Temporal precision and reliability of audience response system based annotation

Jens Edlund[1], Samer Al Moubayed[1], Christina Tånnander[2], Joakim Gustafson[1]

[1]KTH Speech Music and Hearing, Sweden
edlund@speech.kth.se,sameram@kth.se,jocke@speech.kth.se
[2]Swedish Agency for Accessible Media (MTM), Sweden
christina.tannander@mtm.se

Abstract. Manual annotators are often used to label human interaction data. This is associated with high costs and high time consumption, which is one reason annotation by crowd sourcing is increasing. But in crowd sourcing, control over the conditions is largely lost. To get increased throughput with a higher measure of experimental control, we suggest borrowing from the Audience Response Systems used extensively in the film and television industries. We present (a) a cost-efficient setup for rapid, plenary annotation of human spoken interaction data; and (b) a study that quantifies the temporal precision and reliability of annotations made with the system in the auditory perception domain.

Keywords. Index Terms: annotation, speech, spoken dialogue

1 Background

Over the past decades, corpus studies of human interactive behaviour has been increasingly common, for purposes varying from basic research of human-human interaction to analyses undertaken as a starting point in efforts to build humanlike spoken dialogue systems [1]. This has led to a dramatic increase in numbers of data collections of human interactions and in the sheer amounts of data that is captured per hour of interaction in the resulting corpora, as exemplified by massively multimodal corpora such as the AMI Meeting Corpus comprised of a large number of video and audio channels as well as projector output [2] or our own D64 and Spontal corpora [3, 4] combining multiple video and audio channels with motion capture data.

While these corpora are useful, the task of annotating them is forever more daunting. It is becoming impossible to produce annotations in the traditional manner, with one or more highly skilled, highly trained experts spending many times real time for each annotation type – even a simple task such as correcting utterance segmentations that are already largely correct requires 1-3 time real time [5]. For many types of annotation, we may also raise a question related to ecological validity: why should it be so hard to label what people perceive effortlessly in everyday conversation?

The rapidly growing demand for annotation has led to an increasing interest and use of crowdsourcing services such as Amazon Mechanical Turk. But although Mechanical Turk annotations are reported to be cheap, fast and good enough [6],

crowdsourcing means that we relinquish control over the situation in which the annotation is made. To achieve increased throughput while maintaining a higher measure of experimental control, we take the ideas behind the Rapid Prosody Annotation pioneered by Jenifer Cole as our starting point (see [7]), and add a technical solution borrowed from the Audience Response Systems used in the film and television industries. The goal is to have laymen annotators annotate some straightforward event in real time. Apart from being fast and cost effective, it places annotators in a situation similar to that in which they perceive speech on an everyday basis.

We have previously investigated the use web based ARS-based methods for perception experiments [8] and evaluation of children's speech [9]. The present system is currently being tested for speech synthesis evaluation and for annotation of the type of speech events that are sometimes describes as disfluencies in the literature.

Fig. 1. The system: Xbox 360 Controllers and Receivers in a custom made portable case.

2 The ARS-based annotation system

ARS systems are used for screenings of new films and television series. They are typically expensive, and their software proprietary. In order to make an ARS-based system that is affordable for academic research, we built our system (see Figure 1) of Xbox 360 Controllers for Windows. These devices use low-latency wireless communication over a receiver which supports four controllers, but we have been able to robustly use more than one receiver per computer, allowing us to use more frame synchronized controllers simultaneously. Our system uses a standard Windows PC with an external loudspeaker, and a good quality projector for multimodal annotations.

To capture the controller states, we developed a Java library based on the DirectInput Microsoft API. The software automatically captures all controllers connected to the system by querying them for the state of their buttons, triggers and pads at a specified frame rate. The capture software can read the state of all input components on the controller (analogue and digital), but cannot yet utilize the two-way communication afforded by the controller by its rumble motors and diodes.

3 Method

The purpose here is to provide baseline statistics showing what can be expected of such a system. To achieve this goal, we ran a series of five tests with stimuli that are

significantly clearer than most dialogue and human interaction phenomena, which allows us to measure with accuracy the temporal precision of the system (for such stimuli). We then added a sixth, typical annotation task: end-of-utterance detection.

We used 4 groups of subjects (G1-G4), with 8, 8, 7 and 3 participants, respectively, for a total of 26 subjects. The subjects within a group took the test simultaneously, in full view of each other. All subjects were computer science students participating as part of a class. 10 of the subjects were female, 16 male.

Stimuli series S1 through S5 consist of 1 s long beeps at 170 Hz, but their temporal patterns vary, as do the annotators' task. S1 consists of 20 beeps spaced evenly, so that their onset occur regularly every 3.6 seconds. S2 also contains 20 evenly spaced beeps, but at 11.9 seconds interval. S3 contains 40 beeps, spaced evenly as in S1. S4 contains 60 beeps, presented in groups of three within which the beeps are spaced evenly at 3.5 seconds, as in S1. Between groups, the is an 8 second spacing corresponding to the duration an extra beep would have taken up. S5 holds 20 beeps spaced irregularly, at random intervals of up to 10 seconds. S6, finally, contains 20 brief, everyday utterances (e.g. "I drive to work every day", and "Are those new?"), read clearly with audible pauses averaging approximately 1s between each utterance.

The subjects were presented with the stimuli sets in order. For each set, they were told what to expect (i.e. if the clicks would be regular or irregular, and roughly at what intervals). They were also instructed as to what they should react to (click on). For S1, S2, and S5, they we simply asked to click as close to beep onset as possible. For S3, they were told to click as close to every other beep onset as possible. For S4, they were asked to click where the left out fourth beep in every series of three should have been - they were asked to click at something that was not present in the stimuli. They were asked to click as close as possible to the end of each utterance in S6.

4 Results

There were no signs of equipment failure or code malfunction. In group 2, there were 615 instances of double clicks within less than 10 ms. The latter of each of these was removed. Once this was done, between-group differences were negligible, with three groups producing and average number of clicks per stimuli and subject varying between 0.99 and 1.01. The handsets performed uniformly, with average clicks per stimuli and handset varying between 0.99 and 1.03.

Overall, the subjects performed well and did what they were asked to do: subjects produced between 0.98 and 1.03 clicks per stimuli.

All clicks could easily and automatically be related to the stimuli by virtue of appearing shortly after stimuli onset, and long before the next stimuli onset.

For all stimuli series except S4, there is a very clear effect of the onset of the stimuli series. The average response time to the first stimuli in a series is 2-4 times larger than that of any of the remaining 19. We therefore treat the first stimuli in each series as an outlier and remove it from further analysis. Table 1 shows the mean response times to stimuli types. The differences are significant (one-way ANOVA, $P<0.0001$). Table 2 shows the individual differences between groups.

Table 1. Average response times (ms) for all subjects over stimuli types, with standard deviation, counts, and significance at 5 % level versus the rest of all stimuli types.

	Mean	*Stddev.*	*N*
S1	525	162	498
S2	615	260	501
S3	528	226	518
S4	696	1276	518
S5	592	130	495
S6	471	295	485
All	572	571	3015

Table 2. Results of pairwise T-tests. * represents $p < 0.05$, ** $p<0.01$, *** $p<0.001$, and **** $p<0.0001$. The rightmost column shows the group tested against all other groups.

	S1	*S2*	*S3*	*S4*	*S5*	*S6*	*All*
S1	n/a	****	-	**	****	***	*
S2	****	n/a	****	-	*	****	*
S3	-	****	n/a	***	****	***	*
S4	**	-	***	n/a	*	***	****
S5	****	*	****	*	n/a	****	-
S6	***	****	***	***	****	n/a	****

For purposes of using our system for annotation of speech and interaction phenomena, we need an analysis of response time distribution that allows us to predict where, in a continuous data stream, the unknown event that caused a group of subjects to click is most likely to be found. Instead of histograms, we perform these analyses using Kernel Density Estimation (KDE) estimations, which produce an analysis that is similar to a histogram, but produces a continuous, smooth curve.

Table 3. RT means over all subjects and KDE peaks over G1 and all subjects (ms).

	Mean	*G1*	*All*
S1	525	550	520
S2	615	610	560
S3	528	470	490
S4	696	720	710
S5	592	600	590
S6	471	430	450
All	572	550	550

Our ultimate goal is to use subjects' clicks to localize events in streaming data. As a first step, we return to KDE: we build an estimate over an entire stimuli set by adding a narrow width (0.5 s) Gaussian for each click of each subject relative to the start of the session. We estimate the onset of a trigger by finding a peak in the KDE curve for a stimulus and deducting the RT estimate for that stimulus type from the peak position. In the remainder of this analysis, we use reaction time estimates acquired by finding the peak of KDE curves, rather than using averages. Two sets of reaction times are used: for descriptions, we base reaction time estimates on all subjects. For error estimation, we use reaction time estimates based only on G1 (8 subjects), which

is then held out from the testing. For completeness, the relation between the mean reaction times and the KDE reaction time estimates based on G1 and all subjects are presented in Table 3. Table 4 shows the errors for trigger estimates when both RT estimates and KDE curves are based on all participants. Overall differences were tested with a one-way ANOVA ($p < 0.001$). Differences between each set and all other sets were tested with pairwise T-tests and are reported in the table.

Table 4 (left). The mean error (ms; the mean of the absolute difference between actual trigger positions and estimated trigger positions) for each set, with standard deviation and significance. **Table 5 (right).** Mean errors differences between 3, 10, 18 and 26 (all) subjects. Minus signs stand for a decrease in mean error, plus for an increase. Only significant changes are included.

	Mean	*Std. dev.*	*N*	*Sig.*
S1	23	33	19	-
S2	9	11	19	**
S3	28	25	39	-
S4	61	45	19	**
S5	25	22	19	-
S6	73	103	19	***
All	36	51	134	n/a

	3 vs. 10	*10 vs 18*	*18 vs. all/all*
S1	-	-	-
S2			-
S3	-		
S4	-	-	-
S5	+	-	-
S6	-		
All	-	-	-

In order to investigate the effect of the number of subjects used on the reliability of the trigger estimates, we use one group (group 1 of 8 subjects) to estimate RT and then use group 4 (G4, 3 subjects), group 3 and 4 (G34, 10 subjects) and group 2, 3, and 4 (G234, 18 subjects) to estimate trigger times for each trigger in each stimuli set.

One-way ANNOVAs within each stimulus set as well as within all stimuli all show a significant overall effect of number of subjects ($P < 0.0001$ in all cases). The results of T-tests between pairs of increasing numbers of subjects are shown in Table 5. A weak tendency to a lowered error over time can be seen in some of the stimuli sets, but the differences are small compared to the overall variance.

5 Discussion and future work

We have presented a cost efficient and robust system for plenary, simultaneous annotation by multiple subjects under controlled circumstances. In an experiment of 26 subjects responding to six different types of simple, salient auditory stimuli, we have shown that (a) although mean RT varies with stimuli, it can be modelled with relative ease during a single session in the system; that (b) an estimated mean RT for a stimuli type and a group of 10 subjects can be used to identify the position of unknown triggers of the same type with a precision well under a 200 ms for all stimuli types; and that (c) larger subject groups increase the precision of our trigger position estimates.

We note that S4, in which subjects were asked to react to omissions of expected auditory stimuli, predictably shows the lowest precision. We also note with some surprise that S5 – the only proper reaction time test – has a very short response time and an equally high precision, possibly because the subjects were kept alert by the task. Finally we note that over the duration of these tests, we could not find any clear evidence of effects of fatigue or training.

We will attempt to add more game controllers to the system, in order to be able to get more out of one single, controlled test series. We have made successful tests with one more receiver, making the number of possible active controllers twelve. Further technical development includes using of the feedback systems provided in the controllers: they are able to vibrate and have a small number of LEDs.

Regarding actual annotation, we are in the process of annotating language phenomena such as filler pauses, hesitations and repairs. We will follow up on these annotations and compare them, from efficiency and from an accuracy point of view, to traditional expert annotation.

6 Acknowledgements

This work was funded by the GetHomeSafe project (EU 7th Framework STREP 288667) and by the Swedish Research Council (VR) project Introducing interactional phenomena in speech synthesis (2009-4291).

7 References

1. Edlund, J., Gustafson, J., Heldner, M., & Hjalmarsson, A. (2008). Towards human-like spoken dialogue systems. Speech Communication, 50(8-9), 630-645.
2. McCowan, I., Carletta, J., Kraaij, W., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., Post, W., Reidsma, D., & Wellner, P. (2005). The AMI Meeting Corpus. In Proc. of Measuring Behavior 2005, the 5th International Conference onMethods and Techniques in Behavioral Research. Wageningen, Netherlands.
3. Edlund, J., Beskow, J., Elenius, K., Hellmer, K., Strömbergsson, S., & House, D. (2010). Spontal: a Swedish spontaneous dialogue corpus of audio, video and motion capture. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odjik, J., Piperidis, S., Rosner, M., & Tapias, D. (Eds.), Proc. of the Seventh conference on International Language Resources and Evaluation (LREC'10) (pp. 2992 - 2995). Valetta, Malta.
4. Oertel, C., Cummins, F., Campbell, N., Edlund, J., & Wagner, P. (2010). D64: A corpus of richly recorded conversational interaction. In Kipp, M., Martin, J-C., Paggio, P., & Heylen, D. (Eds.), Proceedings of LREC 2010 Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality (pp. 27 - 30). Valetta, Malta.
5. Goldman, J-P. (2011). EasyAlign: a friendly automatic phonetic alignment tool under Praat. In Proc. of Interspeech 2011. Florence, Italy.
6. Novotney, S., & Callison-Burch, C. (2010). Cheap, fast and good enough: automatic speech recognition with non-expert transcription. In Proc. of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT '10).
7. Mo, Y. (2010). Prosody production and perception with conversational speech. Doctoral dissertation, University of Illinois at Urbana-Champaign.
8. Edlund, J., Hjalmarsson, A., & Tånnander, C. (2012). Unconventional methods in perception experiments. In Proc. of Nordic Prosody XI. Tartu, Estonia.
9. Strömbergsson, S., & Tånnander, C. (submitted). Correlates to intelligibility in deviant child speech – comparing clinical evaluations to audience response system-based evaluations by untrained listeners. Submitted to Proc. of Interspeech 2013. Lyon, France.