Proc. INTERSPEECH 2005 Lisbon, 5-9 September (2005).

Results from a Survey of Attendees at ASRU 1997 and 2003

Roger K. Moore

Department of Computer Science
University of Sheffield, Regent Court, 211 Portobello Street, Sheffield, S1 4DP, UK
r.k.moore@dcs.shef.ac.uk

Abstract

In 1997 the author conducted a survey at the IEEE workshop on 'Automatic Speech Recognition and Understanding' (ASRU) in which attendees were offered a set of twelve putative future events to which they were asked to assign a date. Six years later at ASRU'2003, the author repeated the survey with the addition of eight additional items. This paper presents the combined results from both surveys.

1. Introduction

For over 20 years, the IEEE has organised a biennial workshop covering the latest developments in automatic speech recognition and attended by the leading researchers in the field. Many of these meetings have been of high significance; for example, it was at the 1985 workshop entitled 'Frontiers of Speech Recognition' that Fred Jelinek uttered the now immortal phrase *"Every time we fire a phonetician/linguist, the performance of our system goes up"*. By 1995 the series had become known as 'ASRU' - the IEEE workshop on Automatic Speech Recognition and Understanding.

In 1997 the author was involved in the organisation of the ASRU meeting scheduled to take place at Santa Barbara, USA. 1997 had been an exciting year; the US-DARPA and EU-Framework programmes were sponsoring significant research projects, and in April the application world saw the surprise release of the first commercial continuous speech dictation system - Dragon *Naturally Speaking* - closely followed by IBM's competing *ViaVoice* product in June that year. Therefore, in consultation with the other ASRU organisers, it was decided that it would be timely to conduct a survey of workshop attendees in order to gain an insight into the possible future of speech technology.

The 1997 survey was entitled 'Prospects for the Next Millennium', and its construction was interestingly different from previous surveys [1] in that, rather than asking attendees simply to speculate on the future, they were instead offered a set of putative future events to which they were asked to assign a date. This approach meant that it was possible to construct distributions of the responses, and thereby derive useful information such as the mean responses and also the minimum and maximum expected dates associated with each possible event. The results were compiled during the course of the meeting, and the author presented a summary at a special interactive plenary session. However, following much discussion about the possible negative impact of the results on potential funding agencies, it was agreed that survey should not be published in the open literature at the time.

In 2003, the author was contacted by the organisers of the ASRU scheduled to take place in the US Virgin Islands to see if it would be possible to conduct a similar survey – six years on (with the intervening years including significant events such as the 'dot.com' boom/bust and the rise and fall of Lernout & Hauspie). A second survey thus offered the opportunity, not only to compare and contrast two sets of responses on the same set of statements, but also to add new ones. Again the results were compiled during the course of the meeting, and the author presented a summary at a special plenary session under the title 'Speculating on the Future for Automatic Speech Recognition'. A copy of the presentation is lodged on the ASRU'2005 website [2], but the results have not appeared in the open literature – until now.

Therefore, this paper presents a formal record of the results of the 1997 and 2003 surveys of attendees at the IEEE ASRU workshops. It lists the statements used in each survey, and summarises the responses derived from the attendees.

2. The 1997 Survey

Attendees at the 1997 ASRU workshop were presented with a sheet containing twelve statements and the following instruction: *'Insert the year in which you estimate the statement will become true (use "X" to indicate "never")'*. The twelve statements were as follows:

1. More than 50% of new PCs have dictation on them, either at purchase or shortly after.
2. Most telephone Interactive Voice Response systems accept speech input (and more than just digits).
3. TV closed captioning is automatic and pervasive.
4. Voice recognition is commonly available at home (e.g. interactive TV, control of home appliances and home management systems).
5. Automatic airline reservation by voice over the telephone is the norm.
6. It is possible to hold a telephone conversation with an automatic chat-line system for more than 10 minutes without realising it isn't human.
7. Voice-enabled command, control and communication in cars becomes as common as intermittent wiper, power window or power door lock.
8. No more need for speech research.
9. A leading cause of time away from work is being hoarse from talking all the time, and people buy keyboards as an alternative to speaking.
10. Public proceedings (e.g. courts, public inquiries, parliament etc.) are transcribed automatically.
11. First legal case in which a recording of a person's voice is thrown out because it cannot be proved whether a computer or a person said it.
12. Speech recognition accuracy equals that of the average (individual) human transcriber.

Although the survey was anonymous, respondents were also asked to indicate if they were willing to comment on their

answers in the open discussion, in which case they were instructed to provide their name.

The 1997 plenary discussion was recorded.

3. The 2003 Survey

In the 2003 survey, the first twelve statements were *exactly* the same as those posed to the ASRU participants six years earlier. A further eight brought the total up to twenty. The additional statements were either suggested by the ASRU'2003 Technical Committee or derived from the predictions made by Ray Kurzweil [3][4]. The additional eight statements were as follows:

13. The majority of text is created using continuous speech recognition.
14. The majority of automatic speech recognition systems have completely abandoned the n-grams paradigm for language modelling.
15. Telephones are answered by an intelligent answering machine that converses with the calling party to determine the nature and priority of the call.
16. The majority of automatic speech recognition systems have completely abandoned the HMM paradigm for acoustic modelling.
17. Most routine business transactions take place between a human and a virtual personality (including an animated visual presence that looks like a human face).
18. Translating telephones allow two people across the globe to speak to each other even if they do not speak the same language.
19. Most interaction with computing is through gestures and two-way natural-language spoken communication.
20. Pocket-sized listening machines are commonly available for the hearing impaired.

Respondents were again asked if they would be prepared comment on your responses in an open discussion.

4. Results of the Surveys

4.1. Overall Statistics

The overall statistics for both surveys are shown in Table 1. As can be seen, there is remarkable agreement in terms of the number of attendees, the proportion of forms returned and – perhaps more surprisingly – the overall mean (on the twelve common statements). In both surveys, just under one half of the attendees provided responses. The overall number of "never" responses showed a small increase from 1997 to 2003, but for some reason the number of people prepared to comment on their views decreased by a factor of four.

Table 1: Overall statistics.

	2003	**1997**
No. of attendees	222	180
% of forms returned	47%	45%
Overall mean year	2055	2056
% of "Never" responses	24%	17%
No. of attributed responses	4	18

Looking more closely at the overall position on the dates, Figure 1 shows the distribution of the average responses from all the respondents. From this it can be seen that the peak response in 1997 was between "2005" and "2010", whereas in 2003 it was between "2010" and "2015" – i.e. about 10 years on from the date of the survey *in both cases*. This is an effect observed by Ken Church in his Eurospeech'2003 Plenary talk.

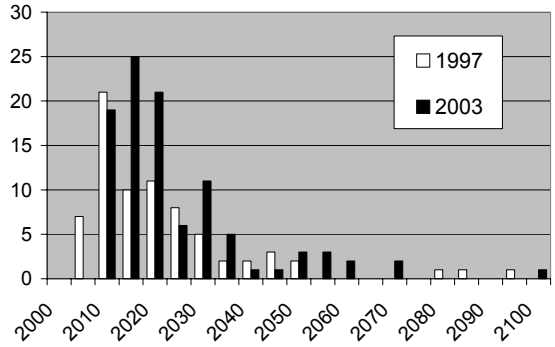

Figure 1: The distribution of the average responses over all the respondents.

Figure 2 shows the mean responses for each individual statement. This distribution is clearly dominated by the responses for statement 8 – *"No more need for speech research"*. The response with the earliest mean date is statement 1 - *"More than 50% of new PCs have dictation on them, either at purchase or shortly after"* - and this was the case for both the 1997 and 2003 surveys. Clearly the statements that were introduced in 2003 were judged as referring to events that were likely to be much further in the future than the bulk of the statements in the original set.

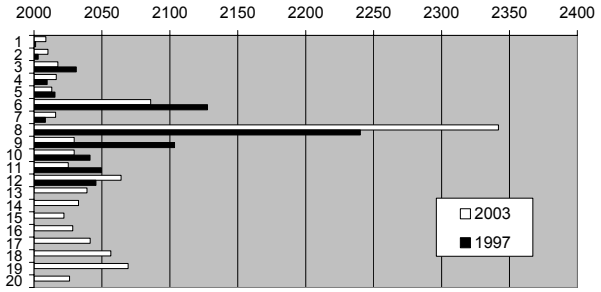

Figure 2: The distribution of the average responses for each individual statement.

The overall result for each individual statement is presented in Table 2. The table gives the mean year, the standard deviation (in years), the minimum date and the maximum date for both the 1997 and 2003 surveys. The percentage of "never" responses is also indicated.

Of particular interest in Table 2 is the column showing the 'minimum' dates. In most cases, these indicate that at least one respondent considered that the events referred to had either occurred already or were soon to take place (the maximum minimum date was 2006!). On several statements, this contrasted sharply with nearly half of the respondents believing that the same events would *never* take place.

Table 2: Statistics for each individual statement.

	Year	Mean	SD	Min	Max	Never
1	*1997*	2001	3	1997	2010	0%
	2003	2009	7	2000	2050	15%
2	*1997*	2003	4	1998	2020	3%
	2003	2010	10	2000	2060	2%
3	*1997*	2031	124	1997	3001	8%
	2003	2018	17	1998	2100	8%
4	*1997*	2010	12	1999	2100	4%
	2003	2016	15	2004	2100	5%
5	*1997*	2015	57	1999	2500	5%
	2003	2013	10	2002	2050	14%
6	*1997*	2128	328	1998	4001	30%
	2003	2086	228	2000	3579	34%
7	*1997*	2008	8	1999	2050	8%
	2003	2016	13	2004	2075	9%
8	*1997*	2240	546	1984	5001	53%
	2003	2342	1308	1981	10K	62%
9	*1997*	2103	287	1998	3020	68%
	2003	2030	31	2006	2150	79%
10	*1997*	2041	128	2000	3001	6%
	2003	2025	26	2006	2150	4%
11	*1997*	2050	167	1990	3000	8%
	2003	2064	29	1995	2150	19%
12	*1997*	2046	124	1997	3001	9%
	2003	2039	222	2005	3827	19%
13	*2003*	2033	48	2000	2300	47%
14	*2003*	2022	39	1995	2200	47%
15	*2003*	2022	25	2000	2150	10%
16	*2003*	2029	33	2005	2200	41%
17	*2003*	2041	66	1994	2500	25%
18	*2003*	2057	116	2000	3000	6%
19	*2003*	2069	225	2004	3827	37%
20	*2003*	2026	32	2001	2275	3%

4.2. Selected Individual Results

It is not possible to discuss the responses to each individual statement in the limited space available here. However, it is feasible to give an idea of the type of analysis that has been performed. For example, Figure 3 shows the distribution of responses for the statement *"More than 50% of new PCs have dictation on them, either at purchase or shortly after"*. The number of "never" responses is indicated by the hashed column at the extreme right-hand end of the horizontal time axis. (The falling hash pattern indicates the response from 2003; rising hash patterns in subsequent Figures indicate the response from 1997.)

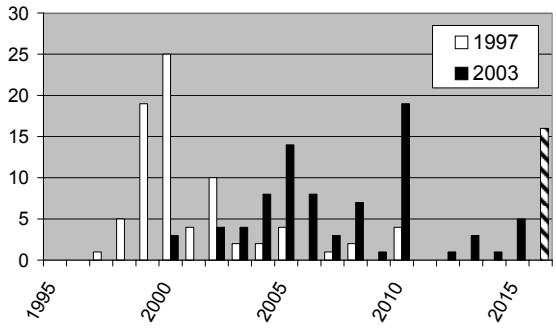

Figure 3: Responses to the statement *"More than 50% of new PCs have dictation on them, either at purchase or shortly after"*.

Clearly Figure 3 illustrates that the average date by which ASRU attendees thought that this statement would become true has shifted into the future by about eight years, and this is confirmed by the statistics shown in Table 2. Table 2 also indicates that not only had there been some increase in uncertainty over time, but a significant number of attendees in 2003 thought that this event would never happen (no-one responded with "never" in 1997).

Discussion of this statement at the 1997 ASRU meeting revealed that many people firmly believed that dictation would become available on most PCs, but there was considerable speculation about how many people would actually choose to use it.

Figure 4 shows the distribution of responses for the statement *"Most telephone Interactive Voice Response systems accept speech input (and more than just digits)"*. These responses show a very similar pattern to the statement above. However, unlike PC-based dictation, very few people thought that speech-enabled IVR would never happen.

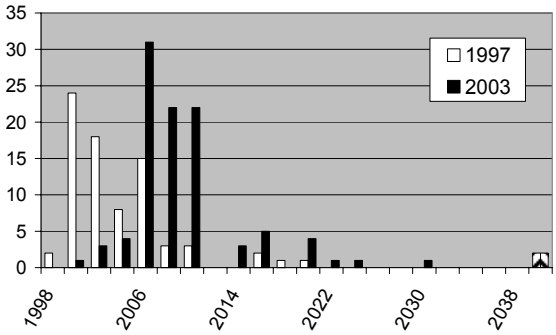

Figure 4: Responses to the statement *"Most telephone Interactive Voice Response systems accept speech input (and more than just digits)"*.

Discussion at the 1997 meeting included the observation that this application would be most likely to happen because of business pressure to lower costs, and that this would lead to the possibility that users might not be given a choice about whether they would use such systems.

A completely different result emerges from the distribution of responses for the statement *"TV closed captioning is automatic and pervasive"* (shown in Figure 5). In this case the average date has come forward from thirty years away to only fifteen, and the confidence has increased significantly (see Table 2).

Although more optimistic, the 2003 responses nevertheless still reflect an appreciation that subtitling involves significant challenges over and above pure transcription.

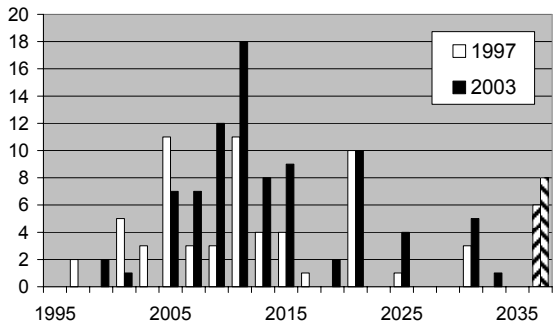

Figure 5: Responses to the statement *"TV closed captioning is automatic and pervasive"*.

Figure 6 shows the distribution of responses for the statement *"Translating telephones allow two people across the globe to speak to each other even if they do not speak the same language"*. This prediction from Ray Kurweil raised considerable interest at ASRU'2003 given the existence of a number of international collaborative research projects on this topic. As can be seen, responses were quite widely distributed, with only a few people thinking that it would never happen.

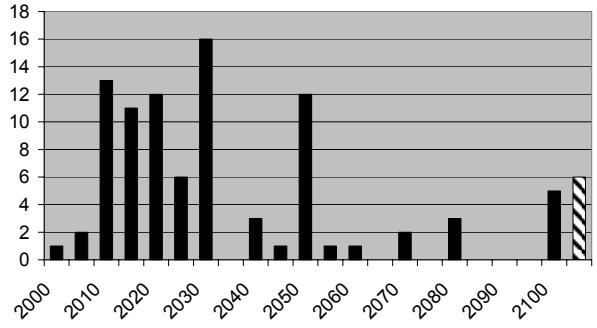

Figure 6: Responses to the statement *"Translating telephones allow two people across the globe to speak to each other even if they do not speak the same language"*.

Finally, the issue that aroused most interest in both 1997 and 2003 was statement *"No more need for speech research"*. Figure 7 illustrates the responses, which are pretty much as one would expect. What was controversial - particularly in 1997 - was the fact that someone had given "1984" as a response (see Table 2), i.e. they judged that there had already been *too much* speech research!

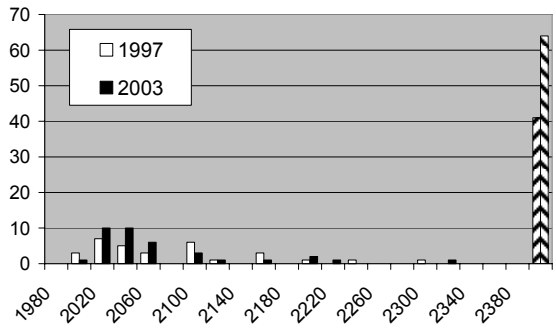

Figure 7: Responses to the statement *"No more need for speech research"*.

5. Conclusions

As Niels Bohr said in 1922 - *"It is hard to predict ... especially the future"*. Nevertheless, the results of these two ASRU surveys do give some insight into the degree of difficulty and chances for success in the field of speech technology, as seen through the eyes of the R&D community.

Overall the results are remarkably consistent between the two surveys. There is strong evidence of psychological quantisation effects and for putting most hard challenges a fixed distance into the future. The 2003 survey was neither more optimistic nor more pessimistic than the 1997 survey. However there was more agreement in 2003, and the responses could be argued to be more realistic. In contrast, people seemed less willing to be associated with their opinions in 2003 than they did six years earlier.

Finally, several respondents took the opportunity to suggest their own statements, for example *"Computers do their own research in human speech recognition by building better humans"*, and one person answered *all* the statements in the 1997 survey with the tongue-in-cheek yet insightful response *"... in 10 years (valid 25 years ago, and will remain valid for the next 25 years)"*!

6. Acknowledgements

The author would like to thank all of the people who participated in the 1997 and 2003 surveys, the organisers of ASRU'97 and ASRU'03, and ELSNET (the European Network of Excellence in Human Language Technologies) who sponsored the author's attendance at ASRU'03 to conduct the survey as part of their 'HLT Roadmap' action.

7. References

[1] Moore, R K., "Twenty Things We Still Don't Know About Speech", *Progress and Prospects of Speech Research and Technology*, H. Niemann and R. de Mori (Eds.), Infix, Germany, 1994.

[2] *IEEE ASRU 2003 Automatic Speech Recognition and Understanding Workshop*, St. Thomas, U.S. Virgin Islands, http://www.asru2003.org/.

[3] Kurzweil, R., *The Age of Intelligent Machines*, MIT Press, 1990.

[4] Kurzweil, R., *The Age of Spiritual Machines*, Phoenix Press, 1999.