4. Conclusions

It will be clear from this draft proposal that many practical matters only have been partly resolved so far. However, there is no better way to learn than by trying out the procedure for various systems in various languages. Rather than wait until the Synthesis workshop in November 1998 in Australia (http://www.itl.atr.co.jp/cocosda/synthesis/3rd_ws.html) to run preliminary tests there, and then to hear the opinions from the experts, it would help to have already run such test beforehand, in one or more countries.

5. References

Benoit, C., Grice, M. & Hazan, V. (1996), "The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences", *Speech Communication* 18, 381-392.

Gibbon, D., Moore, R. & Winski, R. (Eds.) (1997), "Assessment of synthesis systems", Chapter 12 in *Handbook of standards and resources for spoken language systems*, Mouton de Gruyter, Berlin, 481-563.

Goldstein, M. (1995), "Classification of methods used for assessment of text-to-speech systems according to the demands placed on the listener", *Speech Communication* 16, 225-244.

ITU-T, Rec. P.85 (1994), *Subjective performance assessment of the quality of speech voice output devices*, Geneva.

Klaus, H., Fellbaum, K. & Sotscheck, J. (1997), "Auditive Bestimmung und Vergleich der Sprachqualität von Sprachsynthesesystemen für die deutsche Sprache", *Acta Acustica* 83(1), 124-136.

Pols, L.C.W. (1992), "Quality assessment of text-to-speech synthesis by rule", Ch. 13 in: Furui, S. & Sondhi, M.M. (Eds.), *Advances in speech signal processing*, Marcel Dekker, Inc., New York, 387-416.

Pols, L.C.W. & SAM-partners (1992), "Multilingual synthesis evaluation methods", *Proc. ICSLP'92*, Banff, 181-184.

Pols, L.C.W., Van Santen, J.P.H., Abe, M., Black, A., House, D., Liberman, M. & Wu, Z. (1998), "Easy access via a TTS website to mono- and multilingual text-to-speech systems", *Proc. Elsnet in Wonderland*, Soesterberg, the Netherlands, 40-44.

Salza, P.L., Foti, E., Nebbia, L. & Oreglia, M. (1996), "MOS and pair comparison combined methods for quality evaluation of text-to-speech systems", *Acta Acustica* 82, 650-656.

Santen, J.P.H. van (1993), "Perceptual experiments for diagnostic testing of text-to-speech systems", *Computer Speech and Language* 7, 49-100.

Schmidt, M., Fitt, S., Scott, C. & Jack, M. (1993), "Phonetic transcription standards for European names (Onomastica)", *Proc. Eurospeech'93*, Berlin, Vol. 1, 279-282.

Spiegel, M. (1993), "Using the ORATOR® synthesizer for a public reverse-directory service: Design, lessons, and recommendations", *Proc. Eurospeech'93*, Berlin, Vol. 3, 1897-1900.

Sproat, R. (Ed.) (1998), *Multilingual text-to-speech synthesis: The Bell Labs Approach*, Kluwer Academic Publishers, Dordrecht.

Subjects have to judge on 5-point scales the 7 or 8 categories specified below ('Stress' was not included in the original ITU draft), resulting in so-called Mean Opinion Scores (MOS).

categories	scales
- Overall impression	excellent - bad
- Listening effort	complete relaxation - no meaning understood with any feasible effort
- Comprehension	all the time - never
- Articulation	yes, very clear - no, not at all
- Anomalies in pronunciation	no - yes, very annoying
- Speaking rate	much faster - slower than preferred
- Voice pleasantness	very pleasant - very unpleasant
- Anomalies in stress	no - yes, very annoying

Salza et al. (1996) also used this MOS method to evaluate 3 Italian TTS systems, together with a paired comparison method (preference judgment).

The actual formulation of the 7 categories used varies a little bit between the two research groups, but this may, at least partly, be a translation problem. The way the questions are formulated is crucial for the reliability of the answers though. The speaking rate category is kind of an outlier, because both 'too slow' and 'too fast' might be undesirable. Salza et al., probably for that reason, used as scale descriptors: Yes; Yes, but slower than preferred; Yes, but faster than preferred; No, too slow; and No, too fast upon the question "Did you find the speed of delivery of the message appropriate?", rather than the description from 'too fast', via 'preferred', to 'too slow'.

The 7 category descriptors Salza et al. used were:

- How do you rate the sound quality of the voice you have heard? (Global impression)
- How do you rate the degree of effort you had to make to understand the message? (Listening effort)
- Did you find single words hard to understand? (Comprehension problems)
- Did you distinguish the speech sounds clearly? (Speech sound articulation)
- Did you notice any anomalies in the naturalness of sentence pronunciation? (Pronunciation)
- Did you find the the speed of delivery of the message appropriate? (Speaking rate)
- Did you find the voice you heard pleasant? (Voice pleasantness)

It is partly a matter of policy to stick to these categories, or to adapt them to one's own needs. For instance, rather than 'comprehension', one would perhaps prefer 'all words pronounced correctly (in terms of phones and syllabic stress)'. One could also simply ask subjects to judge intelligibility on a scale. Rather than 'anomalies in stress', one could perhaps ask about 'correct word emphasized', or 'bad intonation but correct words and syllables accented'.

A serious technical problem concerns the amount of testing that realistically can be done at the workshop. If sufficient PCs can be made available, we could perform individual testing, otherwise we must work in groups. It is not unrealistic to consider that at least 10 English TTS systems will be available, plus up to 4 TTS systems in 5 or more other languages. To demonstrate a possible design we suppose for a moment that there will be 12 English TTS systems (s1, ..., s12), one or more groups of 12 native English listeners (L1, ..., L12), and that 3 text types (T1, T2, and T3) will be used, of which 48 items each (T1(1), T1(2), ..., T3(47), T3(48)) will be presented. Each group of 12 listeners will perform individual testing for which 12 PCs will have to be available. By creating groups of 12 trials, in which systems and text items are mixed (even the order in which the systems are listened to could be randomized), we achieve an optimally balanced design:

trial	listener				
nr.	L1	L2	L3	…	L12
1	s1,T1(1)	s2,T1(1)	s3,T1(1)	…	s12,T1(1)
2	s2,T1(2)	s3,T1(2)	s4,T1(2)	…	s1,T1(2)
3	s3,T1(3)	s4,T1(3)	s5,T1(3)	…	s2,T1(3)
…	…	…	…	…	…
12	s12,T1(12)	s1,T1(12)	s2,T1(12)	…	s11,T1(12)
13	s1,T1(13)	s2,T1(13)	s3,T1(13)	…	s12,T1(13)
…	…	…	…	…	…
48	s12,T1(48)	s1,T1(48)	s2,T1(48)	…	s11,T1(48)
49	s1,T2(1)	s2,T2(1)	s3,T2(1)	…	s12,T2(1)
…	…	…	…	…	…
144	s12,T3(48)	s1,T3(48)	s2,T3(48)	..	s11,T3(48)

These 144 items, at a tough rate of 20 seconds per trial plus a couple of minutes for instructions, would require one hour. With other groups of 12 listeners, the experiment can be repeated, although preferably with different text items in order to maximize text coverage. For other languages that have only, say, 4 TTS systems, we could create similar matrices. It would have only four columns, again with 3 groups of 48 trials, consisting of 12 (instead of 4) 4 x 4 (instead of 12 x 12) sub-matrices. The number of data points per group of 4 listeners per text type / TTS system combination would again be 48.

Per text type certain tasks have to be performed, either scoring on a small number of semantic scales (which could appropriately be done for newspaper texts), or writing out what has been understood from the spoken text (which would be an appropriate task both for telephone directory entries and for semantically unpredictable sentences), and/or rating the degree to which the actual text (displayed later) matches which what one thought to have heard (which would make most sense for telephone directory entries).

Another political question, one which was raised in Jan van Santen's first proposal, is whether a comparative test could result in an ordered list of systems made public, or whether the results should be kept for private use by the system designers only. The designers would then have collected interesting information to further improve their system, whereas the speech community would have gained experience with the evaluation method as such. The closer systems get to commercial products, the more reluctant one might be to allow complete openness about the test results, unless the procedures were fully standardized and fit optimally into the diverse application domains of the various systems. Apparently we are not yet at that level of agreement.

the input) of this sorted list. This leads to (top 1%) German sentences like:

> *"Nachdem der Vorfall bekannt wurde, seien vier Personen festgenommen worden."*

The *overall word frequency based* procedure selects longer sentences with many high-frequency words and orders the sentences on the basis of the added log frequency of all the words in a sentence. An example of a top 1% Spanish sentence is:

> *"Será ésta la segunda visita de Delors a Santiago en los últimos 18 meses, luego que en marzo de 1993 realizó una gira que incluyó Chile Argentina y México."*

The *overall trigram based* procedure leads to longer sentences containing several rarer words. The sentences are ordered according to the added log frequency of all its letter trigrams. An example of such a top 1% English sentence is:

> *"They are crammed in and around a shabby government building, one of countless facilities around the country that have been hastily converted to orphanages. "*

Once the language, the text sentence and the TTS system are selected, and some other practical details are specified (sampling frequency, coding quality, file type, speaker characteristics), one can ask for the synthesized version of that sentence. By direct system access the audio file becomes available for subsequent listening. Next one can ask for the same sentence to be generated by another system, or one could select another sentence to be generated by the same system. Since the synthesizers generally take some time to generate the sentences and of course the internet is also usually notoriously slow, the LDC system currently has the capability of collecting and temporarily storing these files on their system. Through ftp or otherwise these compressed files then can be downloaded later on to one's own system for side-to-side comparison.

Participants of this First International Conference on Language Resources and Evaluation are encouraged to visit this website in order to try out the possibilities for themselves and to provide the present authors with some feedback.

2.2 Other text types

Given the diversity of applications of TTS systems it should be clear that newspaper texts alone, although challenging in their diversity, length and complexity, are not representative enough to allow judging all aspects of system performance. It still has to be decided whether various application-specific types of texts, such as train table information, telephone directory entries, weather reports, conversational dialogues, or raw e-mail texts, should be added as well. No one would deny that, for instance, proper pronunciation of names and addresses is another challenging task, especially given the many foreign names in most applications (e.g., Spiegel, 1993; Schmidt et al., 1993). However, free availability of representative telephone directories in electronic form for several languages is not yet settled.

Another approach is to generate specific test material, such as nonsense names, or language-specific nonsense syllables of various types, the simplest ones being CV, VC, CVC, and VCV. One could also consider including consonant clusters, as far as appropriate in that language.

Within the European SAM project (Pols & SAM-partners, 1992) an interesting proposal was made to use so-called semantically unpredictable sentences (Benoit et al., 1996). With a reservoir of high-frequent words in several word categories (verbs, nouns, adjectives, prepositions, question adverbials, determiners, and adverbs), plus some 5 fixed grammatical structures, one can automatically generate ever different sentences of the type:

Dutch	*De stoel loopt door een lief huis*
English	*The strong way drank the day*
French	*Tourne la date ou la main*
Italian	*Il piatto apre il pesce che ride*
Swedish	*Hur drack en lukt ett snabbt hus?*

3. Scoring categories

In addition to the text selection, we also have to decide upon the scoring procedure(s).

In a straightforward application- and text-specific side-by-side comparison, one could perhaps limit oneself to a one-level preference judgment, for instance based on 5 randomly selected newspaper sentences. However, most of the time the next question would immediately be "why do you prefer system1 over system2", or "do you prefer system1 also with respect to its prosody, its segmental quality, or its pronunciation of that specific word". Of course, other features, such as choice of speaker and speaking rate, as well as price and availability on a certain platform, will also influence one's choice.

For these and other reasons, the capability of offering a suite of tests and of scoring procedures will probably be unavoidable. System developers and researchers would certainly like to run tests that concentrate not just on global performance but also on specific stages that are generally distinguished in synthesis development, such as:

- text preprocessing
- grapheme-to-phoneme conversion
- phonological rule application
- prosodic phrasing
- duration modelling
- F0 modelling, and
- signal generation techniques.

Over the years a great many different methods have been proposed and used. For some useful overviews, see Pols (1992), Van Santen (1993), Goldstein (1995), or Chapter 12 in the Eagles handbook (Gibbon et al., 1997). Still, we also feel that time has come to perform appropriate product evaluations, whether or not with an emphasis on the specific application in mind, rather than system evaluations for further improvement of specific modules of the system.

Whether we like it or not, the International Telecommunication Union (ITU, the former CCITT), has recently standardized a multidimensional subjective category rating test (ITU-T, 1994) for the subjective performance assessment of the quality of speech voice output systems. Klaus et al. (1997) have used this method to compare 13 German text-to-speech systems, along with a dictation test (in which percentage syllables correct were counted).

The use of large text corpora for evaluating text-to-speech systems

Louis C.W. Pols [1)], Jan P.H. van Santen [2)], Masanobu Abe [3)], Dan Kahn [4)], and Eric Keller [5)]

[1)] Institute of Phonetic Sciences / IFOTT, University of Amsterdam
Herengracht 338, 1016 CG Amsterdam, The Netherlands
[2)] Bell Labs, USA; [3)] NTT, Japan; [4)] E-Speech Corp., USA; [5)] Univ. Lausanne, Switzerland

Abstract

The starting point of this draft proposal for a TTS evaluation procedure is that the procedure should allow participants to a forthcoming synthesis workshop to listen to, and to compare, various TTS systems under fair conditions. This implies the use of input texts that are new to each system, thus preventing any fine-tuning. Furthermore, the same texts (different listeners) should be used for all available systems in one language, to allow direct side-by-side comparison, whereas a variety of text types (newspaper sentences, telephone directory entries, isolated words, semantically unpredictable sentences, etc.) should be used to test the strong and weak points of each system. This requires both a text server and a TTS server via the web, both of which are presently available in prototype form. Perhaps some preliminary testing will already be done before the start of the workshop itself.

1. Introduction

Text-to-speech synthesizers are meant to produce intelligible and high-quality speech either from any text ("open domain synthesis"), or from application-specific texts such as train table information or a telephone directory. Even in the latter case, input domains are quite large, which makes it astounding that most systems so far are demonstrated using only a few carefully prepared sentences.

Following the example of the benchmark tests of the ARPA Spoken Language Program, and based on a proposal made by one of the authors (Jan van Santen) at the 1997 Cocosda meeting in September in Rhodes, Greece, we feel it would be more appropriate to run fair tests based on constantly changing selections (according to certain selection criteria), taken from very large text corpora (preferably in several different languages).

A committee consisting of the above authors is presently trying to design such a TTS evaluation procedure, which will be proposed to the synthesis community. It is the intention to run tests according to the proposed protocol at the third ESCA Synthesis Workshop in Australia (27-29 Nov. 1998), and, if feasible, to actually start the process before that event begins.

The protocol will define various types of text (such as newspaper, telephone directory, single words or names, semantically unpredictable sentences, etc.), as well as selection criteria, such as frequency weighting.

To ensure complete fairness, access to and selection from the text corpora, as well as actual generation of the speech utterances, should be fully automatic, without any manual intervention. This requires, among other things, that in addition to a text server there should also be a TTS web server through which to-be-tested systems are accessible on-line, unless they are locally available. In Section 2 we will give some details about progress we have made so far with the TTS website. For the test runs in Australia we may prefer, for practical reasons, to locally install several systems.

Next, the listening procedure will have to be defined. One of the most critical points of discussion right now concerns the scoring categories to be used (such as overall, as well as specific, quality and intelligibility rating scales). Various possibilities will be presented in Sect. 3.

2. TTS web-site

A prototype TTS website (http://www.ldc.upenn.edu/ltts/) has already been developed at the Linguistic Data Consortium (LDC) by Mark Liberman and Zhibiao Wu; for some more details see Pols et al. (1998).

2.1 Present form

In its preliminary form this website is already accessible and can be used for pilot tests. One can choose from several TTS systems in several languages. One can type in one's own text, or one can choose from a text corpus. Presently newspaper texts in three different languages (English, German, and Spanish) are available via the text server. Hopefully more text corpora will soon be added. One sentence at a time is selected from the newspaper corpus according to certain search criteria. To date several options have been implemented. Search can be based on:

- random selection
- minimum word frequency
- overall word frequency
- overall trigram frequency

The *random* option speaks for itself. The other three are based on word or (letter) trigram frequency counts.

The *minimum word frequency based* procedure selects short sentences made up entirely of common words. First, the frequency of occurrence of each word in the text corpus is determined; then for each sentence the least frequent word is found. All sentences are then sorted according to this least frequent word frequency. Finally a sentence is randomly selected from the top 1, 5, or 10% (specified at