The ISCA Special Interest Group on Speech Synthesis

Nick Campbell [1], *Wolfgang Hess*[2], *Bernd Möbius*[3], *Jan van Santen*[4]

[1] ATR Spoken Language Translation Research Laboratories, Kyoto, Japan
nick@slt.atr.co.jp
[2] Institut für Kommunikationsforschung und Phonetik, University of Bonn, Germany
wgh@ikp.uni-bonn.de
[3] Institute of Natural Language Processing, University of Stuttgart, Germany
moebius@ims.uni-stuttgart.de
[4] Center for Spoken Language Understanding, Oregon Graduate Institute, USA
vansanten@ece.ogi.edu

Abstract

This paper describes the constitution and activities of the ISCA Speech Synthesis Special Interest Group, SynSIG. It summarises past achievements and suggests ways in which future development could be maintained. The aims of the Special Interest Group on Speech Synthesis are to promote the study and diffusion of knowledge about speech synthesis in general, in a number of ways including: dedicated web pages, a mailing list, a bibliographic database, organisation of workshops on specific themes, exchange of students, and helping to co-ordinate sessions on speech synthesis in international conferences and workshops. The international and multi-disciplinary nature of the SIG also provides a means for diffusing information both to and from the different research communities involved in the synthesis of various languages.

1. Introduction

Speech synthesis has moved to the main stage of speech technology, sharing this stage with speech coding, speech recognition, and speaker identification. There has been significant increase in research activities as well as a significant improvement in quality in virtually all aspects of TTS. We also observe an increase in commercial applications of TTS. Special conferences and workshops have been organized (e.g., Autrans 1990, Mohonk 1994, Jenolan Caves 1998).

2. Motivation

In early 1998 Christian Benoît and Gerrit Bloothooft offered support for the creation of Special Interest Groups within (then) ESCA, the European Speech Communication Association. The idea was, and still is, to offer an organizational framework for any specific areas in speech communication that consist of significant groups of interested individuals. Just a few weeks after this initial call by ESCA, Nick Campbell of ATR suggested the formation of a special interest group in speech synthesis in order to take an active part in developing tools and resources for the synthesis community. His early initiative was shortly thereafter joined and intensified by members of the Bell Labs TTS group, and within a few days after the renewed call, 57 speech scientists representing a large number of institutions involved in speech technology and synthesis expressed their interest and support; 20 more individuals and institutions signed up for the SIG during the subsequent weeks.

The intended purpose of the proposed SIG was, according to its constitution, to promote interest in speech synthesis including text-to-speech conversion, concept-to-speech conversion and related disciplines; to provide members of ESCA/ISCA with a special interest in speech synthesis with a means of exchanging news of recent research developments and other matters of interest in speech synthesis; to sponsor meetings and workshops in speech synthesis that appear to be timely and worthwile; and to provide and make available resources relevant to speech synthesis, including text and speech corpora, analysis tools, analysis and generation software, research papers and generated data.

SynSIG, the ESCA (now ISCA) Special Interest Group on Speech Synthesis, was officially approved by the ESCA Board in August 1998. The original steering committee during the SynSIG formation phase was endorsed without formal elections at the Third International Workshop on Speech Synthesis in Jenolan Caves, Australia; since then the SynSIG officers have been: *Chair*: Wolfgang Hess, University of Bonn (Germany); *Vice-Chair and ISCA Liaison*: Nick Campbell, ATR (Japan); *Secretary and Treasurer*: Bernd Möbius, University of Stuttgart (Germany); *Liaison for Evaluation Issues*: Jan van Santen, CSLU/OGI (USA).

SynSIG made its official start on December 2, 1998, when a formal meeting was held at ICSLP-1998 in Sydney. The meeting was attended by some 60 researchers involved or interested in speech synthesis, who agreed upon the following activities.

Evaluation: The broadly stated goal of SynSIG's evaluation activities is that of stimulating evaluations that benefit science and help both individual and business consumers of synthesis to select systems that meet their needs. A preliminary statement and rationale for the evaluation activities of the SIG were made accessible online [1].

Collaborative experiments: develop and make available a setup and design for collaborative international and multilingual experiments.

Resources: disseminate or help disseminate scientific articles, software tools, speech corpora, and other relevant material, including a databank of references related to speech synthesis.

Teaching resources: collection and exchange of tools and resources for teaching purposes.

Student exchanges: organize the exchange of research students to encourage inter-lab communication.

Standardization: Effective communication should be established between the speech synthesis community and other groups and disciplines (e.g., speech coding), in particular in the area of standardization.

Membership is open to anyone with an interest in speech synthesis. SynSIG offers three levels of membership, depending on the individual member's level of interest:

Active members: members who take on an active role in advancing the goals of the SIG;

Helpers: members who agree to lend support to the steering committee on a specific scientific or organizational issue;

Listeners: onlookers and listeners.

Despite the fact that some 60 researchers attended the Sydney meeting, a significant number of people who had responded to the initial SIG proposal could not attend the meeting. Therefore, the steering committee was confirmed for one year. A nomination and election procedure for the future committee will soon be implemented on the SIG's web site. Nominations will be called for in the near future; all members of the SIG will be invited to propose suitable candidates and official elections will be held subsequently, possibly through a WWW interface.

3. Activities

3.1. SynSIG web site

SynSIG maintains a web site [2], also accessible through ISCA's home page [3], to enhance the exchange of news on recent research developments in speech synthesis and to make available relevant resources, in particular databases, corpora, tools, and reference lists. A mailing list (`synsig@slt.atr.co.jp`) was activated in January 1999.

Furthermore, a web page has been set up that is intended to hold contributed modules, interfaces, components and tools and may eventually evolve into a speech synthesis toolkit [4]. The web page calls for input from the speech synthesis community on such critical matters as inter-component interfaces and underlying data structures. It also points out the pioneering efforts by the Festival [5] and MBROLA [6] teams.

Some speech scientists have recently observed that speech synthesis research may be moving in a direction where different organizations specialize on different synthesis components, or modules. If this observation is valid, then it may be critical that there be agreement about how these components should communicate with each other. It may become essential to develop standards for how, for instance, text analysis modules should interface with prosody and how prosody should interface with signal processing.

3.2. Workshops

SynSIG has made and will make active contributions to a number of speech conferences and workshops as well as publications.

1. At Eurospeech-1999 in Budapest three oral (5 presentations each) and three poster sessions (about 15 presentations each) were devoted to speech synthesis or speech generation or both.
2. At ICSLP-2000 in Beijing three oral (7 presentations each) and one poster session (35 presentations) were devoted to speech synthesis or speech generation or both, and many more papers relevant to speech synthesis were presented in other sessions.
3. SynSIG co-sponsored the IEE event 'State-of-the-art Speech: Synthesis', organised by Justin Fackrell and PG E5, in April 2000 at Savoy Place, London.
4. The publication by Springer-Verlag of "Machines Talk: Updates on Speech Synthesis", based on the Third International Workshop on Speech Synthesis in Jenolan Caves, Australia, edited by Nick Campbell, Andy Breen, Jan van Santen, and Julie Vonwiller. There has been considerable delay in the publication of this book due to the fact that all four of the editors changed labs and became otherwise involved shortly afer the conference, but the chapters are now with the publisher and we expect to see the book on the shelves before the start of the next academic year.

Further suggested SynSIG activities include:

5. creating a standard system for marking up text input to speech synthesizers, e.g. a more elaborate version of SABLE [7];
6. getting actively involved in standardization efforts, e.g. XML, VoiceXML, or W3C voice markup languages;
7. establishing a common test-bed to evaluate the output of current speech synthesizers;
8. making available large speech corpora, especially those that are tagged in terms of discourse and prosody features;
9. making available tools and components to facilitate the development of different synthesis methods, following the lead of HTK, MBROLA, and Festival, to speed up the development of systems for new languages and different speaking styles;
10. encouraging research into different speaking styles in order to redefine the tasks expected of a speech synthesiser and to facilitate its development as a communication aid;
11. setting up an online database of papers and references;
12. setting up an online database of applications for speech synthesis and samples of their output;
13. encouraging the integration of speech synthesis research with that of graphic, gestural, and other modalities, in multimodal dialog systems.

3.3. Teaching

During the initial meeting of SynSIG at the ICSLP conference in Sydney 1998 it was suggested by several SIG members that we should devote some of our efforts to enhance the teaching activities at universities and other academic institutions. Although every lecturer has his or her own preferences of teaching style and material, the SIG's intention is to improve courses and

classes on speech synthesis by sharing syllabi, course material, tools, and experience in teaching speech synthesis.

A dedicated web site was set up to bring these ideas and intentions into existence [8]. Its content ranges from a list of academic institutions teaching speech synthesis to various types of material that is shared within the community. Several teachers and institutions have provided the slides that they use in their courses. One highlight is a long list of historical pictures showing important milestones in the development of speech synthesis, such as the famous Voder of Homer Dudley being presented at the 1939 World Fair in New York. Finally there is an extensive annotated list of links to software, tutorials and other resources pertinent to teaching speech synthesis.

The pages are far from being complete. Many more educators and researchers should contribute and provide their course materials. Many thousand downloads within a year show that there is a strong interest in sharing this type of information, especially when setting up a new course in the area of speech synthesis. Additionally, a crosslink with the new Special Interest Group in Education in the field of speech communication—EduSIG [9]—would bundle the efforts and should, therefore, be taken into consideration.

3.4. Evaluation

Until recently the only exposure participants of speech conferences and even speech synthesis workshops were given to TTS systems was in the form of prepared demonstrations, typically played from tape recorders, and it used to be very difficult to estimate the true quality of the systems. Therefore, a major effort was made at the most recent speech synthesis workshop at Jenolan Caves, Australia, in 1998 [11] to provide a presentation format whereby TTS systems were confronted with the same unknown textual materials, which covered newspaper text, semantically unpredictable sentences, and telephone directory listings. Text materials were created by standardized automated methods, based on text corpora owned by the Linguistic Data Consortium (LDC) with no ties to any particular TTS system. The text materials were unknown to the system developers.

For the *newspaper text* two selection methods were applied. The first method was based on word frequency and was intended to guarantee that all words in a selected sentence have a frequency of occurrence in the text corpora that is above a certain threshold. For TTS systems that rely on pronunciation dictionaries sentences of this type should present no major obstacle in terms of grapheme-to-phoneme conversion. The sentence may, however, have a complicated syntactic structure and thus challenge the prosodic components.

The second selection method for newspaper text used sequences of three orthographic characters ("trigrams") as the basic unit and selected sentences with a maximum diversity of trigrams, weighted by frequency of occurrence of these trigrams but without consideration of word frequency. This task was considered as challenging for several TTS components, including grapheme-to-phoneme conversion, acoustic unit selection and quality, and the prosodic components.

Construction of the *semantically unpredictable sentences* (SUS) followed the procedure proposed by Benoît [12, 13]. This method involves common syntactic structures that are paradigmatically filled with words randomly selected from special word lists. Examples for such sentences are, for English,

(1a) The chair ran through the yellow trust.

or equivalently for German,

(1b) Der Stuhl lief durch das gelbe Vertrauen.

This task is designed to primarily challenge the segmental intelligibility of TTS systems.

Subjects were asked to evaluate the systems by answering two types of questions. First, global judgments on dimensions such as naturalness or overall voice quality were provided on quasi-continuous rating scales from poor to excellent. Second, more fine-grained problem areas were rated, such as mispronunciations, wrong syllabic stress, bad durations, inappropriate sentence melody.

The evaluation procedure followed standard experimental designs [14] and had the following properties. To each listener for a given language, the same text items were presented. Each subject listened to each TTS system equally often. Across subjects, each TTS system was presented only once with each text item.

This design prevents as many learning effects as possible. It also provides reliable estimates of system performance in the statistical sense, provided that no interactions between the statistical factors subject and system exist. This turned out to be a theoretical consideration only, because the population of listeners was almost identical to the workshop participants, i.e. the system developers. Even if a conscious bias was avoided by the subjects, familiarity with their own system must have introduced an unavoidable bias.

As many as 68 TTS systems in 18 languages participated in the evaluation session in Jenolan Caves [10]: 16 systems for English (10 American English, 5 British English, 1 Australian English); 10 systems for German; 8 for Spanish (5 Iberian, 3 Mexican); 7 each for French and Japanese; 5 for Mandarin Chinese; 3 each for Dutch and Italian; 2 each for Catalan; 1 each for Basque, Galician, Korean, Portuguese, Romanian, Russian. Multilingual systems were presented by Bell Labs (9 languages), ETI-Eloquence (8), ATR-ITL (5), Telefonica (4), BaBel and OGI (3 each).

The fact that the synthesis researchers were also the evaluators was the one major shortcoming in the Jenolan Caves evaluation session, but it was unavoidable and deliberately taken into account. Given the number of languages involved it would have been a practically impossible logistic task to recruit a sufficient number of "naive" native speakers of all these languages. The procedure should therefore not be considered as a formal evaluation, and to reflect this informality it was decided not to publish system-specific results.

This drawback notwithstanding, the evaluation workshop has succeeded in a number of aspects. First and foremost, valuable experience has been gained on the methodology of speech synthesis evaluation. This judgment applies in particular to the methods used for the selection of the textual test material, and these methods have since been used also on the LDC's web server for the online comparison of TTS systems, which includes 19 TTS research and development sites and 13 languages [15].

Second, software tools for text selection and rule-based construction of test materials as well as the software for the web-based evaluation session has been developed and made publicly available [16]. It is worth noting that, in addition to these software tools, large annotated online text corpora, in conjunction with natural language and speech annotation tools are indispensable resources for text-to-speech evaluation tasks.

Given all these experiences and the practical achievements in terms of tools and software, there is no reason today why any research or development group working on speech synthesis should not offer an interactive, online, real-time demonstration of their TTS system for anybody interested to try out. Most end users are not in a position to conduct large-scale system comparisons. But even informal demonstrations on interactive web sites provide the potential user with a means of assessing and evaluating a system's performance on a task that matches the user's needs.

4. Future directions

While tests of segmental intelligibility provide invaluable information for the engineering of synthesis systems, they produce little information on the acceptability or appropriateness of a given voice or speaking style when it is generated synthetically. We expect that future evaluations will also take into consideration the needs of the task situation and the sensitivities of the listener in order to provide a measure of the acceptability of given synthesis methods and voices in the different application areas.

SynSIG will continue to encourage wider development of speech synthesis methods and applications, balancing the technology-driven advances with application-oriented, needs-based research. As with speech recognition technology, it has been found that the more a task can be defined, the easier it is to achieve truly high-quality performance. It is perhaps time to accept that the rendering of unlimited text into speech is a task that is difficult even for many humans to perform well, and to concentrate instead on providing a quality of speech synthesis that is acceptable to more listeners for domain-specific applications.

SynSIG needs input from a variety of sources in order to achieve its goals; not just from the engineers and scientists who are developing the technology, but also from the users, customers, and developers who have various needs for speech output in their daily lives.

Due to the high interest in this area, and the over-subscribed workshop at Jenolan Caves, SynSIG has a strong reserve of funds for encouraging future developments and facilitating exchanges of data and software. What the organisation needs now is information and advice on the directions in which these funds can best be applied. We look forward to hearing from you and being able to incorporate your contribution.

5. Conclusion

This paper has presented details of the activities of the ISCA Speech Synthesis Special Interest Group, SynSIG. The organisation is still relatively young but has already made several contributions to the various fields of speech synthesis research. A SIG depends on contributions from active and interested members of the community, and should be constantly renewing itself as trends develop and interests change. We encourage interested and active persons to become involved, to take the lead, and to organise more activities and better methods of intormation dissemination so that the science and industry of speech synthesis may be better informed and may work more closely in developing better methods of synthesis, better quality voices, and more varied speaking styles.

6. References

[1] ISCA Special Interest Group on Speech Synthesis, Evaluation issues update, 1998. Available online at [http://www.bell-labs.com/project/tts/ttsws98.html].

[2] ISCA Special Interest Group on Speech Synthesis, Home page, 2001. Available online at [http://www.slt.atr.co.jp/cocosda/synthesis/synsig.html].

[3] International Speech Communication Association, Home page, 2001. Available online at [http://www.isca-speech.org/].

[4] ISCA Special Interest Group on Speech Synthesis, Speech synthesis toolkit, 2000, Available online at [http://www.itl.atr.co.jp/cocosda/synthesis/toolkit.html].

[5] Festival speech synthesis, Home page, 2001, Available online at [http://www.cstr.ed.ac.uk/projects/festival.html].

[6] The MBROLA Project, Home page, 1999, Available online at [http://tcts.fpms.ac.be/synthesis/].

[7] SABLE: A synthesis markup language, Version 1.0, Available online at [http://www.research.att.com/~rws/Sable.v1_0.htm].

[8] ISCA Special Interest Group on Speech Synthesis, Teaching speech synthesis, 2001, Available online at [http://www.ims.uni-stuttgart.de/~moehler/ISCA-SynSIG].

[9] ISCA Special Interest Group on Education in the field of speech communication, Home page, 2001. Available online at [http://www.jiscmail.ac.uk/lists/isca-edusig.html].

[10] van Santen, J. P. H., Pols, L. C. W., Abe, M., Kahn, D., Keller, E., and Vonwiller, J., Report on the Third ESCA TTS Workshop Evaluation Procedure, in *Proceedings of the Third ESCA Workshop on Speech Synthesis (Jenolan Caves, Australia)*, 1998, 329–332.

[11] Breen, A., Campbell, N., van Santen, J., and Vonwiller, J. (eds.), *Proceedings of the Third International Workshop on Speech Synthesis (Jenolan Caves, Australia)*, 1998, Available online at [http://www.itl.atr.co.jp/cocosda/jenolan/index.html].

[12] Benoît, C., An intelligibility test using semantically unpredictable sentences: Towards the quantification of linguistic complexity, *Speech Communication*, 1990, 9:293–304.

[13] Benoît, C., Grice, M., and Hazan, V., The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences, *Speech Communication*, 1996, 18:381–392.

[14] van Santen, J. P. H., Perceptual experiments for diagnostic testing of text-to-speech systems, *Computer Speech and Language*, 1993, 7:49–100.

[15] Linguistic Data Consortium, Interactive speech synthesizer comparison site, 2001. Available online at [http://www.ldc.upenn.edu/ltts/].

[16] van Son, R., Evaluation software, 1999, Available online at [http://fonpc18.hum.uva.nl:4711/].