Detection of Laughter-in-Interaction in Multichannel Close-Talk Microphone Recordings of Meetings

Kornel Laskowski and Tanja Schultz

Cognitive Systems Lab, Universität Karlsruhe, Karlsruhe, Germany
Language Technologies Institute, Carnegie Mellon University, Pittsburgh PA, USA

Abstract. Laughter is a key element of human-human interaction, occurring surprisingly frequently in multi-party conversation. In meetings, laughter accounts for almost 10% of vocalization effort by time, and is known to be relevant for topic segmentation and the automatic characterization of affect. We present a system for the detection of laughter, and its attribution to specific participants, which relies on simultaneously decoding the vocal activity of all participants given multi-channel recordings. The proposed framework allows us to disambiguate laughter and speech not only acoustically, but also by constraining the number of simultaneous speakers and the number of simultaneous laughers independently, since participants tend to take turns speaking but laugh together. We present experiments on 57 hours of meeting data, containing almost 11000 unique instances of laughter.

1 Introduction

Laughter is a key element of human-human interaction, occurring surprisingly frequently in multi-party conversation. In meetings, laughter accounts for almost 10% of vocalization effort by time [1]. It has been identified as potentially relevant to discourse segmentation [2], to inference of humorous intent and detection of interlocutor-specific emotional expression [3], and to classification of perceived emotional valence [4]; several of these tasks call for not only the detection of laughter, but also its correct attribution to specific participants. Laughter is known to lead to the temporary abandonment of turn-taking policy, making its detection relevant in topic change detection [5], important for meeting browsing [6], and potentially instrumental to the identification of conversational hotspots, of which an overwhelming majority is associated with amusement [7].

Laughter detection in meetings has received some attention, beginning with [2] in which farfield group laughter was detected automatically, but not attributed to specific participants. Subsequent research has focused on laughter/speech classification [8, 9] and laughter/non-laughter segmentation [10, 11]. However, in both cases, only a subset of all laughter instances, those not occurring in the temporal proximity of the laugher's speech, was considered. Furthermore, in segmentation work, some form of pre-segmentation was assumed

to have eliminated long stretches of channel inactivity [10, 11]. These measures have led to significantly higher recall and precision rates than would be obtained by a fully automatic segmenter with no a priori channel activity knowledge.

The aim of the current paper is to provide a first fully automatic baseline system for the detection and participant attribution of laughter as it occurs naturally in multiparticipant conversation. While in single-participant recordings laughter can be detected using a speech recognizer augmented with laughter models, in multiparticipant contexts audio must first be segmented and crosstalk from background participants to each channel suppressed. The latter represents a significant challenge for vocal activity detectors in meetings [12]. In constructing the proposed baseline system, we rely on several contrastive aspects of laughter and speech, including acoustics, duration, and the degree of vocalization overlap.

This work begins with a description of the meeting data used in our experiments (Section 2), which was selected to be exactly the same as in previous work [2, 10, 8, 9, 11]. However, our aim is to detect all *laughter-in-interaction*, including laughter which is interspersed among lexical items produced by each participant. We describe our multiparticipant vocal activity model in Section 3 and quantify the performance of its implementation in Section 4. Contrastive experiments are presented in Section 5, leading to a discussion of various aspects of the proposed task. Finally, we compare our findings and observations with those of other authors in Section 6, before concluding in Section 7.

2 Data

As in other work on laughter in naturally occurring meetings [2, 10, 8, 9, 11], we use the ICSI Meeting Corpus [13]. 67 of the 75 meetings in the corpus are of one of three types, `Bed`, `Bmr`, and `Bro`, representing longitudinal recordings of three groups at ICSI. The total number of distinct participants in these three subsets is 23; there are 3 participants who attend both `Bmr` and `Bro` meetings, and only 1 participant who attends both `Bmr` and `Bed` meetings. Importantly, none of the meeting types have a fixed number of participants per meeting, allowing us to demonstrate the applicability of our methods to arbitrary group sizes.

We rely on two reference segmentations of the ICSI corpus, one for speech and one for laughter. The speech segmentation was constructed using the word start and end times from automatic forced alignment, available in the ICSI MRDA Corpus [14]. Inter-word gaps shorter than 0.3 s have been bridged to yield *talkspurts* [15], consisting of one or more words (and/or word fragments); this process, as well as the 0.3 s threshold, has been used extensively in NIST Rich Transcription Meeting Recognition evaluations [16]. The corresponding segmentation of *laugh bouts* [17] has recently been built for this data [1, 18] using the available mark-up in the orthographic transcription and a combination of automatic and manual alignment methods. Intervals during which a participant both laughs and speaks (a phenomenon referred to as "speech-laughs" [19]) have been mapped to speech only, such that the categories of silence $\mathcal{N}$, speech $\mathcal{S}$, and laughter $\mathcal{L}$ are mutually exclusive.

The majority of experiments we present are performed using one type of meeting in the corpus, the `Bmr` meetings. In [2, 8, 9, 11, 10], the first 26 `Bmr` meetings were designated as training data, and the last 3 held out for testing. We retain that division in the current work.

3 Multiparticipant 3-state Vocal Activity Recognition

3.1 Model Topology

Detection in the proposed system consists of Viterbi decoding in a hidden Markov model (HMM) state space which simultaneously describes the state of all K participants to a particular conversation C. Each participant k, $1 \leq k \leq K$, can occupy one of three acoustically distinct (AD) states: speech $\mathcal{S}$, laughter $\mathcal{L}$, and non-vocalization $\mathcal{N}$; where convenient, we will also refer to vocalization $\mathcal{V} \equiv \neg\mathcal{N} \equiv \mathcal{S} \cup \mathcal{L}$. Furthermore, each AD state is implemented by a left-to-right state sequence, enforcing a minimum duration constraint on AD state occupation. A projection of the complete K-participant HMM topology onto the state subspace of any single participant is shown in Figure 1. Each minimum duration constraint T^{Υ}_{min}, $\Upsilon \in \{\mathcal{S}, \mathcal{L}, \mathcal{N}\}$, yields the corresponding number of single-participant topology states per AD state, $N^{\Upsilon}_{min} \equiv \lceil T^{\Upsilon}_{min} / T_{step} \rceil$, where T_{step} is the frame step or shift. As a result, the single-participant state subspace consists of $N = \sum_{\Upsilon} N^{\Upsilon}_{min}$ states.

A consequence of the above is that a multiparticipant conversation C, of K participants, can be in one of N^K states. To render search computationally tractable, we admit only a fraction of these states during decoding, via three constraints: (1) the number of simultaneously speaking participants can be no greater than $K^{\mathcal{S}}_{max}$; (2) the number of simultaneously laughing participants can be no greater than $K^{\mathcal{L}}_{max}$; and (3) the number of participants not in the "default" state $\mathcal{N}^{(0)}$ can be no greater than $K^{\neg\mathcal{N}}_{max}$. The resulting space consists of N_{eff} states, $\{\mathbf{S}_i\}$, with $1 \leq i \leq N_{eff}$. Each state $\mathbf{S}_i$ emits a multi-channel observation with time-independent emission probability b_i.

Transition from state $\mathbf{S}_i$ to state $\mathbf{S}_j$, $\mathbf{S}_i \rightarrow \mathbf{S}_j$, with $1 \leq i \leq N_{eff}$ and $1 \leq j \leq N_{eff}$, is possible provided that for each speaker k, the single-participant transition $\mathbf{S}_i[k] \rightarrow \mathbf{S}_j[k]$ is licensed by Figure 1. An allowed transition $\mathbf{S}_i \rightarrow \mathbf{S}_j$ is taken with time-independent probability $a_{ij} = P(\mathbf{q}_{t+1} = \mathbf{S}_j \,|\, \mathbf{q}_t = \mathbf{S}_i)$, where $\mathbf{q}_t$ is the multiparticipant state of the meeting at time t.

3.2 Acoustic Model

We seek to define the probability density that a particular *multi-channel* observation $\mathbf{X}_t \in \Re^{K \times F}$, where F is the number of features drawn from a single channel in an observation window of T_{size} in duration, is emitted from a *multi-participant* state $\mathbf{S}_i$. The main difficulty is that K, the number of participants, may vary from conversation to conversation, and we wish to avoid having to train

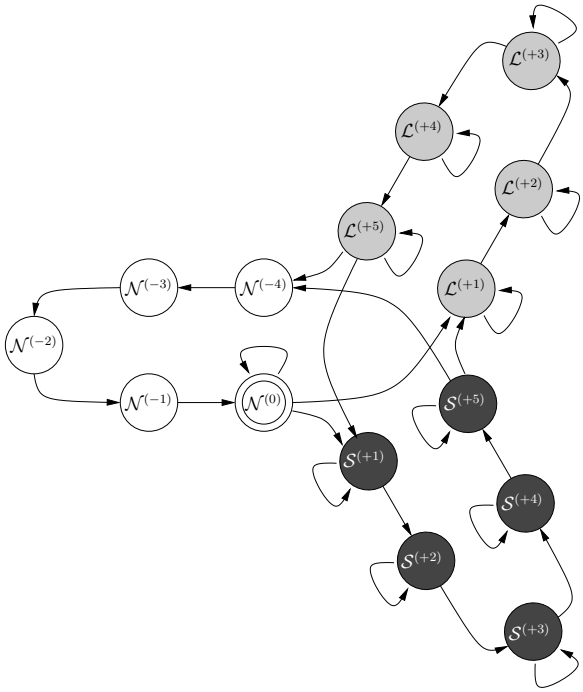

Fig. 1. A projection of the full HMM multiparticipant state space onto the state subspace of a single participant. Shown are three acoustically distinct (AD) states, each duplicated 5 times to illustrate how minimum AD state occupation is enforced. $\mathcal{N}^{(0)}$ represents the default long-time inactive state.

variable-length observation models. We address this difficulty as in [20], by introducing the factorial decomposition $P\left(\mathbf{X}_t \,|\, \mathbf{S}_i\right) = \prod_{k=1}^{K} P\left(\mathbf{X}_t\left[k\right] \,|\, \zeta\left(\mathbf{S}_i, k\right)\right)$. Each factor is a Gaussian mixture model (GMM) likelihood

$$P\left(\mathbf{X}\left[k\right] \,|\, \zeta\left(\mathbf{S}_i, k\right)\right) = \sum_{m=1}^{M} p_{\zeta(i,k),m} P\left(\mathbf{X}\left[k\right] \mid \mathbf{N}\left(\mu_{\zeta(i,k),m}, \sigma^2_{\zeta(i,k),m}\right)\right) \;, \quad (1)$$

where M is the number of GMM components, $\sum_{m=1}^{M} p_{\zeta(i,k),m} = 1$ and $\mathbf{N}\left(\mu, \sigma^2\right)$ is a multivariate, diagonal-covariance Gaussian distribution. The number of dimensions is equal to F, the number of single-channel features computed. $\zeta\left(i,k\right)$ represents the state of the kth close-talk microphone, as explained below.

Although modeling each microphone as being in one of three states is the most natural approach to $\mathcal{N}/\mathcal{S}/\mathcal{L}$ segmentation, efforts in single-participant $\mathcal{N}/\mathcal{S}$ segmentation have extended this model to farfield activity states (ie. [21]). In [22], three states were considered: $\mathcal{S}$, $\mathcal{N}$, and $\mathcal{V}_F$, the latter corresponding to

only farfield speech. We make the corresponding extension here, whereby

$$\zeta(i,k) \equiv \begin{cases} \mathcal{S}, & \text{if } \mathbf{S}_i[k] = \mathcal{S} \\ \mathcal{L}, & \text{if } \mathbf{S}_i[k] = \mathcal{L} \\ \mathcal{V}_F, & \text{if } \mathbf{S}_i[k] = \mathcal{N} \text{ and } \exists j \text{ such that } \mathbf{S}_i[j] \neq \mathcal{N} \\ \mathcal{N}, & \text{if } \mathbf{S}_i[j] = \mathcal{N} \ \forall j \ . \end{cases} \tag{2}$$

As a result, there are 4^K multimicrophone states; however, only 3^K of them correspond to valid conversation states (e.g., all participants cannot be in $\mathcal{V}_F$).

All 4 single-microphone acoustic models are defined over a feature space of $F = 41$ features: log-energy, 13 Mel-frequency cepstral coefficients (MFCCs; excluding $\mathbf{C}_0$), their first- and second-order derivatives, as well as the minimum and maximum normalized log-energy differences (NLEDs). The latter two features were designed for differentiating between nearfield and farfield vocal activity [23]. Using the reference speech and laughter segmentation of all 26 `Bmr` meetings, one Gaussian mixture with $M = 64$ components was trained per model to maximize the class-conditional likelihood of the training data.

3.3 Transition Model

We seek to a time-independent probability that conversation C will transition out of a multiparticipant state $\mathbf{S}_i$ into a multiparticipant state $\mathbf{S}_j$. As with emission probabilities, the fundamental difficulty is the potential for K to not be known, or ever seen in the training material.

Although a full exposition of our transition model considerably exceeds the current space limitations, we mention that the model probabilities are independent both of the identities of all participants and of their assignment to particular channels k, namely that

$$\begin{aligned} a_{ij} &= P(\mathbf{q}_{t+1} = \mathbf{S}_j \,|\, \mathbf{q}_t = \mathbf{S}_i) \\ &= P(\mathbf{R} \cdot \mathbf{q}_{t+1} = \mathbf{R} \cdot \mathbf{S}_j \,|\, \mathbf{R} \cdot \mathbf{q}_t = \mathbf{R} \cdot \mathbf{S}_i) \ . \end{aligned} \tag{3}$$

where $\mathbf{R}$ is an arbitrary $K \times K$ row rotation operator. We refer the reader to [24] for full details of the model, its general training algorithm, and its application.

Here, the transition model probabilities a_{ij} were trained using forced-alignment of the reference 3-way $\mathcal{N}/\mathcal{S}/\mathcal{L}$ multiparticipant segmentation. To achieve this, each frame $\mathbf{q}_t$ was assigned a pseudo-likelihood $P(\mathbf{q}_t|\mathbf{S}_i) = \alpha^d$, where d is the number of mismatched participant states between $\mathbf{q}_t$ and $\mathbf{S}_i$, and α is a small number (10^{-4}). The Viterbi pass was performed with all allowed transitions a_{ij} having a probability of unity (leading to $\sum_i a_{ij} \geq 1$), to not disfavor self-transitions at high fan-out states.

4 Performance of Proposed System

The HMM topology described in Subsection 3.1 was constructed with frame step and size of $T_{step} = T_{size} = 0.1$ seconds, as in our work on $\mathcal{V}/\neg\mathcal{V}$ segmentation

[25]. The minimum duration constraints $\mathbf{T}_{min} \equiv \left(T^{\mathcal{S}}_{min}, T^{\mathcal{L}}_{min}, T^{\mathcal{N}}_{min}\right)$ were set to $(0.2, 0.4, 0.3)$ seconds, leveraging our findings in [25] and [1]. The latter work, in which it was shown that overlap rates are higher for laughter than for speech, has also led us to impose the overlap constraints $\mathbf{K}_{max} \equiv \left(K^{\mathcal{S}}_{max}, K^{\mathcal{L}}_{max}, K^{\neg\mathcal{N}}_{max}\right) = (2, 3, 3)$. System sensitivity to these settings is explored in Section 5.

Table 1. Confusion matrix for 3-way $\mathcal{N}/\mathcal{S}/\mathcal{L}$ participant-state recognition for the system described in Section 3. Reference (Ref) class membership is shown in rows, hypothesized membership in columns. Time is shown in minutes; the total duration of the analyzed audio is 827 minutes. Total reference and hypothesized state occupation (*total*), per state, is given in italics in the last column and row, respectively.

Ref	Hypothesized as			
	$\mathcal{N}$	$\mathcal{S}$	$\mathcal{L}$	*total*
$\mathcal{N}$	685.4	7.8	22.9	*716.2*
$\mathcal{S}$	11.0	79.0	4.5	*94.4*
$\mathcal{L}$	6.5	1.0	9.2	*16.6*
total	*702.9*	*87.8*	*36.6*	

With emission and transition probabilities inferred as described in Subsections 3.2 & 3.3, the system was applied to the 3 `Bmr` meetings in the testset. The resulting confusion matrix is shown in Table 1. As can be seen, the prior distribution over the three classes $\mathcal{N}$, $\mathcal{S}$, and $\mathcal{L}$ (column 5), is significantly unbalanced. Laughter is hypothesized for 9.2 minutes out of the total 16.6 present, yielding a recall of 55.2%. However, laughter is also hypothesized for 22.9 minutes of nearfield silence, pulling precision down to 25.1%. In fact, the largest confusions in the matrix are seen between laughter and nearfield silence. Preliminary analysis suggests that this is due to laughter models capturing participants' breathing. Unvoiced laughter in particular is perceptually similar to exhalation. This suggests that, in future work, voiced and unvoiced laughter should be modeled separately, especially given that unvoiced laughter is overlapped with other unvoiced laughter only infrequently; the same is not true for voiced laughter [18].

5 Contrastive Experiments

In this section, we would like to answer the following questions:

1. *What role do minimum duration constraints play in detecting laughter?*
2. *What role do vocalization overlap constraints play in detecting laughter?*
3. *How does detection performance generalize to unseen datasets?*

We train alternate systems to answer each question, and contrast performance with that of the system from Section 4. Recall, precision, and F-scores of both speech and laughter $\mathcal{V} \equiv \mathcal{S} \cup \mathcal{L}$, of speech alone $\mathcal{S}$, and of laughter alone $\mathcal{L}$, are shown over the full 13.8 hours of test audio.

5.1 Minimum Duration Constraints

To determine the impact of duration modeling on system performance, we train two alternate transition models, differing in the minimum duration constraints $\mathbf{T}_{min} \equiv \left(T^{\mathcal{S}}_{min}, T^{\mathcal{L}}_{min}, T^{\mathcal{N}}_{min}\right)$ from the system described in Section 4. The first of these systems involves a fully-connected (ergodic) HMM topology, on which no minimum duration constraints are imposed (ie. $\mathbf{T}_{min} = (0.1, 0.1, 0.1)$ seconds, given our analysis frame step $T_{step} = 0.1$ s). The second system enforces equal minimum duration constraints of 0.3 s on each of the three AD states, $\mathcal{N}$, $\mathcal{S}$, and $\mathcal{L}$; its $\mathbf{T}_{min}$ is $(0.3, 0.3, 0.3)$ seconds. In every other respect, these two systems are identical to that in Section 4; performance of all three is shown in Table 2.

Table 2. Recall (R), precision (P) and F-score (F) as a function of minimum duration constraints $\mathbf{T}_{min} \equiv \left(T^{\mathcal{S}}_{min}, T^{\mathcal{L}}_{min}, T^{\mathcal{N}}_{min}\right)$. The frame step and frame size are identically 100ms, and the maximum simultaneous vocalization constraints $\mathbf{K}_{max} \equiv \left(K^{\mathcal{S}}_{max}, K^{\mathcal{L}}_{max}, K^{\neg\mathcal{N}}_{max}\right)$ are $(2, 3, 3)$ for all systems shown. Performance is shown for vocalization $\mathcal{V} = \mathcal{S} \cup \mathcal{L}$ (versus $\mathcal{N}$) in columns 2-4, for $\mathcal{S}$ (versus $\neg\mathcal{S} = \mathcal{N} \cup \mathcal{L}$) in columns 5-7, and for $\mathcal{L}$ (versus $\neg\mathcal{L} = \mathcal{N} \cup \mathcal{S}$) in columns 8-10. The system from Section 4 is identified with "§4"; best performance on each metric, across systems, is in bold.

	$\mathbf{T}_{min}$ (s)	$\mathcal{V} \equiv \mathcal{S} \cup \mathcal{L}$			$\mathcal{S}$			$\mathcal{L}$		
		R	P	F	R	P	F	R	P	F
	$(0.1, 0.1, 0.1)$	84.1	72.8	78.1	82.3	89.9	86.0	**55.9**	22.1	31.7
	$(0.3, 0.3, 0.3)$	**84.5**	75.1	**79.5**	**83.7**	**90.4**	**86.9**	54.7	24.2	33.6
§4	$(0.2, 0.4, 0.3)$	84.3	**75.3**	**79.5**	83.6	90.0	86.7	55.2	**25.1**	**34.5**

As the table shows, the system with equal minimum duration constraints of 300 ms on the occupation of each of $\mathcal{N}$, $\mathcal{S}$, and $\mathcal{L}$ outperforms the ergodic system on all measures except recall of laughter, which is lower by 1.2%. In particular, we note a 2.3% increase in $\mathcal{V}$ precision and a 2.1% increase in $\mathcal{L}$ precision. This variation is expected since the non-ergodic system cannot hypothesize spurious single-frame segments, which are unlikely to be vocal productions for physiological reasons. For assessing whether minimum duration constraints discriminate between speech and laughter, the $\mathbf{T}_{min} = (0.3, 0.3, 0.3)$ system is most appropriate because both it and the system in Section 4 allow each participant to be in one of 9 states; in the ergodic system, that number of state is 3. Table 2 shows that both the recall and precision of laughter are higher in the $(0.2, 0.4, 0.3)$ system than in the $(0.3, 0.3, 0.3)$ system, and suggests that minimum duration constraints can be used to advantage when detecting laughter-in-interaction in multi-channel audio.

5.2 Maximum Simultaneous Vocalization Constraints

Second, we assess the impact of limiting the maximum number of participants allowed to simultaneously vocalize by modifying the maximum simultaneous

vocalization constraints $\mathbf{K}_{max} \equiv \left(K^{\mathcal{S}}_{max}, K^{\mathcal{L}}_{max}, K^{\neg\mathcal{N}}_{max}\right)$. For this purpose, we construct 3 alternate systems. The first, whose $\mathbf{K}_{max} = (2, 2, 2)$, allows up to two participants to be in single-participant states other than $\mathcal{N}^{(0)}$, and up to two participants to be simultaneously speaking or laughing. This is a standard extension of our meeting recognition $\mathcal{V}/\neg\mathcal{V}$ segmenter [25]. The second alternate system, whose $\mathbf{K}_{max} = (2, 2, 3)$, adds two additional cases: (1) only two participants speaking and only one participant laughing; and (2) only two participants laughing and one participant speaking. Finally, the third alternate system ($\mathbf{K}_{max} = (3, 2, 3)$) adds the case of only three participants speaking and none laughing. In contrast, the system desribed in Section 4, allows for only three participants laughing and none speaking. The $\mathbf{K}_{max} = (3, 2, 3)$ could be expected to outperform the $\mathbf{K}_{max} = (2, 3, 3)$ system if speech exhibited higher rates of overlap than does laughter. All 4 systems are shown in Table 3.

Table 3. Recall (R), precision (P), and F-score (F) as a function of maximum simultaneous vocalization constraints $\mathbf{K}_{max} \equiv \left(K^{\mathcal{S}}_{max}, K^{\mathcal{L}}_{max}, K^{\neg\mathcal{N}}_{max}\right)$. The frame step and frame size are identically 100ms, and the minimum duration constraints $\mathbf{T}_{min} \equiv \left(T^{\mathcal{S}}_{min}, T^{\mathcal{L}}_{min}, T^{\mathcal{N}}_{min}\right)$ are $(0.2, 0.4, 0.3)$ seconds for all systems shown. Symbols as in Table 2.

	$\mathbf{K}_{max}$	$\mathcal{V} \equiv \mathcal{S} \cup \mathcal{L}$			$\mathcal{S}$			$\mathcal{L}$		
		R	P	F	R	P	F	R	P	F
	$(2, 2, 2)$	80.5	**82.1**	**81.3**	83.3	**90.6**	**86.8**	36.9	**27.8**	31.7
	$(2, 2, 3)$	84.0	76.1	79.9	84.0	89.0	86.4	48.8	24.3	32.4
	$(3, 2, 3)$	84.1	76.1	79.9	**84.2**	88.6	86.4	49.1	24.6	32.8
§4	$(2, 3, 3)$	**84.3**	75.3	79.5	83.6	90.0	86.7	**55.2**	25.1	**34.5**

As Table 3 shows, increasing $K^{\neg\mathcal{N}}_{max}$ from 2 to 3 increases recall but reduces precision; the effect is more dramatic for $\mathcal{L}$ than for $\mathcal{S}$ because more of laughter than of speech occurs in overlap. Allowing a third simultaneous speaker decreases $\mathcal{S}$ precision by 0.4% and increases $\mathcal{S}$ recall by 0.2%. In contrast, allowing a third simultaneous laugher increases $\mathcal{L}$ precision by 0.8%, and at the same time increases $\mathcal{L}$ recall by 6.4%.

5.3 Generalization to Other Data

To close this section, we explore the performance of the system described in Section 4 on several other datasets drawn from the ICSI Meeting Corpus. In Table 4, we show the performance of our system on the `Bro` meetings, of which there are 23, and on the `Bed` meetings, of which there are 15. Both of these sets were completely unseen during development, and consist of 116 and 81 total hours of single-channel audio, respectively.

We note first of all that although $\mathcal{V}$ recall and precision are lower on `Bmr`(test) than on `Bmr`(train) by 0.8% and 0.4%, respectively, the differences are small. This

Table 4. Recall (R), precision (P), and F-score (F) of the system described in Section 4 on different subsets of the ICSI Meeting Corpus. $p_{\mathcal{V}}(\mathcal{L})$ is the proportion of vocalization time spent in laughter. Symbols as in Table 2.

Test data		$p_{\mathcal{V}}(\mathcal{L})$	$\mathcal{V} \equiv \mathcal{S} \cup \mathcal{L}$			$\mathcal{S}$			$\mathcal{L}$		
			R	P	F	R	P	F	R	P	F
`Bmr`	train	10.91	85.1	**75.7**	**80.1**	83.4	89.8	86.5	53.0	19.4	28.4
	test	14.94	84.3	75.3	79.5	83.6	90.0	**86.7**	55.2	**25.1**	**34.5**
`Bro`	(all)	5.94	83.7	73.2	78.1	81.1	**90.6**	85.6	57.8	11.4	19.0
`Bed`	(all)	7.53	**88.5**	65.2	75.1	**84.6**	85.7	85.2	**58.7**	10.0	17.0

suggests that model complexity is low and the system not particularly prone to overfitting. It is more surprising that $\mathcal{V}$ performance on the training data is not higher, and may be indicative of the difficulty of the task.

As can be seen, laughter detection for `Bmr`(test) is better than for `Bmr`(train), and much better in both `Bmr` subsets than for either the `Bed` or `Bro` meetings. It appears that $\mathcal{L}$ precision is strongly correlated ($r = 0.943$) with the proportion of vocalization time spent in laughter ($p_{\mathcal{V}}(\mathcal{L})$ in column 3). Although $p_{\mathcal{V}}(\mathcal{L})$ is higher for `Bed` meetings than for `Bro` meetings, F-scores are higher for the latter for all three of $\mathcal{V}$, $\mathcal{S}$, and $\mathcal{L}$. This is likely attributable to the fact that fewer of the `Bed` meeting participants than of the `Bro` meeting participants are present in the `Bmr` training data (cf. Section 2).

The above findings indicate that the proposed data split [2, 8, 9, 11, 10] is not particularly helpful in predicting laughter detection performance on unseen data. This is because the `Bmr` test meetings contain an atypically high proportion of transcribed laughter, even within the `Bmr` subset, rendering the distribution of vocal activity types more balanced than elsewhere in the corpus, and therefore detection results more optimistic. Further analysis is required to assess the correlation between detectability and factors such as participant identity, laughter quality, and the degree of laughter overlap by time.

6 Qualitative Comparison with Related Work

As mentioned in the Introduction, aspects of laughter detection in meetings have been treated in [2, 8, 9, 11, 10]. Although the goal of each of the aforementioned publications was different from ours, we present several common and differentiating aspects in Table 5.

In the earliest work, [2], the authors dealt with multiple farfield microphones, in an effort to identify simultaneous laughter from the majority of participants present, with no intention of attributing laughter to specific participants. These three aspects make [2] the most dissimilar from among the work cited in Table 5.

Research on laughter/speech classification [8, 9] has assumed the presence of manual pre-segmentation into intervals of approximately 2 s in duration and anticipates balanced priors in the testset. Furthermore, it treats only 47% of the

transcribed laugh bouts, namely those which have been assigned their own utterances by the original ICSI transcription team. Although these conditions are different from the ones faced in the current work, [9] has shown that focusing on only 28% of the transcribed laugh bouts, those considered clearly perceptible, decreases EERs by 4%. This suggests that $\mathcal{N}/\mathcal{S}/\mathcal{L}$ segmentation may benefit by treating different types of laughter differently, especially if applications distinguish among laughter types.

Table 5. Overview of previous research on laughter/speech ($\mathcal{L}/\mathcal{S}$) classification and laughter/non-laughter ($\mathcal{L}/\neg\mathcal{L}$) segmentation, and of the current work, in terms of several differentiating aspects.

Aspect	$\mathcal{L}/\mathcal{S}$ class.		$\mathcal{L}/\neg\mathcal{L}$ segm.			this
	[8]	[9]	[11]	[10]	[2]	work
close-talk microphones	✓	✓	✓	✓		✓
farfield microphones					✓	
single channel at-a-time	✓	✓	✓	✓		
multi-channel at-a-time					✓	✓
participant attribution	✓	✓	✓	✓		✓
only group laughter					✓	
only isolated laughter	✓		✓	✓		
only clear laughter		✓				
rely on pre-segmentation	✓	✓	?			
rely on prior rebalancing	✓	✓	?			
rely on channel exclusion			?	✓		

Research in laughter/non-laughter segmentation [11, 10] is more relevant to the current work. This is not least because, as we have shown, nearfield laughter tends to be confused much more with nearfield silence than with nearfield speech. In spite of this, and despite identical training and testing data, a direct performance comparison with the current work is not possible. [10] assumes the presence of a preliminary (perfect) vocal activity detector which justifies the exclusion of nearfield channels exhibiting prolonged silence during testing. This is effectively a form of pre-segmentation which also achieves prior rebalancing, and the extent to which [10] relies on such exclusion is not documented. Furthermore, contrary to our own unpublished observations, the experiments in [10] recommend a framing policy with a small frame step but a large frame size; in conjunction with the current work, a potential emerging strategy is multipass segmentation in which frame step and frame size decrease and increase, respectively, from one pass to the next.

For completion, it should be noted that low precision continues to be a challenging problem [12] in speech/non-speech segmentation [26, 21, 22], and automatic speech recognition word error rates are currently 2-3% absolute higher with automatically produced segments than with manual segmentation [23, 27,

25]. As our confusion matrix in Section 4 shows, the separation between speech and silence appears to be easier than that between laughter and silence, and laughter segmenters exposed to the full duration of meeting audio are likely to incur more insertions than those exposed only to pre-segmented portions.

7 Conclusions

We have proposed a simultaneous multiparticipant architecture for the detection of laughter in multi-channel close-talk microphone recordings of meetings. The implemented system does not rely on any form of manual pre-segmentation, and achieves laughter recall and precision rates of 55.2% and 25.1%, respectively, on a commonly used 14-hour dataset in which laughter accounts for 2% of time. These figures represent the first baseline results for this task, and the findings indicate that discrimination between nearfield laughter and nearfield silence, rather than between nearfield laughter and nearfield speech, presents the biggest difficulties.

Our experiments suggest that laughter segmentation stands to benefit from contrastive constraints placed on the maximum allowed degree of simultaneous vocalization as well as on minimum allowed state duration. Finally, we have shown that laughter precision throughout the ICSI Meeting Corpus is most strongly a function of the proportion of laughter present, and only second a function of participant novelty.

Acknowledgments

We would like to thank Liz Shriberg for access to the ICSI MRDA Corpus.

References

1. Laskowski, K., Burger, S.: Analysis of the occurrence of laughter in meetings. In: Proc. INTERSPEECH, Antwerpen, Belgium (2007) 1258–1261
2. Kennedy, L., Ellis, D.: Laughter detection in meetings. In: Proc. ICASSP Meeting Recognition Workshop, Montreal, Canada, NIST (2004) 118–121
3. Russell, J., Bachorowski, J.A., Fernandez-Dols, J.M.: Facial and vocal expressions of emotion. Annual Review of Psychology **54** (2003) 329–349
4. Laskowski, K., Burger, S.: Annotation and analysis of emotionally relevant behavior in the ISL Meeting Corpus. In: Proc. LREC, Genoa, Italy (2006)
5. Galley, M., McKeown, K., Fosler-Lussier, E., Jing, H.: Discourse segmentation of multi-party conversation. In: Proc. ACL, Sapporo, Japan (2003) 562–569
6. Banerjee, S., Rose, C., Rudnicky, A.: The necessity of a meeting recording and playback system, and the benefit of topic-level annotations to meeting browsing. In: Proc. INTERACT. Volume 3585 of Springer Lecture Notes in Computer Science., Rome, Italy (2005) 643–656
7. Wrede, B., Shriberg, E.: Spotting "hotspots" in meetings: Human judgments and prosodic cues. In: Proc. EUROSPEECH, Geneva, Switzerland (2003) 2805–2808
8. Truong, K., van Leeuwen, D.: Automatic detection of laughter. In: Proc. INTERSPEECH, Lisbon, Portugal (2005) 485–488

9. Truong, K., van Leeuwen, D.: Automatic discrimination between laughter and speech. Speech Communication **49**(2) (2007) 144–158
10. Knox, M., Mirghafori, N.: Automatic laughter detection using neural networks. In: Proc. INTERSPEECH, Antwerpen, Belgium (2007) 2973–2976
11. Truong, K., van Leeuwen, D.: Evaluating automatic laughter segmentation in meetings using acoustic and acoustics-phonetic features. In: Proc. ICPhS Workshop on The Phonetics of Laughter, Saarbrücken, Germany (2007) 49–53
12. Pfau, T., Ellis, D., Stolcke, A.: Multispeaker speech activity detection for the ICSI Meeting Recorder. In: Proc. ASRU, Madonna di Campiglio, Italy (2001) 107–110
13. Janin, A., et al.: The ICSI Meeting Corpus. In: Proc. ICASSP, Hong Kong, China (2003) 364–367
14. Shriberg, E., Dhillon, R., Bhagat, S., Ang, J., Carvey, H.: The ICSI Meeting Recorder Dialog Act (MRDA) Corpus. In: Proc. SIGdial, Cambridge MA, USA (2004) 97–100
15. Norwine, A., Murphy, O.: Characteristic time intervals in telephonic conversation. Bell System Technical Journal **17** (1938) 281–291
16. Fiscus, J., Ajot, J., Michel, M., Garofolo, J.: The Rich Transcription 2006 Spring Meeting Recognition Evaluation. In: Proc. MLMI. Volume 4299 of Springer Lecture Notes in Computer Science., Bethesda MD, USA (2006) 309–322
17. Bachorowski, J.A., Smoski, M., Owren, M.: The acoustic features of human laughter. J. of Acoustical Society of America **110**(3) (2001) 1581–1597
18. Laskowski, K., Burger, S.: On the correlation between perceptual and contextual aspects of laughter in meetings. In: Proc. ICPhS Workshop on the Phonetics of Laughter, Saarbrücken, Germany (2007)
19. Nwokah, E., Hsu, H.C., Davies, P., Fogel, A.: The integration of laughter and speech in vocal communication: A dynamic systems perspective. J. of Speech, Language & Hearing Research **42** (1999) 880–894
20. Laskowski, K., Schultz, T.: A supervised factorial acoustic model for simultaneous multiparticipant vocal activity detection in close-talk microphone recordings of meetings. Technical Report CMU-LTI-07-017, Carnegie Mellon University, Pittsburgh PA, USA (December 2007)
21. Wrigley, S., Brown, G., Wan, V., Renals, S.: Speech and crosstalk detection in multichannel audio. IEEE Trans. Speech and Audio Proc. **13**(1) (2005) 84–91
22. Huang, Z., Harper, M.: Speech activity detection on multichannels of meetings recordings. In: Proc. MLMI. Volume 3869 of Springer Lecture Notes in Computer Science., Edinburgh, UK (2005) 415–427
23. Boakye, K., Stolcke, A.: Improved speech activity detection using cross-channel features for recognition of multiparty meetings. In: Proc. INTERSPEECH, Pittsburgh PA, USA (2006) 1962–1965
24. Laskowski, K., Schultz, T.: Modeling duration contraints for simultaneous multiparticipant vocal activity detection in meetings. Technical report, Carnegie Mellon University, Pittsburgh PA, USA (February 2008) in preparation.
25. Laskowski, K., Fügen, C., Schultz, T.: Simultaneous multispeaker segmentation for automatic meeting recognition. In: Proc. EUSIPCO, Poznań, Poland (2007) 1294–1298
26. Wrigley, S., Brown, G., Wan, V., Renals, S.: Feature selection for the classification of crosstalk in multi-channel audio. In: Proc. EUROSPEECH, Geneva, Switzerland (2003) 469–472
27. Dines, J., Vepa, J., Hain, T.: The segmentation of multi-channel meeting recordings for automatic speech recognition. In: Proc. INTERSPEECH, Pittsburgh PA, USA (2006) 1213–1216