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

Progress and Prospects for Spoken Language Technology: What Ordinary People Think

Roger K. Moore, Hui Li, Shih-Hao Liao

Speech and Hearing Research Group, Dept. Computer Science, University of Sheffield, UK
r.k.moore@sheffield.ac.uk, ivydream@live.cn, will00319@gmail.com

Abstract

Arguably the most significant milestone (so far) in the spoken language technology field was the appearance in November 2011 of *Siri* - Apple's voice-based 'personal assistant and knowledge navigator' for the iPhone. *Siri* brought the potential of spoken language technology to the attention of the wider general public, and speech finally became "*mainstream*". This meant that ordinary people suddenly had an informed opinion about the merits (or otherwise) of using their voice to access information, send messages and control their smart devices. So, this paper presents the results of two surveys that were conducted in order to find out what ordinary people think about contemporary spoken language technology. The first used a modified version of the surveys conducted every six years at the IEEE ASRU series of workshops, and the second addressed questions about the awareness and usage of speech technology by members of the general public. The overall results suggest that ordinary people are more optimistic than the experts about what spoken language technology might have to offer, but usage patterns reveal that the majority of end users still prefer typing to talking, with accuracy, privacy and online accessibility cited as the main impediments to wider take-up.

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

1. Introduction

Since its early beginnings in the 1950s (or thereabouts), the field of spoken language technology has passed many significant milestones in terms of system performance and market penetration [1, 2, 3, 4]. For example progress has spanned from the publication of the first papers on spoken digit recognition in 1952 [5] and text-to-speech synthesis in 1964 [6], the publication of Bruce Lowerre's PhD thesis on the *HARPY* connected speech recognition system in 1976 [7], the release of Texas Instruments' *Speak-and-Spell* educational toy in 1978, Jim Baker's public demonstration of Dragon's HMM-based isolated word recogniser on a PC at IEEE ICASSP in Boston in 1983, the publication of Kai Fu Lee's Ph.D. thesis on *SPHINX* - "*the first system to demonstrate the feasibility of accurate large-vocabulary speaker-independent continuous speech recognition*" in 1988 [8], to the release of Dragon's *Naturally Speaking* and IBM's *Via Voice* continuous speech recognition products in 1997. See [9] for a comprehensive and up-to-date timeline of key developments in spoken language technology.

However, arguably the most significant milestone (so far) has been the appearance in November 2011 of *Siri* - Apple's voice-based 'personal assistant and knowledge navigator' for the iPhone. *Siri* finally brought the potential of spoken language technology to the attention of the wider general public, and - to use Xuedong 'XD' Huang's catchphrase - "*speech became mainstream*" [10]. Competitor products such as *Google Now* and Microsoft's *Cortana* soon followed, and the consequence has been that, suddenly, ordinary people had an informed opinion about the merits (or otherwise) of attempting to use spoken language technology to access information, send messages and control their smart devices.

This paper presents the results of two surveys that were deployed in order to find out what ordinary people think about contemporary spoken language technology. The first - conducted in 2013 - used a modified form of the surveys conducted every six years at the IEEE series of international workshops on *Automatic Speech Recognition and Understanding* (ASRU) [11, 12, 13]. The second - conducted in 2015 - addressed questions about the awareness and usage of speech technology by members of the general public.

2. The Surveys

2.1. Expert *vs.* non-expert opinion

Every six years since 1997, the first author has conducted a survey of attendees at the IEEE ASRU workshops, the most recent being held in Scottsdale, Arizona, USA (in December 2015). The ASRU surveys are based on a set of 'statements' which describe putative events concerned with spoken language technology, and respondents are asked to estimate the year in which each statement might become true. For example, a typical statement is "*Automatic airline reservation by voice over the telephone is the norm*", and a respondent might supply the answer "2020". The advantage of this approach is that it is possible to construct distributions of the numerical responses and to compute relevant summary statistics (such as the medians, minima and maxima). Respondents are also given the opportunity to answer "*Never*" to any particular statement. Over the years, these surveys have provided an interesting and valuable insight into *expert* opinion (i.e. the views of practitioners in spoken language technology R&D) of progress and prospects in the field.

In 2013 it was decided that it would be interesting to use a modified version of the ASRU-2009 survey to determine what *non-experts* (i.e. members of the general public) think about progress and prospects in spoken language technology. The modifications were necessary because some of the statements (*e.g.* "*The majority of automatic speech recognition systems have completely abandoned the HMM paradigm for acoustic modelling*") were too technical for non-experts, so they were deleted. Other statements were re-worded to make them more generally understandable (*e.g.* "*More than 50% of new PCs have dictation on them . . .* " was replaced with "*More than 50% of new PCs have speech recognition or synthesis software installed . . .* ").

http://dx.doi.org/10.21437/Interspeech.2016-874

Having started with twelve statements in 1997, by 2009 the ASRU survey had grown to twenty-six statements (there were thirty in 2015). Pruning and editing to make them suitable for a more general audience resulted in a slightly reduced set of twenty-one (see Table 1). In addition, respondents to the 2013 survey were asked the following supplementary questions:

- *Age, gender, nationality, educational background?*
- *Are you a technology lover?*
- *Please choose three products or services which you are most willing to use.*
- *Do you have any experience of using speech technology products/services (if yes, what are they)?*
- *Do you feel comfortable in using these products/services?*

The 2013 survey was conducted by the second author (as her final-year undergraduate dissertation project) using *Google Docs*, and it was sent to the 'volunteer list' at the University of Sheffield. A total of 188 responses were received, and the answers to the supplementary questions revealed that over half (57%) of the respondents were under 25, and two-thirds (63%) considered themselves to be technophiles. Also, it was found that, while there was no correlation ($r = -0.01$) between respondents' educational background and their average responses to the twenty-one statements, there was a significant inverse correlation ($r = -0.73$) between the responses and respondents' age; younger respondents being more pessimistic than older respondents (this is in-line with a similar result found for experts [12]). Full results are presented in [14].

As stated above, the aim of the 2013 survey was to compare the opinions of non-experts with experts. So, Table 1 summarises the responses in comparison with those obtained from the recently conducted ASRU-2015 survey [13]. What is immediately apparent is that there is a remarkable degree of agreement between the two sets of responses (correlation $r = 0.85$; Spearman's rank-order correlation $\rho = 0.60$, $p = 0.004$). However, it is interesting to observe that, overall, the non-experts appear to be more optimistic than the experts. For example, the median response across all twenty-one statements was 2030 for the experts, whereas it was 2023 for the non-experts.

The largest differences in opinion occurred for statement #12 "*The majority of text is created using continuous speech recognition*" (2050 for experts, but 2030 for non-experts) and for statement #5 "*Automatic airline reservation by voice over the telephone is the norm*" (ranked 13th by experts, but 5th by non-experts). In both cases, the non-experts have higher expectations than the experts that these events will indeed take place (and sooner rather than later). Interestingly, in both cases, the experts returned a much higher percentage of "*Never*" responses than the non-experts.

One area where both the experts and the non-experts agree is on statements #8 "*No more need for speech research*" and #9 "*A leading cause of time away from work is . . .* " which both received a very high number of "*Never*" responses from each group of respondents. Indeed, the spoken language technology research community may be relieved to learn that non-experts are even more strongly in favour of continued research than they are themselves!

2.2. Awareness and usage

The comparison between the opinions of non-experts and experts is interesting, but it does not give any insight into the degree to which technology such as *Siri* is actually used. Therefore, in 2015 it was decided that it would be useful to conduct

Table 1: *Comparison of responses from the survey of 'experts' (conducted in 2015) and 'non-experts' (conducted in 2013).*

Statement	**Opinion**	**Median**	**"*Never*"**
1. *More than 50% of new PCs have speech recognition or synthesis software installed, either at purchase or shortly after.*	Expert Non-Expert	2016 2010	3% 1%
2. *Most telephone Interactive Voice Response systems accept speech input (and more than just digits).*	Expert Non-Expert	2018 2010	2% 1%
3. *TV closed captioning is automatic and pervasive.*	Expert Non-Expert	2023 2015	5% 5%
4. *Voice recognition is commonly available at home (*e.g. *interactive TV, control of home appliances and home management systems).*	Expert Non-Expert	2022 2020	6% 1%
5. *Automatic airline reservation by voice over the telephone is the norm.*	Expert Non-Expert	2032 2020	41% 7%
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.*	Expert Non-Expert	2035 2025	5% 15%
7. *Voice-enabled command, control and communication in cars becomes as common as intermittent wiper, power window or power door lock.*	Expert Non-Expert	2025 2025	3% 5%
8. *No more need for speech research.*	Expert Non-Expert	"*Never*" "*Never*"	58% 73%
9. *A leading cause of time away from work is being hoarse from talking all the time, and people buy keyboards as an alternative.*	Expert Non-Expert	"*Never*" "*Never*"	76% 61%
10. *Public proceedings (*e.g. *courts, public inquiries, parliament, etc.) are transcribed automatically.*	Expert Non-Expert	2030 2025	0% 4%
11. *Speech recognition accuracy equals that of the average (individual) human transcriber.*	Expert Non-Expert	2030 2030	4% 9%

Statement	Opinion	Median	*"Never"*
12. *The majority of text is created using continuous speech recognition.*	Expert Non-Expert	2050 2030	21% 16%
13. *Telephones are answered by an intelligent answering machine that converses with the calling party to determine the nature and priority of the call.*	Expert Non-Expert	2027 2025	5% 9%
14. *Most routine business transactions take place between a human and a virtual personality (including an animated visual presence that looks like a human face).*	Expert Non-Expert	2040 2035	16% 22%
15. *Translating telephones allow two people across the globe to speak to each other even if they do not speak the same language.*	Expert Non-Expert	2035 2030	0% 6%
16. *Most interaction with computing is through gestures and two-way natural- language spoken communication.*	Expert Non-Expert	2045 2035	15% 18%
17. *Pocket-sized listening machines are commonly available for the hearing impaired.*	Expert Non-Expert	2025 2020	7% 0%
18. *Most information access and search using mobile phones are done through speech recognition and synthesis (*e.g. *web search, SMS).*	Expert Non-Expert	2025 2020	11% 9%
19. *Mobile phones are used to control and monitor home appliances remotely using speech (*e.g. *remote access to DVR, recording programmes, TV).*	Expert Non-Expert	2025 2020	5% 3%
20. *Most multilingual people communicate with each other through speech-to-speech translation at any time using their mobile device.*	Expert Non-Expert	2044 2030	15% 13%
21. *All mobile devices have built-in speech recognition capability.*	Expert Non-Expert	2020 2020	4% 4%

a survey of smartphone users to determine their awareness and usage of spoken language technology. On this occasion, the survey consisted of fifteen straightforward questions:

1. *What is your gender?*
2. *What is your nationality?*
3. *What is your current education level?*
4. *What is your age?*
5. *Do you possess a smartphone device?* (if "No", end of survey)
6. *How competent do you consider yourself with technology?* [very competent, competent, not very competent, not at all competent]
7. *Are you aware of the voice control function on your mobile?* [yes, no]
8. *What is the voice assistant on your mobile system?* [Siri, Google Now, Cortana, something else, don't know]
9. *Have you ever used the speech recognition service on your mobile and how often do you use it?* [several times a day, at least once a day, at least once a week, at least once a month, only a few times, never (go to 13)]
10. *My experience of using speech recognition is ...* [very satisfactory, satisfactory, neutral, unsatisfactory, very unsatisfactory]
11. *How do you get to use the speech recognition on your mobile?* [built in, downloaded]
12. *What kind of voice function do you use on your mobile?* [make calls, open apps, send messages, ask questions, ...]
13. *If you don't use it regularly, what is the main problem?*
14. *Will you continue to use this function, even though you found it is hard to use at the moment?* [yes, maybe, no]
15. *Which mode do you prefer to use on your mobile?* [typing, voice, gesture]

This particular survey was conducted by the third author (as his Masters dissertation project) using *Qualtrics*, and it was advertised around the world using social media such as Facebook, Twitter and LinkedIn. A total of 250 responses were received, 98% of whom owned a smartphone and 92% of whom considered themselves to be competent or very competent with technology. Only 6% of the respondents were not aware of the voice control facility on their device. In terms of the market share for the different systems; 52% of the respondents were using *Siri*, 40% used *Google Now* and 5% used *Cortana*.

The full results are presented in [15], and the main outcomes are summarised in Figures 1, 2 and 3, and Table 2. Of particular interest is the discovery that only a quarter (26%) of the respondents used their voice assistant on a fairly regular (daily or weekly) basis, and that two-thirds (66%) had only tried it once or not at all. Having said that, over half the respondents (54%) reported that they had had a satisfactory experience. Unsurprisingly, the main speech functions were voice search (64%) and voice command (54%), but it was interesting to discover that a good proportion of users (30%) also used it to recognise music. Of the reported problems, having to repeat yourself was judged to be the biggest issue (30%), with the need to be connected to the internet coming second (16%). Rather more concerning is that only just over a third of the respondents (38%) were willing to persevere with a speech interface, and a huge majority (85%) preferred typing over speech or gesture.

These results suggest that, notwithstanding the excitement surrounding the appearance of voice-based personal agents such as *Siri*, *Google Now* and *Cortana*, they nevertheless occupy

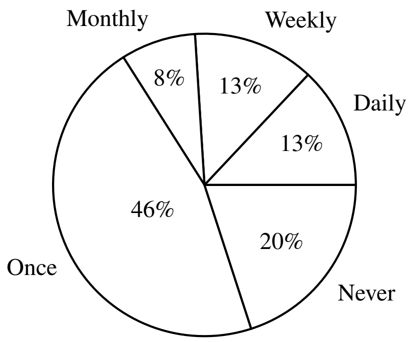

Figure 1: *How often smartphone users make use of automatic speech recognition function(s).*

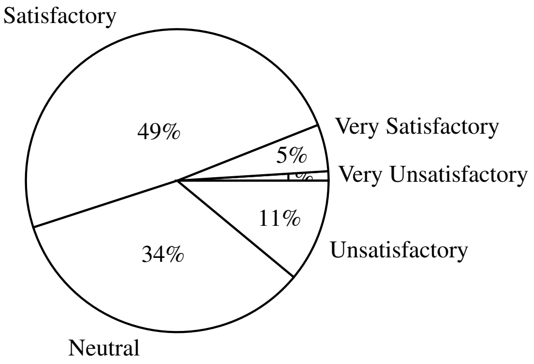

Figure 2: *Smartphone users' experiences using automatic speech recognition.*

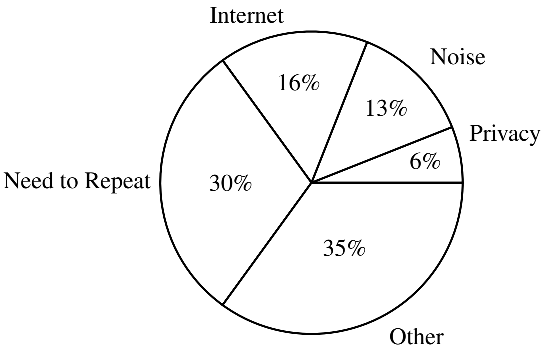

Figure 3: *The main problems encountered by smartphone users attempting to use automatic speech recognition.*

Table 2: *Other outcomes from the survey of smartphone users.*

	Options	**Responses**
Function:	Voice control	54%
	Ask questions	64%
	Voice note	25%
	Recognise music	30%
Continue with speech:	Yes	38%
	Maybe	41%
	No	21%
Preferred mode:	Typing	85%
	Voice	8%
	Gesture	7%

somewhat niche application areas. The potential benefits of hands-free eyes-free operation across more general applications appear to be negated by issues relating to accuracy, privacy and accessibility. Of course, the accuracy of automatic speech recognition, especially in noisy environments, has always been a major research challenge, and recent gains arising from the introduction of 'deep learning' may serve to mitigate some of these problems [16]. Likewise, issues of accessibility are being addressed by the manufacturers, as evidenced by Google's recent announcement (March 2016) that they were investing in faster and more accurate speech recognition that can function without an internet connection.

3. Discussion

Taken together, the two surveys presented herein provide an interesting insight into contemporary views about spoken language technology that are held by individuals who are not experts in the field, but who are actual or potential end-users. The main aim was to discover if ordinary people understand the relative difficulty of different potential applications and whether, despite the tremendous technical progress that has taken place in recent years, systems like *Siri*, *Google Now* and *Cortana* are actually being used in practice. The results confirm the informal impressions gained by talking to general audiences that, although most people are now aware of the technology and have even given it a go, practical usage remains remarkably low. The main exception appears to be users who cannot or will not type, for example people with disability or niche professional application domains such as medical dictation.

Of course, the spoken language technology field is by no means standing still, and progress continues to be made on all fronts. As a result, the two surveys reported here must be interpreted as a snapshot of an underlying trajectory. As mentioned in Section 1, the ASRU surveys have been conducted every six years since 1997, so it has been possible to track the opinions of the speech technology experts for almost twenty years. However, the views of ordinary users have only been collected in the past couple of years. So, it is not possible to say whether and how fast the 85% of users who currently still prefer to type might ultimately be converted to using speech.

Finally, notwithstanding potentially serious impediments to the wholesale use of spoken language technology (such as privacy concerns), there is also a larger question about how far it is possible to go in creating *habitable* language-based interfaces between human beings and 'intelligent' technology [17].

4. Summary and Conclusion

This paper has presented the results of two surveys that were conducted in order to find out what ordinary people think about contemporary spoken language technology. The first used a modified version of the surveys conducted every six years at the IEEE ASRU series of workshops, and the second addressed questions about the awareness and usage of speech technology by members of the general public. The overall results suggest that, as one might expect, ordinary people are more optimistic than the experts about what spoken language technology might have to offer. However, usage patterns reveal that the majority of users still prefer typing to talking.

5. References

[1] S. Furui, “Fifty years of progress in speech and speaker recognition,” *The Journal of the Acoustical Society of America*, vol. 116, no. 4, pp. 2497–2498, 2004.

[2] F. Juang and L. Rabiner, “Automatic speech recognition: A brief history of the technology development,” in *Encyclopedia of Language and Linguistics*. Elsevier, 2005.

[3] D. O’Shaughnessy, “Automatic speech recognition: History, methods and challenges,” *Pattern Recognition*, vol. 41, no. 10, pp. 2965–2979, 2008.

[4] R. Pieraccini, *The Voice in the Machine*. MIT Press, Cambridge, MA, 2012.

[5] K. H. Davis, R. Biddulph, and S. Balashek, “Automatic recognition of spoken digits,” *Journal of the Acoustical Society of America*, vol. 24, pp. 637–642, 1952.

[6] J. N. Holmes, I. G. Mattingly, and J. N. Shearme, “Speech synthesis by rule,” *Language and Speech*, vol. 7, pp. 127–143, 1964.

[7] B. T. Lowerre, “The HARPY speech recognition system,” PhD Dissertation, Carnegie-Mellon University, Pittsburgh, PA, 1976.

[8] K. F. Lee, “Large vocabulary speaker-independent continuous speech recognition: the SPHINX system,” PhD Dissertation, Carnegie-Mellon University, 1988.

[9] R. K. Moore, “The Past, Present and Future of Speech Technology.” [Online]. Available: http://www.dcs.shef.ac.uk/~roger/progress.html

[10] L. Deng and X. Huang, “Challenges in adopting speech recognition,” *Communications of the ACM*, vol. 47, no. 1, pp. 69–75, 2004.

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

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

[13] R. K. Moore and R. Marxer, “Progress and prospects for spoken language technology: results from four sexennial surveys,” in *INTERSPEECH*. San Francisco, CA: ISCA, 2016.

[14] H. Li, “Market/Opinion Survey for Speech Technology Systems,” Unpublished Final-Year Project Dissertation, Dept. Computer Science, University of Sheffield, 2013.

[15] S.-H. Liao, “Awareness and Usage of Speech Technology,” MSc Project Dissertation, Dept. Computer Science, University of Sheffield, 2015.

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

[17] R. K. Moore, “Is spoken language all-or-nothing? Implications for future speech-based human-machine interaction,” in *International Workshop on Spoken Dialogue Systems (IWSDS)*, Saariselkä, Finland, 2016.