Audience response system based annotation of speech

Jens Edlund [1], Samer Al Moubayed [1], Christina Tånnander [2] & Joakim Gustafson [1]
[1] KTH Speech, Music and Hearing, Stockholm, Sweden
[2] Swedish Agency for Accessible Media, MTM, Stockholm, Sweden

Abstract

Manual annotators are often used to label speech. The task is associated with high costs and with great time consumption. We suggest to reach an increased throughput while maintaining a high measure of experimental control by borrowing from the Audience Response Systems used in the film and television industries, and demonstrate a cost-efficient setup for rapid, plenary annotation of phenomena occurring in recorded speech together with some results from studies we have undertaken to quantify the temporal precision and reliability of such annotations.

Introduction

We present a cost-efficient and robust setup for plenary perception experiments based on existing consumer market technologies. In an experiment in which 26 subjects divided in 4 groups react to sets of auditory stimuli by pressing buttons, we show that humans annotating as a group, in real time, can achieve high precision for salient acoustic events, making this a feasible alternative to expert annotations for many tasks. The experimental results validate the technique and quantify the expected precision and reliability of such plenary annotations on different tasks.

Over the past decades, corpus studies of speech and spoken interaction have been increasingly common, for purposes ranging from basic research to analyses undertaken as a starting point in efforts to build humanlike spoken dialogue systems (Edlund et al., 2008). 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 (McCowan et al., 2005) or our own D64 and Spontal corpora (Edlund et al., 2010; Oertel et al., 2010) combining multiple video and audio channels with motion capture data.

While these corpora are useful, the task of annotating them is daunting. It is becoming near impossible to produce annotations in the traditional manner, with one or more highly skilled, highly trained experts spending several times real time or more for each annotation type – even a simple task such as correcting utterance segmentations that are already largely correct requires 1–3 time real time (Goldman, 2011). 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 each everyday conversation?

This 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 (Novotney & Callison-Burch, 2010), crowdsourcing means that we relinquish control over the situation in which the annotation is made.

To achieve increased throughput while maintaining a high measure of experimental control, we merge the ideas behind the Rapid Prosody Annotation pioneered by Jenifer Cole (see Mo, 2010) with a technical solution borrowed from the Audience Response Systems used in the film and television industries with the goal of having laymen annotators annotate some events in real time. Apart from being fast and cost effective, it places annotators in a situation similar how they perceive speech on an everyday basis.

We have previously investigated the use web based ARS-based methods for perception experiments (Edlund et al., 2012) and evaluation of children's speech (Strömbergsson & Tånnander, submitted). The present system has been used for speech synthesis evaluation (Tånnander et al., submitted), and is currently being tested for annotation of the type of speech events that are sometimes describes as disfluencies in the literature (Edlund et al., submitted).

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.

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

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).

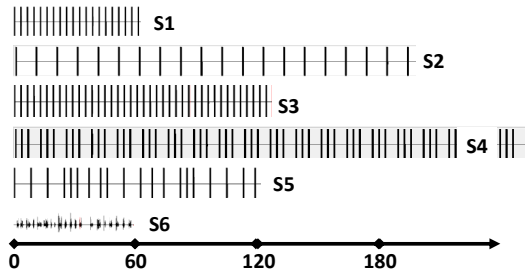

Figure 2. The waveforms of the six stimuli sets S1 – S6 (S6 is not included in the study).

Method

The purpose of this study to provide baseline statistics that show what can be expected of such a system. In order to achieve this goal, we ran a series of tests with stimuli that are significantly clearer than most speech phenomena, allowing us to measure with accuracy the temporal precision of the system (for such stimuli).

We used 4 groups of subjects, 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 (*Figure 2*) all consist of 1 second long beeps at 170 Hz, but their temporal patterns vary, as do the task of the annotators. The first series 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 a larger interval: 11.9 seconds. 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, there is an 8 second spacing corresponding to the duration an extra beep would have taken up: 3.5 seconds for the first spacing, 1 second for the beep, and 3.5 seconds for the second spacing. S5 holds 20 beeps spaced irregularly, at random intervals of up to 10 seconds.

The subjects were presented with the stimuli sets one after the other. 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, starting with a click of their own choice (the first or the second). For S4, they were asked to click where the left out fourth beep in every series of three *should have been*. In other words, they were asked to click at something that was lacking in the stimuli, rather than present.

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.

Overall, the subjects performed well and did what they were asked to do: subjects produced on average 1 click per occasion.

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. Average response times (ms) for all subjects over stimuli types, with standard deviation, counts, and significance at 4 % 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

Table 1 shows the mean response times to stimuli types. The differences are significant (one-way ANOVA, $p < 0.0001$).

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 this analyses using Kernel Density Estimation (KDE) estimations, which produces an analysis that is similar to a histogram, but produces a continuous, smooth curve.

The stimuli sets S1 and S3 (both with a predictable 3.5 second interval, with the task of clicking only every second beep for S3) show similar distributions. S2 (a predictable 11.7 second interval) has a wider distribution, S4 has a very different distribution spanning well over 2 seconds. S5 shows the most narrow distribution of all.

In the remainder of this analysis, we use reaction time estimates acquired by finding the peak of these 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 group 1 (8 subjects), which is then held out from the testing.

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. The resulting curve for S1 along with the wave form of the stimuli is shown in *Figure 3*.

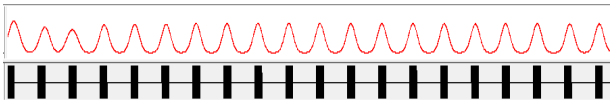

Figure 3. Overview of the waveform of S1 (below) with the KDE estimate based on all all subjects (above).

We estimate the onset of a trigger by finding a peak P in the KDE curve for a stimulus and deducting the RT estimate from that stimulus from the peak position.

Table 2. The mean error in ms (the mean of the absolute difference between actual trigger positions and estimated trigger positions) for each stimuli set.

	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	-

Table 2 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.

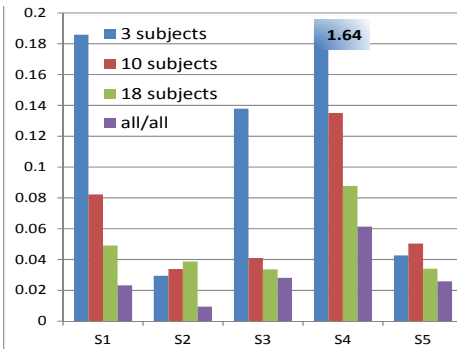

Figure 4. Average errors for S1-S6 for trigger estimates based on the clicks of 3, 10 and 18 subjects and RT estimates based on 8 different subjects, plus estimates based on clicks and RT times from all subjects. Note that two bars are truncated.

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 (G_4, 3 subjects), group 3 and 4 (G_{34}, 10 subjects) and group 2, 3, and 4 (G_{234}, 18 subjects) to estimate trigger times for each trigger in each stimuli set.

Figure 4 shows the mean error for each stimulus set for trigger estimates based on 3, 10, and 18 subjects (all disjoint from G1, the subjects used to estimate RT for each stimuli type), and also for all 28 subjects, using RT estimates from the same 28 subjects.

One-way ANOVAs 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).

Future work

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. Optimally, we would like to be able to run groups of 16 or 24 subjects. Further technical development includes making use of the feedback systems provided in the controllers: they are able to vibrate and have a small number of LEDs that could be used for this.

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.

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).

References

Edlund, J., C. Hjalmarsson & C. Tånnander. 2012. Unconventional methods in perception experiments. In *Proc. of Nordic Prosody XI*. Tartu, Estonia.

Edlund, J., J. Beskow, K. Elenius, K. Hellmer, S. Strömbergsson & D. House. 2010. Spontal: a Swedish spontaneous dialogue corpus of audio, video and motion capture. In: N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, M. Rosner & D. Tapias (eds.), *Proc. of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valetta, Malta, 2992–2995.

Edlund, J., J. Gustafson, M. Heldner & A. Hjalmarsson. 2008. Towards human-like spoken dialogue systems. *Speech Communication, 50*(8–9):630–645.

Edlund, J., S. Strömbergsson & J. Gustafson. Submitted. Audience response system-based annotation of conversational speech phenomena. Submitted to *Proc. of DiSS 2013*. Stockholm.

Goldman, J.-P. 2011. EasyAlign: a friendly automatic phonetic alignment tool under Praat. In: *Proceedings of Interspeech 2011*. Florence, Italy, Ses1-S3:2.

McCowan, I., J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, W. Post, D. Reidsma & P. Wellner. 2005. The AMI Meeting Corpus. In *Proc. of Measuring Behavior 2005, the 5th International Conference on Methods and Techniques in Behavioral Research*. Wageningen, Netherlands.

Mo, Y. 2010. *Prosody production and perception with conversational speech*. Doctoral dissertation, University of Illinois at Urbana-Champaign.

Novotney, S. & C. Callison-Burch. 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)*, 207–205.

Oertel, C., F. Cummins, N. Campbell, J. Edlund & P. Wagner. 2010. D64: A corpus of richly recorded conversational interaction. In: M. Kipp, J.-C. Martin, P. Paggio & D. Heylen (eds.), *Proceedings of LREC 2010 Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*, Valetta, Malta, 27–30.

Strömbergsson, S. & C. Tånnander. 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.

Tånnander, C., J. Edlund & J. Gustafson. Submitted. Audience response system-based evaluation of speech synthesis. Submitted to *Proc. of Interspeech 2013*. Lyon, France.

Proceedings of Fonetik 2013

The XXVIth Annual Phonetics Meeting 12–13 June 2013, Linköping University Linköping, Sweden

Studies in Language and Culture
no. 21

Robert Eklund, editor

Linköping University

Conference website: www.liu.se/ikk/fonetik2013
Proceedings also available at: http://roberteklund.info/conferences/fonetik2013

Cover design and photographs by Robert Eklund
Photo of Claes-Christian Elert taken by Eva Strangert on the occasion of his 80th birthday

Proceedings of Fonetik 2013, the XXVIth Swedish Phonetics Conference
held at Linköping University, 12–13 June 2013
Studies in Language and Culture, no. 21
Editor: Robert Eklund
Department of Culture and Communication
Linköping University
SE-581 83 Linköping, Sweden

ISBN 978-91-7519-582-7
eISBN 978-91-7519-579-7
ISSN 1403-2570

Printed by LiU-Tryck, Linköping, Sweden, 2013