Spoken Language Output: Realising the Vision

Roger K. Moore

20/20 Speech Ltd.
Science Park, Geraldine Road, Malvern, Worcs., WR14 3PS, UK
r.moore@2020speech.com

Abstract

Significant progress has taken place in 'Spoken Language Output' (SLO) R&D, yet there is still some way to go before it becomes a ubiquitous and widely deployed technology. This paper reviews the challenges facing SLO, using 'Technology Roadmapping' (TRM) to identify market drivers and future product concepts. It concludes with a summary of the behaviours that will be required in future SLO systems.

1. Introduction

The past twenty or so years have witnessed considerable progress in both the science and technology of 'spoken language output' (SLO) [1-6]. Advances in areas such as linguistic text analysis, natural language generation, prosodic modelling and digital speech signal processing, coupled with the immense growth in available computational resources, have lead to the successful commercialisation of a range of high-quality text-to-speech (TTS) systems. As in the field of automatic speech recognition (ASR), TTS has benefited from the introduction of a 'data-driven paradigm' in which corpora of annotated speech recordings are used to estimate the parameters of the underpinning model(s) and, in the majority of contemporary systems, to provide an inventory of acoustic segments from which selected units are concatenated together to produce the output speech.

However, whilst current systems are impressive in comparison to their earlier counterparts, the limited number of available voices and accents, and their general lack of 'expressiveness' [7], means that there is still some way to go before SLO becomes a truly ubiquitous and widely deployed technology. As Henton [8] put it recently:

> *"After sixty years of concentrated research and development in speech synthesis and text-to-speech (TTS), our gadgets, gizmos, executive toys and appliances still do not speak to us intelligently."*

Also, whilst the data-driven approach has provided considerable practical benefit, many researchers believe that there is a limit as to how far it can be taken. For example, Keller [7] estimates that the size of database needed to capture 100 different talking styles and 10,000 different voices would be 5,000 Gbytes. From this, he concludes that current technology is too cumbersome and that automatic processing is not up to generating such databases automatically[1].

It is therefore generally agreed that improvements are needed in areas such as speech sound generation, prosody generation, higher-level linguistic processing, text analysis, talker individuality and voice quality [5][6] – the latter being seen as particularly significant commercially:

> *"The undeniable efficiency gains of TTS need to be balanced against consumer perceptions of voice quality."* [10]

Clearly, almost all aspects SLO would seem to be in need of further research, and thus there is interest in agreeing a technical timetable or 'roadmap' of the necessary developments. This issue has been taken up by the 'European Network of Human Language Technologies' (ELSNET) and, since 2000, they have been conducting a roadmapping exercise across the range of 'Human Language Technologies' (HLT) [11-13]. This paper is an attempt to contribute to the debate – starting with a short introduction into to how technology roadmapping is performed in a more formal context.

2. Technology Roadmapping

Technology Roadmapping - 'TRM' - was pioneered by Motorola in the 1980s [14] and is defined as a "needs driven technology planning process". The main principle is that a technical roadmap should be based on 'market pull' rather than 'technology push'. In other words, TRM is not primarily concerned with "where can I go from here?" but with "where do I want to get to, and what's the best way of getting there?"[2]. The results are often presented in the form illustrated in Fig. 1.

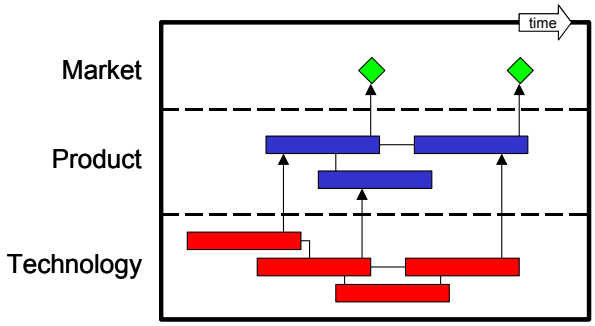

Figure 1: Typical output from a TRM exercise.

The formal TRM process starts with the identification of relevant *market drivers* and *market opportunities*, followed by the *product feature concepts* that could satisfy them. Consideration would then be given to the *technical solutions* that would be required to realise the new products, and then

[1] Interestingly, this parallels the conclusions reached by Moore [9] in analysing the amount of data needed for an ASR system to approach the performance of a human listener.

[2] Readers with a knowledge of ASR will immediately appreciate the analogy with 'dynamic programming' (DP) search and hence the significance of the TRM approach in a complex planning task.

milestones, *resources* and *tasks* would be planned and quantified. The process is facilitated by the creation of two analysis grids – one relating market drivers to product features, and the other relating product features to technical solutions (see Fig. 2).

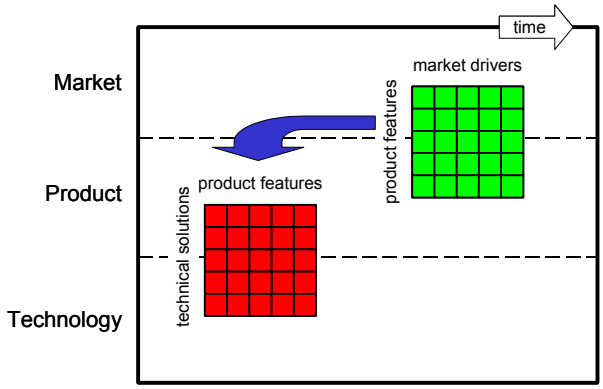

Figure 2: Analysis grids used in a TRM process.

An important aspect of the start of the TRM process is consideration of the 'performance dimensions' that drive product development. Product performance is a fundamental factor that can be used to link the market drivers to technological capability. A product's 'performance envelope' is dictated by a trade-off between market pull (requirements) and technology push (capabilities). This is entirely analogous to the concept of 'capability/requirement profile' developed by Moore [15] in the early 1990s to characterise the performance of speech technology systems.

The next three sections illustrate (albeit in a necessarily superficial way) the consequences of applying the TRM process to obtain an insight into the future of SLO and SLO systems.

3. Future market drivers

Some indication of the market drivers that are anticipated in the near future can be gained from the European Commission's recent call for the 6th Framework 'Information Society Technologies' (IST) programme [16]. Focused on 'ambient intelligence', the call describes a future "in which computers and networks will be integrated into the everyday environment, rendering accessible a multitude of services and applications through easy-to-use human interfaces".

Other market drivers that may have some bearing on future SLO systems undoubtedly include such things as an increasing penetration of mobile devices into environments that can benefit from hands-free eyes-free operation, legislation banning the use of mobile devices whilst driving, the growth of network-based information services, messaging and edutainment, the demand for 'personalised' services, 3G connectivity and, of course, Moore's law [17].

Indeed the latter, coupled with an anticipated growth in bioengineering, has prompted Kurzweil [18] to speculate on the implications of the fact that a PC will have the computational power of the human brain by 2019, and it will be equivalent to 1000 human brains by 2029. From these particular drivers, he predicts, not just the creation of automated systems with human-like characteristics, but also the replacement of human faculties with automated (prosthetic) processes - see the next section.

4. Advanced product features

As an example of a set of advanced product features derived from the future market drivers identified in the previous section, the EU's IST work programme calls for multimodal interfaces using robust dialogue that are natural/intuitive, autonomous, adaptive and multilingual, that recognise emotive user reaction and can handle both unconstrained and ill-formed inputs, and provide an intelligent response in unrestricted domains such as wearable interfaces, intelligent rooms, collaborative working tools and cross-cultural communications.

In the research community, new uses of SLO are envisaged in areas such as language learning [19], virtual reality, language teaching, training in reading, linguistic and psycholinguistic experimentation, and historic reconstruction [20]. In the commercial arena, Philips' [21] speculates on a wide variety of speech-enabled devices ranging from multimedia kiosks, to enhanced jewellery, interactive wallpaper and billboards, magic pens, 'hear me' devices and data zones.

However, as indicated in the previous section, perhaps the most adventurous attempts at predicting future technical capabilities has been those by Kurzweil [18][22]:

- **early 2000s**: "translating telephones allow two people across the globe to speak to each other even if they do not speak the same language; speech-to-text machines translate speech into a visual display for the deaf; telephones are answered by an intelligent answering machine that converses with the calling party to determine the nature and priority of the call"

- **2009**: "the majority of text is created using continuous speech recognition; ubiquitous language user interfaces; most routine business transactions take place between a human and a virtual personality (including an animated visual presence that looks like a human face); pocket-sized reading machines for the visually impaired; listening machines for the deaf; translating telephones commonly used for many language pairs"

- **2019**: "three-dimensional virtual reality displays, embedded in glasses and contact lenses, as well as auditory 'lenses', are used routinely as primary interfaces for communication with other persons, the Web, and virtual reality; most interaction with computing is through gestures and two-way natural-language spoken communication; deaf persons read what other people are saying through their lens displays; the vast majority of transactions include a simulated person; people are beginning to have relationships with automated personalities"

- **2029**: "permanent or removable implants are used to provide input and output between the human user and the world-wide computer network; direct neural pathways have been perfected for high-bandwidth connection to the human brain; a range of neural implants is available to enhance visual and auditory perception and interpretation, memory and reasoning; automated agents are learning on their own; widespread use of all-encompassing visual, auditory and tactile communication using direct neural connections; the majority of communications involving a human is

between a human and a machine; growing discussion on what constitutes being 'human'"

- **2049**: "nanobot swarm projections used to create visual-auditory-tactile projections of people and objects in real reality"
- **2099**: "no longer any clear distinction between humans and computers"

Clearly, regardless of the claimed timelines (and whether one is willing to believe Kurzweil's interpretation of future events), all of the advanced concepts cited in this section present important and worthwhile technical challenges with respect to our current understanding of SLO and SLO systems.

5. Technical challenges

Many researchers have reviewed the key technical challenges facing SLO, and it is interesting to consider them in the context of the requirements presented above. For example, Keller [7] cites the need for advanced spectral synthesis techniques and improved modelling of style and voice. To support this, he sees the need for systematic research on novel signal generation techniques, more sophisticated phonetic and prosodic models, and work on style, voice, language and dialect.

Dutoit [1] suggests that improvements are needed to the underlying speech models, especially in the handling of coarticulatory phenomena. He also looks forward to research on corpus-based *models* of speech segments, as opposed to corpus-based *instances* of segments, and to the establishment of a meaningful relationship between syntax, semantics, pragmatics and prosody. Dutoit also speculates on the introduction of variability (not randomness) with hidden coherence, the need to study speaker and speaking style effects, and the possibility of creating truly multilingual (rather than a collection of monolingual) solutions.

Mobius [23] points out that one of the most serious challenges facing SLO is the systematic treatment of events that are known to have low frequencies of occurrence. He observes that the characteristic of LNRE (large number of rare events) distributions is that the probability of encountering one is high, and that they occur in key SLO processes such as text analysis, syllabification, duration modelling and acoustic unit inventories.

Huckvale [24] recently observed that SLO now has two distinct goals: to understand how humans talk, and to simulate a talking process. For simulation systems, he identifies the need for more natural corpora, higher-level linguistic descriptions, richer phonetic labelling, advances in machine learning, more perceptually-based unit selection, and the introduction of mechanisms for extrapolation and interpolation. For computational models of human talkers, he calls for better modelling of the vocal tract, SLO with hearing, with proprioceptive feedback, with the ability to mimic what it hears and monitor its own performance, an innate ability to learn, to want to learn and to want to communicate.

Meanwhile, Chen [25] points out that prosody prediction has another dimension for tonal languages.

6. Behavioural challenges

However, whilst it is informative to enumerate the technical challenges that derive from the advanced features outlined in the earlier sections, a more productive exercise might be to focus, not on the technical shortfalls *per se*, but on the advanced *behaviours* that will be required in future SLO systems. For example, in order to produce *believable* behaviour that is appropriate to a suitably individualised 'communicative agent', future systems need to be able to:

- **talk 'clearly'**: Dynamic content is, by its very nature, somewhat unpredictable, and thus harder for a user to recognise in a noisy environment. Model-based SLO systems have been shown to be more intelligible and comprehensible than concatenative systems [26], but neither approach has addressed the classic 'hypo-hyper' behaviour of a human talker [27].
- **talk 'intelligently'**: The performance of current SLO systems is severely limited by the fact that they do not understand what they are saying [24]. This means that many key behavioural features are either absent, or even worse, misrepresented to the human listener. SLO should be seen as a combination of natural language generation and speech synthesis [2] in which speech is generated from an abstract representation of concepts [28] rather than from text. The issues are thus what concepts to include, how to realise them in words and what intonation to use in the context of the past discourse and the listener's goals and background.
- **talk 'expressively'**: Motivational and emotional states are key drivers of intelligent behaviour [29]. However, in order to express a large number of emotional states with a natural-sounding voice, either model-based techniques need to become more natural-sounding, or the selection-based techniques must become more flexible [30].
- **talk 'appropriately'**: SLO systems participating in dialogue must be able to select and organise content as part of a larger discourse structure, and convey this structure, as well as the content, to the user(s) [2]. The language used should be appropriate to spoken situations, e.g. using tailored responses [31] and terminology that is familiar to the user. Also, human factors choices need to be made about the appropriate output modes and media [32].
- **talk 'realistically'**: Natural sounding output would seem to be an incontrovertible goal. However, user interactions with an anthropomorphic agent face has been reported to take more effort [33], and eye-blinking has been shown to have a great affect on a user's appraisal of an agent's capabilities and can thus be misleading [34]. On the other hand, it is well established that users behave more cooperatively when interacting with a clearly automated system, and this difference can be conditioned purely by the sound of the voice [35].

7. Conclusion

Finally, everyone is agreed that measuring progress is a key aspect that underpins future development [1-4]. However,

whilst there are some standard protocols in place [36], there remains some question marks over the adequacy of given tests [37] as diagnostic tools. These problems will only get worse, as SLO systems become embedded in real-time interactive applications.

8. References

[1] Dutoit, T. *An Introduction to Text-to-Speech Synthesis*, Kluwer Academic Publishers, 1997.

[2] McKeown, K. R. and Moore, J. D., "Spoken Language Generation", *Survey of the State of the Art in Human Language Technology*, Cole, R. et al (eds.), Cambridge University Press & Giardini, 1997.

[3] Sproat, R. (ed), *Multilingual Text-to-speech Synthesis: The Bell Labs Approach*, Kluwer Academic Publishers, 1997.

[4] Keller, E., Bailly, G, Monaghan, A., Terken, J. and Huckvale, M. (eds.), *Improvements in Speech Synthesis*, Wiley & Sons, Chichester, UK, 2001.

[5] Holmes, J. and Holmes, W. *Speech Synthesis and Recognition*, Taylor & Francis, 2001.

[6] Furui, S., Digital Speech Processing, Synthesis, and Recognition, Marcel Decker Inc., 2001.

[7] Keller, E., "Towards Greater Naturalness: Future Directions of Research in Speech Synthesis", *Improvements in Speech Synthesis*, Keller, E., Bailly, G, Monaghan, A., Terken, J. and Huckvale, M. (eds.), Wiley & Sons, Chichester, UK, 2001.

[8] Henton, C. G., "Fiction and Reality of TTS", *Speech Technology Magazine*, Vol.7, No.1, Jan./Feb. 2002.

[9] Moore, R. K., "A Comparison of the Data requirements of Automatic Speech Recognition Systems and Human Listeners", *submitted to EUROSPEECH'03*, Geneva, 2003.

[10] "Text-to-Speech: How Do You Speak To Your Customers?", *Datamonitor Market Study Report*, Ref. BFTC0570, Oct. 2001.

[11] "HLT Roadmap", *ELSNET*, http://www.elsnet.org/roadmap.html

[12] "The ELSNET Road map for Human Language Technologies", Nov. 2002, http://www.elsnet.org/dox/dfki2002.pdf

[13] Bernsen, N. O., "Speech Related Technologies: Where will the Field Go in 10 Years?", http://www.elsnet.org/dox/rm-bernsen-v2.pdf

[14] Willyard, C. H. and McClees, C. W., "Motorola's Technology Roadmap Process", *Research Management*, pp.13-19, Sept.-Oct. 1987.

[15] Moore R. K., "Users guide", *EAGLES Handbook of Standards and Resources for Spoken Language Systems*, Gibbon, D., Moore, R. K. and Winsky, R. (eds.), Mouton de Gruyter, pp.1-28, 1997.

[16] *Information Society Technologies Workprogramme for 2003-2004*, European Commission, http://www.cordis.lu/ist

[17] Moore, G.E., "Progress in digital integrated electronics", *Proc. IEEE International Electron Devices Meeting*, pp.11-13. 1975.

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

[19] Keller, E. and Zellner-Keller, B., "Speech Synthesis in Language Learning: Challenges and Opportunities", *Proc. Workshop on Integrating Speech Technology in (Language) Learning (InSTIL2000)*, Dundee, Scotland, Aug. 2000.

[20] Keller, E. and Zellner-Keller, B., "New Uses for Speech Synthesis", *The AGORA Newsletter*, Vol.6, No.5, Summer 2000.

[21] "Vision of the Future", *Philips*, 1997. http://www.design.philips.com/vof/toc1/home.htm,

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

[23] Möbius, B., "Rare Events and Closed Domains: Two Delicate Concepts in Speech Synthesis", *Proc. 4th ISCA Workshop on Speech Synthesis*, Scotland, 2001.

[24] Huckvale, M., "Speech Synthesis, Speech Simulation and Speech Science", *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, Denver, Sept. 2002.

[25] Chen, F., "Issues in Speech Synthesis for tonal Languages", *IBM Research Report RC22389 (C0204-002)*, Apr. 2002.

[26] "DECtalk is the Most Intelligible TTS System in Noise", http://www.fonix.com/downloads/dectalk/manuals/dectalk.pdf

[27] Lindblom, B., "Explaining Phonetic Variation: A Sketch of the H&H Theory", *Speech Production and Speech Modeling*, Hardcastle & Marchal (eds.), Kluwer, pp.403-439, 1990.

[28] Young, S. J. and Fallside, F., "Speech Synthesis from Concept: a Method for Speech Output from Information Systems", *J. Acoust. Soc. Amer.*, Vol.66, No.3, pp.685-695, 1979.

[29] Cañamero, D., "Modeling Motivations and Emotions as a Basis for Intelligent Behavior", *Proc. Agents'97*, 1997.

[30] Schröder, M. "Emotional Speech Synthesis: A Review", *Proc. EUROSPEECH'01*, pp.561-564, Aalborg, Denmark, 2001.

[31] Walker, M. A., Whittaker, S., Stent, A., Maloor, P., Moore, J. D., Johnston, M. and Vasireddy, G., "Speech-Plans: Generating Evaluative Responses in Spoken Dialogue", *Proc. Int. Conf. On Natural Language Generation*, 2002.

[32] Maybury, M. T., "Multimedia Interaction for the New Millenium", *Proc. EUROSPEECH*, Budapest, 1999.

[33] Brennan, S. E. and Ohaeri, J.O., "Effect of Message Style on Users' Attribution Toward Agents", *Proc. CHI'94 Conference Companion Human Factors in Computing Systems*, ACM Press, pp.281-282, 1995.

[34] King, W. J. and Ohya, J., "The Representation of Agents: Anthropomorphism, Agency and Intelligence", *Electronic Proceedings CHI'96*, 1996. http://www.acm.org/sigchi/chi96/proceedings/shortpap/King/kw_txt.htm/

[35] Moore, R. K. and Morris, A., "Experiences Collecting Genuine Spoken Enquiries using WOZ Techniques", *Proc. 5th DARPA Workshop on Speech and Natural Language*, New York, 1992.

[36] "Assessment of Synthesis Systems", *The Eagles Handbook Of Standards And Resources For Spoken Language Systems*, Gibbon, D., Moore, R. K. and Winsky, R. (eds.), pp.481-563, Mouton de Gruyter, 1997.

[37] Vazquez Alvarez, Y. and Huckvale, M., "The Reliability of the ITU-T P.85 Standard for the Evaluation of Text-to-Speech Systems", *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, Denver, Sept. 2002.