FINDING EMOTIONALLY INVOLVED SPEECH USING IMPLICITLY PROXIMITY-ANNOTATED LAUGHTER

Kornel Laskowski

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

ABSTRACT

Browsing through collections of audio recordings of conversation nominally relies on the processing of participants' lexical productions. The evolving verbal and non-verbal context of those productions, likely indicative of the degree of participant involvement, is often ignored. The present work explores the relevance of laughter to the retrieval of conversation intervals in which the speech of one or more participants is prosodically or pragmatically marked as involved. Experiments indicate that the relevance of laughter depends on its temporal distance to the laugher's speech. The results suggest that in order to be pertinent to downstream emotion recognition applications, laughter detection systems must first and foremost detect that laughter which is most temporally proximate to the laugher's speech.

Index Terms— Emotion detection, Laughter modeling, Vocal activity modeling, Speech retrieval, Meetings.

1. INTRODUCTION

The ability to index and search through audio recordings of multiparty conversation, and meetings in particular, is becoming increasingly important as the number and size of such collections grow. At the current time, search through meeting audio is almost entirely restricted to that through the automatically or manually prepared transcripts of what is said, through textual sources derived from them, or through textual artifacts prepared by the participants themselves [1].

A potentially important capability of conversational audio browsing is that of identifying intervals based not on what participants say, but on how emotionally involved they are while speaking. The analogy of radio summaries of sporting events illustrates that conversational involvement may be worth retrieving to potential users. Previous work on involved speech in multiparty meetings has shown that it is correlated with prosodic variables [2] and manually annotated dialog act types [3], but no system based on such information has been put forward for its detection and indexing.

To the author's best knowledge, the only published detector for involved speech in meetings, whose accuracy significantly exceeds majority class guessing, is one that identifies *intervals containing* involved speech [4]. Its performance relies on the detection of *co-located laughter*, defined here to be laughter, from any participant, found in the temporal vicinity of that speech. Performance has been shown to be highest when only *laughter produced simultaneously with speech* by the same participant, known as speech-laughter [5], is considered. This presents a problem for fully automatic system development. A recent study of the occurrence of laughter in meetings [6] has shown that speech-laughs account for less than 4% of all laughter by time, and therefore less than 0.4% of all vocalization effort by time. Although the detection of laughter is currently gaining attention [7, 8], this prior makes the successful acoustic treatment of speech-laughs in the near term unlikely.

The current work explores the detection of intervals containing involved speech based on only the *non-speech laughter* produced by meeting participants. Using data described in Section 2, the results of [4] are first duplicated in Section 3 as a baseline. Section 4 then proposes a framework for the analysis of speech and laughter collocation. The experiments in Section 5 indicate that of all non-speech laughter, voiced laughter which is most proximate to the laugher's speech is most salient to the proposed task, and that annotation with temporal distance improves retrieval scores. These findings, summarized in Section 6, present an important opportunity for future acoustic laughter detection efforts. At the present time, detection has focused on those instances of laughter which are transcribed as isolated utterances [7, 8].

2. DATA

The experiments presented rely on the ICSI Meeting Corpus [9] which consists of 75 meetings, each with 3 to 9 participants. The meetings would have occurred even if they were not recorded; together they comprise 63 hours of longitudinal interaction. The corpus is accompanied by orthographic transcription, lexical item forced alignment, and dialog act (DA) annotation [10]. It includes, per DA, an attribute signaling that the speaker is "more involved (emotionally or 'interactively') [and/or that] there is a higher degree of interaction by participants who are trying to get the floor" [11]. Also available is a segmentation of laughter and its classification into voiced and unvoiced bouts [12]. As proposed in [4], the meetings in the corpus are separated into TRAINSET (of 29 meetings), DEVSET (of 31), and EVALSET (of 15).

For each meeting, three binary human-generated segmentations are prepared; reference as opposed to automatically inferred segmentations are used in order to quantify the amount of information which is present, rather than the amount that is detectable using current state-of-the-art technologies. These three segmentations are: (1) speech $\mathcal{S}$, from forced alignment of manually transcribed words [10]; (2) voiced laughter $\mathcal{L}_V$, from [6, 12]; and (3) involved speech $\mathcal{S}_I$, formed by allowing each lexical item to inherit the involvement attribute of the dialog act it belongs to. It should be noted that $\mathcal{S}$ and $\mathcal{L}_V$ are not disjoint (due to speech-laughs).

3. BASELINE

All systems in this work model participant-attributed vocal activity for the detection of intervals containing $\mathcal{S}_I$, regardless of which participant produces the latter. The interval size is 60 seconds, informed by observations in the original description of involvement in this data [2]. To generate a sufficient amount of exemplars for training, development, and testing, an interval step of 15 seconds is used. For each interval, the baseline system [4] extracts a feature vector $\mathbf{f}$ which consists of the proportion of interval duration for which each participant vocalizes, given the $\mathcal{L}_V$ segmentation, and sorted by decreasing magnitude. The vector is padded with zeros to a length of 9, which is relevant for those meetings that contain fewer than 9 participants (the maximum in the ICSI data).

For each interval's $\mathbf{f}$, a support vector machine trained using TRAINSET yields a hypothesized label ("containing $\mathcal{S}_I$" or "not containing $\mathcal{S}_I$"). Forward feature selection is performed by maximizing either classification accuracy or F-score, the unweighted harmonic mean of recall and precision, using DEVSET. This leads to two systems relying on potentially dissimilar features; where this is the case, accuracies and F-scores are reported only for that of the two systems which has been optimized for the measure in question.

The performance of the baseline is shown in Table 1. Features drawn from laughter segmentations, and from logical intersections with laughter segmentations, lead to F-scores exceeding 50% (except the intersection of unvoiced laughter ($\mathcal{L}_U$) with speech, $\mathcal{L}_U \cap \mathcal{S}$, which is near-empty). Unvoiced laughter appears to be much less relevant to this task than does voiced laughter, and features drawn from the latter often outperform those drawn from all laughter. This is felicitous, since voiced laughter is acoustically easier to detect [13] than unvoiced laughter (and hence than all $\mathcal{L}$).

Feature Set	Accuracy, %		F-Score, %	
	DEVSET	EVALSET	DEVSET	EVALSET
guess, priors	*60.9*	*61.2*	—	—
guess, major.	*72.9*	*73.7*	—	—
$\mathbf{f}\,(\mathcal{S})$	*74.4*	*75.3*	34.4	28.0
$\mathbf{f}\,(\mathcal{L})$	*80.4*	*80.8*	64.6	64.8
$\mathbf{f}\,(\mathcal{L} \cap \mathcal{S})$	*81.2*	*83.3*	68.9	70.6
$\mathbf{f}\,(\mathcal{L} \cap \neg\mathcal{S})$	*80.4*	*80.8*	64.6	64.8
$\mathbf{f}\,(\mathcal{L}_V)$	*81.5*	*81.6*	65.1	64.3
$\mathbf{f}\,(\mathcal{L}_V \cap \mathcal{S})$	*82.9*	*85.6*	69.5	67.1
$\mathbf{f}\,(\mathcal{L}_V \cap \neg\mathcal{S})$	*81.5*	*81.4*	65.0	67.1
$\mathbf{f}\,(\mathcal{L}_U)$	*76.4*	*77.4*	56.3	55.1
$\mathbf{f}\,(\mathcal{L}_U \cap \mathcal{S})$	73.7	72.6	27.6	21.9
$\mathbf{f}\,(\mathcal{L}_U \cap \neg\mathcal{S})$	76.4	77.4	56.3	55.1

Table 1. Baseline system accuracies and F-scores for retrieval of 60-second meeting intervals containing involved speech, based on features drawn from logical combinations of the speech $\mathcal{S}$ and laughter $\mathcal{L}$ reference segmentations. Numbers in italics are from [4].

Table 1 also shows that speech-laughs, $\mathcal{L}_V \cap \mathcal{S}$, lead to the best classification accuracies and F-scores. This is unsurprising, and such laughter is likely not only to be informative of intervals containing involved speech but also to identify the involved participant. However, current laughter detection systems do not consider speech-laughs, and are unlikely to do so in the near term owing to that laughter type's infrequence [12]; most such systems focus on a two-way distinction between speech and laughter, or the two-way distinction between laughter and silence [7, 8], and are limited to bouts that have been transcribed as separate utterances, suggesting that those bouts may be temporally distant from the laugher's speech.

The remainder of this work focuses on the much more realistic expectation of acceptable detection of voiced laughter *not* produced simultaneously with speech, i.e. $\mathcal{L}_V \cap \neg\mathcal{S}$. The experiments rely on the reference segmentation of this subset of all laughter, produced using the $\mathcal{L}_V$ and $\mathcal{S}$ segmentation described in Section 2.

4. AN ANALYSIS FRAMEWORK

This work qualifies voiced laughter instants by measuring their temporal proximity to the *laugher's own* speech. This is achieved by masking out (i.e. discarding) voiced laughter which falls outside of a particular proximity range relative to that speech. Masks are defined using a segmentation Υ, logical AND ($\Upsilon' = \Upsilon_1 \cap \Upsilon_2$) and complement ($\Upsilon' = \neg\Upsilon$) operations, and *temporal extension* $\Upsilon' = \sigma(\tau_L, \Upsilon, \tau_R)$. The latter consists of pre-padding "on" intervals in Υ with τ_L seconds and post-padding them with τ_R seconds.

Figure 1 depicts the process of isolating laughter which occurs at least τ_L^A seconds and at most τ_L^B seconds prior to the laugher's speech, as well as at least τ_R^A seconds and at most τ_R^B seconds anterior to the laugher's speech. The starting point is a speech segmentation $\mathcal{S}$ and a voiced laughter segmentation $\mathcal{L}_V$, shown in panel (a). In panel (b), the temporal extension operator σ is applied to $\mathcal{S}$, with arguments τ_L^A and τ_R^A, to produce the intermediate segmentation Υ_1; its complement is shown as Υ_2, and identifies instants that occur at least τ_L^A seconds before and at least τ_R^A seconds after any speech from the depicted participant. Panel (c) depicts the construction of Υ_3, similar to Υ_1, but with arguments τ_L^B and τ_R^B; Υ_3 identifies those instants that occur at most τ_L^B seconds before and at most τ_R^B seconds after any speech from the same participant. Panel (d) shows the voiced laughter segmentation $\mathcal{L}_V$, before and after logical AND with Υ_4, which is the logical AND of Υ_2 and Υ_3. The resulting Υ_5 identifies the remaining laughter as specified by the arguments τ_L^A, τ_R^A, τ_L^B, and τ_R^B.

Three families of masks are proposed: Υ^{slice}, Υ^{prox}, and Υ^{dist}.

4.1. Masks Υ^{slice}

These masks are at most 1 s in duration. There are 3 subfamilies of these masks, depending on their location relative to laugher's speech:

- *pre-talkspurt masks*, $\Upsilon_{pre}^{slice}(\tau) = \neg\sigma(\tau - 1, \mathcal{S}, \tau) \cap \sigma(\tau, \mathcal{S}, 0)$, consisting of slices of up to 1 second in duration, at least $\tau - 1$ seconds before subsequent speech, at most τ seconds before subsequent speech, and at least τ seconds after antecedent speech;
- *post-talkspurt masks*, $\Upsilon_{post}^{slice}(\tau) = \neg\sigma(\tau, \mathcal{S}, \tau - 1) \cap \sigma(0, \mathcal{S}, \tau)$, consisting of slices of up to 1 second in duration, at least $\tau - 1$ seconds after antecedent speech, at most τ seconds after antecedent speech, and at least τ seconds before subsequent speech; and
- *inter-talkspurt masks*, $\Upsilon_{inter}^{slice}(\tau) = \neg\sigma(\tau - 1, \mathcal{S}, \tau - 1) \cap \sigma(\tau, \mathcal{S}, 0) \cap \sigma(0, \mathcal{S}, \tau)$, consisting of slices of 1 second in duration, at least $\tau - 1$ seconds after antecedent speech, at most τ seconds after antecedent speech, at least $\tau - 1$ seconds before subsequent speech, and at most τ seconds before subsequent speech. The latter category consists of all those slices that are equally proximate to antecedent and subsequent talkspurts.

From these three subfamilies, 2 additional families of masks are derived by composition.

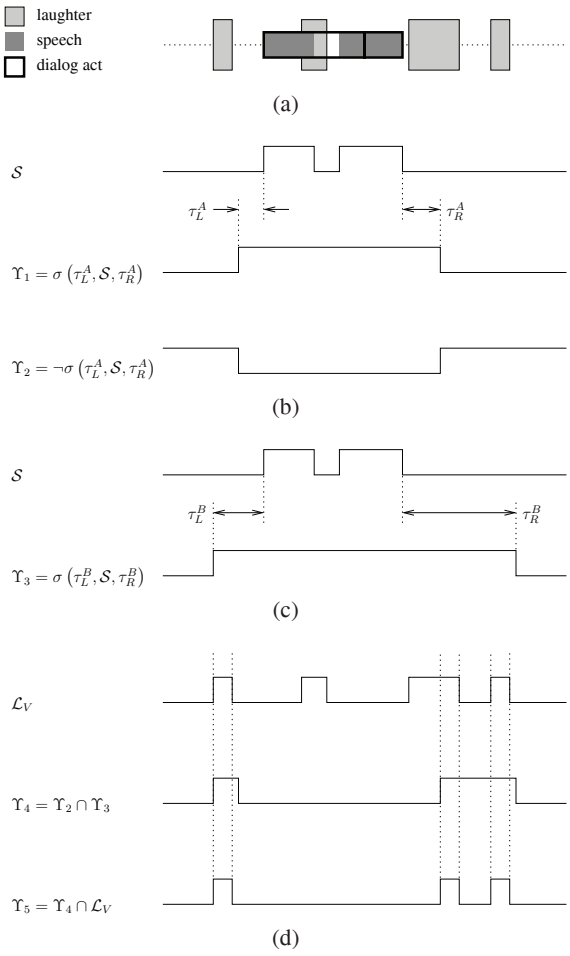

Fig. 1. Masking voiced laughter $\mathcal{L}_V$ with a mask constructed using speech $\mathcal{S}$; time τ is shown from left to right. The process of arriving at the final binary trajectory in panel (d), from the given speech and laughter segmentations for a particular participant shown in panel (a), is as described in the text.

4.2. Masks Υ^{prox}

The first derived family includes instants from the edge of proximate speech to τ seconds away from it, defining a cumulative mask Υ^{prox}. Three subfamilies are considered:

- *pre-talkspurt masks*, $\Upsilon^{prox}_{pre}(\tau) = \bigcup_{\tau'=1}^{\tau} \Upsilon^{slice}_{pre}(\tau')$;
- *post-talkspurt masks*, $\Upsilon^{prox}_{post}(\tau) = \bigcup_{\tau'=1}^{\tau} \Upsilon^{slice}_{post}(\tau')$; and
- *inter-talkspurt masks*, $\Upsilon^{prox}_{inter}(\tau) = \bigcup_{\tau'=1}^{\tau} \Upsilon^{slice}_{inter}(\tau')$.

4.3. Masks Υ^{dist}

The second derived family of masks consists of compositions of Υ^{slice} which extend from τ seconds from the edge of proximate speech to 10 seconds away:

- *pre-talkspurt masks*, $\Upsilon^{dist}_{pre}(\tau) = \bigcup_{\tau'=\tau}^{10} \Upsilon^{slice}_{pre}(\tau')$;
- *post-talkspurt masks*, $\Upsilon^{dist}_{post}(\tau) = \bigcup_{\tau'=\tau}^{10} \Upsilon^{slice}_{post}(\tau')$; and
- *inter-talkspurt masks*, $\Upsilon^{dist}_{inter}(\tau) = \bigcup_{\tau'=\tau}^{10} \Upsilon^{slice}_{inter}(\tau')$.

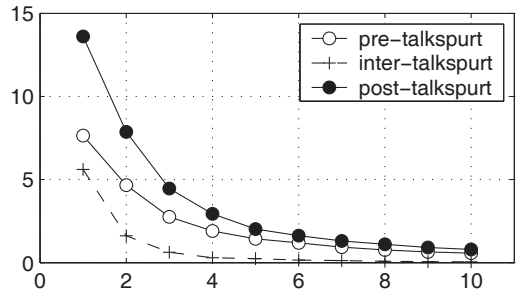

Fig. 2. Proportion (in %, along the y-axis) of voiced laughter by time, per mask $\Upsilon^{slice}_{\alpha}(\tau)$; τ shown in seconds along the x-axis. $\alpha \in \{pre, inter, post\}$.

5. EXPERIMENTS

Figure 2 shows the proportion of voiced laughter found within each Υ^{slice} mask, as indexed by temporal distance τ from speech, for all three subfamilies of masks (pre, $post$, $inter$). As can be seen, the occurrence of voiced laughter decreases exponentially with temporal distance away from proximate speech. There is also far more voiced laughter following speech than preceding it. It appears likely that that for $\tau > 10$ seconds the exponential trend observed for all three of pre-talkspurt, post-talkspurt, and inter-talkspurt voiced laughter continues, making the amount of voiced non-speech laughter beyond $\tau = 10$ too sparse for modeling. In the remainder of this paper, only the slices $\tau \in [1, 10]$ are considered; together, they account for 68% of all voiced non-speech laughter by time.

With this in mind, the experiments of Section 3 are repeated using only the subset of voiced laughter given by $\Upsilon' = \Upsilon^{slice}_{\alpha}(\tau) \cap \mathcal{L}_V$, for each of ten masks indexed by $\tau \in [1, 2, \cdots, 10]$ seconds and each of the three subfamilies of masks $\alpha \in \{pre, post, inter\}$. Classification accuracy for EVALSET is shown in panel (a) of Figure 3. It is evident that the most relevant voiced non-speech laughter is that found in immediate proximity to laugher's speech; F-scores which lie above 50% are to be found for all three of pre-talkspurt, post-talkspurt, and inter-talkspurt contexts for $\tau = 1$, and only for the post-talkspurt context for $\tau = 2$. As $\tau \longrightarrow 10$ seconds, the accuracies for voiced laughter approach those obtained with majority class guessing (i.e., that no interval contains involved speech $calS_I$).

It also appears that, at least for $\tau < 4$, post-talkspurt voiced laughter is more relevant than pre-talkspurt voiced laughter, and that for $\tau = 1$ inter-talkspurt voiced laughter is most relevant. The latter may be due to the fact that inter-talkspurt laughter is much more likely to "bleed" from or into the laugher's ongoing verbal production, making that production involved and potentially marking the current interval as containing involved speech $\mathcal{S}_I$. Other laughter is arguably less likely to have affected speech production.

The performance of the system using voiced laughter which is found *at most* τ seconds away from the laugher's closest talkspurt, identified by the first derived family of masks (Υ^{prox}) defined in Subsection 4.2, is shown in panel (b) of Figure 3. It can be seen that classification accuracy for pre-talkspurt and post-talkspurt voiced laughter never exceeds that for all voiced non-speech laughter, at any value of the threshold τ past which all voiced laughter is discarded. Post-talkspurt laughter appears to be more relevant than pre-talkspurt laughter. Second, except for voiced pre-talkspurt laughter at small values of τ, the accuracy actually decreases as more and more distant voiced non-speech laughter is considered.

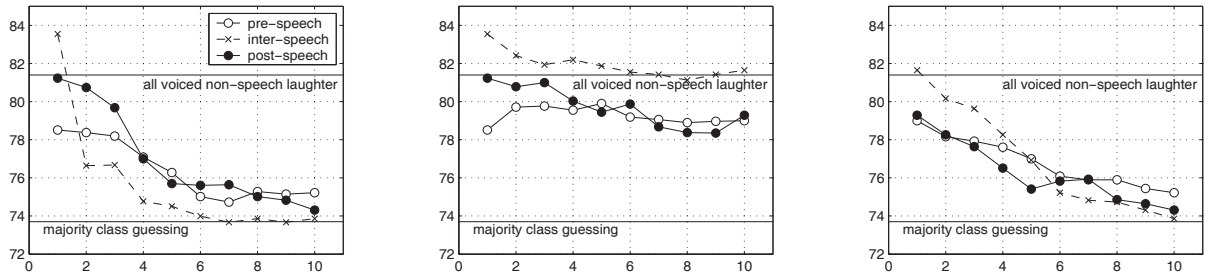

Fig. 3. Classification accuracy (in %, along the y-axis) as a function of τ (in seconds, along the x-axis), for features $\mathbf{f}$ extracted from: (a) laughter in 1-second slices τ seconds away from speech, using $\Upsilon_\alpha(\tau) \cap \mathcal{L}_V$; (b) laughter in τ-second slices immediately proximate to speech, using $\Upsilon_\alpha^{prox}(\tau) \cap \mathcal{L}_V$; and (c) laughter in $10-\tau$-second slices τ seconds away from speech, using $\Upsilon_\alpha^{dist}(\tau) \cap \mathcal{L}_V$.

Panel (c) of Figure 3 shows results for voiced laughter which is found *at least* τ seconds away from the laugher's closest talkspurt, given by intersection with the second derived family of masks (Υ^{dist}) of Subsection 4.3. Classification accuracies appear to fall steeply towards the accuracy achieved by majority class guessing.

As a final experiment, the feature selection scheme is exposed to features drawn from all of the 1-second-slice subsegmentations $\Upsilon_\alpha^{slice}(\tau) \cap \mathcal{L}_V$. Doing so entails feature selection in a space not of 9 features, but of $9 \times 3 \times 10$ features, namely $\mathbf{F} \equiv \bigcup_\alpha \bigcup_{\tau=1}^{10} \mathbf{f}\left(\Upsilon_\alpha^{slice}(\tau) \cap \mathcal{L}_V\right)$. To a certain degree, this approach is tantamount to annotating slices of laughter of up to 1 second in duration with temporal proximity to and co-orientation with the laugher's nearest talkspurt. Accuracies for a system modeling $\mathbf{F}$ are 83.2% and 84.4% for DEVSET and EVALSET, respectively; the corresponding F-scores are 71.2% and 70.3%, respectively. These numbers should be compared to those for $\mathbf{f}(\mathcal{L}_V \cap \mathcal{S})$ in Table 1. For EVALSET, this classification accuracy is a 3% absolute improvement over the unannotated, complete voiced non-speech laughter segmentation $\mathcal{L}_V$, representing a 16% relative reduction of classification error. Notably, F-scores are higher than for the speech-laugh segmentation ($\mathcal{L}_V \cap \mathcal{S}$) which has been shown to be more difficult to produce automatically [13].

6. CONCLUSIONS

This work has explored the relevance of voiced non-speech laughter to the retrieval of intervals containing involved speech. A novel means of studying conversational phenomena was proposed, via speech-segmentation-defined masking operators. Experiments showed that laughter which is temporally closest to the laugher's speech is most indicative of co-located involved speech from any participant to the conversation, and that laughter following speech is more relevant than laughter preceding speech. Conditioning voiced non-speech laughter on its temporal proximity to the laugher's speech resulted in a 3.3% absolute improvement in both classification accuracy and F-score, representing a 16% relative reduction of classification error over a system with no such annotation. These are crucial findings, of which future laughter detection efforts should take note when considering utility to downstream applications.

7. ACKNOWLEDGMENTS

The author would like to thank Liz Shriberg for access to the ICSI MRDA Corpus (release `icsi_mrda+hs_corpus_050512`).

8. REFERENCES

[1] S. Tucker and S. Whittaker, "Accessing multimodal meeting data: Systems, problems and possibilities," in *Proc. MLMI*, 2004, pp. 245–252.

[2] B. Wrede and E. Shriberg, "Spotting "hot spots" in meetings: Human judgments and prosodic cues," in *Proc. EUROSPEECH*, 2003, pp. 2805–2808.

[3] B. Wrede and E. Shriberg, "The relationship between dialogue acts and hot spots in meetings," in *Proc. ASRU*, 2003, pp. 180–185.

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

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

[6] K. Laskowski and S. Burger, "Analysis of the occurrence of laughter in meetings," in *Proc. INTERSPEECH*, 2007, pp. 1258–1261.

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

[8] M. Knox, N. Morgan, and N. Mirghafori, "Getting the last laugh: Automatic laughter segmentation in meetings," in *Proc. INTERSPEECH*, 2008, pp. 797–800.

[9] A. Janin et al., "The ICSI Meeting Corpus," in *Proc. ICASSP*, 2003, pp. 364–367.

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

[11] B. Wrede and E. Shriberg, "Reliability analysis for hot spot annotations in the MRDA Corpus," *Internal document*, ICSI, Berkeley CA, USA, 2005.

[12] 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*, 2007, pp. 55–60.

[13] K. Laskowski, "Contrasting emotion-bearing laughter types in multiparticipant vocal activity detection for meetings," in *Proc. ICASSP*, 2009, pp. 4765–4768.