter is responsible for a significant amount of vocal activity overlap in meetings, and that transitioning out of laughter overlap is much less likely than out of speech overlap. We have quantified these effects in meetings, for the first time, in terms of probabilistic transition constraints on the evolution of conversations involving arbitrary numbers of participants.

7. Acknowledgements

We would like to thank our annotators Matthew Bell and Jörg Brunstein. We would also like to thank Liz Shriberg for investing time to explain the ICSI MRDA corpus and for continued encouragement in our study of laughter, and Alan Black for useful discussion. 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`).

8. References

[1] Shriberg, E., "Spontaneous Speech: How People Really Talk, and Why Engineers Should Care", Proc. EUROSPEECH, pp1781–1784, Lisbon, Portugal, 2005.

[2] Russell, J. et al, "Facial and Vocal Expressions of Emotion", Ann. Rev. Psychol. 54:329–349, 2003.

[3] Laskowski, K. and Burger, S., "Annotation and Analysis of Emotionally Relevant Behavior in the ISL Meeting Corpus", Proc. LREC, Genoa, Italy, 2006.

[4] Truong, K. and van Leeuwen, D., "Automatic Discrimination between Laughter and Speech", Speech Communication 49:144–158, 2007.

[5] Kennedy, L. and Ellis, D., "Laughter Detection in Meetings", Proc. ICASSP Meeting Recognition Workshop, pp118–121, Montreal, Canada, 2004.

[6] Sacks, H. et al, "A Simplest Semantics for the Organization of Turn-Taking for Conversation", Language 50(4):696–735, 1974.

[7] Provine, R., "Contagious Laughter: Laughter is a Sufficient Stimulus for Laughs and Smiles", Bull. Psychonomic Soc 30(1):1–4, 1992.

[8] Bachorowski, J.-A. et al, "The Acoustic Features of Human Laughter", J. Acoust. Soc. Amer. 110(3):1581–1597, 2001.

[9] Jefferson, G., "A Technique for Inviting Laughter and its Subsequent Acceptance Declination", in Everyday Language: Studies in Ethnomethodology (Psathas, G., ed), Irvinton Publishers, pp79–96, 1979.

[10] Shriberg, E. et al, "Observations on Overlap: Findings and Implications for Automatic Processing of Multi-Party Conversation", Proc. EUROSPEECH Vol.2, pp1359–1362, Aalborg, Denmark, 2001.

[11] Trouvain, J., "Phonetic Aspects of "Speech-Laughs"", Proc. ORAGE, pp634–639, Aix-en-Provence, France, 2001.

[12] Janin, A. et al, "The ICSI Meeting Corpus", Proc. ICASSP Vol.1, pp364–367, Hong Kong, China, 2003.

[13] Shriberg, E. et al, "The ICSI Meeting Recorder Dialog Act (MRDA) Corpus", Proc. SIGdial, pp97–100, Cambridge MA, USA, 2004.

[14] Laskowski, K. and Schultz, T., "Modeling Vocal Interaction for Segmentation in Meeting Recognition", Proc. MLMI, Brno, Czech Republic, 2007.