Progress and Prospects for Speech Technology: Results from Three Sexennial Surveys

Roger K. Moore

Speech and Hearing Group, Dept. Computer Science, University of Sheffield, UK
r.k.moore@dcs.shef.ac.uk

Abstract

In 1997, and again in 2003, the author was invited to conduct a survey at the IEEE workshop on 'Automatic Speech Recognition and Understanding' (ASRU) in which attendees were offered a set of statements about putative future events relating to progress in various aspects of speech technology R&D. The task of the respondents was to assign a date to each possible event. The 1997 and 2003 results were published at INTERSPEECH 2005 in Lisbon. Six years later, the author was invited by the organisers of ASRU'2009 to repeat the survey for a third time, and this paper presents the combined results from all three 1997, 2003 and 2009 surveys. The overall conclusion is that, over the twelve year period progress is perceived as slow, and the future appears to be generally no nearer than it has been in the past. However, on a positive note, the 2009 survey confirmed that the market for speech technology applications on mobile devices would be highly attractive over the next ten or so years.

Index Terms: speech recognition, speech synthesis, survey of progress, future predictions

1. Introduction

Since the early 1980s, the IEEE has organised a biennial workshop covering the latest developments in automatic speech recognition and attended by the leading researchers in the field. 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. 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 conducted by the author [1] in that, rather than asking attendees simply to speculate on the future, they were instead offered a set of statements describing 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, minimum and maximum dates associated with each statement. The results were compiled during the course of the meeting, and the author presented a summary at a special interactive plenary session. Overall, the 1997 results were fairly negative. So, following much discussion about the possible impact on potential funding agencies, it was agreed that the outcome of the 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 U.S. Virgin Islands to see if it would be possible to conduct a similar survey – six years on. A second survey 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 from attendees 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*'. Overall, the 2003 survey concluded that the 2003 respondents were neither more optimistic nor more pessimistic than the 1997 respondents, that there was more agreement in 2003 than in 1997, but that people seemed less willing to be associated with their opinions in 2003 than they did in 1997. The combined results of the 1997 and 2003 surveys were published at INTERSPEECH in 2005 [2],

In 2009, the author was once more contacted by the organisers of ASRU (this time scheduled to take place in Merano, Italy) and was requested to conduct a third - by now 'sexennial' - survey. On this occasion the survey was conducted on-line and in advance, and the author presented a verbal summary of the outcome at the workshop under the title '*Progress and Prospects for Speech Technology*'.

This paper presents a formal record of the results of all three surveys from the 1997, 2003 and 2009 IEEE ASRU workshops, and draws conclusions spanning the twelve year period.

2. The Three Surveys

2.1. 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.*

2.2. The 2003 survey

In the 2003 survey, the first twelve statements were identical to those presented to the 1997 ASRU participants. However, the opportunity was taken to add a further eight statements, either from suggestions provided by the ASRU'2003 Technical Committee or derived from predictions made by Ray Kurzweil in his two '*The Age of …*' books [3][4]. The eight additional 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.*

2.3. The 2009 survey

In the 2009 survey, the first twenty statements were identical to those posed to the 2003 ASRU participants. This meant that the first twelve statements were identical to those posed to both the 2003 and 1997 ASRU participants. A further six statements (which primarily related to mobile devices and applications) were suggested by the ASRU'2009 Technical Committee. The six additional statements were as follows:

21. *Most information access and search using mobile phones are done through speech recognition and synthesis (e.g., web search, SMS).*
22. *Mobile phones are used to control and monitor home appliances remotely using speech (e.g., remote access to DVR, recording programs, TV).*
23. *Most multilingual people communicate with each other through speech to speech translation at any time using their mobile device.*
24. *Number of speech-enabled applications created within the mobile ecosystem (e.g., Apple store, RIM, Android, etc) reaches 1 million.*
25. *Mobile speech applications generate a $10 billion in revenue.*
26. *All mobile devices have built-in speech recognition capability.*

In addition to the twenty-six statements to which respondents were asked to provide a date, the 2009 on-line survey also posed a small number of questions relating to the individuals themselves. For example, respondents were asked to state whether their responses related to a specific language market or to a specific geographic area, whether they had participated in the previous surveys, how many years they had been involved in speech technology, their career status, and to which professional organisation (if any) they belonged.

3. Results

Although information about the respondents was only available from the 2009 survey, it is interesting to see the profile that emerged. For example, of the 127 people who took part, 83% were members of ISCA and 54% were members of the IEEE. Rather surprisingly, only 5% of the 2009 respondents had participated in the earlier surveys. Unsurprisingly, the majority of respondents were based at universities and considered themselves to be researchers; 35% were students, but 55% had been employed in a company.

Of particular interest in the 2009 survey was the possibility of correlating each individual's median response with the length of time they had spent in the speech technology field (ranging from a few months to over 40 years). The hypothesis was that the younger respondents might be the most optimistic, and that the older respondents might be the most pessimistic. However, the rather surprising outcome was that there was *no* correlation between the two (ρ = -0.005). Some of the younger respondents were quite pessimistic, and the longest-standing members of the speech technology community were revealed to be among the most optimistic.

3.1. Overall responses

The overall statistics for the three surveys (based on responses to the first twelve statements) are shown in Table 1. As can be seen, the number of respondents has grown steadily, but the willingness of individuals to be associated with their opinions exhibits a significant dip in 2003 (possibly as a direct consequence of the post-'dot.com' boom/bust period).

In terms of overall progress in the field, if it were perceived as being steady and continuous, then it would be expected that the average responses to the twelve statements that were common to all three surveys would be approximately the same. However, as can be seen in Table 1, the median values are consistently eight to ten years apart. This suggests that progress is perceived by the community as being somewhat slower than real-time.

Table 1. *Overall statistics from the three surveys (based on responses to the first 12 statements).*

	1997	**2003**	**2009**
Respondents:	81	105	127
Overall Median:	2010	2020	2028
"*Never*"s:	17%	22%	22%
Named responses:	22%	4%	21%

The distributions of responses over the twelve common statements are presented in Figure 1. As can be seen, each of the three distributions is somewhat Poisson shaped and, unsurprisingly, there is a clear tendency for the year 2050 to be a popular choice in all three surveys (a natural quantisation effect that can be interpreted as "*sometime in the distant future*"). Overall, the three distributions are similar in shape, but shifted in time with respect to each other.

The columns at the right-hand end of Figure 1 correspond to the response "*never*" which, for the twelve common statements, is consistently at a reasonably low level compared to the dated responses.

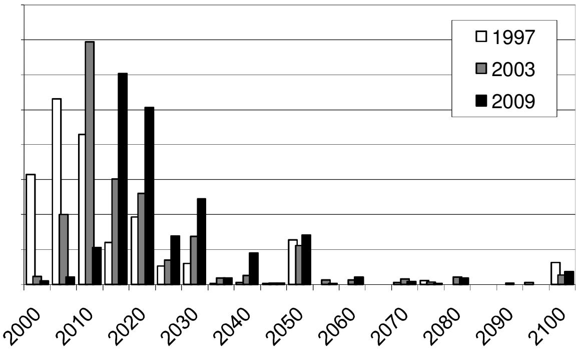

Figure 1. *Distribution of the average responses over all respondents (based on responses to the first 12 statements).*

3.2. Responses for the first 12 statements

The detailed numerical results for the first twelve statements (which appeared in all three 1997, 2003 and 2009 surveys) are presented in Table 2.

Table 2. *Statistics for the first twelve statements.*

	Year	Median	SD	Min	Max	Never
1	*1997*	2000	3	1997	2010	0%
	2003	2010	7	2000	2050	15%
	2009	2015	7	2000	2050	6%
2	*1997*	2002	4	1998	2020	3%
	2003	2008	10	2000	2060	2%
	2009	2015	109	2001	3220	2%
3	*1997*	2010	124	1997	3001	8%
	2003	2012	17	1998	2100	8%
	2009	2020	12	2000	2080	13%
4	*1997*	2007	12	1999	2100	4%
	2003	2011	15	2004	2100	5%
	2009	2020	11	2010	2070	10%
5	*1997*	2007	57	1999	2500	5%
	2003	2010	10	2002	2050	14%
	2009	2022	10	2001	2080	37%
6	*1997*	2050	328	1998	4001	30%
	2003	2050	228	2000	3579	34%
	2009	2050	110	2010	3000	36%
7	*1997*	2007	8	1999	2050	8%
	2003	2012	13	2004	2075	9%
	2009	2020	93	2009	3000	13%
8	*1997*	“never”	546	1984	5001	53%
	2003	“never”	1308	1981	10K	62%
	2009	“never”	93	2009	3000	79%
9	*1997*	“never”	287	1998	3020	68%
	2003	“never”	31	2006	2150	79%
	2009	“never”	185	1990	2080	85%
10	*1997*	2020	128	2000	3001	6%
	2003	2020	26	2006	2150	4%
	2009	2030	98	2009	3000	16%
11	*1997*	2020	167	1990	3000	8%
	2003	2020	29	1995	2150	19%
	2009	2025	13	2000	2080	18%
12	*1997*	2020	124	1997	3001	9%
	2003	2030	222	2005	3827	19%
	2009	2035	310	2010	5000	19%

Table 2 gives the median year, the standard deviation (in years), the minimum date, the maximum date and the percentage of “*never*” responses. The median was chosen (rather than the average) in order to take into account the number of “*never*” responses.

The general trend observed in Table 2 reflects the overall forward shift illustrated in Figure 1. The single exception to this is statement #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”*) which, probably by virtue of it being judged to be some way in the future, seems to be fairly stable.

Of the others, several (e.g. statements #5, #8, #9 and #10) show an increasing tendency to the response “*never*”. In the case of statement #5 (“*Automatic airline reservation by voice over the telephone is the norm*”) - a popular objective for speech technology research in the 1990s - such a judgement is almost certainly being made on the basis of the growth of easy-to-use web interfaces for travel planning and booking.

Statement #9 (“*A leading cause of time away from work is being hoarse from talking all the time …*”) is probably judged as being rather ‘tongue-in-cheek’, and the extreme accuracy requirements for statement #10 (“*Public proceedings … are transcribed automatically*”) are probably perceived as being out of reach.

3.3. Responses for the eight intermediate statements

The overall results for the eight intermediate statements (which appeared in the 2003 and 2009 surveys) are presented in Table 3.

Table 3. *Statistics for the eight intermediate statements.*

	Year	Median	SD	Min	Max	Never
13	*2003*	2100	48	2000	2300	47%
	2009	“never”	184	2010	3000	56%
14	*2003*	2100	39	1995	2200	47%
	2009	2045	160	2009	3009	35%
15	*2003*	2015	25	2000	2150	10%
	2009	2020	93	2004	3000	8%
16	*2003*	2040	33	2005	2200	41%
	2009	2033	150	2013	3009	29%
17	*2003*	2043	66	1994	2500	25%
	2009	2060	118	2012	3000	44%
18	*2003*	2030	116	2000	3000	6%
	2009	2040	94	2009	3000	11%
19	*2003*	2053	225	2004	3827	37%
	2009	2100	25	1960	2140	48%
20	*2003*	2020	32	2001	2275	3%
	2009	2020	31	2009	2300	2%

Of these, the two technical statements #14 (“*The majority of automatic speech recognition systems have completely abandoned the n-grams paradigm for language modelling*”) and #16 (“*The majority of automatic speech recognition systems have completely abandoned the HMM paradigm for acoustic modelling*”) show increasing support, while statement #17 (“*Most routine business transactions take place between a human and a virtual personality …*”) appears to be judged as being increasingly unlikely.

The responses for statement #20 (“*Pocket-sized listening machines are commonly available for the hearing impaired*”) are quite stable and indicate a strong likelihood that such devices will be realised in about ten year’s time.

3.4. Responses for the six final statements

The overall results for the final six statements (which appeared for the first time in the 2009 survey) are presented in Table 4.

Table 4. *Statistics for the six final statements.*

	Year	Median	SD	Min	Max	Never
21	*2009*	2025	144	2010	3000	26%
22	*2009*	2020	11	2009	2090	15%
23	*2009*	2060	114	2014	3000	40%
24	*2009*	2020	16	2009	2100	6%
25	*2009*	2020	17	2010	2105	8%
26	*2009*	2019	15	2000	2100	11%

As stated in Section 2.3 above, the six statements added to the 2009 survey were all concerned with mobile/portable devices and applications. Looking at the results shown in Table 4, it is interesting to note that statements #24 ("*Number of speech-enabled applications created within the mobile ecosystem … reaches 1 million*") and #26 ("*All mobile devices have built-in speech recognition capability*") elicited positive responses, whereas statement #23 ("*Most multilingual people communicate with each other through speech to speech translation at any time using their mobile device*") was judged to be the most unlikely to take place.

4. Discussion

Analysing the data overall, it is possible to observe some interesting and relevant trends. For example, the statements that were judged to be most *likely* to come to pass were statement #1 ("*More than 50% of new PCs have dictation on them …*") and statement #2 ("*Most telephone Interactive Voice Response systems accept speech input …*"). Indeed, it can be argued that both of these statements are probably already true; for example, automatic speech recognition has been part of the Windows OS on PCs since the launch of Vista in 2007, and utility companies, cinemas and other service providers have deployed quite sophisticated IVR systems in recent years.

Of those statements which were judged to be most *unlikely* to take place, it is interesting to note that statement #13 ("*The majority of text is created using continuous speech recognition*") and statement #19 ("*Most interaction with computing is through gestures and two-way natural-language spoken communication*") were both predictions made by Ray Kurzweil [3][4]. Kurzweil suggested that these events would become true in 2009 and 2019 respectively, but the results of the surveys reported here indicate a large disagreement by the speech technology community who suggest that statement #13 falls into the category of "*never*" and statement #19 is judged to be possible, but about 100 years into the future.

Statement #18 ("*Translating telephones allow two people across the globe to speak to each other even if they do not speak the same language*") was another of Kurzweil's predictions. He proposed that this would be true by the early 2000s, whereas both the 2003 and 2009 surveys suggest that it *will* happen, but not for another twenty or thirty years.

Among the statements with the most stable responses (i.e. with no significant change in opinions between surveys) were statement #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*") and statement #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*"). Both of these reflect a view that there will come a time when the quality of speech technology will have reached a stage of development where confusion could arise between natural and automated systems.

Two statements stand out as having been judged to take longer than at first thought: statement #17 ("*Most routine business transactions take place between a human and a virtual personality …*") and statement #19 ("*Most interaction with computing is through gestures and two-way natural-language spoken communication*"). This would seem to suggest that the community is not convinced that intelligent speech-based interfaces using conversational agents/avatars will become as pervasive as the proponents of those technologies might hope.

5. Summary and Conclusion

This paper has presented a formal record of the results of three surveys conducted for the 1997, 2003 and 2009 IEEE ASRU workshops. Based on a set of 26 statements describing putative future events, an average of around 100 respondents drawn from the community of leading speech technology researchers provided their opinions on when such events might become true on three different occasions at six yearly intervals. The outcome was a set of distributions that give some insight into the community's view of progress and prospects in the speech technology field.

Overall, it can be concluded that the future appears to be no nearer than it has been in the past! While a few statements were judged to be likely to become true in the near future, the majority continue to be judged to be some way in the future.

Respondents seem to have recovered their willingness to be associated with their opinions, and it was interesting to discover that there is no correlation between an individual's optimism/pessimism and the length of time that they have spent in the speech technology field.

In terms of applications, it appears that speech technology on mobile devices was almost uniformly judged to be realisable in around 10 years time. However, the opinion on classic applications (such as dictating text) was that they might never happen. In terms of research, there appears to be increasing support for changing the current paradigms based on n-grams for language modelling (statement #14) and HMMs for acoustic modelling (statement #16) but, of course, the survey does not reveal what those new paradigms might be. Also, researchers will be encouraged to see a continuing increase in "*never*" responses for statement #8: "*No more need for speech research*"!

Finally, although these surveys appear to draw strong interest from the R&D community, they are somewhat limited in their measurement methodology. Perhaps this could be addressed in any future endeavour, e.g. for ASRU 2015!

6. Acknowledgements

The author would like to thank the organisers of the three ASRU workshops at which the surveys were carried out, and of course the individuals who agreed to participate on each occasion.

7. References

[1] R. K. Moore, "Twenty things we still don't know about speech," in *Progress and Prospects of Speech Research and Technology*, H. Niemann and R. deMori, Eds. Germany: Infix, 1994.

[2] R. K. Moore, "Results from a survey of attendees at ASRU 1997 and 2003," in *INTERSPEECH* Lisbon, 2005.

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

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