Speech Output Systems Assessment: Following the Jenolan Synthesis Evaluation Workshop

Nick Campbell
ATR Interpreting Telecommunications Research Labs
Kyoto, 619-02 Japan. (nick@itl.atr.co.jp)

Abstract

In November 1998, COCOSDA, in conjunction with ESCA, organised a Synthesis Workshop where the focus was on the evaluation of speech synthesis systems. This paper reports on that workshop and suggests some conclusions that can be drawn from the experience. It proposes that whereas this kind of on-site perceptual evaluation of synthesis can be extremely useful, there is a need for Internet-based tools and procedures for automatic or distributed methods of evaluation.

A subsidiary theme of the paper is the evaluation of large-corpus concatenative speech synthesis, which presents a special case since the use of unmodified speech waveform segments can render voice quality identical to that of the source speaker, but can also result in noise at segment boundaries and distorted prosody if appropriate units are not selected.

1 Introduction

A grep of the still incomplete COCOSDA speech-synthesis references page [1] yields 153 papers having the words 'evaluation' or 'assessment' in the title (e.g. [2-8]). There are many more papers referenced that are implicitly devoted to the evaluation of speech synthesis quality. It is a subject rich in both commercial and research interest.

The ESCA/COCOSDA Speech Synthesis Workshop held last year at Jenolan Caves near Sydney in Australia was the third in the ESCA Tutorial-Workshop series on speech synthesis, but was the first to replace the traditional 'Tutorial Day' with an 'Evaluation Day'. This change reflects both the maturity of speech synthesis as a discipline and the desire of researchers working in the field to have a common ground for comparing the performance of speech synthesis systems. That the workshop was 50% oversubscribed can be taken as a further indication of the interest in this topic. The Jenolan Evaluations will be reported in full in the forthcoming book from the workshop, so only highlights of the discussion will be summarised briefly below (but see also [9, 10] for further comments and analysis).

The rest of this paper will explore some of the ideas arising out of that workshop, and will present some suggestions for formalising and automating further evaluations, with particularly focus on large-corpus concatenative synthesis systems. It ends with a call for contributions of tools and procedures that can be made more widely available through the COCOSDA and ESCA-SynSig web pages to facilitate the testing and comparison of speech synthesis systems and components.

note: There is continuing debate in COCOSDA about the distinctions in meaning between the words 'assessment' and 'evaluation' but they will be used interchangeably in this paper.

2 The Jenolan Evaluation

Details of the assessment procedure are available on the Internet at [10] and were presented at the May 1998 meeting of Oriental COCOSDA in Tsukuba [11].

The goals of the Jenolan Evaluation (summarised from the post-workshop COCOSDA web pages) were:

1. to obtain a thorough and honest impression of the current state of the art,
2. to provide feedback to system developers and researchers about their systems,
3. to encourage "honest demonstrations" of current speech-synthesis quality, and
4. to do this under the structured format of a formal evaluation in order to learn more about the practical issues of speech synthesis evaluation.

i.e., it was not our goal to decide 'Which system might be "best"', but rather to explore 'Which approaches to which TTS functions look promising'.

Since the exercise did not meet the criteria for a formal evaluation (i.e., the listeners were unrepresentative, potentially biased, and too few), it was a condition of the evaluation (announced beforehand) that no public disclosure would be made of individual results that might identify the systems concerned. The effect of the evaluation was to increase the awareness

of the participating members and to encourage discussion of future assessment techniques. The experience of listening to so many different systems under identical test conditions was certainly a beneficial one, appreciated by all who took part.

The evaluation was open to full text-to-speech systems only, and was limited to structured, formal, reading-style, tasks (i.e., telephone listings, newspaper texts, and semantically-anomalous sentences). It was agreed at the workshop that future evaluations should include concept- and document-to-speech, as well as dialogue speech, rather than limiting the synthesis to just 'readings' of disjoint sentences. The use of relatively context-independent text data does test many aspects of (particularly segmental) intelligibility, but may not be representative of the majority of tasks required of speech synthesisers in the coming decade.

It was also agreed that in future not only 'whole-system' but also 'individual component' evaluations should be performed, so that modularity can be encouraged (once agreements can be reached on the necessary per-component I/O standards) and to facilitate further inter-organizational collaboration on the development and testing of integrated and distributed synthesis systems. There were several experimental synthesis systems that could not take part in the Jenolan evaluation because they lacked one or other essential component.

Unfortunately, there was insufficient time allowed for on-site processing of the data files into test sets for the evaluation, so not all systems were able to demonstrate the full extent of their capabilities, in spite of having submitted the necessary speech files in the required formats. With respect to availability and preparation of the data, therefore, future evaluations will probably require that all speech files be made accessible over the Internet (or will produce a CD-ROM set containing the speech files and software) for independent replication and validation of the structured listening experiments.

Even wider use may be made of the Internet, to carry out the listening tests themselves. Recent developments eliciting web-based responses for perception experiments [12] have received extremely high returns (numbering several thousand) and indicate promising results with respect to statistical reliability. The current LDC synthesis evaluation site for example [13] could be extended in this way for further testing and comparison of synthesis results, once controls have been established for verification of the responses collected over the Internet.

3 Concatenative Speech Synthesis

There is a growing interest in the use of large corpora for speech synthesis. For example, the ASA-EAA joint meeting in Berlin earlier this year [14, 15] included three sessions on data-based speech synthesis (1pSCb, 2aSCa, 2pSCa: 35 papers in all). Many of the systems presented in these sessions used 'unit selection' methods to determine an optimal sequence of speech sounds from a natural-speech source corpus, concatenating them with minimal signal processing to produce the required synthetic speech.

With the development of such concatenative synthesis methods, new criteria will be needed for the assessment of the resulting synthesis. Previously, only two dimensions, 'naturalness' and 'intelligibility' were tested, but naturalness of the speech quality in some concatenative systems can almost be taken for granted, since the original voice is re-used untouched. A third dimension of evaluation needs to be established: 'Suitability', but the criteria by which to judge the suitability of e.g., speaker selection, speaking style, tone-of-voice, or emotional colouring, require more information to be given about the context and situation of the utterance than is normally provided for 'reading-machine' tests of synthesis.

With the measurement of suitability as a goal for future evaluations, we can concentrate as a present task on facilitating and automating measures of naturalness and intelligibility. For example, signal-based characteristics should be amenable to objective measures of quality; the effect of noise at concatenation boundaries, resulting from sub-optimal unit joins, can perhaps be evaluated in this way, but not the effects of deviation of prosodic contours from their predicted targets.

3.1 Signal Quality: noise

There are two main causes of distortion to the signal quality in concatenative speech synthesis: 'warping' of the speech due to signal processing applied to modify the prosody, or 'noise' arising from poorly-aligned waveforms or parameter discontinuities at the unit boundaries.

Warping may require extensive human listening to assess, for while some methods of signal modification are perceptually less noticeable than others, the effects of most can vary from speaker to speaker and according to degree (and direction) of change. Furthermore, the degree of tolerance to degraded voice quality is also very application-specific.

Concatenation noise, however, can be measured by objective distance measures if the thresholds and

correlations with human perceptual scores can be determined. Figure 1 shows results of a MOS (mean opinion score) evaluation of CHATR [16] using different methods of signal processing. With such test scores as targets, we have been able to find correlations between physical characteristics of the speech signal and the subjective MOS scores, enabling prediction of the latter from the former (see [17, 18] for further discussion).

3.2 Prosody & Intonation

The difficulty in evaluation of prosody in synthesised speech lies in the discrimination of 'intended meaning' for an utterance, since there can often be a large number of different prosodic realisations which although equally 'natural' may signal different and possibly misleading interpretations for a given word sequence. This problem can be illustrated by 'focus' differences: a single sentence in isolation (e.g. "I'll call you on Wednesday") can carry focal prominence on potentially any word (in this example, even on the preposition, as in the case of "I'll call you 'on', but not 'before', Wednesday") depending on the intended meaning of the utterance in context. Without a clear disambiguating context to signal the intended variant (as provided by a preceding question, a subsequent reply, a clarification clause, etc.,), the listener trying to evaluate the prosody may be unable to judge the suitability of the realised interpretation. The synthesised utterance may be perfectly intelligible, and may also sound completely natural, but at the same time the prosody may be inappropriate to convey the intended meaning of the utterance in context, and therefore wrong.

Physical measures of prosodic suitability can be difficult because an appropriate response in a different pitch range may result in a large difference in absolute values (e.g., of fundamental frequency, power, or duration) when to a human listener the prosody is functionally equivalent. We as humans have no difficulty in finding the equivalence in meaning between high- and low-pitched speech (e.g., between children and adults) in spite of the real physical differences.

3.3 Speaker Characteristics

Evaluating the suitability of a speaker or a voice-quality can also be subjective and task-dependent. As with the judgement of beauty, different people can have very strong feelings for or against a particular characteristic as a matter of personal taste without being able to justify or explain their reactions.

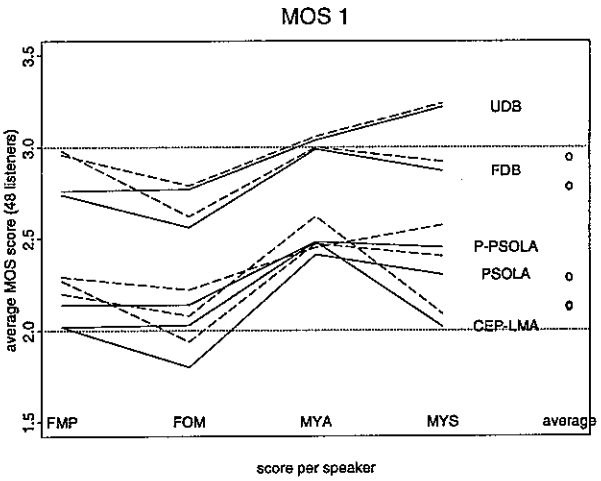

Fig. 1 Results from a MOS Test of 4 voices (2 male, 2 female) from CHATR. 12 sentences were synthesised using 5 different methods and evaluated by 48 naive listeners. The dotted lines show the scores of a second test, circles show average scores. Methods UDB and FDB use no signal processing.

In order to assess the suitability of a voice for a given task, we therefore need to poll many individuals in order to form a comprehensive opinion. However, because the use of human subjects in controlled perceptual tests can be both time-consuming and expensive, we need to develop methods of processing results from large numbers of less controlled responses instead. These can easily be collected using the Internet.

There is a lack of consistency in Internet-based response collection, due to the fact that some individuals may respond many times, and due to the widely differing subject backgrounds, equipment and conditions under which the tests are 'administered'. For this reason, there has been reluctance to use such methods, but as with large-corpus analysis, techniques of filtering, normalisation and validation must be developed.

4 Use of Non-speech Corpora

The potential of a speech corpus for synthesis can be measured off-line by comparing the distribution of speech units contained in the corpus with the distribution of sounds in a corpus of data representative of the language or tasks required of the synthesiser.

We have performed experiments based on a million-word corpus of written texts, using the pre-processing modules of the synthesiser to obtain statistics on the distribution of speech sounds (and their prosodic characteristics) in order to validate natural-speech corpora in terms of appropriate coverage [19]. By

incorporating prosodic annotations as part of the phone text sequence, we enable both variabilities to be calculated simultaneously. Using likelihood-based functions to compute the probable collocations, we are then able to predict weaknesses in the corpus without the need for extensive synthesis of speech.

5 Discussion

This paper has presented the need for objective methods of quality assessment and for high-volume collection of subjective scores. Many laboratories around the world are working on similar tasks for the evaluation of speech synthesis systems, but the needs of each are subtly different. In the past, these differences have hindered the co-operative co-ordination of assessment between different research labs (with the notable exception of the French under the AUPELF initiative), but with the ESCA-COCOSDA workshop, significant barriers to mutual development have been overcome. We have founded a set of evaluation procedures that, however primitive, will form the basis for further mutual evaluations.

The next step is to set up facilities for the sharing, testing, and development of tools and resources for component-based evaluations. Following from this will come the development of co-operative interfaces. There has in the past been little incentive for a synthesis developer to open up interfaces between component modules, but the recent development of open-architecture systems such as Festival and MBROLA means that more work can be performed without the need for full-systems to be developed for a given language. Databases and components are being shared, and the development of a synthesiser for a new language can now be completed in a matter of weeks. Testing it may take longer.

6 Conclusion

Without measurements, research cannot proceed in a scientific manner, but in the case of speech synthesis the question of 'what to measure' is still open. Intelligibility and naturalness form the two main criteria for judgement, but in the case of concatenated speech the naturalness of the voice can be extremely high, whereas the naturalness of the utterance can still be difficult to judge.

There is a practical need to establish common standards of reference and quantitative comparison. This can be done most efficiently through use of Internet-based facilities. The role of COCOSDA is to promote international cooperation and to further the coordination of research and development by offering such facilities and know-how.

Our next step is to establish a base for sharing test data and tools (though not necessarily source code) and to encourage labs around the world to publish interface specs and standards for their modules.

References

[1] http://www.itl.atr.co.jp/cocosda/synthesis/refs.html

[2] Akers, G. & Lennig, M. (1985), "Intonation in text-to-speech synthesis: Evaluation of algorithms", J. Acoust. Soc. Amer. 77, 2157-2165.

[3] Bernstein, J. (1982), "Evaluating synthetic speech", Proc. NBS Workshop on Standardization for Speech I/O Technology, Gaithersburg, 87-91.

[4] Bezooijen, R. van & Pols, L. C. W. (1990), "Evaluating text-to-speech systems: Some methodological aspects", Speech Communication, 9(4), 263-270.

[5] Pols, L. C. W. (1989), "Improving synthetic speech quality by systematic evaluation", Proc. of the ESCA Tutorial Day, Vol. Noordwijkerhout, 3-11.

[6] Pavlovic, C.V., Rossi, M. & Espesser, R. (1991), "Definition of assessment methodology for overall quality of synthetic speech", Stage Report So.3, Year 2 Interim Report, SAM-UCL-G003, ESPRIT Project 2589 (SAM).

[7] Pisoni, D.B., Greene, B.G. & Logan, J.S. (1989), "An overview of ten years of research on the perception of synthetic speech", Proc. ESCA Workshop on Speech Input/Output Assessment and Speech Databases. Noordwijkerhout, The Netherlands, 1.1.1-1.1.4.

[8] Spiegel, M.F., Altom, M.J., Macchi, M.J. & Wallace, K. (1990), "Comprehensive assessment of the telephone intelligibility of synthesized and natural speech", Speech Communication 9(4), 279-291.

[9] www.itl.atr.co.jp/cocosda/synthesis/evaltalk.html

[10] www.itl.atr.co.jp/cocosda/synthesis/3rd_ws.html

[11] Campbell, N. "The COCOSDA/LDC Speech Synthesis Evaluation Facilities", in Proc Oriental COCOSDA, May 1998.

[12] http://bluebacks.hip.atr.co.jp/ja/index.html

[13] http://www.ldc.upenn.edu/ltts

[14] JASA Vol 105, No 2, Pt 2, February 1999.

[15] Acta Acustica Vol 85, E-21-466, Supplement 1. 1999.

[16] http://www.itl.atr.co.jp/chatr

[17] Chen, J. D, & Campbell, N., "Objective Distance Measures for Assessing Concatenative Speech Synthesis", Proc IEICE SP, May 1999.

[18] Ding, W., Fujisawa, K., Campbell, N. "Improving Speech Synthesis of CHATR using a perceptual discontinuity function and constraints of prosodic modification" Proc ESCA/COCOSDA Synthesis W/S, November 1998.

[19] Campbell, N., & Saenko, E., "Factors to Consider in the Design of an Optimal Speech Corpus for Concatenative Speech Synthesis". Proc ASJ, Mar 1999.