Detecting Attempts at Humor in Multiparty Meetings

Kornel Laskowski
Language Technologies Institute
Carnegie Mellon University
Pittsburgh PA, USA
kornel@cs.cmu.edu

***Abstract*—Systems designed for the automatic summarization of meetings have considered the propositional content of contributions by each speaker, but not the explicit techniques that speakers use to downgrade the perceived seriousness of those contributions. We analyze one such technique, namely attempts at humor. We find that speech spent on attempts at humor is rare by time but that it correlates strongly with laughter, which is more frequent. Contextual features describing the temporal and multiparticipant distribution of manually transcribed laughter yield error rates for the detection of attempts at humor which are 4 times lower than those obtained using oracle lexical information. Furthermore, we show that similar performance can be achieved by considering only the speaker's laughter, indicating that meeting participants explicitly signal their attempts at humor by laughing themselves. Finally, we present evidence which suggests that, on small time scales, the production of attempts at humor and their ratification via laughter often involves only two participants, belying the allegedly multiparty nature of the interaction.**

***Keywords*-humor; laughter; conversation; modeling; meetings.**

I. Introduction

Human understanding of multiparty conversation relies on a wide range of linguistic and para-linguistic abilities. Duplication of these abilities by machine has been driven largely by the desire to automatically distill or summarize recordings of meetings [1]. Systems built for this task have considered the propositional content of speaker contributions, but have neglected that speakers may be qualifying those contributions in ways which are important to deciphering their intent. At one end of the spectrum, speakers can be deceptive [2], i.e. they may covertly exert effort to impart as true or as serious information which is neither. At the other end, they may overtly exert effort to impart as unserious information which might otherwise be taken at face value [3]. Both types of speaker behavior have clear implications not only for automatic summarization but also for many other systems deployed in naturally occurring conversational settings.

This work considers the detection of the latter type of behavior, namely of those verbal productions which constitute attempts at unseriousness or humor. Regardless of the type of conversation, humor plays a socially cohesive role, offering opportunity for the expression of solidarity and alignment [3]. As a result, its detection may be crucial to the automatic inference of group identity and of the construction, maintenance, and dissolution of interpersonal relationships. In conversations whose main aim is the collaborative creation and/or dissemination of information, such as work-oriented meetings, humor-bearing talk is reported to arise at specific instants of the interaction [4], and may therefore be important to the automatic segmentation of meeting records at the turn, topic, or meta-conversation levels [5].

Although humor, and the laughter that often ensues, has been extensively studied in the social sciences [6], [7], [3], in computational settings its treatment has been largely limited to the construction by synthetic agents of humor-bearing language [8]. Computational modeling of how such language is *deployed* by humans has received almost no attention; an earlier attempt to find humor in the same data as used in this work was briefly reported on in [9], with results "not distinguishable from chance". Work on meetings which is related to our proposed task includes description of emotionally relevant behavior [10], the detection of sentiment [11], and the detection of involvement [12].

In contrast, a considerable body of research exists on the acoustic detection of laughter in meetings [13], [14], [15], [16], [17], whose co-occurrence with humor-bearing talk appears self-evident but which, to our knowledge, has never been measured. This measurement, via a system which predicts attempts at humor from surrounding laughter, is the main goal of the current work. We employ a large corpus of naturally-occurring multiparty meetings, described in Section II, and first present and benchmark (in Sections III and IV-A) the performance of a system which jointly segments and classifies the human-transcribed words into mutually exclusive dialog act (DA) types, of which one models attempts at humor. This system achieves significantly above-chance performance, in contrast to [9]. We then go on to show that a model of the laughter context, derived from manually segmented laughter instances from all participants and described in Section IV-B, yields a system whose miss rate is lower by 72% relative. System combination is presented in Sections IV-C and IV-D, and yields additional improvement. Analysis, discussion, and conclusions of these findings are provided in Sections V, VI and VII, respectively.

II. DATA

The data used in this work is the ICSI Meeting Corpus, consisting of 75 longitudinal recordings of naturally occurring meetings by several groups at ICSI [18], [19]. We rely on the previously published split of this data into a TRAINSET of 51 meetings, and a DEVSET and a TESTSET of 11 meetings each.

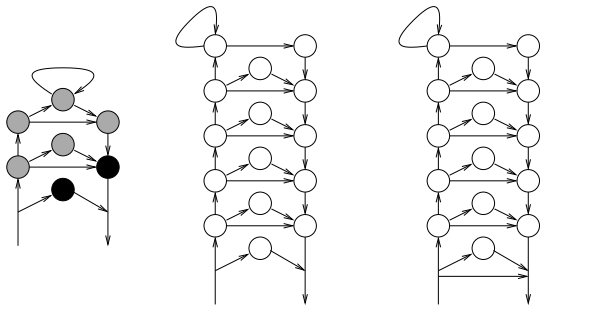
The meetings are provided with forced alignment of all words spoken, as well as with DA annotation. The systems constructed in this work consider eight dialog act types (namely: floor grabbers `fg`, floor holders `fh`, holds `h`, backchannels `b`, acknowledgments `bk`, asserts `aa`, questions `q`, and statements `s`), but separate from questions and statements those DAs that have been additionally annotated as utterances of humorous or sarcastic nature (`j`). These are modeled as a separate DA. We note that a joke may consist of multiple consecutive `j` DAs. The proportion of these 9 DA types in all three data sets is shown in Table I. `j` DAs, which we refer to as attempts at humor, account for 0.53–0.73% of speaking time.

Table I
PROPORTION BY TIME, IN %, OF SPEECH IMPLEMENTING 9 DA TYPES, FOR ALL THREE DATASETS; ATTEMPTS AT HUMOR (`j`) SHOWN IN BOLD.

DA Type	TRAINSET	DEVSET	EVALSET
`aa`	1.17	1.13	1.10
`b`	2.86	2.65	2.83
`bk`	1.41	1.41	1.48
`fg`	0.55	0.58	0.62
`fh`	2.32	2.29	3.00
`h`	0.21	0.36	0.26
`j`	**0.73**	**0.53**	**0.62**
`q`	6.47	7.41	7.86
`s`	84.28	83.63	82.24

III. BASELINE

Our baseline system is a hidden Markov model (HMM) Viterbi decoder, with a frame size and step of 100 ms. The HMM topology is constructed in a hierarchical fashion as follows. Each of the nine types of DAs is modeled using an identical DA network. A DA network consists of one subnetwork modeling a non-DA-terminal stretch of contiguous speech, which we refer to as a talkspurt fragment (TSF), one subnetwork modeling intra-DA non-speech (GAP), and one subnetwork modeling a DA-terminal TSF. Non-DA-terminal TSF subnetworks may be visited repeatedly per DA. The proposed TSF and GAP subnetworks are shown in panels (a) and (b) of Figure 1, respectively, and are explained in more detail in our earlier work [20]. `q` and `s` DA networks both have two additional alternative DA-terminating TSF subnetworks, modeling DA termination due to abandonment and interruption. Transitions between the 9 DA networks are mediated via inter-DA GAP subnetworks, shown in panel (c) of Figure 1, which are identical to intra-DA GAP subnetworks except that they may have zero duration.



(a) TSF (b) intra-DA GAP (c) inter-DA GAP

Figure 1. Subnetworks in an HMM topology for conversational speech, with a frame step of 100 ms; states shown in white denote non-speech. In (a), egress states, optionally punctuation-bearing, are shown in black. Note that (b) and (c) are identical, except that inter-DA GAPs may have zero duration.

The observables in each speech state of the full topology are the left and right bigram of the word whose temporal support coincides with the 100 ms duration of the state in question. We model these bigrams as follows. Where words are separated by more than 0.7 s of non-speech, we insert a `SIL` token; this value was found to be optimal in TRAINSET for unconditionally deciding whether an inter-word gap is also a inter-DA gap (cf. [21]). Then, for each DA type d, $1 \leq d \leq 8$, we form the set $\mathcal{U}_d$ of unigrams from TRAINSET. These are sorted by frequency of occurrence, and those unigrams whose probability of occurrence exceeds 0.1% are placed in $\mathcal{U}'_d$. Unigrams not found in the union $\mathcal{U}' \equiv \cup_d \mathcal{U}'_d$ are mapped to the token `UNK1`. Following this mapping, we form for each DA type d the set $\mathcal{B}_d$ of bigrams, some of which may contain `UNK1` and/or `SIL`. As for unigrams, those bigrams in $\mathcal{B}_d$ whose probability of occurrence exceeds 0.1% are placed in $\mathcal{B}'_d$, and those not found in the resulting union $\mathcal{B}' \equiv \mathcal{B}'_d$ are mapped to `UNK2`. $\mathcal{B}'$, together with the `UNK2` token, comprises our lexical feature space. The resulting model was found to perform slightly better [20] than the hidden event language model on a more standard 5-DA-type classification task [21].

IV. EXPERIMENTS

We present the results of several experiments, involving functional components of the baseline system and of systems which rely on the surrounding multiparticipant laughter and speech activity contexts; we also present the performance of a system combining all three sources of information. Systems are trained using TRAINSET, with model parameters optimized using DEVSET. We show the performance of the DEVSET-optimized systems on EVALSET as a measure of generalization to unseen data, alongside the DEVSET numbers.

Performance is measured using the commonly-employed detection error (ERR), which is the sum of the miss rate

(MS, true negatives normalized by all positives) and the false alarm rate (FA, false positives normalized by all negatives). We report only the MS and FA rates for model parameter combinations at which the lowest ERR was observed for each system. Assessment of performance using all observed MS and FA pairs, via receiver operating characteristic curves, is presented in Section IV-E.

A. Baseline Performance

The baseline system, described in Section III, consists of an HMM topology which licenses only a subset of possible state transitions, a transition probability model which assigns probabilities to those transitions, and an emission probability model which describes the likelihood of observing specific bigrams at each state. We present its performance, as well that of systems with some of these components ablated, in Table II.

The first line in Table II is for a system T0 whose transition probabilities are equiprobable and whose emission probabilities of speech are unity for TSF states and zero for GAP states. We present this system only to contrast it with system T1 (on the second line), which is identical but whose transition probabilities are inferred from TRAINSET. Low error rates for system T1 relative to that of T0 would indicate that j DAs can be predicted from talkspurt duration and from the sequencing of j DAs with respect to other DA types. As can be seen, the miss rates achieved by both T0 and T1 are quite high, indicating that j talk cannot be predicted in this way.

Table II
DETECTION PERFORMANCE, IN %, OF THE TOPOLOGY WITH EQUIPROBABLE TRANSITION PROBABILITIES (T0), THE TOPOLOGY WITH TRANSITION PROBABILITIES TRAINED USING TRAINSET (T1), THE LEXICAL EMISSION MODEL ALONE (LEX W/O T), AND THE LEXICAL EMISSION MODEL WITH BOTH T0 AND T1 (LEX W/ T0 AND LEX W/ T1, RESPECTIVELY). FA IS THE FALSE ALARM RATE, MS IS THE MISS RATE, AND ERR = FA + MS.

System	DEVSET			EVALSET		
	FA	MS	ERR	FA	MS	ERR
T0	8.1	90.6	98.7	8.3	92.5	100.7
T1	0.3	96.7	97.0	0.2	94.0	94.2
LEX w/o T	53.6	32.8	86.4	53.7	32.9	86.6
LEX w/ T0	40.2	42.9	83.1	40.5	44.2	84.7
LEX w/ T1	12.7	67.0	79.6	12.8	70.5	83.3

The third line (LEX w/o T) shows the performance of the bigram emission probability model, with no topological constraints (i.e. an ergodic HMM topology). We stress that the model relies on transcribed rather than automatically recognized words, and therefore represents the maximum achievable performance. Lines 4 and 5 in Table II show the performance of systems for which the bigram emission probabilities are embedded in the proposed topology, with equiprobable transition probabilities (T0) and transition probabilities inferred from TRAINSET (T1), respectively. Although performance is significantly above random guessing (cf. also Section IV-E), these three systems demonstrate that the bigram features do not discriminate very successfully between attempts at humor and other DA types (in contrast to their utility for discriminating among non-j DAs [20]). In the rest of this work, the system "LEX w/ T1" in line 5 of the table will be referred to simply as LEX.

B. The Laughter Context

We now turn to an alternative source of information, that of the laughter context in which DAs are produced. We anticipate that subsequent laughter is a strong predictor of whether specific DAs are attempts at humor or not. Table III lists the proportion of vocalizing time which is spent on the production of laughter, for TRAINSET, DEVSET, and EVALSET. In all three sets, laughter is significantly more frequent by time than are verbalized attempts at humor (cf. Table I).

Table III
PROPORTION BY VOCALIZING TIME, IN %, OF LAUGHTER, FOR ALL THREE DATASETS.

Laughter	TRAINSET	DEVSET	EVALSET
all, $\mathcal{L}$	10.0	8.9	10.0
voiced, $\mathcal{L}_V$	7.6	6.3	6.6
unvoiced, $\mathcal{L}_U$	2.4	2.6	3.4

We propose to model the laughter context in the following way. For each instant t, the participant we are currently decoding (the "target participant") is in a specific HMM state. For speech states, i.e. those implementing a TSF, we rank the remaining participants by the amount of laughter they produce over a $[-5, +5]$-second context around t. We tessellate this context with 0.5-second windows, and then extract the proportion of time each of the top-3-ranked interlocutors laughs in each window; we also include the same features for the target participant. Additionally, for each of the 3 interlocutors, we extract whether they are laughing or not at instant t. This leads to 21 features from each of the 3 most-laughing interlocutors, and 20 features from the participant currently being decoded, for a total of 83 features. An example of context tessellation and feature extraction is shown in Figure 2.

We rotate the computed features using linear discriminant analysis (LDA), and model their emission probability for each state with a Gaussian mixture (GMM); the number of LDA discriminants and GMM components, as well as the weight of the resulting emission model log-likelihood relative to the transition log-probability, are optimized to minimize ERR on DEVSET. The results are shown in Table IV. As can be seen, laughter context features are significantly better than lexical features; on DEVSET, they yield error rates which are only a quarter of the error rate achieved with our baseline system, due mostly to much lower miss rates.

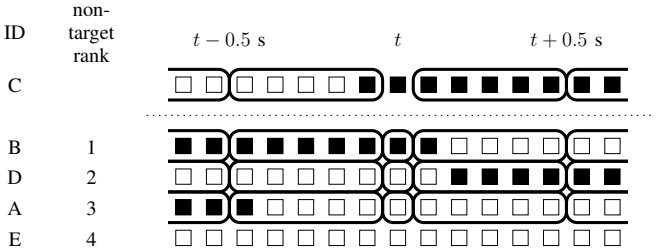


Figure 2. An example of interlocutor rotation and feature extraction at instant t, with time depicted from left to right, when decoding participant C. Non-target participants A, B, D, and E have been ranked according to their amount of laughter in the local neighborhood, shown as black squares. Windows for which features are extracted are shown as ovals; a single mean laughter activity posterior is computed for each, out to $t-5.0$ s and $t+0.5$ s (only windows near t are shown).

Table IV
DETECTION PERFORMANCE, IN %, OF SEVERAL SYSTEMS EMPLOYING TOPOLOGY T1, THE LAUGHTER $\mathcal{L}$ CONTEXT AND THE SPEECH $\mathcal{S}$ CONTEXT; FA IS THE FALSE ALARM RATE, MS IS THE MISS RATE, AND ERR = FA + MS.

System	DEVSET			EVALSET		
	FA	MS	ERR	FA	MS	ERR
$\mathcal{L}$	14.0	5.3	19.3	15.6	8.1	23.7
$\mathcal{L}_V$	8.7	16.0	24.7	9.5	9.9	19.4
$\mathcal{L}_U$	12.4	21.2	33.6	13.8	17.4	31.2
$\mathcal{L}_V$ Ⓜ $\mathcal{L}_U$	7.4	15.7	23.1	8.0	13.6	21.7
$\mathcal{L}_V$ Ⓕ $\mathcal{L}_U$	14.2	6.6	20.8	15.7	7.0	22.7
$\mathcal{L}_V$ Ⓒ $\mathcal{L}_U$	14.0	6.3	20.3	15.1	8.3	23.3
$\mathcal{S}$	7.5	47.4	54.9	8.6	62.8	71.4
$\mathcal{L}$ Ⓜ $\mathcal{S}$	9.7	6.6	16.3	11.0	8.4	19.4
$\mathcal{L}$ Ⓕ $\mathcal{S}$	6.0	17.8	23.8	6.8	21.6	28.4
$\mathcal{L}$ Ⓒ $\mathcal{S}$	6.0	16.0	22.0	6.4	17.8	24.2

We duplicate these experiments using only voiced laughter ($\mathcal{L}_V$) and only unvoiced laughter ($\mathcal{L}_U$), defined as involving and not involving periodic glottal excitation [22], respectively, in lines 2 and 3 of the table. We make this distinction because, for the detection of involved speech (most of it judged amused), it was reported that excluding unvoiced laughter leads to improved performance [12]. However, as Table IV shows, on the current task and on DEVSET data it appears that both types of laughter are important, with unvoiced laughter less relevant than voiced laughter. Model-space and feature-space combination of $\mathcal{L}_V$ and $\mathcal{L}_U$ features, denoted Ⓜ and Ⓕ, respectively, in Table IV, offer performance which is better than either laughter type alone but not better than all laughter $\mathcal{L}$, indicating that the voicing distinction hurts performance on this task. We note that model-space and feature-space combinations can in theory involve up to 6 interlocutors, since interlocutors are ranked independently for $\mathcal{L}_V$ and $\mathcal{L}_U$ feature computation according to their amount of $\mathcal{L}_V$ and $\mathcal{L}_U$ time (cf. Figure 2). The alternative *feature-computation-space combination*, using $\mathcal{L}$ to rank interlocutors first (regardless of relative quantity of $\mathcal{L}_V$ and $\mathcal{L}_U$), and only then extracting $\mathcal{L}_V$ and $\mathcal{L}_U$ features (and thus involving exactly 3 interlocutors), is shown in line 6 of the table as Ⓒ, and also does not outperform modeling all $\mathcal{L}$ in a single model.

Although laughter appears to be almost as relevant to humor detection in EVALSET as in DEVSET, several of the above mentioned trends do not generalize to this set. In particular, voiced laughter does appear to be better than all laughter. As a result, model-space, feature-space, and feature-computation-space combinations perform better than does all laughter, but in this case not better than only voiced laughter. We suspect that these differences between DEVSET and EVALSET performance are due to the complexity of the feature extraction regions employed here, which were proposed for modeling speech context in [20] and have not been re-optimized for the current task.

C. The Speech Context

We apply the same feature extraction method as was applied to the multiparticipant laughter context in the previous section, to the multiparticipant speech $\mathcal{S}$ context. The results are shown in the bottom four lines of Table IV. As could be expected, speech context features are not as good at predicting attempts at humor as laughter context features; nevertheless, they are significantly better than lexical features (cf. Table II). Although the difference is much more dramatic for DEVSET, a similar albeit weaker trend is observed for EVALSET. We attribute the difference in performance for the two datasets, as in Section IV-B, to the complexity of the feature extraction process; although appropriate for improving precision for the 8 non-`j` DA types [20], a simpler set of context features may be more appropriate to reducing the `j` false alarms.

The table also shows the three types of system combination described in Section IV-B, this time of the $\mathcal{L}$ and $\mathcal{S}$ systems. Although feature-space (Ⓕ) and feature-computation-space (Ⓒ) combinations lead to higher error rates, model-space combination (Ⓜ) improves the performance over the laughter-only $\mathcal{L}$ system. This is observed for EVALSET also, and to a similar extent. The speech context appears to offer complimentary information for predicting attempts at humor.

D. Augmenting the Baseline

We now combine all three of the LEX system (cf. Table II), the $\mathcal{L}$ system (cf. Table IV), and the $\mathcal{S}$ system (cf. Table IV) to yield their model-space combination ALL = LEX Ⓜ $\mathcal{L}$ Ⓜ $\mathcal{S}$. The results, shown in Table V, demonstrate a relative reduction in error, over the best system $\mathcal{L}$, of 23% for DEVSET and 18% for EVALSET. At the minimum ERR point, the $\mathcal{S}$ and LEX systems appear to lower the false alarm rate otherwise incurred by the $\mathcal{L}$ system alone.

E. Receiver Operating Characteristics

To describe system performance at locations other than the ERR minimum, we present receiver operating characteristic

Table V
DETECTION PERFORMANCE, IN %, OF THE MODEL-SPACE COMBINATION (ALL) OF THREE SYSTEMS EMPLOYING TOPOLOGY T1; FA IS THE FALSE ALARM RATE, MS IS THE MISS RATE, AND ERR = FA + MS.

System	DEVSET			EVALSET		
	FA	MS	ERR	FA	MS	ERR
LEX	12.7	67.0	79.6	12.8	70.5	83.3
$\mathcal{S}$	7.5	47.4	54.9	8.6	62.8	71.4
$\mathcal{L}$	14.0	5.3	19.3	15.6	8.1	23.7
ALL	7.7	7.2	14.8	8.3	11.0	19.4

curves in Figure 3. Points forming each system curve are the convex hull of FA and MS error pairs observed during DEVSET tuning of the systems described in Sections III, IV-B, IV-C and IV-D. Also shown are the line of no discrimination and the equal error line for which FA = MS.

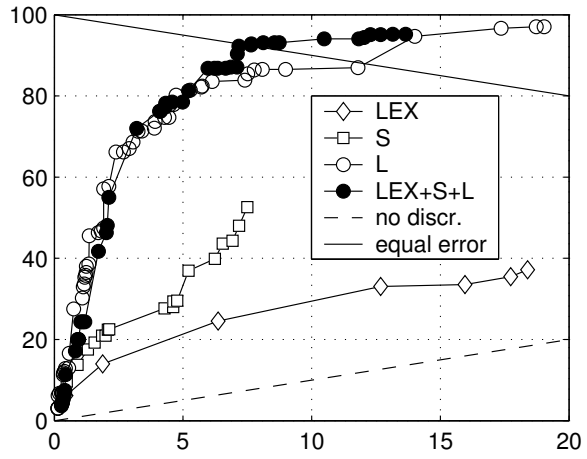


Figure 3. Receiver operating characteristic curves for 4 systems employing topology T1, produced using DEVSET; also shown is the line of no discrimination and the equal error rate line. False positive and true positive rates shown in % along the x- and y- axes, respectively.

As can be seen, lexical features offer performance above the line of no discrimination over the whole range observed, but performance is much poorer than for any other system explored in this work. Laughter context offers significantly better performance, approximately quadrupling the lexical system true positive rate at the same false positive rates. Speech context features yield performance which is intermediate between lexical features and laughter context features. The model-space combination of all three systems follows the $\mathcal{L}$-only curve at low false positive rates, but near the equal error rate point achieves miss rates and false alarm rates which are both approximately 5% absolute lower than for the $\mathcal{L}$-only system.

V. MODEL ANALYSIS

We now turn to an analysis of what the laughter context models actually learn. The LDA transform applied to our raw feature space makes the models used in classification difficult to interpret visually; therefore, we retrain a single-Gaussian model on the raw, untransformed features for the analysis in this section.

Figure 4 shows the laughter context emission probability for DA terminating at time t. The temporal distribution of laughter from the interlocutor who laughs the most in the $[t-5, t+5]$-second window is given in panel (a) for DAs labeled as an attempt at humor, and in panel (b) for all other DA types. Similarly, panels (c) and (d) show the same distributions as (a) and (b), respectively, for the interlocutor who laughs the second most in the $[t-5, t+5]$-second window. As can be seen, attempts at humor are quite different from DAs not so labeled, in terms of how much the two most laughing interlocutors laugh. The peaks in the distributions of (a) and (c) occur just after completion of the humorous DA, with the most laughing interlocutor being more likely to laugh than not. However, panels (a) and (c) also show that interlocutors laugh a significant amount prior to humorous DA completion.

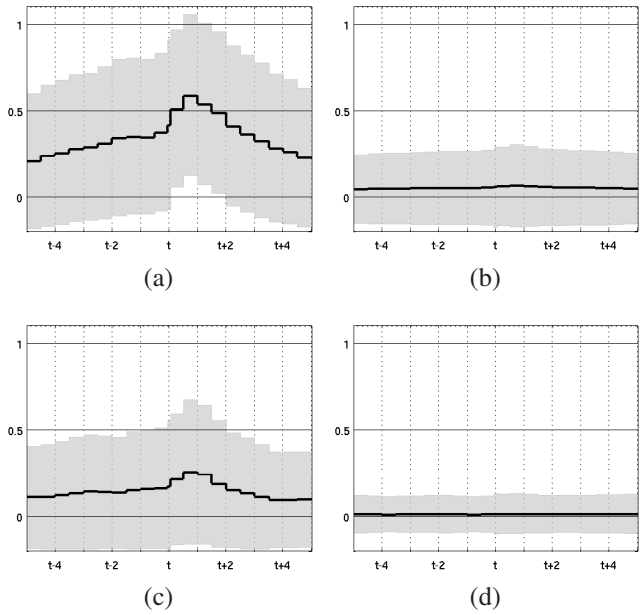


Figure 4. Single-Gaussian emission probabilities, in speech states completing DAs, of the raw laughter context produced by the most laughing interlocutor (panels (a) and (b)) and the second-most laughing interlocutor (panels (c) and (d)). Panels (a) and (c) pertain to humorous DAs, while panels (b) and (d) pertain to a model which, for the purposes of analysis, was trained on all other DA types. The x-axis shows time in seconds; probabilities in $[0, 1]$ are shown along the y-axis, with the mean in black and the gray area showing one standard deviation away from the mean.

In Figure 5 we depict the laughter context emission probability for a participant terminating a DA at time t, for laughter from him-/her- self only. Somewhat surprisingly, the amount of laughter produced by the participant completing a humourous DA is almost as high as for the most laughing interlocutor, and significantly higher than for

the second most laughing interlocutor. This suggests that laughter, like speech, may occur predominantly in dyads. Closer inspection of Figure 4(a) and Figure 5(a) indicates that the temporal distribution of laughter for the teller of the humorous DA is almost identical to that of the most laughing interlocutor, everywhere in the $[-5, +5]$-second context of that DA's completion except during the interval during which the teller is preoccupied with speaking – from approximately $t - 3$ to t seconds.

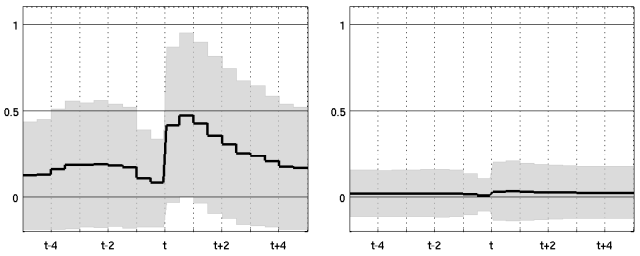


Figure 5. Single-Gaussian emission probabilities, in speech states completing DAs, of the raw laughter context produced by the participant completing the DA. Panel (a) pertains to humorous DAs while panel (b) pertains to a model trained on all other DA types. Axes as in Figure 4.

VI. Discussion

Although it would have been desirable to compare automatic j detection performance with the agreement achieved by the original labelers, meeting excerpts for which multiple labeler annotations are available in our corpus include only 6 DAs annotated as attempts at humor. This precludes such a comparison, which awaits future work.

In this section, we instead provide an alternative assessment of automatic performance by exposing the decoder to human-generated DA boundaries, and reoptimizing model parameters for this new condition. This allows for assessment of performance in terms of the number of DAs correctly classified as attempts at humor, rather than in terms of speaking time. Additionally, we explore the performance of a decoder exposed to only the speaker's laughter context, and enumerate several observations regarding the sequencing of attempts at humor with both laughter and with other participants' j productions.

A. Alternative Performance Assessment

The results presented so far have been frame-level detection errors. This is appropriate in the proposed setting, since the decoder is not exposed to reference DA segmentation and must explicitly segment talkspurts while at the same time classifying them into DA types. However, a frame-level detection error obscures *how many* attempts at humor are detected, as opposed to by how much they are missegmented. To shed light on this issue, we expose the decoder to reference DA boundaries by disallowing DA-terminal frames from aligning to non-DA-terminal topology states and vice versa. The best Viterbi path can then be scored as we have done previously, at the frame-level, as well as at the DA level, allowing for comparison between the two metrics.

Our results are shown in Table VI. The first panel duplicates the results from Table V for ease of comparison. In the second panel, we retain the original model parameters but additionally force-align DA boundaries during decoding. ERR minima for all three of LEX, $\mathcal{L}$, and $\mathcal{S}$ systems, as well as for their model-space combination, are higher in this condition than when DA boundaries are unknown, due largely to significantly higher miss rates. This suggests that the decoder oversegments j productions and correctly classifies some of the shorter segments as j, but that when inserting DA boundaries is not allowed it classifies the resulting longer j segments as not-j. A possible explanation is that DAs labeled as j may exhibit intention to amuse during only a fraction of their duration, or that laughter, which dominates performance, identifies only the tail end of such DAs; further analysis is required to assess the extent to which this might be the case.

The third panel in Table VI shows performance when parameters of the three individual systems and of their model-space combination are re-optimized using DevSet for the condition in which DA boundaries are known, and scoring at the DA level rather than at the frame level. Relative to the second panel, DA-level error rates are smaller for both DevSet (except for the $\mathcal{S}$ system) and EvalSet (except the $\mathcal{L}$ system).

Table VI
Frame-level detection performance in %, for three systems and their model-space combination (ALL), without DA boundary information, with DA boundary information, and with DA boundary information and reoptimized model parameters, the latter scored at the DA-level. FA is the false alarm rate, MS is the miss rate, and ERR = FA + MS.

System	DevSet			EvalSet		
	FA	MS	ERR	FA	MS	ERR
As in Table V:						
LEX	12.7	67.0	79.6	12.8	70.5	83.3
$\mathcal{L}$	14.0	5.3	19.3	15.6	8.1	23.7
$\mathcal{S}$	7.5	47.4	54.9	8.6	62.8	71.4
ALL	7.7	7.2	14.8	8.3	11.0	19.4
Adding DA boundary information:						
LEX	4.7	79.6	84.3	5.1	77.1	82.2
$\mathcal{L}$	7.3	26.5	33.8	8.7	18.2	26.9
$\mathcal{S}$	6.5	48.6	55.1	5.5	66.5	72.0
ALL	7.0	20.1	27.1	5.8	24.0	29.8
Scoring at the DA-level:						
LEX	7.9	66.7	74.5	8.1	61.9	70.0
$\mathcal{L}$	14.4	12.9	27.3	14.9	19.1	34.0
$\mathcal{S}$	7.3	50.5	57.8	8.1	60.0	68.1
ALL	7.6	16.1	23.7	8.1	20.0	28.1

B. Laughter as Invitation to Laugh

The analysis in Section V, and in particular Figure 5, indicates that those attempting humor themselves laugh. This

is especially true immediately following DA completion. In light of this, we construct an alternative system which excludes laughter from interlocutors and uses only the laughter-context from the participant whose DA productions we are decoding. Detection scores when using this system, with parameters reoptimized for this new task, are shown as $\mathcal{L}'$ in Table VII. Performance is lower than when other laughers are considered, but only by 6.3% on unseen EVALSET data, and still considerably lower than when either lexical or speech context features are employed instead.

Table VII
DETECTION PERFORMANCE, IN %, FOR THE SYSTEM RELYING ON LAUGHTER CONTEXT FROM THE SPEAKER AND THE THREE MOST LAUGHING INTERLOCUTORS ($\mathcal{L}$, AS IN TABLE IV), AS WELL AS AN ALTERNATIVE SYSTEM RELYING ON LAUGHTER CONTEXT FROM THE SPEAKER ONLY ($\mathcal{L}'$). FA IS THE FALSE ALARM RATE, MS IS THE MISS RATE, AND ERR = FA + MS.

System	DEVSET			EVALSET		
	FA	MS	ERR	FA	MS	ERR
$\mathcal{L}$	14.0	5.3	19.3	15.6	8.1	23.7
$\mathcal{L}'$	8.7	20.3	28.9	8.5	22.4	31.0

These results indicate that those making attempts at humor communicate their intent by laughing themselves, signaling to interlocutors that it is appropriate for them to take up laughter. Although this finding corroborates qualitative studies in the literature [3], the observed level of performance on our meeting data is surprising. It suggests that meetings may provide an environment in which producers of j DAs deliberately perform additional work (by deploying laughter) to limit the potential ambiguity of their intent, more so than in non-work-oriented conversation.

Furthermore, because producers of j DAs are more likely to laugh following DA completion than any but one other interlocutor (cf. Figures 4 and 5), such DAs appear to be directed at specific other participants rather than the group as a whole. As additional evidence of the dyadic nature of j talk we show its temporal distribution for participants completing a j DA, their most j-talkative interlocutor, and their second-most j-talkative interlocutor, in panels (a), (b), and (c), respectively, of Figure 6.

As can be seen, inside of the $[-5, +5]$-second context of a terminating j DA, the second-most j-talkative interlocutor is unlikely to be active, i.e. at most two participants produce j talk in any 10 second window centered on a j-terminal DA boundary. Furthermore, the most j-talkative interlocutor has two maxima, at approximately $t - 3$ s and $t + 2.5$ s, indicating that they are likely to be producing j talk before the current DA, following the current DA, or both. Finally, the interval between maxima in the probability of j talk from different participants appears to have a most likely value of approximately 3 seconds. This regularity, which is not explicitly modeled in our systems, may improve the detection of attempts at humor in future systems.

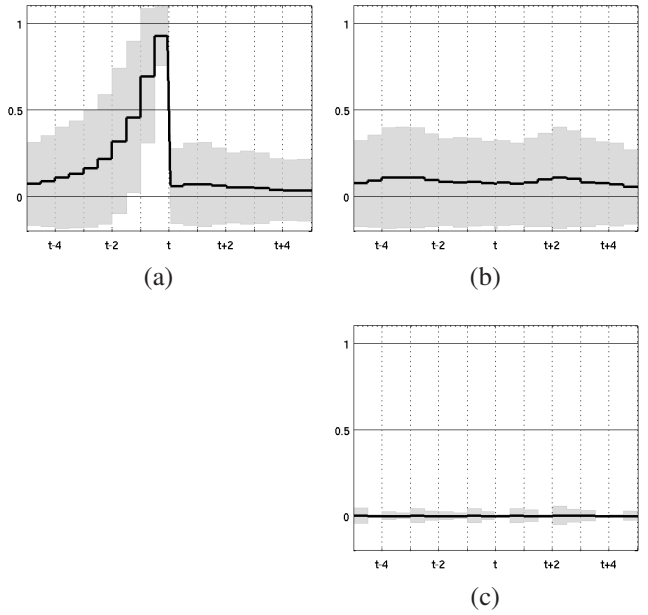


Figure 6. Single-Gaussian emission probabilities, for speech states completing j DAs, of the raw j-speech context produced by the participant completing the DA, in panel (a), her or his most-j-speaking interlocutor, in panel (b), and her or his second-most-j-speaking interlocutor, in panel (c). Axes as in Figure 4.

VII. CONCLUSIONS

We have proposed a system for the detection of attempts at humor in multiparty meetings, a phenomenon which accounts for approximately 0.6% of speaking time. Our experiments indicate that traditional lexical features perform relatively poorly on this task. In contrast, contextual features describing the temporal, multi-participant distribution of proximate laughter achieve detection error rates on unseen data of 23.7% by time, representing a 72% relative reduction of error over the oracle lexical baseline. Model-space combination of lexical and laughter-context features, as well as speech-context features, reduces detection errors on unseen data by an additional 18% relative. The best performing systems achieve a detection error of 19.4% by time and 28.1% by DA count.

Detailed analysis shows that the speaker's own laughter is quite indicative of whether she or he is attempting to be humorous. Although this corroborates well-established findings in the literature with regard to conversation in general, the strength of this cue in our meeting data is surprising: participants appear to exert significant effort to disambiguate their intended degree of seriousness. Additionally, we have found that humor-bearing talk induces dyadic interaction: it appears to lead to more laughter than from the speaker in only one other participant, and, similarly, to subsequent humor-bearing talk from only one other participant. These findings suggest that the detailed study of the sequence of attempts at humor and their ratification through laughter

should be studied for pairs of participants, even in unconstrained and explicitly unpaired multiparty settings.

ACKNOWLEDGMENTS

We would like to thank Liz Shriberg for access to the ICSI MRDA data and for helpful discussion during the preparation of this manuscript.

REFERENCES

[1] K. Zechner, "Automatic summarization of open-domain multiparty dialogues in diverse genres," in *Computational Linguistics*, vol. 28, no. 4, 2002, pp. 447–485.

[2] J. Hirschberg, S. Benus, J. Brenier, F. Enos, S. Friedman, S. Gilman, C. Girand, M. Graciarena, A. Katho, L. Michaelis, B. Pellom, E. Shriberg, and A. Stolcke, "Distinguishing deceptive from non-deceptive speech," in *Proc. EUROSPEECH*, Lisboa, Portugal, 2005, pp. 1833–1836.

[3] P. Glenn, *Laughter in interaction*. Cambridge University Press, 2003.

[4] H. Kangasharju and T. Nikko, "Emotions in organizations: Joint laughter in workplace meetings," in *J. of Business Communication*, vol. 46, no. 1, 2009, pp. 100–119.

[5] M. Galley, K. McKeown, E. Fosler-Lussier, and H. Jing, "Discourse segmentation of multi-party conversation," in *Proc. ACL*, Sapporo, Japan, 2003, pp. 562–569.

[6] G. Jefferson, H. Sacks, and E. Schegloff, "Preliminary notes on the sequential organization of laughter," in *Pragmatics Microfiche*. Cambridge University Press, 1977.

[7] S. Attardo, "Humor," in *Handbook of Pragmatics*. John Benjamins, 1996, pp. 1–17.

[8] A. Nijholt, "Observations on humor act construction," in *Proc. Cybernetics and Systems*, Wien, Austria, 2004, pp. 91–96.

[9] A. Clark and A. Popescu-Belis, "Multi-level dialogue act tags," in *Proc. SIGdial*, Boston MA, USA, 2004, pp. 163–170.

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

[11] S. Somasundaran, J. Ruppenhofer, and J. Wiebe, "Detecting arguing and sentiment in meetings," in *Proc. SIGdial*, Anwerpen, Belgium, 2007, pp. 26–34.

[12] K. Laskowski, "Modeling vocal interaction for text-independent detection of involvement hotspots in multi-party meetings," in *Proc. SLT*, Goa, India, 2008, pp. 81–84.

[13] L. Kennedy and D. Ellis, "Laughter detection in meetings," in *Proc. ICASSP Meeting Recognition Workshop*, Montreal, Canada, 2004, pp. 118–121.

[14] A. Ito, X. Wang, M. Suzuki, and S. Makino, "Smile and laughter recognition using speech processing and face recognition from conversation video," in *Proc. Cyberworlds (CW)*, Singapore, 2005.

[15] K. Truong and D. van Leeuwen, "Evaluating automatic laughter segmentation in meetings using acoustic and acoustic-phonetic features," *Proc. ICPhS Workshop Phonetics of Laughter*, pp. 49–53, 2007.

[16] M. Knox and N. Mirghafori, "Automatic laughter detection using neural networks," in *Proc. INTERSPEECH*, Antwerpen, Belgium, 2007, pp. 2973–2976.

[17] K. Laskowski and T. Schultz, "Detection of laughter-in-interaction in multichannel close-talk microphone recordings of meetings," in *Proc. MLMI*, ser. Springer LNCS **5237**, Utrecht, The Netherlands, 2008, pp. 149–160.

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

[19] E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, and H. Carvey, "The ICSI Meeting Recorder Dialog Act (MRDA) Corpus," in *Proc. SIGdial*, Cambridge MA, USA, 2004, pp. 97–100.

[20] K. Laskowski and E. Shriberg, "Modeling other talkers for improved dialog act recognition in meetings," in *Proc. INTERSPEECH*, Brighton, UK, 2009.

[21] J. Ang, Y. Liu, and E. Shriberg, "Automatic dialog act segmentation and classification in multiparty meetings," in *Proc. ICASSP*, Philadelphia PA, USA, 2005, pp. 1061–1064.

[22] K. Laskowski and S. Burger, "On the correlation between perceptual and contextual aspects of laughter in meetings," in *Proc. ICPhS WS on Phonetics of Laughter*, Saarbrücken, Germany, 2007, pp. 55–60.