AN ACOUSTIC ANALYSIS OF SHARED ENJOYMENT IN ECA INTERACTIONS OF CHILDREN WITH AUTISM

Theodora Chaspari, Emily Mower Provost, Athanasios Katsamanis, Shrikanth Narayanan

University of Southern California (USC)
Signal Analysis and Interpretation Laboratory, USC, Los Angeles, California, USA
chaspari@usc.edu, emprovost@umich.edu, nkatsam@sipi.usc.edu, shri@sipi.usc.edu

ABSTRACT

The quality of shared enjoyment in interactions is a key aspect related to Autism Spectrum Disorders (ASD). This paper discusses two types of enjoyment: the first refers to humorous events and is associated with one's positive affective state and the second is used to facilitate social interactions between people. These types of shared enjoyment are objectively specified by their proximity to a voiced and unvoiced laughter instance, respectively. The goal of this work is to study the acoustic differences of areas surrounding the two kinds of shared enjoyment instances, called "social zones", using data collected from children with autism, and their parents, interacting with an Embodied Conversational Agent (ECA). A classification task was performed to predict whether a "social zone" surrounds a voiced or an unvoiced laughter instance. Our results indicate that humorous events are more easily recognized than events acting as social facilitators and that related speech patterns vary more across children compared to other interlocutors.

Index Terms— Social interactions, shared enjoyment, laughter, autism, embodied conversational agent

1. INTRODUCTION

Social interaction deficits are one of the core symptoms of people diagnosed with Autism Spectrum Disorders (ASD). Children with autism take fewer social initiatives and have difficulty engaging in interactions. The "pleasure in interactive participation or conversation [1]", known as "shared enjoyment", has been codified as a main aspect of rating social interaction in autism [2].

Laughter is an objective expression of enjoyment. It can be divided into two categories: voiced and unvoiced. Voiced laughter is caused by quasi-periodic vibrations of the vocal folds and includes mostly chuckles and giggles. Unvoiced laughter is more atonal and results from fricative excitation. It includes open-mouth, pant-like sounds, closed-mouth grunts and nasal snorts [3]. In a relevant study, it is hypothesized that voiced laughter is strongly associated with positive affect, whereas unvoiced laughter acts as a social facilitator by supporting aspects of social communication. It was also shown that children with autism produce a higher percentage of voiced than unvoiced laughter [4].

Children with autism have difficulty maintaining social interaction. Consequently, a strong understanding of the interaction between speech and voiced/unvoiced laughter may provide therapists with a greater understanding of a child's communicative and social abilities. This paper addresses the effect of the two laughter categories on the surrounding speech. We hypothesize that if the laughter types serve different social functions, the speech surrounding different laughter instances should exhibit different feature-level patterns. The objective analysis of vocal cues existing in these areas can afford us new insights into the nature of social interactions in children with autism.

The data used in this paper were recorded from interactions between a child, his parent, and an Embodied Conversational Agent (ECA). ECAs provide a context for the elicitation of social communicative behavior in child-machine interactions [5]. The Rachel system, developed at the University of Southern California [6, 7], is a platform for the collection of emotional, social and communicative behavior observations between a child, the ECA agent, and the child's parent (Section 2).

In this paper we differentiate between areas of hypothesized shared enjoyment specified by proximity to voiced and/or unvoiced laughter. We define these laughter-proximal areas as "social zones". Given that voiced and unvoiced laughter serve different social roles, our hypothesis is that the speech in social zones can be used to classify the type of laughter (voiced/unvoiced) that either precedes or follows the speech incident, because the speech proximal to these two types of laughter should maintain echoes of the social function. We test this assumption by automatically classifying speech regions surrounding laughter instances to two classes: whether they are around a voiced or an unvoiced laughter instance.

Our results demonstrate that speech patterns of shared enjoyment instances are different according to whether the enjoyment instances correspond to the two types of laughter. They also indicate that humorous events are easier to recognize than social facilitators, showing that the first kind is more prominent in these social interactions. The slightly greater variance of the results across the children implies a difference in the amount of children's social engagement in shared enjoyment interactions.

2. DESCRIPTION OF DATA

The data utilized in this study come from the "Rachel ECA Interaction Corpus" [6], which contains experiments designed to promote social interactions of children with autism and highlight their emotional reasoning abilities. It consists of four sessions in which the child interacts with Rachel and his/her parent. Each session was recorded with a smart-room setup consisting of three Sony High-Definition cameras and two shot-gun microphones. The ECA was controlled using the Wizard of Oz paradigm, in which a hidden experimenter uses a graphical user interface (GUI) to chose the ECAs actions and utterances.

This paper includes data from the pilot studies conducted for two children. The first child was a 12-year old boy with an expres-

Thanks to NSF for funding.

Table 1: Number and duration of laughter instances

	S1 experiments		S2 experiments	
	voiced	unvoiced	voiced	unvoiced
total	166	66	95	15
non-speech-overlap	126	63	69	13
mean duration (sec)	3.34	2.84	2.87	1.69

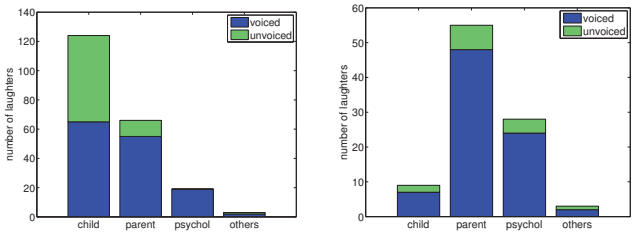
sive language score of 6 years, 7 months. The second child was a 6-year-old boy with an expressive language score of 2 years, 9 months. The interactions of these children with Rachel, their parents and the psychologist will be referred as "subject 1 (S1) experiments" and "subject 2 (S2) experiments" respectively. A more detailed description of the experimental conditions can be found in [7].

3. ANALYSIS OF SHARED ENJOYMENT INSTANCES

Shared enjoyment instances were tagged manually by one transcriber, who marked the starting and ending time for each instance, the identity of the individual(s) who laughed (child, parent, psychologist, other people -mainly Rachel wizards-), whether or not laughter overlapped with speech, and the kind of event (humorous/voiced laughter or social facilitator/unvoiced laughter). We believe that one transcriber is sufficient since these types of events are easily differentiated using audio information.

There were a total of 232 and 110 laughter instances in the S1 and S2 data, respectively. Table 1 shows the number of instances overlapping with speech and their duration and Fig. 1 displays a detailed distribution of laughter among individuals. It is shown that there are differences in the content of social interactions between S1 and S2 experiments and among people participating in them. Age difference between the two subjects could be a main reason for this variability, which is also reflected in our results (Sections 5, 6).

Inspection of various shared enjoyment events in our data led to a number of interesting observations. In most cases, it was easy to identify the particular social role that was served each time, i.e., either an expression of positive affect in a humorous situation or an effort to facilitate social interaction. For example, a lot of subject's 1 voiced laughter happened while he was playing with his parent and apparently enjoying the interaction. On the other hand, he often expressed unvoiced laughter right after Rachel asked him a question, seemingly trying to show his unfamiliarity with the unconventional interface. We also noticed that there were often short pauses before or after unvoiced laughters for all interlocutors, suggesting that this kind of laughter was an effort to "smoothen" the social interaction.



(a) S1 (b) S2

Fig. 1: Distribution of laughter instances across child, parent, psychologist and other people (usually wizards) participating in the subject 1 (S1) and the subject 2 (S2) experiments.

4. METHODS

Motivated by our observations and the study presented in [4] we investigate whether we can establish a clear link between the general affect in laughter-proximal regions with the type of the corresponding laughter, namely voiced or unvoiced. More specifically, we hypothesize that we can determine the type of laughter from the speech region that precedes or follows, i.e., the corresponding social zone. In our current work, we just focus on the acoustic properties of these zones while in the future we plan to also exploit lexical and visual information as well.

To test this hypothesis we performed two classification experiments based on appropriately selected acoustic features. In the first experiment we classified the type of laughter (voiced or unvoiced) just based on acoustic information extracted from the laughter region. In the second experiment we tried to predict the type of laughter from the surrounding social zones. The results indicate that the difference between the two types of laughter is "echoed" in the surrounding areas.

Table 2: Number of before- and after- laughter regions for each social zone size (szs) in seconds.

szs	S1 experiments				S2 experiments			
	before		after		before		after	
	voiced	unvoiced	voiced	unvoiced	voiced	unvoiced	voiced	unvoiced
2	158	65	164	59	94	14	94	14
4	133	49	133	49	91	12	91	12
6	117	41	117	41	88	12	89	12
8	115	39	115	39	83	11	82	12
10	108	35	107	36	82	10	81	11

4.1. Social Zones

The hypothesis of our study is that speech patterns surrounding social events are indicative of the nature of social interaction. We isolated social zones by manually detecting laughter instances and selecting the regions before and after these instances, which will be referred as "before-laughter" and "after-laughter" regions. The time duration of these regions was 2, 4, 6, 8 and 10 seconds. These values were found empirically to be indicative of the content of the laughter incident. It is reasonable to expect that the laughter's effect on speech dissipates beyond 10 seconds.

In case of two consecutive laughter instances, if there was an overlap between the after-laughter region of the first and the before-laughter region of the second, then each region would be extended to no more than the middle of this overlapping area (Fig. 2). For a specific social zone size, if the length of a region was smaller than the size of the next smaller social zone, the region would not be taken into account. The number of regions for each social zone size is shown in Table 2.

4.2. Feature Extraction

The audio sampling frequency was 16kHz. The features were extracted using Praat [8] with 0.04sec frame size and 0.01sec step and consist of pitch, intensity and the first 13 Mel Filterbank Coefficients

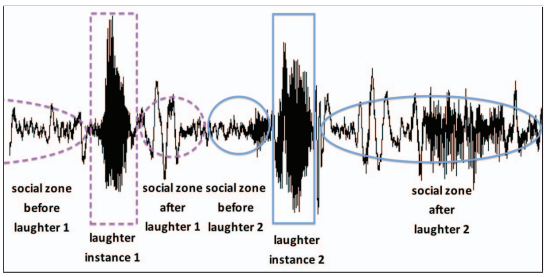


Fig. 2: Example of two laughter instances and their "social zones".

(MFB). These features were shown to be effective in studies of emotion classification [9] and laughter detection [10]. We computed statistical functionals over the features in non-overlapping windows of 0.5, 1.0, 1.5, 2 and 2.5 seconds. The statistical functionals are: mean, range, standard deviation, skewness, kurtosis, 25% quantile, 75% quantile and quantile range. The original feature set is 120-dimensional. We reduced the dimensionality of our features by applying Principal Component Analysis (PCA) and then selecting an appropriate subset of the resulting dimensions as it will be further described in Sec. 5.

4.3. Classification

We used a Naive Bayes classifier in both classification tasks due to the small size of our corpus. The experiments were done with data from all interlocutors over the two different parent-child pairs using a leave-one-instance-out cross-validation. The first experiment included only laughter instances not overlapping with speech, while this was not a constraint for the second one. PCA coefficients were computed over the training data of each fold. The test set was normalized using the mean and variance of the train set and projected over the components identified during training.

5. EXPERIMENTS

To reduce the dimensionality of the original feature space and preserve information relevant to our task we used Principal Component Analysis. By PCA we project the features on a new coordinate system, defined by an ordered set of axes (components) $\{\alpha_i : i = 1 \dots 120\}$, so that the projection on the first axis will have the maximum variance, the projection on the second axis the second maximum variance, etc. Intuitively, we would expect that most of the variability in our acoustic features comes from differences in the spoken content or in speaker characteristics. Based on that, we can assume that the most significant PCA components, i.e., those corresponding to the highest variance, will mainly capture utterance and speaker dependent information, while socially relevant information will probably be represented by the least significant dimensions.

To investigate this assumption, we studied the laughter type classification based on the around-laughter regions when using different subsets of PCA components to project our data on. More specifically, for each classification experiment we used a different subset $S_k, k = \{1 \dots 23\}$ of 10 principal components of consecutive order, defined as $S_k = \{\alpha_i : i = 5(k-1)+1 \dots 5k+5\}$. We then performed Naive Bayes classification with leave one-instance-out cross-validation. The results are presented in Fig. 3 when using the before-laughter social zones. Same trends are observed for the after-laughter regions. Projection on the least significant subsets, i.e., those corresponding to less variance, yields better classification results. This suggests that shared enjoyment-relevant information may be better represented by the higher-order principal components. We also see that the lower-order dimensions can be more discriminative than the intermediate ones, yet less effective than the higher-order dimensions. These findings indicate that shared enjoyment information is probably responsible for a relatively small amount of speech acoustics variance, but may also be conveyed by other more variant channels. Based on these observations, the following classification experiments are performed on data projected onto the 10 highest-order PCA components, i.e. $S_{23} = \{\alpha_{110} \dots \alpha_{120}\}$.

The goal of the first classification experiment was to identify the type of laughter (voiced vs. unvoiced) given an unlabeled pre-segmented laughter instance. The success percentage when using a 2.5 second window was 85.2% and 96.5% for S1 and S2 experiments

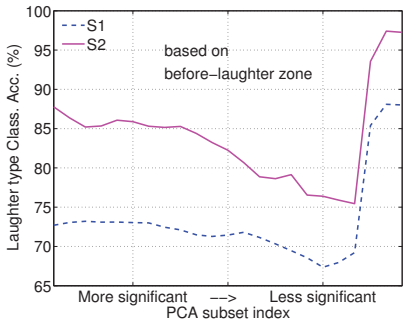


Fig. 3: Mean laughter type classification accuracy (over all window lengths and social zone sizes) when using before-laughter regions for acoustic feature projections on 10-dimensional PCA subsets with decreasing significance.

Table 3: Success percentages for voiced/unvoiced laughter classification over all window lengths (in seconds)

	baseline	window length (sec) 0.5	1.0	1.5	2.0	2.5
S1	66.7	74.3	67.7	86.8	85.6	85.2
S2	84.15	94.8	92.6	98.9	99.3	96.5

(Table 3). Larger windows produce in general better results, which implies that these statistical features are more reliably estimated over larger windows. This might also suggest that information over a long-time period can better capture the social content than over a short-time period. Voiced and unvoiced laughter incidents from S2 data seem to be more differentiable than from S1 data. This may be due to the difference in the amount of data between the two subjects. The age difference of the children might also cause this variability of the results. The corresponding results when using mean pitch values were on the baseline, which might come from the skewness of data towards voiced instances, but could also suggest that automatic classification between these two types of laughter is not trivial.

In the second experiment we classified whether an around laughter region occurred before/after a voiced or an unvoiced laughter instance given the human-specified labels. For 1.5 second window and 10 seconds social zone, before- and after- laughter regions were correctly recognized by 95.9% and 91.9% for S1 experiments and 98.1% and 98.8% for S2 experiments (Table 4). The results from S2 experiments are higher, which indicates that S2 data are more differentiable, but due to the small number of samples generalization in larger datasets is not guaranteed. We reported the results for 0.5, 1.0 and 1.5 second windows for visual clarity and because we didn't observe further improvement after this point. With the increase of zone size there is a saturation of the recognition measures, which suggests that the useful data predicting the kind of social interaction exists within a 10 second time interval. We also found that the classification accuracy calculated using speech preceding a laughter instance does not differ much from that based on speech data following a laughter instance. This suggests that equal amount of social information is spread before and after a laughter incident.

The measure of accuracy in all experiments is the weighted percentage and the baseline is noted in the corresponding tables (3, 4).

6. DISCUSSION

The classification results suggest that acoustic features of social zones are discriminative of the type of shared enjoyment interaction

Table 4: Success percentages of before-/after- laughter regions over all types of laughter (voiced/unvoiced) for different window lengths (win.len.) and social zone sizes.

(a) S1 experiments

		before laughter social zone size (sec)					after laughter social zone size (sec)				
		2	4	6	8	10	2	4	6	8	10
baseline		70.9	73.1	74.1	74.7	75.5	73.5	73.1	74.1	74.7	74.8
win.	0.5	81.1	85	93.9	88.3	95.3	82	89.6	85.6	91	89.3
len.	1	91.2	84.6	88.6	88.9	92.5	86.6	90.3	86.4	86.1	83.5
(sec)	1.5	85.5	93.6	84.6	88.4	95.9	89.2	82.1	74.4	89.8	91.9

(b) S2 experiments

		before laughter social zone size (sec)					after laughter social zone size (sec)				
		2	4	6	8	10	2	4	6	8	10
baseline		87.1	88.4	88	88.3	89.1	87	88.4	88.1	87.2	88
win.	0.5	94.5	99	98.4	98.4	98.3	91.2	99	97.2	95.5	96.9
len.	1	94.3	97.9	99.6	98.9	99.3	98.7	99.2	99.1	94.7	99
(sec)	1.5	97.6	99	97.2	97.9	98.1	96.7	99.7	98.3	97.3	98.8

that takes place. We will further see how social zones are affected depending on the type of shared enjoyment interaction and on the personal characteristic's of each child.

We examined if/how our classification results differ according to the identity of interlocutors. Seven possible groups of people were considered: child/ parent/ psychologist/ others laughing alone ("ch", "pa", "ps", "oth",), the child laughing at the same time with his/her parent ("ch-pa"), the parent laughing with the psychologist ("pa-ps") and the parent laughing with other people ("pa-oth"). For each group we computed the mean number of times (across all window lengths and for social zone of size 2 seconds) that the corresponding before- and after- laughter regions are classified correctly. Both total laughter instances and correctly classified instances seem similar for the parents and the psychologist accross S1 and S2 experiments. However this is not the case for the children, as they differ in the amount of times they are getting engaged in shared enjoyment interactions. Also regarding around laughter regions belonging to child, S1 seems to be more poorly recognized than S2. This might indicate a difference between the two children in handling shared enjoyment events, but the small amount of data prohibits us from further generalization. The detailed results are shown in Table 5. Similar trends were also followed by the remaining sizes of social zones.

In Table 5 we also see that regions surrounding voiced laughter instances have better classification results than regions around unvoiced laughter. This indicates that shared enjoyment instances reflecting humorous events are more prominent than instances used to negotiate social subtleties.

7. CONCLUSIONS AND FUTURE WORK

This study provides a novel vocal analysis of the indicators of shared enjoyment in ECA interactions of children with autism. The results suggest that acoustic patterns of social zones are indicative of the type of role that a social event serves and reflect the amount of an interlocutor's social engagement to shared enjoyment interactions.

One limitation of this paper is that it relies on a small amount of data. Future work includes extending this study to a larger group of subjects, data for which have already been collected. The complete dataset will allow us to proceed to more complicated modeling of social zones and examine whether acoustic features can be used as a quantitative measure of a child's affective state and social response.

Future work will also examine if the identity of the person speaking and the speech content in an around-laughter region is indicative of the kind of shared enjoyment event that precedes/follows.

Table 5: Mean number of correctly classified before-/after-laughter areas for all interlocutors: child(ch), parent(pa), psychologist(ps), others (oth), child-parent (ch-pa), parent-psychologist (pa-ps), parent-others (pa-oth). Mean is computed across all window lengths for 2 second social zone size. "Total" denotes the total number of around laughter regions for each type of laugher and each interlocutor and "corerct" is the mean number of correctly classified areas.

(a) regions around voiced laughter instances

			ch	pa	ps	oth	ch-pa	pa-ps	pa-oth
S1	before	total	61	54	18	5	15	5	0
		correct	54.25	51.75	17.75	4.75	14.25	5	-
	after	total	65	54	18	5	17	5	0
		correct	58.75	50.5	17.5	5	17	5	-
S2	before	total	5	53	18	5	2	10	1
		correct	5	52.75	17	4	2	10	0.5
	after	total	5	54	17	5	2	10	1
		correct	5	51.5	17	4.5	2	9.25	1

(b) regions around unvoiced laughter instances

			ch	pa	ps	oth	ch-pa	pa-ps
S1	before	total	52	4	0	1	8	0
		correct	36.75	2.5	-	0.75	4.75	-
	after	total	47	4	0	1	7	0
		correct	30.5	3.25	-	1	2.75	-
S2	before	total	2	8	2	1	0	1
		correct	1.75	6.25	1.75	1	-	0.75
	after	total	1	8	3	1	0	1
		correct	0.5	6	2.5	1	-	0.75

8. REFERENCES

[1] S. R. McConnell, "Interventions to facilitate social interaction for young children with autism: Review of available research and recommendations for educational intervention and future research," *Journal of Autism and Developmental Disorders*, vol. 32, pp. 351–372, 2002.

[2] C. Lord, L. Lambrecht, and E. H. Cook et. al., "Autism diagnostic observation schedule - generic: A standard measure of social and communication deficits associated with the spectrum of autism," *Journal of Autism and Developmental Disorders*, vol. 30, pp. 205–223, 2000.

[3] K. Laskowski and S. Burger, "On the correlation between perceptual and contextual aspects of laughter in meetings," *Proc. ICPhS Workshop on the Phonetics of Laughter, Saarbrucken, Germany*, August 2007.

[4] W. J. Hudenko and W. Stone, "Laughter differs in children with autism: an acoustic analysis of laughs produced by children with and without the disorder.," *Journal of Autism and Developmental Disorders*, vol. 39, pp. 1392–1400, 2009.

[5] S. S. Narayanan and A. Potamianos, "Creating conversational interfaces for children," *IEEE Trans. on Speech and Audio Processing*, vol. 10, pp. 65–78, 2002.

[6] E. Mower, M. Black, E. Flores, M. Williams, and S. Narayanan, "Rachel: Design of an emotionally targeted interactive agent for children with autism," *ICME, Barcelona, Spain*, July 2011.

[7] E. Mower, C.C. Lee, J. Gibson, T. Chaspari, M. Williams, and S. Narayanan, "Analyzing the nature of eca interactions in children with autism," *Interspeech, Florence, Italy*, 2011.

[8] P. P. G. Boersma, "Praat, a system for doing phonetics by computer," *Glot international*, vol. 5, pp. 341–345, 2002.

[9] E. Mower and S. Narayanan, "A hierarchical static-dynamic framework for emotion classification," *ICASSP, Prague, Czech Republic*, 2011.

[10] M. T. Knox and N. Mirghafori, "Automatic laughter detection using neural networks," *Interspeech, Antwerp, Belgium*, pp. 2973–2976, 2007.