
Proceedings Paper:
Moore, R.K. orcid.org/0000-0003-0065-3311 and Marxer, R. (2023) Progress and prospects for spoken language technology: Results from five sexennial surveys. In: Proceedings of INTERSPEECH 2023. INTERSPEECH 2023, 20-24 Aug 2023, Dublin, Ireland. International Speech Communication Association (ISCA) , pp. 401-405.

https://doi.org/10.21437/interspeech.2023-235

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

Roger K. Moore[1]*, Ricard Marxer*[2]

[1]University of Sheffield, UK
[2]Université de Toulon, France

r.k.moore@sheffield.ac.uk, ricard.marxer@lis-lab.fr

Abstract

Every six years (since 1997), a survey has been conducted at the IEEE workshop on *Automatic Speech Recognition and Understanding* (ASRU). The aim has been to gain an insight into the research community's perspective on the 'progress and prospects' for spoken language technology. These surveys have been based on a set of 'statements' describing possible scenarios, and respondents are asked to estimate the year (in the future or in the past) when each given scenario might be realised. A number of the statements have appeared in multiple surveys, hence it has been possible to track changes in opinion over time. This paper presents the combined results from five such surveys, the most recent of which was conducted at ASRU-2021. The latest results reveal a dramatic increase in optimism.

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

1. Introduction

Every six years since 1997, a survey has been conducted among the attendees at the IEEE workshop on *Automatic Speech Recognition and Understanding* (ASRU). The aim has been to ascertain the research community's collective perspective on the 'progress and prospects' for 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 describes a possible scenario. Respondents are then asked to estimate the year (in the future or in the past) in which each given scenario might be realised - or they may respond with "*never*". The advantage of this approach is that it facilitates a quantitative analyse of the responses. Also, since a subset of the statements has been the same for each survey, it is possible to track how experts' opinions have changed over time. The most recent (and fifth) survey was conducted at ASRU-2021.

2. The Five Surveys

2.1. The 1997 survey

The first survey - entitled *Prospects for the Next Millennium* - was conducted 25 years ago at the 1997 ASRU workshop (held in Santa Barbara, USA). Attendees were requested to associate a year with each of 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 a summary was presented at a special interactive plenary session. Overall, the 1997 results were surprisingly negative. So, after consideration of the possible impact on 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 original twelve statements were supplemented 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):

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.**

It was found that the results for the twelve repeated state-

ments were remarkably consistent between the two surveys although, on average, the distributions of responses had shifted six years into the future. In other words, the projected scenarios were seen as being no closer in 2003 than they had in 1997. For the eight new statements, respondents appeared to be uniformly pessimistic about how long they would take to be realised.

Nevertheless, in 2003 the workshop attendees felt more secure in voicing their opinions. So the results of both the 2003 and 1997 surveys were published at INTERSPEECH-2005 [3].

2.3. The 2009 survey

In 2009, the survey was conducted on-line and in advance, and the outcome was presented at the ASRU workshop (held in Merano, Italy). Six additional statements (primarily relating to mobile devices and applications) were included:

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 2009, it was again concluded that the future still appeared to be no nearer than it had been in the past [4]. 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. The 2009 survey also revealed that there was no correlation between a respondent's optimism/pessimism and the length of time that they had spent in the field.

2.4. The 2015 survey

The ASRU-2015 workshop (held in Scottsdale, Arizona, USA) provided an opportunity to conduct a fourth survey. Four further statements were added (primarily relating to social agents and 'deep learning' [5]), bringing the total to thirty [6]:

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.*

One interesting outcome from the ASRU-2015 survey was that, for the first time, four of the statements (#4, #10, #18 and #27) received 0% "*Never*"s, reflecting a growing confidence that spoken language technology was maturing (e.g. as evidenced by the launch of Apple's *Siri* in 2011).

2.5. The 2021 survey

The most recent (and fifth) version of the survey was conducted at ASRU-2021 (hosted by colleagues from Cartagena in Colombia, but held on-line due to the restrictions imposed by the COVID-19 global pandemic). No additional statements were added, but it was anticipated that the results could be of special interest due to the dramatic increase in spoken language technology products and services since 2015 – especially the surprise emergence of 'smart speakers' (such as Amazon *Echo*, Google *Home*, etc.) and the continuing impact of deep learning.

However, despite the large number of (virtual) attendees at ASRU-2021, it proved surprisingly difficult to attract respondents to the survey. Nevertheless, a sufficient number of (named) individuals did submit responses, and some interesting trends were again observed.

3. Analysis of the Five Surveys

3.1. Overall results

The combined results for all five surveys (based on responses to the first twelve statements) are shown in Table 1. The main outcome is that the ASRU-2021 survey showed not only a significant drop in the number of "*Never*" responses, but also a dramatic increase in the proportion of respondents who were willing to be associated with their opinions. Both of these changes would suggest that the respondents were quite diligent in their responses.

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

	1997	**2003**	**2009**	**2015**	**2021**
No. of Respondents:	81	105	127	61	21
Overall Median:	2010	2020	2028	2025	2030
Relative to Survey:	+13	+17	+19	+10	+9
"*Never*"s:	17%	22%	28%	17%	2%
Named Responses:	22%	4%	21%	11%	62%

The distribution of responses averaged over the first twelve statements *relative to the year each survey was conducted* is illustrated in Figure 1. It can be seen that the results from the latest ASRU-2021 survey show a significant shift to the left, indicating that a growing number of the original statements have become, or are about to become, realised in practice. As seen previously, the results also show a clear quantisation effect whereby respondents have a natural tendency to assign particular round-numbered dates in their responses.

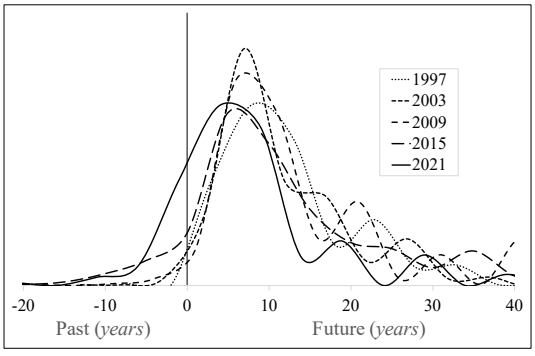

Figure 1: *The distribution of responses from the five surveys relative to the year each survey was conducted (based on responses to statements 1-12).*

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

	Survey	Median	Rel.	"*Never*"
1	'1997'	2000	+3	0%
	'2003'	2010	+7	15%
	'2009'	2015	+6	6%
	'2015'	2016	+1	3%
	'2021'	2020	-1	0%
2	'1997'	2002	+5	3%
	'2003'	2008	+5	2%
	'2009'	2015	+6	2%
	'2015'	2018	+3	2%
	'2021'	2021	0	0%
3	'1997'	2010	+3	8%
	'2003'	2012	+9	8%
	'2009'	2020	+11	13%
	'2015'	2023	+8	5%
	'2021'	2025	+4	0%
4	'1997'	2007	+10	4%
	'2003'	2011	+8	5%
	'2009'	2020	+11	10%
	'2015'	2022	+7	6%
	'2021'	2022	+1	0%
5	'1997'	2007	+10	5%
	'2003'	2010	+7	14%
	'2009'	2022	+13	37%
	'2015'	2032	+17	41%
	'2021'	2030	+9	5%
6	'1997'	2050	+53	30%
	'2003'	2050	+47	34%
	'2009'	2050	+41	36%
	'2015'	2035	+20	5%
	'2021'	2030	+9	0%
7	'1997'	2007	+10	8%
	'2003'	2012	+9	9%
	'2009'	2020	+11	10%
	'2015'	2025	+10	3%
	'2021'	2030	+9	0%
8	'1997'	Never	$+\infty$	53%
	'2003'	Never	$+\infty$	62%
	'2009'	Never	$+\infty$	79%
	'2015'	Never	$+\infty$	58%
	'2021'	2100	+79	14%
9	'1997'	Never	$+\infty$	68%
	'2003'	Never	$+\infty$	79%
	'2009'	Never	$+\infty$	85%
	'2015'	Never	$+\infty$	76%
	'2021'	2070	+49	10%
10	'1997'	2020	+23	6%
	'2003'	2020	+17	4%
	'2009'	2030	+21	16%
	'2015'	2030	+15	0%
	'2021'	2030	+9	0%
11	'1997'	2020	+23	8%
	'2003'	2020	+17	19%
	'2009'	2025	+16	18%
	'2015'	2035	+20	9%
	'2021'	2025	+4	0%
12	'1997'	2020	+23	9%
	'2003'	2030	+27	19%
	'2009'	2035	+26	19%
	'2015'	2030	+15	4%
	'2021'	2030	+9	0%

3.2. Results for statements 1-12

The detailed responses for the first twelve statements are shown in Table 2. As expected, #4 "*Voice recognition is commonly available at home*" was acknowledged as being a reality in 2021. Meanwhile, although #5 "*Automatic airline reservation by voice over the telephone is the norm*" was judged in previous surveys as receding into the future, the ASRU-2021 respondents suggested that, while still several years away, it was now much more likely to happen. Likewise, #3 "*TV closed captioning is automatic and pervasive*" was now seen as being only a couple of years away, and nobody thought that it would never happen.

Interestingly, #6 "*It is possible to hold a telephone conversation with an automatic chat-line system ...*" and #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*" were originally judged as being somewhat far-fetched. However, presumably because of the growing effectiveness of 'large language models' (LLMs) [7] and 'deep-fake' speech synthesis [8], the ASRU-2021 respondents showed a dramatic increase in expectation (albeit a still a few years away).

Also judged as still challenging (and $\sim$10 years away) was #7 "*Voice-enabled command, control and communication in cars becomes as common as ...*" and #12 "*Speech recognition accuracy equals that of the average human transcriber*". Again, no-one thought these two scenarios would never happen.

Finally, perhaps the most dramatic change from the previous surveys was that ASRU-21 respondents no longer judged #8 "*No more need for speech research*" or #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*" as being totally unlikely!

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 [1, 2] (see Section 2.2). As in the previous surveys, #20 "*Pocket-sized listening machines are commonly available for the hearing impaired*" was still seen as the most likely to be realised. On the other hand, #15 "*Telephones are answered by an intelligent answering machine ...*" and #18 "*Translating telephones allow two people across the globe to speak to each other*", both of which Kurzweil predicted to appear in the 2000s, were still judged to be a decade away in 2021.

Even further in the future, #13 "*The majority of text is created using continuous speech recognition*" and #19 "*Most interaction with computing is through gestures and two-way natural-language ...*" nevertheless showed a dramatic drop in "*Never*"s. Interestingly, #17 "*Most routine business transactions take place between a human and a virtual personality ...*" was seen as further away (although more likely) in 2021.

Of particular interest are the radical changes in responses for the two statements not based on predictions made by Kurzweil: #14 "*The majority of automatic speech recognition systems have completely abandoned the n-grams paradigm ...*" and #16 "*The majority of automatic speech recognition systems have completely abandoned the HMM paradigm ...*". Up to 2009 almost half the respondents thought it inconceivable that these standard techniques would be replaced, but at ASRU-2021 no-one thought that would never happen, and the dates by which it would occur were brought forward significantly.

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

	Survey	Median	Rel.	*"Never"*
13	Kurzweil	"2009"	-	-
	'2003'	2100	+97	47%
	'2009'	Never	$+\infty$	56%
	'2015'	2050	+35	21%
	'2021'	2050	+29	5%
14	'2003'	2100	+97	47%
	'2009'	2045	+36	35%
	'2015'	2030	+15	9%
	'2021'	2030	+9	0%
15	Kurzweil	"2000s"	-	-
	'2003'	2015	+12	10%
	'2009'	2020	+11	8%
	'2015'	2027	+12	5%
	'2021'	2030	+9	0%
16	'2003'	2040	+37	41%
	'2009'	2033	+24	29%
	'2015'	2025	+10	9%
	'2021'	2025	+4	0%
17	Kurzweil	"2009"	-	-
	'2003'	2043	+40	25%
	'2009'	2060	+51	44%
	'2015'	2040	+25	16%
	'2021'	2070	+49	5%
18	Kurzweil	"2000s"	-	-
	'2003'	2030	+27	6%
	'2009'	2040	+31	11%
	'2015'	2035	+20	0%
	'2021'	2034	+13	0%
19	Kurzweil	"2019"	-	-
	'2003'	2053	+50	37%
	'2009'	2100	+91	48%
	'2015'	2045	+30	15%
	'2021'	2050	+29	5%
20	Kurzweil	"2019"	-	-
	'2003'	2020	+17	3%
	'2009'	2020	+11	2%
	'2015'	2025	+10	7%
	'2021'	2026	+5	0%

3.4. Results for statements 21-30

The detailed responses for the statements added in 2009 are shown in Table 4. Of particular interest here is that #21 "*Most information access and search using mobile phones are done through speech recognition and synthesis*" and #23 "*Most multilingual people communicate with each other through speech to speech translation*" are still seen as $\sim$20 years away, but with far fewer "*Never*"s than before. The three statements #22 "*Mobile phones are used to control and monitor home appliances remotely using speech*", #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 judged to still be a couple of years away, but the respondents were now confident that they will happen.

Finally, the responses for the statements added in 2015 are also shown in Table 4, and the inclusion of the results for the ASRU-2021 survey means that it is now possible to observe the perceived trend for these statements. The most interesting result here is that the respondents were completely confident that #30 "*DNNs replace all of the major components in a spoken language technology system*" will happen, despite being a couple of years further away than estimated in 2015. Likewise, while #27 "*Conversational interaction with autonomous social agents (such as robots) is commonplace in the home*" and #29 "*Spoken language technology can translate a voice from one language to another as well as a human interpreter*" were still thought to be $\sim$10-15 years away, respondents were nevertheless confident it will happen. On the other hand, #28 "*Speech replaces text-based web search*" was seen to be a long way off, and some respondents were sceptical about whether it would ever happen.

Table 4: *Results for statements 21-30.*

	Survey	Median	Rel.	*"Never"*
21	'2009'	2025	+16	26%
	'2015'	2025	+10	11%
	'2021'	2042	+21	5%
22	'2009'	2020	+11	15%
	'2015'	2025	+10	5%
	'2021'	2025	+4	0%
23	'2009'	2060	+51	40%
	'2015'	2044	+29	15%
	'2021'	2037	+16	5%
24	'2009'	2020	+11	6%
	'2015'	2022	+7	4%
	'2021'	2025	+4	0%
25	'2009'	2020	+11	8%
	'2015'	2025	+10	8%
	'2021'	2028	+7	10%
26	'2009'	2019	+10	11%
	'2015'	2020	+5	4%
	'2021'	2025	+4	0%
27	'2015'	2035	+20	0%
	'2021'	2032	+11	0%
28	'2015'	2061	+46	44%
	'2021'	2050	+29	10%
29	'2015'	2050	+35	15%
	'2021'	2036	+15	0%
30	'2015'	2022	+7	20%
	'2021'	2025	+4	0%

4. Concluding Remarks

This paper presents a formal record of the outcomes of five surveys conducted at the 1997, 2003, 2009, 2015 and 2021 IEEE *Automatic Speech Recognition and Understanding* (ASRU) workshops. It is acknowledged that some of the figures may have limited statistical significance due to the relatively low numbers of respondents. Nevertheless, the results do provide a useful insight into important trends that are taking place in the field of spoken language processing.

The latest survey clearly reveals a growing confidence that spoken language technology has reached a degree of maturity, and that many applications originally judged as unlikely to be realised, are now thought to be within the realms of the possible. As before, 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-2027 – especially as, by then, $\sim$40% of the statements are predicted to have been realised.

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, M. Abdel-Rahman, 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.

[6] R. K. Moore and R. Marxer, "Progress and prospects for spoken language technology: results from four sexennial surveys," in *INTERSPEECH*, San Francisco, CA, 2016, pp. 3012–3016.

[7] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, *Improving language understanding by generative pre-training*. OpenAI, 2018. [Online]. Available: https://cdn.openai.com/research-covers/language-unsupervised/language{_}understanding{_}paper.pdf

[8] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: a generative model for raw audio," in *9th ISCA Workshop on Speech Synthesis*, 2016. [Online]. Available: https://arxiv.org/abs/1609.03499