WHERE IS THE INFORMATION IN SPEECH? (and to what extent can it be modelled in synthesis?)

Nick Campbell
ATR Interpreting Telecommunications Research Labs.
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02, JAPAN
nick@itl.atr.co.jp, www.itl.atr.co.jp/chatr

ABSTRACT

Different kinds of speech information are meaningfully used in human communication. This paper attempts to show how they can be modelled in speech synthesis and suggests that many conventional synthesis methods may fail to take into account the subtleties of human speech variation. It argues that modelling of voice quality should be the next main goal for speech synthesis technology, and proposes that evaluations of synthesis technology should aim to include a Turing component, which measures the ability of each system to perform on a range of human-speech features.

1 INTRODUCTION

When two or more people talk together, they exchange more than simply linguistic information. Even over a telephone, the listener can estimate the speaker's mood, sincerity, urgency, etc., in order to form an interpretation of the intended meaning of an utterance.

Speech synthesis attempts to model the relevant information in human speech in order to produce intelligible output for human listeners, but implicit in the design of current speech synthesisers are many assumptions about the needs of synthesis and about which parts of the speech signal are to be considered relevant. This paper argues that whereas early speech synthesisers were designed primarily to be 'reading machines', with the emphasis on intelligibility, the coming generation should be thought of as 'speaking machines' instead, with more emphasis on naturalness of speech production. This change however, with its requirements for more 'spontaneous' and interactive speaking styles, will place heavier demands on the synthetic voice quality than the original reading task required.

When the task of synthesis was limited to rendering unknown text, the speech needed to be intelligible but not necessarily natural, and in particular it was not expected to have a personality or to show emotions or humour. However, with the growth of applications in interpreting telecommunications and information technology, especially with the recent proliferation of internet information, there is an increasing need for more expressive voices which can differentiate mood as well as signify content.

Because of the changing nature of the texts to be rendered in speech, simply reading from top to bottom may no longer be appropriate (the 'bottom' of an html document can be very difficult to find!) and an interactive question-and-answer style of accessing their content may prove necessary. The input requirements for such forms of interaction are already being investigated, and texts are being annotated with markup to indicate content and style, but the voice quality of many current synthesisers may not be appropriate for such casual interactive use.

2 READING & SPEAKING

T. P. Barnwell [1], a student of Dennis Klatt's at MIT almost 30 years ago described speech synthesis technology in the context of 'reading machines'. Allen [2] and Klatt himself [5] were at the time tackling the problems of synthesis from unrestricted text, and ten years later suggested ideas for text-to-speech synthesis [3] which developed into MITalk [4] and set the standards for many to follow. It wasn't until 5 years later, in 1992, that we started to hear of 'Talking Machines' [6] after the first ESCA Speech Synthesis workshop at Autrans.

Yet even today, in Jenolan, more than 25 years after the start of electronic speech synthesis, where we finally have the first common synthesis evaluation taking place, we are still in danger of perceiving the speech synthesis process as one of mechanical 'reading'. The 'speaking' aspect of synthesis has tended to be approached more from the signal [7, 8, 9, 10] than from a dialogue [11, 12] or emotional [13, 14] point-of-view. As a result, there has been a strong bias towards segmental intelligibility [23, 15, 16, 17, 18] rather than naturalness [24] or voice quality [25].

2.1 Media Conversion

Text is a two-dimensional medium, but speech is one-dimensional. When composing a text, the author has past and coming paragraphs to consider, and carefully constructs the sequence of words in the knowledge that the reader will typically look at more than one line at a time, scanning back and forth over the page of text to absorb the information. Very few authors compose written texts to be read aloud and, as a result, the written sentences tend to be long and convoluted, with syntax and layout performing the work of prosody.

Even something as apparently simple as a directory listing (in UNIX or DOS) makes use of the visual element:

```
bash$ ls /var
./          adm@        locks@      pid/        rwho@
../         lib/        log/        preserve/   spool/
X11R6/      lock/       man/        run/        tmp/
```

The alphabetical ordering is in columns from top to bottom of the page, but the text itself is generated and printed to the screen from left-to-right, in rows, which if rendered directly into speech, would lose much of their visual information.

2.2 Talking

Speech, on the other hand, has evolved for more direct forms of communication, with the full bandwidth of the audio channel available. Its information includes not just segmental details about the text, but also much about the speaker. Age, sex, health and well-being are signalled, as well as attitude, mood, and focus or 'intention'. If we compare the same content, for example a news item, in its read form, its formal spoken or broadcast form, and its informal conversational forms, differences are obvious not only in lexis, word-order, chunking and prominence relations, but also in the mood of the speaker and in the tone of voice.

The brunt of our technology so far may have been devoted to an unnatural task: that of converting between the media without apparent loss of information; but this is a task that is probably difficult even for most humans to do well. So perhaps it is time to start viewing synthesis as a component in systems for 'talking' or 'speaking *about*' information rather than simply for rendering text through the medium of speech?

Text to speech conversion requires more than just reading, but the technology for text explanation is still in its infancy. We can learn much from the information retrieval and text abstracting disciplines, but we must map from a text-based format to a speech-friendly format ourselves. But when that techology is developed, will the synthesised voices be ready for the conversational mode?

3 VOICE FONTS

Speech recognition is still very far from speech understanding. Speech, or spoken language, encodes much more information than just the word sequence, so if synthesis is to become closer to speech, then what are the attributes that need to be modelled?

Apart from the known prosodic aspects, one apparent difference between human speech and speech synthesis is laughter. We use laughter often in social interaction, to reduce tension and express pleasure. Other non-speech sounds such as clucking of the tongue, smacking of the lips, tutting, and inhalation of breath, are similarly used for meaningful effect. Although frequent in speech, very few of these para-linguistic signals are in the repertoire of most synthesisers.

3.1 Engines & Data

By effectively labelling the relevant features in speech, large-corpus concatenative synthesis techniques ('Chunk'n'Chink" [20]) move the speech knowledge out of the synthesiser and into the data. Through an indexing of the segments in a natural speech corpus they enable selection of (in the extreme case, raw) waveform samples that can be concatenated to form novel utterances. Such corpora can contain many kinds of speech noise.

The separation of speech-knowledge from generation technology results in generic, multi-speaker, multi-language, synthesis engines, but at the cost of extremely large source-data requirements. However, because most of the knowledge about the speech is encapsulated in the labelling, the engine is reduced to a simple index-and-retrieval system.

The question of an optimal design for such synthesis unit inventories can be rephrased as a question concerning the adequate representation of speech for synthesis. This is obviously dependent on the uses to which the synthesis is to be put, and we should perhaps distinguish between those cases where the synthesis is used to represent the output from a mechanical system and cases where it is used to represent the speech of a human being. In the former, a mechanical and unexpressive voice quality may be preferred, but these may be quite undesirable characteristics in the latter[1].

3.2 Mood & Personality

The development of switchable databases, or 'voice fonts', marks the beginning of control over the third

[1]It is a matter of opinion whether mechanical-sounding synthesis is even desirable for 'system output' but that point will not be addressed further in this paper, except to say that it may account for the lack of acceptance or widespread use amongst the general public for synthesis technology

main area of speech information: phonation style, or personality and mood. Phonetic production and prosodic variation, are already well controlled in synthesis, but not enough is yet known about the differences in perceived meaning or intention when the same word sequence and prosodic contours are realised in e.g., a harsher or breathier speaking style to express approval or discontent.

Iida [27] has shown that listeners can distinguish at levels significantly better than chance between different emotions in speech synthesised by concatenating segments taken from speech databases having different emotional characteristics, even when the text content and prosody of the utterance is emotionally neutral. This confirms a perceptually relevant element in the acoustics of the speech segments that is independent of phonemic or prosodic attributes.

Trends in concatenative synthesis [28, 29, 30, 31, 32] are towards bigger unit databases, shifting the knowledge from the synthesiser into the data. By accessing the data directly, rather than modelling its variation by rule, we open up the possibility of encoding richer types of speech variation, such as dialectal, emotional, or laryngeal differences, without requiring explicit control over their realisation.

CHATR now has more than a hundred voices, in six languages, that can be accessed by the same method of indexing and retrieval. Advances in prosodic labelling and direct feature-based unit selection [33, 34] enable us to almost eliminate the prosody-prediction modules which were its most language-dependent part. The knowledge about the speaker and the speaking style is encoded directly in the speech corpus, accessible via the index, and with appropriate input can be reproduced for concatenation into novel utterances. The cost in memory size is being more than matched by advances in the hardware.

4 EVALUATION

Evaluation should be classificatory as well as quantitative, taking into account usability criteria as well as effectiveness measures. A recent report on Industry Standard Usability Testing [35], based on an ISO standard [36], suggests criteria for measuring effectiveness, efficiency, and satisfaction for a specific context of use.

Tests of human speaking ability are often based on the degree to which people can render a range of different text types intelligible, but tests of speech performance from a computer need not simulate those we use for humans. The Jenolan synthesis evaluation was limited to 'TTS' systems and developers, but rendering text will probably not be the main use for synthesis technology in the future. Perhaps instead of a 'bake-off' where all synthsisers compete on identical tasks, assuming a single common usage, we might benefit more from designing evaluation methodologies that encourage each system to clarify its individual strengths and define its preferred areas of application, placing more emphasis on differences of use than on similarities.

In humans, the ability to use a range of speaking styles to convey mood and affect is developed almost from birth, and tested daily in interpersonal interactions. Speech synthesis used to represent human speech, whether as a prosthetic or in a commercial application such as telephone-shopping or interpreting telephony, should be tested for its ability to represent the full range of human communication and not just on ability to convey the linguistic aspects of the message. If future synthesis evaluations are to grade voices, perhaps we should start by evaluating the extent to which those voices can portray the depth of subtlety that a human is capable of. This would be a test not just of the synthesis algorithms, but also of the adequacy of database, its labelling and the input specification.

The Turing test may ultimately be the best form of evaluation for some forms of speech synthesis. If a human listener believes that another human is speaking, then the system can be said to have passed this test. Many present synthesis systems might pass such a test if the amount or type of speech could be constrained, but probably none would be able to exceed even a minute of free conversation. So the more interesting question for the current technology is: what limitations can we reasonably put on the Turing test to make it a useful measure of synthesis quality?

5 CONCLUSION

This paper has presented a personal view of the current needs of speech synthesis and suggested some directions for future research and evaluation.

One of the goals of the Jenolan Synthesis Workshop was to encourage standards for the evaluation of speech synthesis and synthesis systems. An announcement related to the Jenolan evaluation stated that "in the typical case the many voices all ride on top of exactly the same software, and hence are *not really different after all*" (my italics). With more than a hundred speech corpora processed for CHATR, no two voices are the same. Judging them is like judging people; we can favour one voice over another but we cannot say that A has a 'better' voice than B.

CHATR is clearly not the 'typical case', but we claim that the ability of a synthesiser to produce an utterance in a voice with a particular quality, or with a speaking style that matches a particlar content, should be considered as important for successful communication (and perhaps as difficult a problem) as the conversion of text into speech.

References

[1] T. P. Barnwell. An algorithm for segmental durations in a reading machine context. Technical Report 479, MIT, Research Laboratory of Electronics, 1971.

[2] Allen, J. (1973), "Speech synthesis from unrestricted text", In: J.L. Flanagan and L.R. Rabiner (Eds.), Speech synthesis, Dowden, Hutchinson & Ross, Inc., Stroudsberg, Penn., 416-428.

[3] Allen, J. (1981), "Linguistic-based algorithms offer practical text-to-speech systems", Speech Techn. 1, nr 1, 12-16.

[4] Allen, J., Hunnicutt, M. S. & Klatt, D.H. (1987), "From text to speech. The MITalk system", Cambridge University Press, Cambridge UK, 216 pages.

[5] Klatt, D.H. (1987), "Review of text-to-speech conversion for English", J. Acoust. Soc. Amer. 82(3), 737-793.

[6] Bailly, G. & Benoit, C. (Eds.), (1992), "Talking Machines, Theories, Models and Designs", Elsevier Science Publishers B. V., North-Holland.

[7] Rabiner, L.R. (1968), "Speech synthesis by rule: An acoustic domain approach", Bell System Techn. J. 47, 17-37.

[8] Coker, C.H. (1973), "Synthesis by rule from articulatory parameters", In: J.L. Flanagan and L.R. Rabiner (Eds.), Speech synthesis, Dowden, Hutchinson & Ross, Inc., Stroudsberg, Penn., 396-399.

[9] Klatt, D.H. (1980), "Software for a cascade/parallel formant synthesizer", J. Acoust. Soc. Amer. 67, 971-995.

[10] Pascal, D. & Combescure, P. (1988), "Evaluation de la qualite de la transmission vocale", L'Echo des Recherches, No. 132, 31-40.

[11] Delogu, C., Di Carlo, A., Sementina, C. & Stecconi, S. (1993), "A methodology for evaluating human-machine spoken language interaction", Proc. Eurospeech'93, Berlin.

[12] Morton, K. & Tatham, M. (1993), "Speech synthesis in dialogue systems", Proc. Eurospeech'93, Berlin, 905-908.

[13] Cahn, Janet E. Generating Expression in Synthesized Speech. Master's Thesis, Massachusetts Institute of Technology. May, 1989.

[14] Cahn, Janet E. The Generation of Affect in Synthesized Speech. In Journal of the American Voice I/O Society, Volume 8. July, 1990. Pages 1-19.

[15] Pols, L.C.W. (1991), "Quality assessment of text-to-speech synthesis-by-rule", In: S. Furui & M.M. Sondhi (Eds.), Advances in Speech Signal Processing, Marcel Dekker Inc., Chapter 13, 387-416.

[16] Greenspan, S.L., Bennett, R.W. & Syrdal, A.K. (1989), "A study of two standard speech intelligibility measures", J. Acoust Soc. Amer. 85, Suppl. 1, S43 (A).

[17] Picone, J., Goudie-Marshall, K.M., Doddington, G.R. & Fisher, W. (1986), "Automatic text alignment for speech systems evaluation", IEEE Trans. ASSP-34(4), 780-784.

[18] Pols, L. C. W. & SAM-partners (1992), "Multi-lingual synthesis evaluation methods", Proc. ICSLP 92, Vol. 1, Banff, Canada, 181-184.

[19] Benoit, C. (1990), "An intelligibility test using semantically unpredictable sentences: Towards the quantification of linguistic complexity", Speech Communication 9(4), 293-304.

[20] Gerard Bailly, personal communication.

[21] Campbell, W. N., "Foreign-language Speech Synthesis", this volume.

[22] Boxelaar, G.W. & Pols, L.C.W. (1985), "State-of-the-art report about intelligibility evaluation of speech coding and text-to-speech synthesis systems", IFA report 79, 20 pag.

[23] Carlson, R., Granstrvm, B. & Nord, L. (1990), "Segmental intelligibility of synthetic and natural speech in real and nonsense words", Proc. ICSLP'90, Kobe, Vol. 2, 989-992.

[24] Espesser, R., Rossi, M. & Pavlovic, C.V. (1988), "The relationship between acceptability, intelligibility, and naturalness in text-to-speech synthesis systems", Travaux de l'Institut de Phonetique d'Aix 12, 89-103.

[25] Carlson, R., Granstrvm, B. & Karlsson, I. (1991), "Experiments with voice modelling in speech synthesis", Speech Communication 10, 481-490.

[26] Pratt, R.L. (1986), "On the intelligibility of synthetic speech", Proc. Inst. of Acoustics, Vol. 8, Part 7, 183-192.

[27] Iida, A., Campbell, N., Iga, S., Higuchi, I, & Yasumura, M., "Acoustic nature and perceptual testing of a corpus of emotional speech", Proc ICSLP-98, forthcoming.

[28] Olive, J.P. (1977), "Rule synthesis of speech from dyadic units", Proc. IEEE-ICASSP77, 568-570.

[29] Olive, J.P. (1980), "A scheme for concatenating units for speech synthesis", Proc. IEEE-ICASSP80, 568-571.

[30] Sagisaka, Y. (1988), "Speech synthesis by rule using an optimal selection of nonuniform synthesis units", Proc. IEEE-ICASSP88, 679-682.

[31] Dutoit, T., Leich, H., "MBR-PSOLA : Text-To-Speech Synthesis based on an MBE Re-Synthesis of the Segments Database", Speech Communication, Elsevier Publisher, November 1993, vol. 13, n?3-4.

[32] Campbell, W. N., "CHATR: A High-Definition Speech Re-Sequencing System", Proc 3rd ASA/ASJ Joint Meeting, 1223-1228, Hawaii, 1996(12).

[33] Silverman et al. (1992), "TOBI: A standard for labeling English prosody", Proc. ICSLP'92, Banff, 867-870.

[34] Campbell, W. N., Processing a speech corpus for CHATR synthesis, Proc ICSP-97, Korea.

[35] Bevan, N., Report on "Usability as Procurement Criteria for Software Workshop", organised by NIST, Gaithersburg, 9-10 September 1998.

[36] ISO 9241-11 (1998). Guidance on usability. ISO (Available from national standards bodies. Contact details at http://www.iso.ch/addresse/membodies.html)