ISCA Special Session: Hot Topics in Speech Synthesis

Gerard Bailly, Nick Campbell, and Bernd Mobius***

Institut de la Communication Parlee, INPG/Universite Stendhal, Grenoble, France
*ATR Human Information Science Labs, Keihanna Science City, Kyoto, Japan
**Institute of Natural Language Processing, University of Stuttgart, Germany
bailly@icp.inpg.fr, nick@atr.co.jp, bernd.mobius@ims.uni-stuttgart.de

Abstract

What are the Hot Topics for speech synthesis? How will they differ in 5-years time? ISCA's SynSIG presents a few suggestions. This paper attempts to identify the top five hot topics, based not on an analysis of what is being presented at current workshops and conferences, but rather on an analysis of what is NOT. It will be accompanied by results from a questionnaire polling SynSIG members' views and opinions.

1. Introduction

In a recent issue of the IEEE magazine Spectrum [1], there was an article by R. W. Lucky entitled "What were we thinking?", in which the author questioned the prevailing models of thought which led to the design of some major technologies. One example was 'video-on-demand', another, the mainframe computer. In the light of present knowledge, he argues that both were misguided because they were based upon principles, held to be fundamental at the time, but which were later replaced by unforeseen technological developments.

The key concept of video-on-demand was the central server, which provided all the information, and which would be managed by the resource holder. This became redundant with the growth of the internet and its distributed knowledge sources that could readily be accessed by the individual users without any central control. The mainframe computer, similarly, was replaced by a host of personal computers in a process that was almost completely unanticipated by the major companies. The prevailing 'logic' was that such distributed computing was 'wasteful of resources' and 'inefficient'. The article concluded with a question: What will people twenty years from now be looking back on with amusement? What aspects of the technology do we take so much for granted that they are blinding us from a vision of the way forward?

A recurring theme in the article is that of 'centralisation', which is contrasted with "distributed, user-empowered solutions"; i.e., how to avoid, or at least recognise, the monolithic concepts which for the engineers of the time seemed so basic and fundamental that 'they could no longer see the wood for the trees'. In the area of telephony, for example, the predominant thinking was of business use. Hence the adherence to wire, and a failure in many parts of the industry to anticipate the development of personal telephones with so much computing power, and add-ons such as cameras and flash memory, that have recently become so 'indispensable'. There was no concept at the time of a phone that could also be a computer, or a computer that could also be a phone, though in retrospect, the evolutional merging of these technologies seems both obvious and natural. However, the idea of 'phone-as-entertainment' was probably never seriously considered as an active research area.

The goal of this paper, and of SynSIG, is to question some of the fundamentals of speech synthesis so that we might become more able to make the imaginative jumps required to anticipate some of the future surprises. Ten years ago, for example, it was considered unthinkable to use raw waveform in any but the most primitive of 'station-announcement'-type synthesis systems. Now, the leading producers are all 'going concatenative'. But does this mean that we must throw away all the old thinking to jump on the concatenative bandwagon? Perhaps not. There is still, for example, much need for an understanding of the articulatory systems, though not perhaps in the way that the original researchers of those techniques anticipated.

What other forms of knowledge might we be leaving unexploited? What aspects of speech synthesis do we take for granted as unchangeable? Given the rapid rate of change of the technology, it would be foolish to attempt to predict what synthesis research will be like in ten years from now. Instead, we will limit our speculation to a five-year time-frame: What predictions can we make about the key research issues for the year 2008?

We will present some more complete answers to that question in the oral presentation of this paper at the ISCA Hot Topics Special Session, based on the results of a questionnaire circulated among speech synthesis researchers and SynSIG members. At the risk of being proven immediately wrong, and as a preliminary to that questionnaire, we present here as an example some of our own thoughts about the current state of the speech synthesis art, and some speculations on how it may evolve in the coming years.

2. A brief history of ISCA's SynSIG

The International Speech Communication Association [2] (ISCA) maintains a core of special interest groups (SIGs) to encourage interest and activity in specific areas of research by means of specialist workshops, special conference sessions, dedicated web pages, discussion and mailing lists. The SIGs are required to make available to the members and to ISCA text and speech corpora, analysis tools, analysis and generation software, research papers, and generated data.

The Special Interest Group on Speech Synthesis (SynSIG [3]) was the first ISCA SIG, founded in 1998, and arising from activities of the COCOSDA Working Group on Speech Synthesis, formed in 1991, at the Chaviari COCOSDA meeting after Eurospeech in northern Italy. SynSIG was joined shortly afterwards by AVISA [4], an ISCA SIG which

focuses on the closely-related area of audio-visual speech processing.

The primary goals of SynSIG are to identify the user needs of each community and to provide education, resources, reference materials, and training so that the science of speech synthesis can be encouraged to develop in directions that were not previously anticipated. The key engineering issues may already be being tackled by the manufacturers, but in times of increasing pressure on research funding, it is the duty of volunteer groups to provide the energy for more basic fundamental research, and for the integration of technologies that may otherwise be seen as unrelated.

The first international research workshop related to SynSIG activities was the Speech Input/Output Assessment and Speech Databases Research Workshop, held in 1989 at Noordwijkerhout in the Netherlands under the organisation of Prof L.C.W.Pols. This formed the impetus for COCOSDA, the International Coordinating Committee for Speech Input/Output Databases and Evaluation, which after an initial meeting in Kobe (following ICSLP-90) was formally initiated at the Chaviari meeting mentioned above. The Synthesis, Recognition, and Corpora working groups convened for the first time at the COCOSDA meeting following the Banff ICSLP in 1992.

The first 'ESCA Speech Synthesis Workshop' was held in Autrans, France, in September 1990, hosted by Benoit and Bailly, and was attended by more than 80 researchers from around the world. The '2nd Speech Synthesis Workshop' was held four years later at the Mohonk Mountain House in New Paltz, NY, USA, in September 1994. The 'ESCA/COCOSDA 3rd International Workshop on Speech Synthesis' was held at Jenolan, in the Blue Mountains of Australia in November 1998. The '4th ISCA Tutorial and Research Workshop on Speech Synthesis' was held at the Atholl Palace Hotel in Perthshire Scotland, August 29th - September 1st, 2001. The most recent meeting was an IEEE Workshop on 'Speech Synthesis' held as a satellite event of ICSLP 2002 at Santa Monica, CA, USA, (note: CD-Rom proceedings only – no page numbers! (the problem of page-numbering and future citations may become a hot issue, but not just for synthesis)).

As with COCOSDA, the facilities for SynSIG were hosted initially by the Advanced Telecommunications Research Institute (ATR [6]) in Japan, first by the Interpreting Telecommunications Research Labs (ITL), then by its successor the Spoken Language Translation Research Laboratories (SLT) which supplied web services, data storage, and a mailing list. Due to administrative changes at the ATR labs, the ITL domain no longer exists, and these services are now being transferred to a machine owned by ISCA, which has recently offered to provide computer resources for a new mailing list and for the more permanent archiving of synthesis samples, tools, components, and resources. Work has started on this new facelift for SynSIG, but there is still a need for more volunteer labour for the maintenance and upkeep of the facilities.

In addition to the mailing list and web pages, SynSIG maintains a list of references to publications related to speech synthesis, an Education page, and a Synthesis Evaluation site (in conjunction with the LDC [7]). New resources are always welcome, as are additions and updates to those already in use.

3. The Hot Issues in Speech Synthesis

There will surely be no unanimous agreement as to what are the top 'Hot Issues' in current speech synthesis research because there are almost as many different synthesisers and synthesis methods as there are different needs for synthesis itself. However, we tentatively put forward the following as the top-five hot issues, in the knowledge that the forthcoming results of our questionnaire may reveal a completely different set of issues than these we mention below:

We note first that *'Evaluation'* must appear high on any list. This topic is perhaps the thorniest of the synthesis issues. Undoubtedly, users need some form of comparison between systems, and some baseline reference by which to compare them, but since the needs of speech synthesis are so varied, the evaluation criteria must either be so general that they apply well to none, or so specific that comparison between different systems becomes infeasible. The Jenolan Workshop was devoted to this issue, but although all present considered it worthwhile to be able to compare current synthesisers on a common test-set of sentences, no tangible results (other than a realisation of the difficulty of the problem) remain from that experience. Whereas speech recognition results can to a certain extent be objectively measured, the multi-faceted and subjective experience of listening to speech synthesis requires a more complex and extensive set of standards and references.

Extension of synthesisers is the second current priority area. The recent surge in research activity arising from military needs in the Middle East has resulted in rapid development of devices for ‘interrogation and refugee assistance’, and multilingual speech synthesis was the subject of several research papers presented at the IEEE workshop. The porting of a synthesiser for use in another language is not a new issue, but sharing of components still remains an unsolved problem. Few systems are yet capable of a mix-and-match exchange of modules because of internal software dependencies and the lack of a standard interface specification between e.g., text-processing, prosody processing, and waveform generation modules. There is no consensus yet for an Open Standards Initiative in speech synthesis, perhaps because the industry is now in a competitive commercial rather than a basic research stage..

We place *Emotion* third on the list because of growing recent interest, but note here that ‘emotion’ may not be the best term for this genre, preferring instead *“Expression”* (there is a special session on the ‘Synthesis of Expressive Speech’ at the present conference), since it may be of more use for speech communication to model the paralinguistic rather than the extralinguistic states and events (i.e., the states and attitudes that a speaker intends to express or reveal, rather than the emotional states that the speaker is subject to at the time). Prosody control has long been an important issue within the field of speech synthesis, but as examples at the Santa Monica workshop illustrated, the prosody of many current synthesisers falls well short of being able to reproduce the variations required for emotional speech. The issue of voice quality control is also now arising as an area that needs to be addressed if synthesis is to express paralinguistic information. There is a Eurospeech 2003 Satellite Workshop and ISCA ITRW ‘VOQUAL’ devoted to this topic.

Multimodal aspects of speech synthesis is another area of fast recent growth. The integration of voice, gesture, eye gaze and facial expression reflects the fact that speech by itself is an impoverished form of communication (this in spite of the popularity of the telephone for both business and personal use) and that speech accompanied by visual information provides a more robust medium of expression, particularly in noisy environments. Moreover face-to-face communication is an essential means of constructing a mutual belief space between a conversational agent and the user. Certainly speech synthesis should benefit from the maturity of facial animation, which is already able to face Turing tests. We hope to see an integration of SynSIG and AVISA arising out of this development so that the visual and spoken information sources can be better merged in future technology.

Finally, we note that *Input* to the synthesiser is of vital importance. Raw text alone does not adequately specify the appropriate or intended paralinguistic or semantic-pragmatic interpretation of its linguistic content. VXML (voice XML) and similar mark-up descriptions will be increasingly necessary as the abilities of synthesisers develop in the direction of interactive speech. The term 'text-to-speech' is often thought of as synonymous with 'speech synthesis', yet this is but one example of synthesis applications. As synthesisers evolve from 'reading machines' to 'talking machines', there will be an increasing need to specify the intended interpretation of each utterance as part of its input description, perhaps even by voice. 'AI' is no longer a popular term, and 'text understanding' has shown distinct limitations in the area of conversational speech synthesis. Annotated input to a synthesiser would allow a finer specification of speaking style and of the intended interpretation of a message. Such input is a prerequisite for speech-to-speech applications where not only linguistic but also paralinguistic information should be translated and rendered properly. Proposals have already been put forward for menu-driven interfaces that allow for the switching of speaker, language, emotion, and speaking-style with automatic, or semi-automatic adjustments to the mark-up of the input.

4. Hot Topics for the Year 2008 (?)

While we believe that the above issues will still perhaps be as relevant in five-years time as they are now, we tentatively propose the following as the Hot Topics for the year 2008.

4.1. Entertainment

Entertainment is a major business area, and also the sink into which all grand ideas are swallowed. There is no common agreement as to who was the inventor of television, but it is doubtful whether he or she could have had the imagination to predict the current uses of that technology. The evolution of the television program has followed the path of maximum popularity, and we find that in almost every country the television is being used primarily for entertainment, with the priorities for education and information-provision being much lower than were originally imagined. So how could speech synthesis be used for entertainment of the masses?

There have been suggestions that the computer voice might trigger images or emotions in the listener as does classical music; and proposals for use of a voice synthesiser in yoga, or in art, especially music, where a voice parameters mixer (perhaps like a 70's Moog synthesiser) would be used in much the same way that Karlheinz Stockhausen used vocal tract configurations and the sounds resulting from these instead of raw moog tones in his compositions. Towa Tei might have set a precedent for this, with his use of CHATR's children voices in the 'house' CD album "Last Century Modern", released in 1999.

4.2. Extensibility

The ability to personalise a speech synthesiser may become an important requirement in the coming years. For example, to teach it to speak with your own voice or that of a well-known personality, to customise or select between different speaking styles, languages, moods, and emotions, and to program the synthesiser for specific messages, much as we now program telephone answering machines with personalised messages. With distributed synthesis services, the programming and the data need not be in the device, but on a remote server, and making use of several distinct modular components for message composition, translation, mood interpretation, voice-colouring, and affect-overlay. Wireless interconnection and remote messaging will reduce the device-dependency, allowing greater use of large databases (both of text and of voice) and of custom-modules for message-specific services. It would require only a standardisation of interfaces between the basic synthesis modules for various present synthesisers to be amalgamated, allowing the customer a free choice in selecting the design that best suits the needs of any occasion.

4.3. Expressiveness

We expect that as the output quality of speech synthesis improves, it will be required more often to replace the human voice in many everyday situations, and to express personality as well as content. Prosthetic devices, games, interpretation, customer-care, information services, remote messaging, robots, toys, and even home-automation; all require more than the simple imparting of novel information, and will need ways to interact with a human in both verbal and non-verbal forms.

If a synthesiser is to be used in seamless or unobtrusive conversational interactions with a human interlocutor, then there will be a need for the expression of personal attitudes, moods, and interest, and more use will be made of non-lexical sounds such as 'grunts', fillers, and laughter. In such cases, the key difference lies in the degree of interaction with the listener and reaction to the contexts of the discourse. Humans raise their voices both to show anger and to adapt to a noisy environment. They whisper when the content is confidential. Conversation is an interactive two-way process, with the listener also taking an active part in the discourse. The synthesiser that takes the part of a human will be required to express personal feelings and attitudes that are perhaps more in the domain of psychology than linguistics.

4.4. Education

We speculate that greater use of speech synthesis will be made in education and training, and particularly in foreign-language training, where the human learner has the natural ability to overcome the artificial failings of the synthesised speech. For example, a parametric synthesiser may have perfect prosody but terrible voice-quality, or a concatenative synthesiser perfect voice-quality but unnatural jumps in

prosody; both failings will be naturally overcome if a human tries to mimic the synthesized voice.

4.5. Elements

Moving away from engineering and application issues and back to basic science, we postulate next that the issue of basic units for speech synthesis will once again rise to the fore. Perhaps because of the influence of dictionaries in the front-end process, we tend to think first of phonemes as the basic units for speech synthesis, but the syllable, the articulatory gestural unit, and even the phrase have also been put forward as possible alternatives. If speech synthesisers are to be driven by voice input (e.g. using cheap labour to generate expensive entertainment) then the granularity of the unit that is used for selection may be better determined by spectral or articulatory characteristics than by phone-based definitions.

4.6. Evaluation

This important topic will remain hot for many years to come. The main focus of the Jenolan speech synthesis workshop was on the evaluation of current speech synthesisers. It was motivated by a reaction against the common practice, which is perhaps encouraged by the current format of scientific presentation of results, in which researchers show only the best performance of their synthesisers, in support of their proposed improvements, without revealing the true baseline performance of these systems. The workshop resulted in a better understanding within the industry (unfortunately, basic science is not yet well represented at synthesis workshops) of the strengths and weaknesses of the individual systems, although there was no dissemination of this understanding to researchers who were unable to attend the workshop because of a reluctance of many of the contributors to reveal the true performance of their systems on a common subset of test data.

4.7. Excitement

Finally, with a wish rather than a prediction; we call to mind the excitement that a parent has when hearing a child first speak. Or the excitement of hearing a new singer whose voice carries such feeling that the song is an instant hit. Wouldn't it be refreshing if the sound of a synthesized voice were to excite such emotions? Speech research in general will benefit from, and researchers should be more aware of, the cognitive research and neuropsychological studies that identify the circuits and processes involved in multimodal communication, that in turn trigger somato-sensorial stimuli and emotional content. We hope that the young researchers in this active field will take the initiative, making imaginative leaps rather than small steps, to lead the way forward to new uses and novel configurations of speech synthesisers that haven't yet been imagined.

5. Discussion

Returning to the IEEE Spectrum article mentioned in the introduction, Lucky concluded that "engineers like to build cathedrals' (citing Raymond [8]), but points out that 'the world often favours a bustling bazaar', noting especially that "the fact that large-scale evolvable systems can actually pull together instead of diverging into chaos is fascinating and conceptually important". We might infer that in our context of speech synthesis research, he would ask how the focus can be shifted from the monolithic to the distributed paradigm, and how distributed components might be integrated in a 'bottom-up' way, to produce a talking machine, the design of which is in the hands of the users, rather than the creators.

Is a speech synthesiser really too small a system to be evolvable in a distributed way? Perhaps so, if we limit ourselves only to the narrow definition of a speech simulator, but not at all if we consider the wider scope of applications that might use speech in an advanced media society. We hope that future researchers will not be blinded by the task constraints and limited assumptions of early synthesis designers, and that they will have the freshness of mind to see how spoken language is used in the wide variety of everyday interactions, and that within the next five years, there will be systems capable of calming a crying baby, shouting at a noisy child, chatting up an attractive other, apologising to an angry spouse, joking with a friend, evading a sales-person, reading to an elderly uncle, and singing to a grandmother. Many of our users want to do this, but no longer have control of their own natural voices. Many of our clients would be happy to have these abilities in their products. And it would be fun if we could do it!

We should begin to think not just about the languages and dialects that are the focus of current research, but also about the roles that the speaker might be required to play in a discourse, whether that speaker be a robot, a watch, a telephone, a car, or whatever. We have shown by our current technology that we are able to model speech as a part of 'language as system'; next, we might turn our attention to its role in 'language in use'.

6. Conclusions

This paper has presented a few subjective and probably biased opinions about the state of current and future speech synthesis. It does not aim to predict the future, but rather to be provocative in order to foster discussion about these and related issues in the hope that some original initiatives might be started a bit earlier than otherwise. The oral presentation of this paper will be accompanied with the results of a questionnaire polling the opinions of the wider community.

Acknowledgements

The authors would like to take this opportunity to remember Christian Benoit, the original motivator for the SIG activities, to acknowledge ISCA for encouragement, and especially to thank Valerie Hazan, the SIG Coordinator, for organising this special session at Eurospeech 2003.

References

[1] Lucky, R. W., *What were we thinking?*, p.56, IEEE Spectrum, Reflections, May 2003.
[2] ISCA: www.isca-speech.org
[3] SynSIG: www.slt.atr.co.jp/cocosda/synsig (will change)
[4] AVISA www.uws.edu.au/marcs/AVISA
[5] COCOSDA: www.cocosda.org
[6] ATR: www.atr.co.jp
[7] LDC: www.ldc.upenn.edu
[8] Raymond, E.S., *The Cathedral and the Bazaar,* O'Reilly, 2001.