INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Progress and Prospects for Spoken Language Technology: Results from Four Sexennial Surveys

Roger K. Moore, Ricard Marxer

Speech and Hearing Research Group, Dept. Computer Science, University of Sheffield, UK
r.k.moore@sheffield.ac.uk, r.marxer@sheffield.ac.uk

Abstract

Since 1997, a survey has been conducted every six years at the IEEE workshop on *Automatic Speech Recognition and Understanding* (ASRU) in order to ascertain the research community's perspective on future progress and prospects in spoken language technology. These surveys have been based on a set of 'statements', each of which portray a possible future scenario, and respondents are asked to estimate the year in which each given scenario might become true. Many of the statements have appeared in several of the surveys, hence it is possible to track changes in opinion over time. This paper presents the combined results of all four surveys, the most recent of which was conducted at ASRU-2015. The results give an insight into the key trends that are taking place in the spoken language technology field, and reveal the realism that pervades the research community. They also suggest that there is growing confidence that some of the scenarios will indeed be realised at some point in the future.

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

1. Introduction

Every six year since 1997, a survey has been conducted at the IEEE workshop on *Automatic Speech Recognition and Understanding* (ASRU) in order to ascertain the research community's perspective on future progress and prospects in spoken language technology. Unlike surveys where respondents are asked to suggest possible future events, the ASRU surveys are based on a set of 'statements', each of which portray a possible scenario. Respondents are then asked to estimate the year in which each given scenario might become true (or respond with "*never*"). The advantage of this approach is that it is possible to perform a quantitative analyse of the responses and to determine useful summary measures such as the median, minimum and maximum dates associated with each statement. Also, since a subset of the statements have been the same on each occasion, it is possible to track whether the experts' opinions changed over time.

2. The Four Surveys

2.1. The 1997 survey

The first survey - entitled *Prospects for the Next Millennium* - was conducted at the 1997 ASRU workshop (held in Santa Barbara, USA). Attendees were presented with a sheet containing the following twelve statements:

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*

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 surprisingly negative. So, following a lively discussion about the possible impact on potential funding agencies, it was agreed that the outcomes from the survey should *not* be published in the open literature!

2.2. The 2003 survey

Six years later, it was felt that it would be appropriate to conduct a follow-up survey at ASRU-2003 (held in the US Virgin Islands). The Technical Committee decided that it would be interesting to supplement the original twelve statements with eight further statements that reflected contemporary issues. In particular, inspiration was taken from predictions made by Ray Kurzweil in his two '*The Age of ...*' books [1, 2] (marked with a '*' below). 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.**

http://dx.doi.org/10.21437/Interspeech.2016-948

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.**

On this occasion, the workshop attendees felt more secure in voicing their opinions, so the results of both the 1997 and 2003 surveys were published at INTERSPEECH-2005 [3]. Overall, it was found that the results were remarkably consistent between the two surveys although, on average, the distributions of responses had shifted six years into the future.

2.3. The 2009 survey

In 2009, the ASRU organising committee again felt that repeating the survey would provide a useful update on the research community's opinions. On this occasion the survey was conducted on-line and in advance, and the outcome was presented by the first author at the workshop (held in Merano, Italy). Six additional statements (primarily relating to mobile devices and applications) were included in the ASRU-2009 survey:

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.*

A comparison of the results from the 1997, 2003 and 2009 surveys was published at INTERSPEECH-2011 [4]. Overall, it was concluded that the future still appeared to be no nearer than it had been in the past. While a few statements were judged as likely to become true in the near term, the majority continued to be assessed as being some way off. For the statements relating to speech technology on mobile devices, the results suggested that they would be realisable around the year 2020. However, the consolidated opinion on classic applications (such as dictating text) was that they might never happen.

2.4. The 2015 survey

The ASRU-2015 workshop (held in Scottsdale, Arizona, USA) provided an opportunity to conduct the fourth 'sexennial' survey. Four further statements were added (primarily relating to robots and deep learning), bringing the total to thirty. The four additional statements were as follows:

27. *Conversational interaction with autonomous social agents (such as robots) is commonplace in the home.*
28. *Speech replaces text-based web search.*
29. *Spoken language technology can translate a voice from one language to another as well as a human interpreter.*
30. *DNNs replace all of the major components in a spoken language technology system.*

The results of the 2015 survey were compiled with those of the three previous surveys, and an analysis of all four is presented below.

3. Analysis of the Four Surveys

3.1. Overall results

The combined results for all four surveys (based on responses to the first twelve statements) are shown in Table 1. The main outcome is that the previously observed increasing trend for the overall median appears to have been reversed, and the scenarios represented by the first twelve statements are now judged, on average, to be ten (rather than twenty) years away. This, coupled with the reduction in "*Never*"s in 2015, suggests increasing confidence in the research community that some of these scenarios will indeed be realised at some point in the future.

Table 1: *Overall results from the four surveys (based on responses to statements 1-12).*

	1997	**2003**	**2009**	**2015**
No. of Respondants:	81	105	127	61
Overall Median:	2010	2020	2028	2025
Relative to Survey:	+13	+17	+19	+10
"*Never*"s:	17%	22%	28%	17%
Named Responses:	22%	4%	21%	11%

The results shown in Table 1 also reveal that the number of respondents willing to participate in the most recent survey was rather low in comparison to the previous surveys. This is surprising given (a) the record number of attendees at ASRU-2015, (b) the promotion of the survey through social media, and (c) the straightforward design of the on-line survey. Whether this indicates a reluctance on the part of the respondents to voice their opinion (as was the case in the 1997 survey), or simply the competing demands of contemporary working life, is hard to determine. On the other hand, many more respondents in the 2015 survey took the opportunity to provide insightful free-text comments to justify their judgements, and some of these are reported below.

Figure 1 shows the distribution of responses averaged over the first twelve statements. It can be seen that the overall shapes have a high degree of similarity (albeit shifted along the time axis), and there is a clear quantisation effect giving rise to a peak at the year 2050 across all four surveys.

3.2. Results for statements 1-12

The detailed responses for the first twelve statements are shown in Table 2. As can be seen, responses for some of the statements are reasonably stable over all four surveys. For example, statement #2 "*Most telephone Interactive Voice Response systems accept speech input*" is consistently judged as very likely

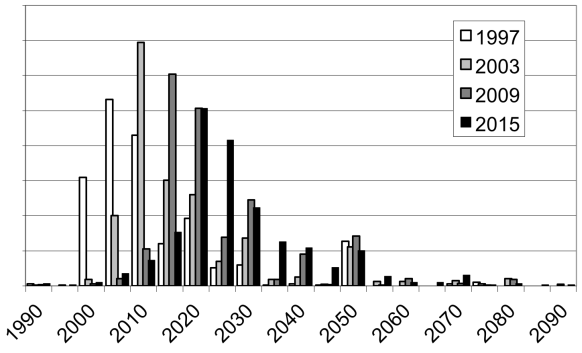

Figure 1: *Distribution of the average responses over all respondents (based on responses to statements 1-12).*

to become true in the next couple of years, and statement #8 "*No more need for speech research*" is consistently judged as being very unlikely (although one respondent noted that funding agencies appear to be adopting this view). Another statement judged reasonably consistently across all four surveys is #10 "*Public proceedings are transcribed automatically*". In this case, the application scenario is assessed as quite challenging due to the need for a high degree of reliability in the transcripts, but it is expected to become true in around 15 years time and, crucially, no-one in the 2015 survey thought that it would never happen.

A number of other interesting trends emerge from the results depicted in Table 2. For example, although the median dates for statements #3 "*TV closed captioning is automatic and pervasive*", #4 "*Voice recognition is commonly available at home*" and #7 "*Voice-enabled command, control and communication in cars becomes as common as . . .* " are slowly receding into the future, they are nevertheless judged as likely to happen once the relevant technical challenges (such as overlapping speech and competing noises) have been overcome. Indeed, several respondents cited the appearance of *Amazon Echo* as evidence that progress is being made. In contrast, the median date for statement #5 "*Automatic airline reservation by voice over the telephone is the norm*" is not only receding into the future, but respondents judged the chances of this taking place as less-and-less likely, citing web-based interfaces as a more user-friendly alternative for this particular task.

Two of the statements explicitly refer to spoken language technology achieving human-level performance - #6 "*It is possible to hold a telephone conversation with an automatic chatline system . . .* " and #12 "*Speech recognition accuracy equals that of the average human transcriber*". As one might expect, both have been consistently judged as challenging, with #6 seen as being more difficult to achieve than #12. However, both saw a dramatic drop in the proportion of "*Never*"s in the 2015 survey, with respondents noting that some companies are already claiming to have achieved human-level performance on particular tasks.

3.3. Results for statements 13-20

The detailed responses for the statements added in 2003 are shown in Table 3. Of particular interest in this group are those which were based on predictions made by Ray Kurzweil (see Section 2.2). Clearly, all the scenarios in this group are judged by the ASRU respondents as challenging (compared to the first twelve statements), and consistently much more challenging than as predicted by Kurzweil. For example, statement #13 "*The majority of text is created using continuous speech recognition*" was predicted by Kurzweil to have taken place by the year 2009. However, whilst the ASRU experts originally judged this to be something that might never happen, in the 2015 survey it was no longer viewed as impossible to achieve, but certainly very hard. One respondent noted that the problem is not only dependent on the accuracy of the automatic speech recognition, but that users are often not skilled in formulating their utterances clearly and precisely.

Table 2: *Results for statements 1-12.*

	Survey	**Median**	**Rel.**	**Min.**	**Max.**	**"*Never*"**
1	'1997'	2000	+3	1997	2010	0%
	'2003'	2010	+7	2000	2050	15%
	'2009'	2015	+6	2000	2050	6%
	'2015'	2016	+1	2002	2050	3%
2	'1997'	2002	+5	1998	2020	3%
	'2003'	2008	+5	2000	2060	2%
	'2009'	2015	+6	2001	3220	2%
	'2015'	2018	+3	1998	2040	2%
3	'1997'	2010	+3	1997	3001	8%
	'2003'	2012	+9	1998	2100	8%
	'2009'	2020	+11	2000	2080	13%
	'2015'	2023	+8	1984	2067	5%
4	'1997'	2007	+10	1999	2100	4%
	'2003'	2011	+8	2004	2100	5%
	'2009'	2020	+11	2010	2070	10%
	'2015'	2022	+7	2002	2060	6%
5	'1997'	2007	+10	1999	2500	5%
	'2003'	2010	+7	2002	2050	14%
	'2009'	2022	+13	2001	2080	37%
	'2015'	2032	+17	2003	2044	41%
6	'1997'	2050	+53	1998	4001	30%
	'2003'	2050	+47	2000	3579	34%
	'2009'	2050	+41	2010	3000	36%
	'2015'	2035	+20	2016	2150	5%
7	'1997'	2007	+10	1999	2050	8%
	'2003'	2012	+9	2004	2075	9%
	'2009'	2020	+11	2009	3000	10%
	'2015'	2025	+10	2016	2066	3%
8	'1997'	Never	$+\infty$	1984	5001	53%
	'2003'	Never	$+\infty$	1981	10000	62%
	'2009'	Never	$+\infty$	2020	3000	79%
	'2015'	Never	$+\infty$	2025	2150	58%
9	'1997'	Never	$+\infty$	1998	3020	68%
	'2003'	Never	$+\infty$	2006	2150	79%
	'2009'	Never	$+\infty$	1990	2080	85%
	'2015'	Never	$+\infty$	1958	2088	76%
10	'1997'	2020	+23	2000	3001	6%
	'2003'	2020	+17	2006	2150	4%
	'2009'	2030	+21	2009	3000	16%
	'2015'	2030	+15	2000	2097	0%
11	'1997'	2020	+23	1990	3000	8%
	'2003'	2020	+17	1995	2150	19%
	'2009'	2025	+16	2000	2080	18%
	'2015'	2035	+20	2004	2150	9%
12	'1997'	2020	+23	1997	3001	9%
	'2003'	2030	+27	2005	3827	19%
	'2009'	2035	+26	2010	5000	19%
	'2015'	2030	+15	2013	2136	4%

Table 3: *Results for statements 13-20.*

	Survey	Median	Rel.	Min.	Max.	"*Never*"
13	Kurzweil	"2009"	-	-	-	-
	'2003'	2100	+97	2005	2300	47%
	'2009'	Never	$+\infty$	2010	3000	56%
	'2015'	2050	+35	2015	2150	21%
14	'2003'	2100	+97	1995	2200	47%
	'2009'	2045	+36	2009	3009	35%
	'2015'	2030	+15	2017	2073	9%
15	Kurzweil	"2000s"	-	-	-	-
	'2003'	2015	+12	2000	2150	10%
	'2009'	2020	+11	2004	3000	8%
	'2015'	2027	+12	2011	2150	5%
16	'2003'	2040	+37	2005	2200	41%
	'2009'	2033	+24	2013	3009	29%
	'2015'	2025	+10	2016	2088	9%
17	Kurzweil	"2009"	-	-	-	-
	'2003'	2043	+40	1994	2500	25%
	'2009'	2060	+51	2012	3000	44%
	'2015'	2040	+25	2012	2150	16%
18	Kurzweil	"2000s"	-	-	-	-
	'2003'	2030	+27	2000	3000	6%
	'2009'	2040	+31	2009	3000	11%
	'2015'	2035	+20	2012	2121	0%
19	Kurzweil	"2019"	-	-	-	-
	'2003'	2053	+50	2004	3827	37%
	'2009'	2100	+91	1960	2140	48%
	'2015'	2045	+30	2019	2138	15%
20	Kurzweil	"2019"	-	-	-	-
	'2003'	2020	+17	2001	2275	3%
	'2009'	2020	+11	2009	2300	2%
	'2015'	2025	+10	1970	2075	7%

Of the other trends that can be found in Table 3, it is interesting to see that most of the statements were judged in the 2015 survey as getting closer. For example, although statement #19 "*Most interaction with computing is through gestures and two-way natural-language spoken communication*" was still judged in 2015 to be quite difficult (with a median of 2045), the probability that it would eventually become true increased significantly (down from 48% to 15%). Also interesting is that none of the respondents to the 2015 survey said that statement #18 "*Translating telephones allow two people across the globe to speak to each other*" would never be true (possibly influenced by the appearance of *Skype Translator* in 2014).

3.4. Results for statements 21-26

The detailed responses for the statements added in 2009 are shown in Table 4. Of particular interest here is the dramatic drop in "*Never*"s for statement #23 "*Most multilingual people communicate with each other through speech to speech translation*", and that #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*" were assessed as being only five years away.

3.5. Results for statements 27-30

Finally, the responses for the statements added in 2015 are shown in Table 5. The most interesting result here is that statement #28 "*Speech replaces text-based web search*" is judged as being very unlikely.

Table 4: *Results for statements 21-26.*

	Survey	Median	Rel.	Min.	Max.	"*Never*"
21	'2009'	2025	+16	2010	3000	26%
	'2015'	2025	+10	2012	2058	11%
22	'2009'	2020	+11	2009	2090	15%
	'2015'	2025	+10	2012	2050	5%
23	'2009'	2060	+51	2014	3000	40%
	'2015'	2044	+29	2015	2126	15%
24	'2009'	2020	+11	2009	2100	6%
	'2015'	2022	+7	2011	2075	4%
25	'2009'	2020	+11	2010	2100	8%
	'2015'	2025	+10	2015	2060	8%
26	'2009'	2019	+10	2000	2100	11%
	'2015'	2020	+5	2001	2049	4%

Table 5: *Results for statements 27-30.*

	Survey	Median	Rel.	Min.	Max.	"*Never*"
27	'2015'	2035	+20	2018	2120	0%
28	'2015'	2061	+46	2015	2150	44%
29	'2015'	2050	+35	2021	2150	15%
30	'2015'	2022	+7	2014	2113	20%

4. Concluding Remarks

This paper presents a formal record of the outcomes of four surveys conducted at the 1997, 2003, 2009 and 2015 IEEE ASRU workshops. Of course, some of the figures may have limited statistical significance due to the relatively low numbers of respondents and the possibility that some individuals may have given whimsical rather than informed opinions. Nevertheless, the results do seem to provide a useful insight into important trends that are taking place in the field of spoken language processing, and reveal the realism that pervades the research community. For example, a common theme noted by many respondents is that although speech technology may indeed become available for certain applications, it might not be used in practice due to the availability of alternative interface technologies that may offer significant benefits in terms of performance or privacy.

One interesting outcome from the ASRU-2015 survey is that, for the first time, four of the statements - #4 "*Voice recognition is commonly available at home*", #10 "*Public proceedings are transcribed automatically*", #18 "*Translating telephones allow two people across the globe to speak to each other*" and #27 "*Conversational interaction with autonomous robots is commonplace in the home*" - received 0% "*Never*"s. This would seem to reflect a growing confidence in the research community that spoken language technology is not only maturing, but that it is also finding applications that are able to capitalise on the unique benefits offered by speech-based interaction. It is hypothesised by the authors that this increase in confidence may be driven by two key developments that took place between the 2009 and 2015 surveys: the launch of *Siri* (in 2011) and the growing impact of 'deep learning' [5].

It will be very interesting to see how these trends translate into responses to the next survey which, according to the sexennial pattern established thus far, should be scheduled to take place at ASRU-2021.

5. References

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

[2] ——, *The Age of Spiritual Machines*. Phoenix Press, 1999.

[3] R. K. Moore, "Results from a survey of attendees at ASRU 1997 and 2003," in *INTERSPEECH*. Lisbon, Portugal: ISCA, 2005, pp. 117–120.

[4] ——, "Progress and prospects for speech technology: Results from three sexennial surveys," in *INTERSPEECH*. Florence, Italy: ISCA, 2011, pp. 1533–1536.

[5] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.