A Comparative Cross-Domain Study of the Occurrence of Laughter in Meeting and Seminar Corpora

Susanne Burger[1], Kornel Laskowski[1,2] and Matthias Wölfel[2]

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

Abstract

Laughter is an intrinsic component of human-human interaction, and current automatic speech understanding paradigms stand to gain significantly from its detection and modeling. In the current work, we produce a manual segmentation of laughter in a large corpus of interactive multi-party seminars, which promises to be a valuable resource for acoustic modeling purposes. More importantly, we quantify the occurrence of laughter in this new domain, and contrast our observations with findings for laughter in multi-party meetings. Our analyses show that, with respect to the majority of measures we explore, the occurrence of laughter in both domains is quite similar.

1. Introduction

Laughter is an intrinsic component of human-human interaction. In multi-party conversational settings, it has been shown to correlate with perceived emotional valence in participants (Laskowski and Burger, 2006), and has generally been hypothesized as a strategic means of affecting others (Russel et al., 2003). Furthermore, when ascribed to specific participants, the amount and distribution of laughter appears to be indicative of social hierarchy (Leffler et al., 1982). The study of these and related effects calls for a detailed segmentation and annotation of laughter in large multi-party conversational corpora which are currently becoming available. Recent work on laughter in the domain of meetings (Laskowski and Burger, 2007) has attempted to quantify the occurrence of laughter in a large corpus of natural, multi-party conversations. It was found that approximately 10%of vocalization effort is spent on laughter, as opposed to speech, and that laughs produced without voicing form a minority of laughter by time in this domain. Additionally, rates of overlap for laughter were shown to be significantly higher than those for speech, and vocalization produced in high degrees of overlap is an order of magnitude more likely to be laughter than speech. A first goal of the current work is to determine whether the above findings generalize to natural multiparty conversation domains other than meetings, and to data recorded elsewhere (the data used in (Laskowski and Burger, 2007) was recorded at a single site). We propose to do this by studying a multi-site corpus of interactive seminars. Partitioning findings into domain-independent and domain-dependent categories is intended to support our second goal, that of characterizing interactive seminars vis-a-vis meetings. Finally, we anticipate that the manual segmentation of laughter in our corpus of seminars will be of use for acoustic modeling.

2. Data

The current study is based on 25 interactive seminars which were recorded in 2006 under the European project CHIL, *Computers in the Human Interaction Loop* . The intent of the recordings originally was to support the Rich Transcription Meeting Recognition (RT) and Classification of Events, Activities and Relationships (CLEAR) evaluations organized by the National Institute of Standards and Technology (NIST) in 2007.

The seminars were held in English, and were recorded at five different sites around the globe (Greece, Italy, Spain, Germany and the United States). Each seminar was attended by three to five participants, gathered around a table. Typically, one participant gave a presentation, during which the other participants interrupted freely in order to ask questions, make comments or give suggestions. This frequently led to open discussion with degrees of interaction similar to those observed in meetings (Burger, 2008).

The 25 seminars, which we refer to as the CHIL06 data, are on average 33 minutes long, and together comprise 13 hours and 52 minutes. A total of 71 individuals originating from 17 different countries spoke in the corpus. As a result, most of the English is accented, with the biggest groups being Spaniards (23%), Italians (15%), and Greeks and Germans (each 14%). Here, we use only their close-talking microphone recordings. The data has been previously transcribed at the orthographic level, which included the annotation of laughter and other events (coughing, filled pauses, breaks, repetitions etc).

In preparation for the NIST RT and CLEAR evaluations, the 25 seminars were split into two subcorpora. The first, `CHIL06_1`, consisted of the first seminar collected at each of the 5 recording sites. NIST denoted these seminars, in their entirety, as development data (accordingly, `CHIL06_1` has been referred to as `rt07s_dev` in the RT community). From the remaining 20 seminars, which we denote as `CHIL06_2`, NIST selected 40 5-minute excerpts to be used as `rt07s_eval`, the unseen evaluation data. The excerpts were intended to cover a balanced assortment of seminar phases, including openings, lecture-like portions, coffee breaks, question-and-answer portions, and closings. The eight excerpts identified as coffee breaks formed the evaluation material for the cbreak task, while the remaining 32 excerpts formed the evaluation material for the lectmtg ("lecture meeting") task. We note that the CHIL06_2 half of CHIL06 is significantly larger than the 40 excerpts selected by NIST. The relevant divisions of the corpus are shown for completion in Figure 1.

To contrast our analysis in the seminar domain, we make use of previous work (Laskowski and Burger, 2007) on the ICSI Meeting Corpus (Janin et al., 2003). This corpus consists of 75 unscripted, naturally occurring meetings, amounting to over 66 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).

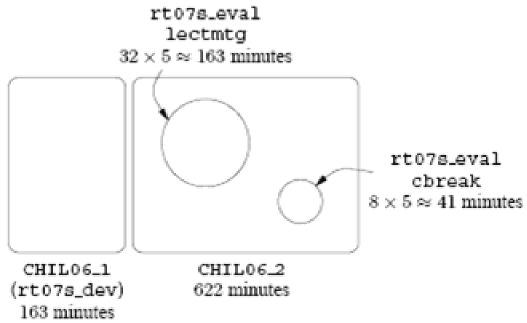

Figure 1: Partitioning of the CHIL06 data into two halves, CHIL06_1 and CHIL06_2; the first half was used in its entirety by NIST as rt07s_dev. 32 5-minute lectmtg excerpts and 8 5-minute cbreak excerpts were drawn from the second half to comprise rt07s_eval.

3. Laugh Bout Segmentation

As mentioned in Section 2, the orthographic transcriptions which accompany the CHIL06 corpus contain mark-up for laughter. The original transcription team had used the token <Laugh>, placing it among word tokens in a manner resembling as closely as possible the sequence of vocal productions. For instances of "laughed speech" (Nwokah et al., 1999), the annotators had inserted <Laugh> after the last laughed word; "laughed speech" was additionally annotated as "hard to understand" if the laughter affected speech intelligibility. Importantly, laughter boundaries in the original transcription effort were not timestamped (although a portion of such timestamps could be inferred from utterance endpoints, in cases where laughter was adjacent to utterance beginnings and/or ends).

As a result, in this work, the near-field audio channels of the complete CHIL06 corpus have been revisited by several annotators in order to timestamp, verify and augment the laughter mark-up present in the original orthographic transcriptions. In listening to the audio, the annotators also checked for laughter instances which had been missed in the orthographic transcription pass. Laughter boundaries were delineated as suggested in (Bachorowski et al., 2001), where laughter is considered as occurring in *bouts*. Each bout consists of one or more *calls*; in contrast to (Bachorowski et al., 2001), we treat audible laughter-related respiration following a bout, and in some cases preceding it, to be part of the bout. In particular, this includes the so-called "recovery exhalation".

In addition to locating the start and end times of each bout, the annotators were asked to manually classify the bout as one of VOICED, UNVOICED, or TALKING, with TALKING taking precedence over VOICED, and VOICED taking precedence over UNVOICED. TALKING ("laughed speech") was defined as laughter that occurs concurrently with speech activity from the laugher, including concurrence with whispered speech and filled pauses. Voicing in laughter was determined as follows: a bout was considered VOICED as a whole if voicing was present at any time during the bout. Otherwise, the bout was considered UNVOICED. A general rule for this distinction which we have found to be useful is that if the gender of the laugher can be inferred from the bout alone, then the bout is likely to be VOICED. Conversely, if the laugher cannot be identified as male or female from the bout alone, then the bout is likely to be UNVOICED.

We estimate the total time spent on this annotation effort to be of the order of 250 hours. The original orthographic transcriptions for all of CHIL06 contained 1381 <Laugh> tokens. The first laughter segmentation and annotation pass, as described above, was performed by one of four annotators and resulted in an 8.7% relative increase in the number of laughter instances, to 1502. A second and final segmentation and annotation pass, performed by the first author, led to a further 4.9% relative increase to 1576 bouts. Across the two passes, the number of TALKING and UNVOICED bouts decreased by 12 and 34, respectively, while the number of VOICED bouts increased by 116; these absolute numbers represent 0.7%, 2.2%, and 7.4%, respectively, of the final total.

4. Talkspurt Segmentation

To contrast the occurrence of laughter with that of speech, we employ a *talkspurt* (Norwine and Murphy, 1938) segmentation produced using forced alignment of speech audio to the lexical items in the orthographic transcription. Both complete words and word fragments were aligned.

Alignment is performed using the *Janus Recognition Toolkit* (JRTk) with a single front-end; the configuration is identical to the warped-MVDR(30) front-end system used in our NIST RT-07s submission (Wölfel et al., 2007). In summary, the front-end computes warped-MVDR spectral envelopes (Wölfel and McDonough, 2005) for 16ms frames every 10ms. The 4000 context-dependent codebooks, with up to 64 diagonal-covariance Gaussians, were trained on approximately 100 hours of audio consisting of the ICSI, NIST, and CMU meeting corpora, the *Translanguage English Database* (TED) lecture corpus, and the CHIL lecture and seminar corpus (Mostefa et al., 2007). Discriminative training with a maximum mutual information criterion was used in the final iteration. During forced alignment, we first perform supervised adaptation of the acoustic models using model-space maximum likelihood linear regression, feature space adaptation, and vocal tract length normalization; labels are written out in a second pass.

5. Comparative Analysis

We now proceed to describe the distribution of laughter in the CHIL06 corpus, in terms of overall quantity, quantity per participant, the use of voicing in laughter, bout duration, and static and dynamic overlap characteristics. We contrast our findings with similar measures for speech in the same data, as well as with our findings in the domain of meetings. For convenience, we employ the symbols $\mathcal{L}$ for the laughter segmentation produced in Section 3, $\mathcal{S}$ for the speech segmentation produced in Section 4, $\mathcal{L}_V$ for the subset of $\mathcal{L}$ annotated as either TALKING or VOICED, and $\mathcal{L}_U$ for the subset of $\mathcal{L}$ annotated as UNVOICED. Note that $\mathcal{L} = \mathcal{L}_V \cup \mathcal{L}_U$. We define as the *talk-time* $T_{\mathcal{S}}^{r,j}$ the total duration of all talkspurts produced by participant j in seminar r. Similarly, $T_{\mathcal{L}}^{r,j}$ is the *laugh-time* of participant j in seminar r, and is computed by summing the durations of laugh bouts. We also define vocalization $\mathcal{V} \equiv \mathcal{S} \cup \mathcal{L}$, and note that the corresponding *vocalization-time* $T_{\mathcal{V}}^{r,j}$ need not equal $T_{\mathcal{S}}^{r,j} + T_{\mathcal{L}}^{r,j}$, since a single participant can produce speech and laughter concurrently. Finally, we denote as $T^{r,j}$ the participation time of participant j in seminar r, and assume this quantity to be equal to T^r, the duration of seminar r.

5.1. Quantity

The CHIL06 corpus contains 1576 distinct bouts of laughter, of which 15% have been annotated as TALKING, 59% as VOICED, and the remaining 26% as UNVOICED. In time, these bouts represent 8.3 minutes, 28.9 minutes, and 8.4 minutes, respectively, for a total of 45.7 minutes of segmented laughter. UNVOICED laughter represents 18.5% of this total, which is slightly lower than that found in the ICSI Meeting Corpus (25.6%).

A relatively large number of participants in the CHIL06 corpus laughs extremely infrequently, as is shown in Figure 2. Bars represent the proportion of participation time that are spent in laughter annotated as one of TALKING, VOICED, and UNVOICED, or in speech (excluding TALKING laughter). For example, the proportion $p_{\mathcal{S}}^j$ of speech for a participant j in the corpus is given by

$$p_{\mathcal{S}}^j = \frac{\sum_{r=1}^{R} T_{\mathcal{S}}^{r,j}}{\sum_{r=1}^{R} T^{r,j}} \quad (1)$$

Participants are ordered from left to right in Figure 2 with increasing $p_{\mathcal{L}}^j$. Both UNVOICED laughter, shown in white in the figure, and TALKING laughter, shown in light gray, are produced by only a minority of participants. However, as for meetings, laugh-time does not appear to be correlated with vocalization-time.

5.2. Duration

Laugh bout duration is shown in Figure 3, for the complete CHIL06 corpus. It can be seen that bouts annotated as VOICED are on average longer than bouts annotated as UNVOICED, an observation which mirrors findings in the meeting domain. TALKING bouts are longer, with a most likely duration of 1.4 seconds.

In the top left panel of Figure 4, we show the normalized distribution of durations of all bouts, irrespective of their TALKING / VOICED / UNVOICED label. The most likely duration is just under 1 second. Also shown is the normalized distribution of talkspurt durations, whose most likely value is somewhat higher than that for bouts. The top right panel of the same figure demonstrates that the most likely interval between any two bouts from the same participant is approximately 1 minute. This value is significantly higher than the most likely interval duration between two talkspurts from the same participant.

The bottom two panels of Figure 4 show the normalized distribution of the durations of contiguous intervals of laughter, in which abutting or overlapping bouts from po-

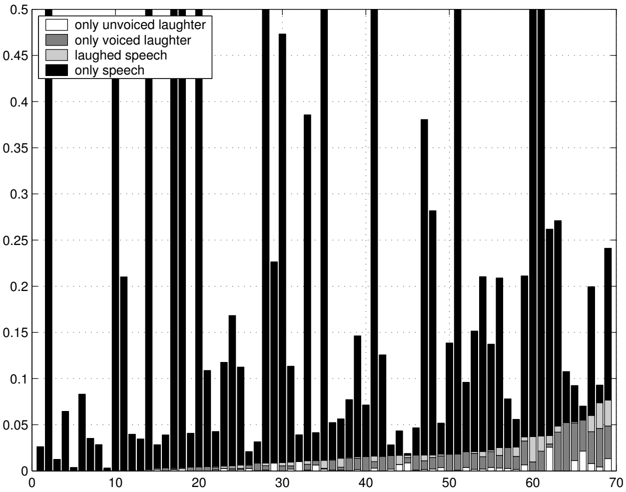

Figure 2: Proportion of participation time spent in TALKING laughter , in VOICED laughter, in UNVOICED laughter and in speech (excluding TALKING laughter) for all 69 participants appearing in CHIL06_1 and rt07s_eval::lectmtg (there are 2 CHIL06 participants which do not appear in these subsets). Participants are ordered by increasing proportion $p_{\mathcal{L}}$ of laugh-time.

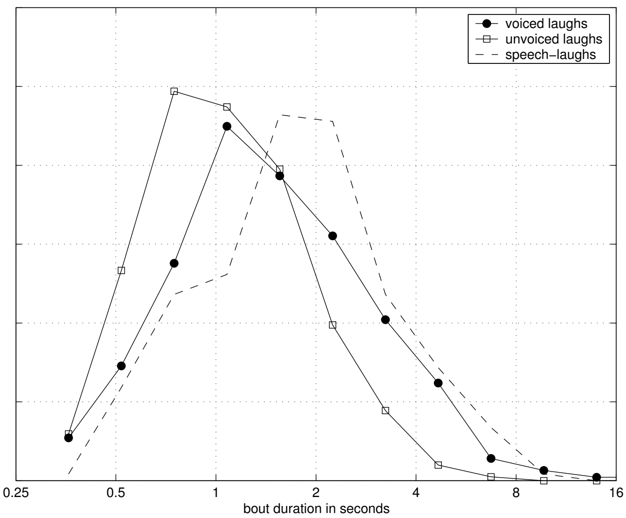

Figure 3: Normalized distributions of duration in seconds for VOICED laughter bouts, UNVOICED laughter bouts, and TALKING bouts, in the entire CHIL06 corpus.

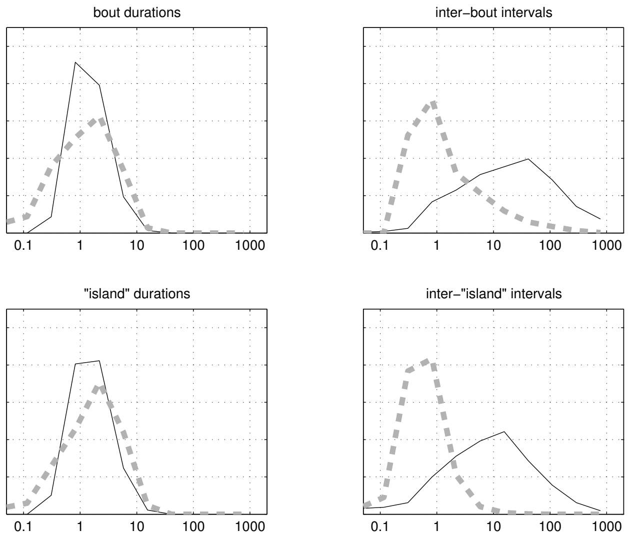

Figure 4: Normalized distributions of the durations of (*top left*) individual laugh bouts; (*top right*) intervals between laugh bouts produced by the same participant; (*bottom left*) multiparticipant laugh bout "islands" (see text); and (*bottom right*) intervals between any two consecutive laugh bouts, for `rt07s_dev` and `rt07s_eval::lectmtg`. Normalized distributions for talkspurts and talkspurt islands are shown in dashed gray. The x-axis represents time in seconds.

tentially different participants are merged into laugh bout "islands". As observed for meetings, this distribution does not differ significantly from the distribution of individual bout durations, suggesting that multiparticipant laughter occurs either in near-perfect synchrony, or in isolation. By contrast, the difference between the normalized distribution of talkspurts and that of talkspurt "islands" is more apparent. This is because talkspurts are generally produced in alternation by different participants, often with overlap during speaker change. The bottom right panel shows the distribution of inter-"island" intervals. It can be seen that the most likely duration between any two participants' laugh bouts is much shorter than between two bouts produced by the same participant (top right panel). These findings concerning the duration of laugh bouts, laugh bout "islands", and the intervals between them are quantitatively similar to our findings for meetings (Laskowski and Burger, 2007).

5.3. Laughter Overlap

Next, we compute the amount of laughter overlap in the `CHIL06` data. We note that in conversational settings, higher levels of speech overlap are indicative of more spontaneous, unstructured interactions. Higher levels of laughter overlap are indicative of simultaneous multiparticipant involvement. Laughter overlap levels are expected to be significantly higher than those characteristic of speech, in any natural domain, since participants tend to take turns speaking but not laughing.

For clarity of exposition, we define a quantity $T_\alpha^{r,*}$, which is the total seminar time during which at least one participant produces vocalization type α, which can be laughter $\mathcal{L}$, speech $\mathcal{S}$, etc. For a dataset consisting of R seminars, the quantity

$$T_\alpha^* \;=\; \sum_{r=1}^{R} T_\alpha^{r,*} \tag{2}$$

represents the time, accumulated over all R seminars, in which at least one participant vocalizes. We also define a quantity $T_\alpha^{r,k}$, which is the time during which the kth participant vocalizes, of K_r participants in seminar r. The sum

$$T_\alpha \;=\; \sum_{r=1}^{R} \sum_{k=1}^{K_r} T_\alpha^{r,k} \tag{3}$$

represents the talk-time of all participants in a corpus of R seminars. The two quantities in Equations 2&3 can be combined to yield a *compression ratio*

$$c_\alpha \;\equiv\; \frac{T_\alpha}{T_\alpha^*} \tag{4}$$

which expresses the predominance of overlap, ie. the ratio of the amount of time spent in vocalization type α, over all participants, to the total amount of seminar time in which that vocalization is produced. c_α must be greater or equal to 1 (no simultaneous vocalization at all); higher degrees of overlap yield higher compression ratios.

Table 1 shows an analysis of overlap for the `rt07s_dev` (`CHIL06_1`) dataset, for the `rt07s_eval::lectmtg` dataset, and for the ICSI Meeting Corpus for comparison. We first compare the proportion of vocal effort spent on laughing, $\frac{T_\mathcal{L}}{T_{\mathcal{L}\cup\mathcal{S}}}$; this quantity is 3.8%, 10.2%, and 9.4%, respectively, for the three datasets. This indicates that the `rt07s_eval::lectmtg` data is similar to meeting data from the point of view of amount and distributionof laughter, and that `rt07s_dev` contains significantly less laughter. We believe that this difference is due to the fact that the `CHIL06_1` seminars were the first of a series to be recorded at each site in 2006, and accordingly resemble CHIL lectures, collected in 2004 and 2005 (Mostefa et al., 2007).

Second, we compare the compression ratios for speech $\mathcal{S}$ and for laughter $\mathcal{L}$, across the three datasets. Irrespective of its overall amount, laughter in all three exhibits high compression ratios, in the range 1.46—1.71; these are significantly higher than the computed compression ratios of speech (1.04—1.08). Closer inspection of the relative proportion of vocalization in specific degrees of overlap reveals that approximately one quarter of meeting time in which laughter occurs is spent in 2-participant laughter. The higher proportions of 3-participant and $\geq$3-participant laughter in the ICSI Meeting Corpus probably reflects the typically larger numbers of participants per meeting. Overlap breakdown for speech $\mathcal{S}$ suggests that the ICSI Meeting Corpus is more interactive than seminar data, consisting of more overlap (speaker changes, interruptions, and backchannels).

Finally, we note that overlap for $\mathcal{V} \equiv \mathcal{S} \cup \mathcal{L}$ shows higher overlap rates than speech alone. This indicates that speech from one participant is frequently produced simultaneously with laughter from others. "Laughed speech", $\mathcal{S} \cap \mathcal{L}$, exhibits relatively low overlap rates, which suggests that typically at most one participant is producing it at a time. In

Vocalization Type α	Vocalizing Time (min)					
	T_α	c_α	number of simultaneously vocalizing participants			
			1	2	3	≥4
`rt07s_dev` *(total duration: 163.1 min)*						
$\mathcal{S}$	131.0	1.037	96.7	3.1	0.2	0.0
$\mathcal{L}$	5.1	1.5	64.0	25.3	9.5	1.2
$\mathcal{S} \cup \mathcal{L}$	133.4	1.050	95.6	3.8	0.5	0.1
$\mathcal{S} \cap \mathcal{L}$	2.5	1.316	74.0	21.4	3.5	1.1
`rt07s_eval::lectmtg` *(tot. duration: 163.60 min)*						
$\mathcal{S}$	120.6	1.062	94.2	5.5	0.3	0.0
$\mathcal{L}$	13.6	1.462	66.5	24.0	6.9	2.6
$\mathcal{S} \cup \mathcal{L}$	132.8	1.127	89.6	8.5	1.4	0.5
$\mathcal{S} \cap \mathcal{L}$	1.4	1.077	95.7	4.3	0.0	0.0
ICSI Meeting Corpus *(total duration: 3978.4 min)*						
$\mathcal{S}$	3249.7	1.081	92.4	7.1	0.5	0.0
$\mathcal{L}$	332.3	1.705	60.4	21.8	9.5	8.3
$\mathcal{S} \cup \mathcal{L}$	3550.2	1.155	88.5	9.0	1.7	0.9
$\mathcal{S} \cap \mathcal{L}$	16.3	1.025	97.9	2.1	0.0	0.0

Table 1: Overlap for speech ($\mathcal{S}$), and all laughter ($\mathcal{L} \equiv \mathcal{L}_V \cup \mathcal{L}_U$), their union, and their intersection, in `rt07s_dev`, `rt07s_eval::lectmtg`, and the ICSI Meeting Corpus. Column 2 ($T\alpha$) shows the total α-vocalization time, summed over all participants in all meetings. Column 3 shows the compression ratio, and columns 4 through 7 show a breakdown of time in which at least one participant α-vocalizes by the number of participants α-vocalizing simultaneously.

Vocalization Type α	Vocalizing Time (min)					
	T_α	c_α	number of simultaneously vocalizing participants			
			1	2	3	≥4
`rt07s_dev` *(total duration: 163.1 min)*						
$\mathcal{L}$	5.1	1.5	64.0	25.3	9.5	1.2
$\mathcal{L}_V$	4.5	1.45	63.6	27.2	8.0	1.2
$\mathcal{L}_U$	0.5	1.0	100.0	0.0	0.0	0.0
`rt07s_eval::lectmtg` *(tot. duration: 163.6 min)*						
$\mathcal{L}$	13.6	1.46	66.5	24.0	6.9	2.6
$\mathcal{L}_V$	11.5	1.46	66.9	24.0	6.8	2.3
$\mathcal{L}_U$	2.0	1.05	95.0	5.0	0.0	0.0
`rt07s_eval::cbreak` *(total duration: 41.0 min)*						
$\mathcal{L}$	5.6	1.51	65.3	22.5	10.1	2.1
$\mathcal{L}_V$	4.5	1.50	64.8	21.5	11.8	1.9
$\mathcal{L}_U$	1.1	1.1	92.6	7.4	0.0	0.0
`rt07s_eval::`all *(total duration: 622.2 min)*						
$\mathcal{L}$	40.7	1.49	69.9	21.2	6.5	2.5
$\mathcal{L}_V$	32.7	1.46	70.1	21.1	6.8	2.1
$\mathcal{L}_U$	7.9	1.18	95.1	4.4	0.5	0.0
ICSI Meeting Corpus *(total duration: 3978.4 min)*						
$\mathcal{L}$	332.3	1.71	60.4	21.8	9.5	8.3
$\mathcal{L}_V$	247.0	1.66	61.6	21.9	9.4	7.1
$\mathcal{L}_U$	85.3	1.13	88.7	10.2	0.9	0.2

Table 2: Overlap for voiced laughter ($\mathcal{L}_V$), unvoiced laughter ($\mathcal{L}_U$), and all laughter ($\mathcal{L} \equiv \mathcal{L}_V \cup \mathcal{L}_U$), in `rt07s_dev`, `rt07s_eval::lectmtg`, `rt07s_eval::cbreak`, all of `CHIL06_2`, and the ICSI Meeting Corpus. Column descriptions as in Table 1.

these ways, seminar and meeting data appear to be broadly similar.

Table 2 gives a similar breakdown, this time to contrast voiced and unvoiced laughter; we compare not only `rt07s_dev`, `rt07s_eval::lectmtg`, and meeting data, but also the coffee break portion of `rt07s_eval`, `rt07s_eval::cbreak`, and the entire `CHIL06_2`. The results show that laughter has compression ratios in the range 1.46—1.71 across multiple datasets, with meetings exhibiting the highest ratios. Among meetings, the highest ratios can be found in the coffee break subset, `rt07s_eval::cbreak`, which agrees with intuition.

In all subsets, voiced laughter $\mathcal{L}_V$ exhibits significantly higher compression ratios than unvoiced laughter $\mathcal{L}_U$; those of the latter are similar to compression ratios of speech (cf. Figure 1, above). Unvoiced laughter almost never occurs in more-than-two participant overlap; in meetings, where its relative proportion for overlap degrees of 3 or more is greatest, it accounts for only 1.1% of all meeting time during which laughter occurs. However, unvoiced laughter is frequently accompanied by voiced laughter from other participants. For datasets in which unvoiced laughter does not occur in overlap degrees of 3 or more, it nevertheless affects the relative overlap proportions of all laughter ($\mathcal{L} \equiv \mathcal{L}_V \cup \mathcal{L}_U$), when combined with voiced laughter ($\mathcal{L}_V$).

5.4. Laughter Overlap Dynamics

Having investigated the degree of overlap for speech, voiced laughter, and unvoiced laughter, we turn to an analysis of how overlap *arises* for both laughter modes, as well as for speech. We do this by treating a seminar involving K participants as a stochastic process, whose multiparticipant vocalization state $\mathbf{q}_t$ is a K-element vector. When considering vocalization type α, each participant can be in either α or $\neg\alpha$ at time t, leading to a space of $N = 2^K$ multiparticipant states. We assume the process if 1st order Markovian, and that the probability of transition from a state $\mathbf{S}_i$ at time t to a state $\mathbf{S}_j$ at time $t+1$ is a function only of the number of participants simultaneously vocalizing in states $\mathbf{S}_i$ and $\mathbf{S}_j$, as well as of the number of same participants continuing to vocalize at $t+1$. This leads to the Extended Degree of Overlap (EDO) model (Laskowski and Schultz, 2007).

$$\begin{aligned} & P\left(\mathbf{q}_{t+1} = \mathbf{S}_j \,|\, \mathbf{q}_t = \mathbf{S}_i \,,\, \mathbf{q}_{t-1} = \mathbf{S}_h \,,\, \ldots\right) \\ &= P\left(\mathbf{q}_{t+1} = \mathbf{S}_j \,|\, \mathbf{q}_t = \mathbf{S}_i\right) \qquad (5) \\ &\propto P\left(\|\mathbf{q}_{t+1}\| = n_j \,,\, \|\mathbf{q}_{t+1} \cdot \mathbf{q}_t\| = o_{ij} \,|\, \|\mathbf{q}_t\| = n_i\right) \end{aligned}$$

where $\|\mathbf{q}_t\|$ represents the number of participants vocalizing at time t, and $\|\mathbf{q}_t \cdot \mathbf{q}_{t+1}\|$ represents the number of participants vocalizing at time t and continuing to vocalizing

EDO Transition			chil06_1 (rt07s_dev)		chil06_2				ICSI Meeting Corpus	
					rt07s_eval			(all)		
					lectmtg		cbreak			
n_i	o_{ij}	n_j	$\mathcal{S}$	$\mathcal{L}$	$\mathcal{S}$	$\mathcal{L}$	$\mathcal{L}$	$\mathcal{L}$	$\mathcal{S}$	$\mathcal{L}$
0	0	0	*57.26*	99.41	*63.60*	97.40	95.94	98.75	*46.73*	98.76
0	0	1	*40.70*	0.56	*33.71*	2.26	3.28	1.10	*48.94*	1.04
0	0	2	*1.92*	0.03	*2.51*	0.31	0.68	0.13	*4.14*	0.15
0	0	≥3	*0.12*	0.00	*0.17*	0.02	0.10	0.02	*0.19*	0.04
1	0	0	*6.52*	27.39	*8.42*	24.20	24.26	26.95	*8.23*	29.36
1	1	1	*89.34*	61.46	*84.27*	64.97	62.62	63.06	*82.59*	57.69
1	1	2	*2.78*	7.64	*4.69*	8.92	10.49	7.14	*6.02*	8.56
1	1	≥3	*0.13*	1.59	*0.20*	1.15	0.66	1.34	*0.39*	2.34
2	0	0	*4.93*	6.19	*3.94*	5.81	5.26	7.24	*3.54*	8.77
2	1	1	*48.01*	22.12	*47.17*	23.23	28.57	25.39	*46.81*	26.57
2	2	2	*37.95*	60.18	*40.11*	60.65	51.88	55.46	*39.26*	46.97
2	2	≥3	*3.25*	10.62	*2.73*	9.68	11.28	9.80	*4.22*	13.47
≥3	0	0	*2.04*	0.00	*0.84*	0.00	4.48	1.70	*1.25*	1.47
≥3	1	1	*17.35*	5.08	*18.49*	7.04	5.97	6.17	*20.42*	6.71
≥3	2	2	*35.71*	25.42	*43.70*	22.54	16.42	21.28	*41.10*	17.42
≥3	≥3	≥3	*36.73*	69.49	*29.41*	69.72	73.13	69.79	*27.84*	70.87

Table 3: Select EDO transitions (n_i, o_{ij} , n_j), and their values as inferred from the speech ($\mathcal{S}$, shown in italics for clarity) and laughter ($\mathcal{L}$) segmentations, for several partitions of the CHIL06 corpus and for the ICSI Meeting Corpus; the frame step and size are 500 ms.

at time $t + 1$. The shorthand $n_i \equiv \|\mathbf{S}_i\|$ and $n_j \equiv \|\mathbf{S}_j\|$ are the numbers of vocalizing participants in states $\mathbf{S}_i$ and $\mathbf{S}_j$, respectively, and $o_{ij} \leq \min(n_i, n_j)$ is the number of same participants who vocalize in both states.

A comparison of voiced laughter, unvoiced laughter, and speech using this model involves first discretizing the corresponding segmentation with a particular frame size and frame step (Laskowski and Schultz, 2007), and then using the discretized segmentation to infer the transition probabilities of the model in Equation 5. Here, we use a frame size and frame step identically equal to 0.5 s. The inferred probabilities are shown in Table 3.

The table shows that there are minor differences among the datasets in the probability that laughter begin ((0, 0, $\neq$ 0) EDO transitions); transitions into exactly one person laughing (n_j = 1) from none are highest in rt07s eval::cbreak. Such transitions are approximately 3 times more likely in this subset than in meetings, or in the entire CHIL06_2 set. In contrast, they are only half as likely in CHIL06_1 as in CHIL06_2. Once exactly one participant is laughing (n_i = 1), the probability that they are joined by a second laugher within 500ms is similar across the datasets, 7.14 - 10.49%; the highest number appears for rt07s eval::cbreak. The probability that they are joined by two participants within 500ms is highest in the ICSI meetings (2.34%), where it is about twice as high as in the other datasets shown.

When two participants are laughing, the probability that they are joined by a third laugher within 500ms is 13.47% for ICSI meetings, and at most 10.62% in the other datasets. However, for all datasets, the most likely transition is for the two laughers to still be the only ones laughing 500ms later. Similarly, when three participants are laughing, the most likely EDO transition is to the same state, where all three continue to laugh. This contrasts with speaking: the most likely transition when three participants are simultaneously speaking is for one of them to stop (except for CHIL06_1 where all three continuing to speak is approximately as likely). Identically, the most likely transition when two participants are simultaneously speaking is for one participant to stop within 500ms.

Although subtle differences in inferred probabilities exist, the general trends observed in meetings appear to hold for seminars. We suspect that the differences that do exist are due to the smaller number of participants in the CHIL06 seminars than in the ICSI meetings.

6. Conclusions

We have constructed a laughter segmentation for the CHIL06 seminar corpus, consisting of 1576 distinct bouts and amounting to 45.7 minutes of close-talk laughter. The latter represents a significant amount of training material for acoustic model training. Laughter in seminars is similar to laughter in meetings in that:

1) laughter not containing voicing represents a minority of all laughter;
2) participants vary widely by laugh-time proportion;
3) the most likely bout duration is approximately 1 second;
4) when laughter re-occurs, the most likely inter-bout interval is approximately 1 minute;
5) compression ratios for laughter are approximately 1.5 (and only negligibly above 1 for speech); and
6) laughter and speech differ significantly in the probability of entry into and egress out of multiparticipant overlap.

The observed differences include:

A) the proportion of vocalization effort spent on laughter is

3.8% for CHIL06_1, whereas it is approximately 10% for both CHIL06_2 and meetings;
B) "laughed speech" represents a higher proportion of all laughter in CHIL06_1 than in either CHIL06_2 or meetings; and
C) although speech overlap in CHIL06_1 is more rare than in both CHIL06_2 and meetings, it is more likely to continue than in either of the latter.
These observations suggest that, from a vocal interaction point of view, the CHIL06_2 subset of our seminar corpus is more similar to meetings than to the CHIL06_1 subset.

7. Acknowledgments

We would like to thank our annotators Mattthew Bell, Brian Anna, Joseph P. Fridy, and Brett Nelson.

8. References

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

S. Burger. 2008. The CHIL RT07 Evaluation Data. In *CLEAR 2007 and RT 2007*, volume 4625 of *Lecture Notes in Computer Science*, pages 390–400, Baltimore MD, USA. Springer-Verlag Berlin Heidelberg.

CHIL Consortium. http://chil.server.de.

CLEAR Evaluation. http:clear-evaluation.org.

NIST Rich Transcription (RT) Meeting Evaluation. http://nist.gov/speech/tests/rt.

A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. 2003. The ICSI Meeting Corpus. In *Proc. ICASSP*, pages 364–367, Hong Kong, China.

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

K. Laskowski and S. Burger. 2007. Analysis of the occurrence of laughter in meetings. In *Proc. INTERSPEECH*, Antwerpen, Belgium, September.

K. Laskowski and T. Schultz. 2007. Modeling vocal interaction for segmentation in meeting recognition. In *Proc. MLMI*, Brno, Czech Republic.

A. Leffler, D. Gillespie, and J. Conaty. 1982. The effects of status differentiation on nonverbal behavior. *Social Psychology Quaterly*, 45(3):153–161.

D. Mostefa, N. Moreau, K. Choukri, G. Potamianos, S. Chu, A. Tyagi, J. Casas, J. Turmo, L. Cristoforetti, F. Tobia, A. Pnevmatikakis, V. Mylonakis, F. Talantzis, S. Burger, R. Stiefelhagen, K. Bernardin, and C. Rochet. 2007. The chil audiovisual corpus for lecture and meeting analysis inside smart rooms. *Journal for Language Resources and Evaluation*, 41:389–407.

A.C. Norwine and O.J. Murphy. 1938. Characteristic time intervals in telephonic conversation. *The Bell System Technical Journal*, 17:281–291.

E. Nwokah, H.-C. Hsu, P. Davies, and A. Fogel. 1999. The integration of laughter and speech in vocal communication: a dynamic systems perspective. *J. Speech, Language and Hearing Research*, 42:880–84.

J. Russel, J.-A. Bachorowski, and J.-M. Fernandez-Dols. 2003. Facial and vocal expressions of emotion. *Ann. Rev. Psychol.*, 54:329–349.

M. Wölfel and J.W. McDonough. 2005. Minimum variance distortionless response spectral estimation, review and refinements. *IEEE Signal Processing Magazine*, 22(5):117–126, Sept.

M. Wölfel, S. Stüker, and F. Kraft. 2007. The ISL RT-07 speech-to-text system. *In Proc. of the Rich Transcription 2007 Meeting Recognition Evaluation Workshop (RT-07), Baltimore, USA*.