ON THE CORRELATION BETWEEN PERCEPTUAL AND CONTEXTUAL ASPECTS OF LAUGHTER

Author Name1 and Author Name2

Institute, School, Address
author1@domain1,author2@domain2

ABSTRACT

We have analyzed over 13000 bouts of laughter, in over 65 hours of unscripted, naturally occurring multi-party meetings, to identify discriminative contexts of voiced and unvoiced laughter. Our results show that, in meetings, laughter is quite frequent, accounting for almost 10% of all vocal activity effort by time. Approximately a third of all laughter is unvoiced, but meeting participants vary extensively in how often they employ voicing during laughter. In spite of this variability, laughter appears to exhibit robust temporal characteristics. Voiced laughs are on average longer than unvoiced laughs, and appear to correlate with temporally adjacent voiced laughter from other participants, as well as with speech from the laugher. Unvoiced laughter appears to occur independently of vocal activity from other participants.

1. INTRODUCTION

In recent years, the availability of large multiparty corpora of naturally occurring meetings [1][2] has shifted attention to previously little-explored, natural human-human interaction behaviors [3]. A non-verbal phenomenon belonging to this class is laughter, which has been hypothesized as a strategic means of affecting interlocutors, as well as a signal of various human emotions [4].

Recently, we produced an annotation of perceived emotional valence in speakers in the ISL Meeting Corpus[5]. We showed that instances of isolated laughter were strongly predictive of positive valence. In a subsequent multi-site evaluation of automatic emotional valence classification within the CHIL project, we found that transcribed laughter is in general much more indicative of perceived positive valence than any other grouping of spectral, prosodic, contextual, or lexical features. Three-way classification of speaker contributions into negative, neutral and positive valence classes (with neutral valence accounting for 80% of the contributions), using the presence of transcribed laughter as the only feature, resulted in an accuracy of 91.2%. The combination of other features led to an accuracy of only 87% (similar results were produced on this data by [6]). A combination of all features, including presence of transcribed laughter, produced an accuracy of 91.4%, only marginally better than laughter alone.

Although these results show that the presence of laughter, as detected by human annotators, was the single most useful feature for automatic valence classification, laughter and positive valence are not completely correlated in the ISL Meeting Corpus. We are ultimately interested in the ability to determine, automatically, whether a particular laugh conveys information about the laughter's valence to an outside observer. The current work is a preliminary step in that effort, in which we characterize laughter along two separate dimensions. First, we determine whether each laugh is voiced or unvoiced. Previous work with this distinction in other domains has shown that voiced laughter may be used strategically in conversation [4], in an effort to modify the affect of possibly specific interlocutors.

Second, we attempt to characterize the temporal context of voiced and unvoiced laughter within the multiparticipant vocal activity on-off pattern of a conversation. The study of laughter in sequence with spontaneous speech has been treated by traditional conversation analysis [7], although not in a quantitative, easily exploitable fashion. Laughter has also been shown to evoke laughing in listeners [8], in this way differing from speech. In particular, laughers do not take turns laughing in the same way that speakers take turns speaking. Vocal activity context therefore appears to provide important cues as to whether ongoing vocal activity is laughter or speech. In the current work, we attempt to determine whether context also disambiguates between voiced and unvoiced laughter.

2. DATA

To study the pragmatics of laughter, we use the relatively large ICSI Meeting Corpus [2]. This corpus consists of 75 unscripted, naturally occurring meetings, amounting to over 71 hours of recording time. Each meeting contains between 3 and 10 participants wearing individual head-mounted microphones, drawn from a pool of 53 unique speakers (13 female, 40 male).

In this section, we describe the process we followed to produce, for each meeting and for each participant: (1) a talk spurt segmentation; (2) a voiced laugh bout segmentation; and (3) an unvoiced laugh bout segmentation.

We note that each meeting recording contains a ritualized interval of read speech, a subtask referred to as `Digits`, which we have analyzed but excluded from the final segmentations. The temporal distribution of vocal activity in these intervals is markedly different from that in natural conversation. Excluding them limits the total meeting time to 66.3 hours.

2.1. Talk Spurt Segmentation

Talk spurt segmentation was produced using the word-level forced alignments in the ICSI Dialog Act (MRDA) Corpus [9]. While 500 ms was used as the minimum inter-spurt duration in [10], we use a 300 ms threshold. This value has recently been adopted for the purposes of building speech activity detection references in the NIST Rich Transcription Meeting Recognition evaluations.

2.2. Selection of Annotated Laughter Instances

Laughter is annotated in the ICSI Meeting Corpus orthographic transcriptions (`.stm`) in two ways. First, discrete events are annotated as `VocalSound` instances, and appear interspersed among lexical

Freq Rank	Token Count	VocalSound Description	Used here
1	11515	laugh	√
2	7091	breath	
3	4589	inbreath	
4	2223	mouth	
5	970	breath-laugh	√
11	97	laugh-breath	√
46	6	cough-laugh	√
63	3	laugh, "hmmph"	√
69	3	breath while smiling	
75	2	very long laugh	√

Table 1. Top 5 most frequently occurring VocalSound types in the ICSI Meeting Corpus, and the next 5 most frequently occurring types relevant to laughter.

items. Their location among such items is indicative of their temporal extent. We show a small subset of VocalSound types in Table 1. As can be seen, the VocalSound type laugh is the most frequently annotated non-verbal vocal production. The second type of laughter-relevant annotation found in the corpus, Comment, describes events of extended duration which often cannot be uniquely localized between specific lexical items. In particular, this annotation covers the phenomenon of "laughed speech" [11]. We list the top five most frequently occurring Comment descriptions pertaining to laughter in Table 2. As with VocalSound descriptions, there is a large number of very rich laughter annotations each of which occurs only once or twice.

Freq Rank	Token Count	Comment Description
2	980	while laughing
16	59	while smiling
44	13	last two words while laughing
125	4	last word while laughing
145	3	vocal gesture, a mock laugh

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

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

2.3. Laugh Bout Segmentation

Our strategy for producing accurate endpoints for the identified laughter instances consisted of a mix of automatic and manual methods.

Of the 12570 non-farfield VocalSound instances, 11845 were adjacent on both the left and the right to either a time-stamped .stm segment boundary, or a lexical item. We were thus able to automatically deduce start and end times for 87% of the laughter instances treated in this work.

The remaining 725 non-farfield VocalSound instances were not adjacent to an available timestamp on either or both of the left and the right. These instances were segmented manually, by listening to the entire .stm utterance containing them[1]. The segmentation of all 13295 VocalSound instances was subsequently checked by at least one annotator.

All of the 1108 Comment instances were segmented manually. Manual segmentation of the 725 non-farfield VocalSound instances and of all of the Comment instances took approximately one hour per 100 instances, for a total of 18 hours. A quarter of the manually segmented Comment instances were then checked by one of the authors, which took 4 hours.

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

2.4. Laugh Bout Voicing Classification

In the last preprocessing task, we classified each laughter instance as either voiced or unvoiced. Additionally, this step involved checking the automatically or manually generated segmentation of each instance, as well as confirming that the event was in fact laughter. In making the latter assessment, we discarded only those instances for which we felt we had ample counter-evidence; in the absence of counter-evidence or agreement, we retained the original ICSI assessment of the acoustic event as laughter.

Our distinction of voiced versus unvoiced with respect to laughter was made according to [4]. Voiced laughter, like voiced speech, occurs when the energy source is quasi-periodic vocal-fold vibration. This class includes melodic, "song-like" bouts, as well as most chuckles and giggles. Unvoiced laughter results from fricative excitation, and is analogous to whispered speech. It includes open-mouth, pant-like sounds, as well as closed-mouth grunts and nasal snorts. Additionally, we decided that bouts consisting of both voiced and unvoiced calls should receive the voiced label when taken together. Instances of "laughed speech" were automatically assigned the voiced label.

Endpoint verification and voicing classification were performed simultaneously. Annotators were shown all the close-talk channels per meeting in parallel, with each segmented instance of laughter already identified, together with it's original ICSI VocalSound. They were able to select and listen to each instance on its foreground channel, the same time interval on any of the remaining channels, and the instance's temporal context on the foreground and remaining channels[2]. Annotators were encouraged to insert ad-hoc comments in addition to their voiced/unvoiced classification.

Annotation took approximately 1.25 times realtime. 58 meetings were labeled by one of two annotators, 14 were labeled by one annotator and were then checked by the other, and 3 were independently labeled by both annotators. We estimate the total effort for this activity to be 110 hours. Finally, all laughter instances which received a comment during classification were subsequently listened to by both authors, an additional effort of 2×12 hours.

Interlabeler agreement on the classification of voicing was computed using the three meetings which were labeled independently by both annotators, Bmr016, Bmr018 and Bmr019. Agreement was between 88–91%, and chance-corrected κ-values for the three meetings fell in the range 0.76–0.79. This is lower than we expected, having had assumed that assessment of voicing is not a very subjective task. Inspection of the disagreements revealed that they occurred for VocalSound instances whose endpoints had been inferred from inaccurate forced alignment timestamps of the adjacent words. In many cases the annotators had optionally labeled the presence of speech in instance segments; since commented cases were revisited by both authors, a portion of the disagreement cases were resolved.

[1] We used the freely available Audacity© for this task. Only the foreground channel for each laughter instance was inspected.

[2] We used our in-house multichannel annotation tool TransEdit for this task.

In the remainder, we kept the voicing label of the annotator who had classified laughter in a larger portion of the meetings.

3. ANALYSIS

In this section, we describe the results of our investigations into the differences between voiced and unvoiced bouts of laughter, in terms of total time spent in laughter, bout duration, and multiparticipant vocal activity context.

3.1. Quantity

Of the 13259 bouts identified in the previous section, 38% were labeled as unvoiced while 61% were labeled as voiced. We are also interested in the total proportion of time spent laughing. For each participant, and for each of voiced and unvoiced laughter categories, we sum the time spent laughing, and normalize this quantity by the total time of meetings attended by that participants. Since a given participant may not have been present for the entirety of each meeting, the results we show represent ceiling numbers.

We found that the average participant spends 1.0% of their total meeting time in voiced laughter, and 0.4% of their total meeting time in unvoiced laughter. For contrast, the average participant spends 13.4% of their total meeting time on speaking. Laughter, whether voiced or unvoiced, therefore accounts for 9.4% of all vocalization effort. In Figure 1 we show that the time spent laughing and the proportion of voiced to unvoiced laughter vary considerably from participant to participant.

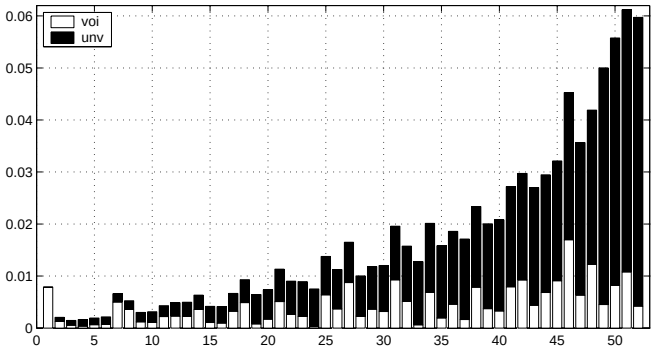

Fig. 1. Proportion of total recorded time per participant spent in voiced and in unvoiced laughter. Participants are shown in order of ascending proportion of voiced laughter.

Visually, there appears to be only a very weak correlation between the amount of individual participants' voiced laughter and their amount of unvoiced laughter. The majority of participants appears capable of both modes of laughter production.

3.2. Duration

Next, we analyze the durations of bouts to determine whether there is a difference for voiced and unvoiced laughter. The results are shown in Figure 2. Although bout durations vary much less than talkspurt durations, the modes for all three of voiced laughter bouts, unvoiced laughter bouts, and talkspurts fall between approximately 1 second and 1.5 seconds. Voiced bouts appear to be slightly longer than unvoiced bouts.

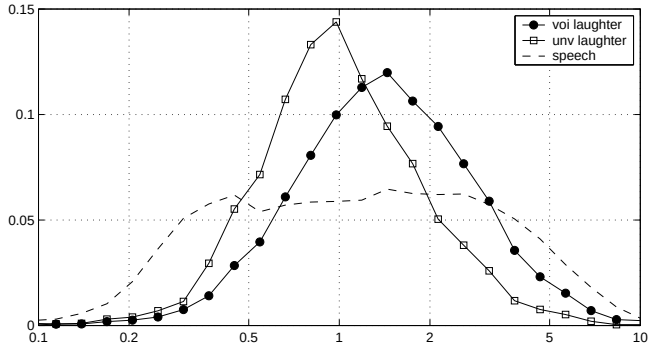

Fig. 2. Normalized distributions of duration for voiced laughter bouts, unvoiced laughter bouts, and talkspurts.

3.3. Interaction

Finally, and perhaps most importantly, we attempt to analyze local (short-time) differences in conversational context for the voiced and unvoiced laughter. For each bout, we study only the vocal interaction context; in particular, we ignore the specific words spoken and focus only on whether each participant is silent, laughing (in either voiced or unvoiced mode), speaking, or both.

We accomplish this analysis in a time-synchronous fashion as follows, accumulating statistics over all meetings. For every meeting, we begin with the reference on-off patterns, corresponing to speech (Subsection 2.1), for each of K participants present in that meeting. We discretize these patterns using 1-second non-overlapping windows, declaring a participant k as speaking for all of frame t if the reference shows participant k as speaking for at least 10% of frame t. We do the same with the on-off voiced laughter segmentation and the on-off unvoiced laughter segmentation, producing for each meeting 3 binary-value matrices of size $K \times T$, where T is the number of 1-second non-overlapping frames (ie. the duration of the meeting in seconds).

For each meeting, when laughter (either voiced or unvoiced) is produced at time $1 \leq t \leq T$ by participant $1 \leq k \leq K$, we inspect whether participant k is are also laughing at time $t-1$ and at time $t+1$. Each frame of laughter is then binned into one of three categories: (1) laughing at time t but not at time $t-1$; (2) laughing at time t but not at time $t+1$; and (3) laughing at frame $t-1$, t, and $t+1$. These bins correspond to laugh initiation, laugh termination, and laugh continuation.

Rather than trying to identify discriminative contexts for voiced and unvoiced laughter by hand, we have chosen to allow a simple machine learning formalism, a decision tree, to learn these context automatically. We describe each frame of laughter, produced by participant k at time t, using the following features: the number of other participants producing speech at times $t-1$, t, and $t+1$; the number of other participants producing voiced laughter at times $t-1$, t, and $t+1$; the number of other participants producing unvoiced laughter at times $t-11$, t, and $t+1$. We use these 9 features to characterize laughter frames in all three of laugh initiation, laugh termination, and laugh continuation. Additionally, for laugh initiation and termination, we also include a binary feature specifying whether the laugher is speaking at time $t-1$ and $t+1$, respectively.

For each of the three categories of laugh initiation, laugh termination, and laugh continuation, we train a separate decision tree to predict whether a particular frame of laughter is voiced or un-

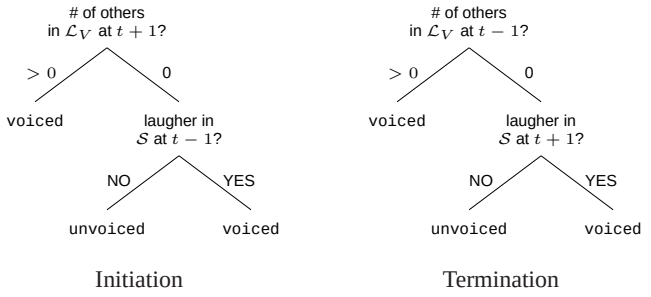

Fig. 3. Automatically identified decision trees for detecting voiced versus unvoiced laughter based on multiparticipant vocal activity context. States $\mathcal{L}_V$ and $\mathcal{S}$ represent voiced laughter and speech, respectively.

voiced, based only on its vocal activity context. We have chosen decision trees as the partitioning hyperplanes they find are orthogonal, making them particularly easy to interpret; overfitting is also easily controlled. Here, we have required that each hypothesized tree node account for at least 1000 exemplars. We note that we are not interested in the absolute classification accuracy of each tree, only in whether a tree survives pruning (ie. its statistical significance).

For continuation frames, no meaningful distinction was hypothesized between voiced and unvoiced laughter. For initiation and termination frames, we show the inferred classification trees in Figure 3. It is surprising that the two trees are symmetric. In classifying a frame which initiates a bout, the most useful contextual feature, of those studied here, is whether others *will be* laughing at $t+1$. In classifying a frame which terminates a bout, the most useful feature is whether others *were* lauging at $t-1$. Similarly, the next most useful feature is temporally adjacent speech by the laugher. For bout-initiating frames, if the laugher *was* speaking at $t-1$ then the inferred decision tree predicts that the laugher is currently producing voiced laughter. Symmetrically for bout-terminating frames, if the laugher *will be* speaking at $t+1$ then the prediction is also that the laugher is currently producing voiced laughter.

4. CONCLUSIONS & FUTURE WORK

We have produced a complete voiced and unvoiced laughter segmentation for the entire ICSI Meeting Corpus, including isolated instances as well as instances of laughter cooccuring with the laugher's speech. We have shown that on average, voiced laughter accounts for 32% of all observed laughter in this corpus, but that participants vary widely in their use of voicing while laughing. Most importantly, we have shown that in spite of inter-participant differences, voiced and unvoiced laughs are correlated with different vocal interaction contexts. Voiced laughter seems to differ from unvoiced laughter in that voiced laughter from other participants follows its initiation and precedes its termination. Voiced laughter also seems more interdependent with the laugher's speech; in cases where laughter follows speech or precedes laugher's speech, it is more likely to be voiced than unvoiced.

We intend to apply these observations to the construction of hidden Markov model topologies for finding single- and multi-participant laughter. We are also interested in determining whether vocal activity context affects the emotional valence of laughers as perceived by outside observers. Our long term goal is to be able to automatically determine which laught bouts are predictive of emotional valence.

5. ACKNOWLEDGMENTS

We would like to thank our annotators Annotator1 and Annotator2. We would also like to thank Acknowledgee3 for investing time to explain the ICSI MRDA corpus and for continued encouragement in our study of laughter, and Acknowledgee4 for useful discussion. This work was funded in part by the FundingBody under ProjectAcronym, ProjectName (`http://project@website`).

6. REFERENCES

[1] S. Burger, V. MacLaren, and H. Yu, "The ISL meeting corpus: The impact of meeting type on speech style," in *Proc. ICSLP*, Denver CO, USA, 2006.

[2] A. Janin et al, "The ICSI meeting corpus," in *Proc. ICASSP*, Hong Kong, China, 2003, vol. 1, pp. 364–367.

[3] E. Shriberg, "Spontaneous speech: How people really talk, and why engineers should care," in *Proc. INTERSPEECH*, Lisbon, Portugal, 2005, pp. 1781–1784.

[4] M. J. Owren and J.-A. Bachorowski, J., "Reconsidering the evolution of nonlinguistic communication: The case of laughter," *Journal of Nonverbal Behavior*, vol. 27, no. 3, pp. 183–199, 2003.

[5] AuthorName1 and AuthorName2, "Papertitle," in *Proc. Conference*, 2006.

[6] D. Neiberg, K. Elenius, and K. Laskowski, "Emotion recognition in spontaneous speech using GMMs," in *Proc. INTERSPEECH*, Pittsburgh PA, USA, 2006, pp. 809–812.

[7] G. Jefferson, *Everyday Language: Studies in Ethnomethodology*, chapter A Technique for Inviting Laughter and its Subsequent Acceptance Declination, pp. 79–96, Irvington Publishers, 1979.

[8] R. R. Provine, "Contagious laughter: Laughter is a sufficient stimulus for laughs and smiles," *Bulletin of the Psychonomic Society*, , no. 30, pp. 1–4, 1992.

[9] E. Shriberg et al, "The ICSI meeting recorder dialog act (MRDA) corpus," in *Proc. SIGdial*, Cambridge MA, USA, 2004, pp. 97–100.

[10] E. Shriberg, A. Stolcke, and D. Baron, "Observations on overlap: Findings and implications for automatic processing of multi-party conversation," in *Proc. EUROSPEECH*, Aalborg, Denmark, 2001, pp. 1359–1362.

[11] J. Trouvain, "Phonetic aspects of "speech-laughs"," in *Conference on Orality and Gestuality (ORAGE)*, Aix-en-Provence, France, 2001, pp. 634–639.