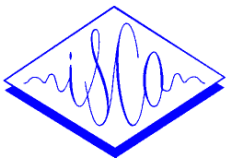

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

MULTI-LINGUAL SYNTHESIS EVALUATION METHODS

Louis C.W. Pols and SAM-partners

Institute of Phonetic Sciences, Univ. of Amsterdam
Herengracht 338, 1016 CG Amsterdam, The Netherlands

ABSTRACT

In the past few years a suite of tests has been developed, tested and implemented within the Esprit-SAM project, in order to evaluate the performance of rule synthesizers in many different languages. Since, up to now, no single rule synthesizer has a fully acceptable segmental intelligibility, nor a reasonable prosody, systematic diagnostic evaluation and comparative tests remain necessary. The SAM segmental test, consisting of CV, VC, and VCV nonsense words, following the phonotactic constraints per language, is a proper means for that. Other word types, for instance including consonant clusters, can easily be added. Next there is the SUS test (semantically unpredictable sentences) in which five common grammatical structures are defined, as well as a list of words per word type, allowing one to generate ever-different test material. We also defined an overall quality test by using either a 20-points categorical estimation procedure, or a magnitude estimation procedure, where the subjects adjust the length of a line segment according to their quality judgment. We are developing prosodic tests in which the form and the function of prosodic characteristics in various word and sentence types are evaluated according to their appropriateness. The frequently neglected variability over listeners and between tests is also studied, as well as ways to measure speech quality in an objective way (by using physical means instead of listeners' judgments). Various test procedures have been implemented on a PC in a software package called SOAP.

10.21437/ICSLP.1992-34

1. INTRODUCTION

In March 1992 the formal three-year contract period of the European collaborative Esprit project SAM (multi-lingual speech input/output assessment, methodology and standardisation) ended. A final report about all aspects of the project was produced, together with several detailed documents describing software, databases, and analysis results. Work is in progress to compile the most relevant information into a book, in order to make the outcomes of this project more accessible than through these reports alone. The present conference is another proper occasion to report about some major outcomes of this SAM initiative [31]. In this presentation I will limit myself to synthesis evaluation.

As has already been expressed before [8, 12, 27, 28, 29, 32], the principle idea is to produce a suite of tests at various levels of interest (from segmental intelligibility to overall quality judgment, from prosodic functionality in a laboratory task to field tests).

Each test should be clearly defined in terms of its goal(s) and content, without being completely fixed in its realization, in order to prevent selective tuning by the designer. The structure of the test should be well defined, but the actual content may vary. This makes it also possible to provide tests that are (up to a certain level) comparable over languages. A nice example of this is the use of semantically unpredictable sentences: A number of frequently used grammatical structures have been selected, such as an imperative form (verb - det - noun - conjunction - det - noun: '*Draw the house and the fact*'). Per word class a lexicon is collected, from which ever-different sentences can be generated. The tests preferably should also have diagnostic capabilities.

2. SEGMENTAL INTELLIGIBILITY

So far none of the presently available text-to-speech (TtS) systems has a phoneme-intelligibility score close to that for natural speech. Therefore, unavoidably, we still have to give attention to this topic. For this evaluation job we prefer the use of nonsense words and an open response task [32]. This certainly is highly artificial, but provides excellent diagnostic material.

In its simplest form we propose to use lists of CV, VC, and VCV words, in which all reasonably frequent initial, medial, and final consonants in a specific language are combined with three vowels, such as /i, u, a/, taking into account the phonotactic constraints. For some representative results for Swedish we refer to [11].

For diphone-based systems, this word set may be not very representative, since only a very small sub-set of all units is actually evaluated this way. The consonant intelligibility, in that case, is more indicative of average performance than of individual diphone units. If the latter information is required, one could use a set of CVC, VCV, CVVC, and VCCV words in which all CV-, VC-, VV-, and CC-diphone units are represented [36].

If the concatenative units are even more diverse (up to the level where they include word boundary effects, such as in the system of Olive [25]), one could use, as van Santen [34] did, the 'greedy algorithm' to select the smallest sub-set of words in which (almost) all units are present at least once. My only objection to this very elegant approach is that (sequences of) real words are used, such as '*A rhetorical headsail for the alumnis*'. These words have a certain internal predictive power, especially the multi-syllabic ones, which may hide the poor intelligibility of certain units if presented otherwise. Actually, van Santen did not study intelligibility directly with this material, but asked subjects to highlight problem words.

Especially for allophone-based formant synthesizers, but also for diphone-based systems using only real diphone units, *consonant clusters* are a challenge, and represent a next level of segmental complexity worth to be evaluated in detail. For some 7 European languages the SAM partners studied and collected the phonotactic constraints per language. This information is compiled in the form of matrices specifying the allowable phoneme combinations, including consonant clusters. Software was developed [22] to be able to generate lists of nonsense words of one or several word types (such as CV or CCCVCC) according to these constraints. On the basis of this type of data Jekosch developed the cluster-identification tests [21]. The scoring module for such an experiment becomes rather complex. Spiegel et al. [37] use a somewhat similar approach, but combine meaningful and non-

sense words, with and without initial and final consonant clusters in a fixed set of 312 words.

Reverse telephone directory (which name + address belongs to this telephone number?) is a potentially very interesting application of rule synthesis. Supposing a proper grapheme-to-phoneme conversion (which is certainly not a simple problem, especially considering the many names with a foreign origin), intelligibility is again mainly defined by segmental characteristics. So, generating names in an intelligible way is a challenging task for each synthesizer, and transcribing such synthesized names is a challenging segmental task for each listener. Van Santen [34] gives a solution for scoring the detailed results of such an orthographic name transcription task. A much more global score would indicate whether the synthesized version of a certain (known) name sounds acceptable or not, after all there are quite a few names for which there is no uniquely best pronunciation [38].

3. WORD AND SENTENCE INTELLIGIBILITY

Any level of intelligibility evaluation, beyond the segmental level, implies the use of sentence-type context. Apart from more global aspects, such as naturalness and acceptability (see below), one could argue that only a proper intelligibility at word and sentence level will make TtS generally acceptable and useful. However, because of the unavoidable redundancy in simple natural sentences, one easily gets a 100% score with such test material, even for rather poor systems. That is why several more-demanding tests have been developed.

One is the fixed set of 100 semantically anomalous, so-called Haskins sentences [24]. All these English sentences have the same grammatical structure (e.g. The late voice knew the table). SAM extended this idea into a multilingual variant, using several (presently 5) different grammatical structures, and a lexicon for each word class from which words are selected, thus creating new sentences all the time. Presently the words in the lexicons are all high frequent and (in principle) mono-syllabic, but one could also decide to make this a variable. Appropriate software is developed and lexicons in six different languages have been collected [4] to produce lists of such semantically unpredictable sentences (SUS). For various systems in various languages meanwhile SUS results have been achieved, see for example [5]. At this level one can also expect to find intelligibility results depended on the syntactico-prosodic parser and the intonation grammar used in the synthesizer [3].

It is worthwhile to mention a few other test procedures that make use of short sentences. One is the sentence verification task, in which subjects have to decide as fast as possible whether short (three- or six-word) sentences are true statements [23]. Although this is a comprehension task, it requires of course intelligibility as well. One can also use noise masking, such as in speech audiometric tests, to derive an equivalent signal-to-noise ratio. Much more diagnostic is the minimal pairs intelligibility test as used by v. Santen [34]. He creates minimal pair nonsense sentences of the type 'The horrid courts scorch/score a revolution'. In a two-alternatives forced-choice task the listener has to choose one of the two alternative sentences.

4. OVERALL QUALITY

For comparing overall performance of systems, or for comparison of (specific details of) algorithms, an overall judgment is very appropriate. Several methodologies have been used for that purpose. One could think of a preference judgment by paired comparison, or a free scale score (magnitude estimation), the use of a fixed scale of say 7, 10, or 20 points (categorical estimation), or the reaction time paradigm. One could also ask subjects to make qualified judgments on a number of scales (semantic scaling). After extensive comparative testing, SAM recommends to use either a free scale magnitude estimation of quality, or a categorical judgment on a 20-points scale. Also for these two procedures the required software has been developed. Actually, the adjustment of a line segment on the screen by mouse control, according to the judged quality, appeared to work very satisfactory [33].

The procedures have been used for the evaluation of a number of systems in a number of different languages including (masked) natural speech, see for instance [15].

5. PROSODY TEST

With the improved performance of syntactico-prosodic parsers and the use of synthetic speech in information-retrieval and dialogue systems, the possibilities and the needs for improved prosody (prosodic phrasing, accentuation, intonation, segmental duration) have grown. The segmental quality of rule-synthesized speech will hopefully soon have reached such a high level of acceptability, that it will make the need for good prosody even more apparent. Although prosody-driven speech recognizers are still a fiction, good prosody for rule synthesizers is now already a very active research field with good progress, which means that also evaluation methods should become available [10].

Although the 5 different grammatical structures, used for the semantically unpredictable sentences, are generally not yet correctly interpreted as such by most present-day rule synthesizers (apart perhaps from the distinction between declarative and interrogative sentence, because of the period or the question mark), that sentence material is too simple as prosodic test material. The functionality of certain prosodic characteristics can only be judged in its proper context.

That is why SAM ran a preliminary experiment together with partners from the Esprit-SUNDIAL project in which specific human-machine dialogue situations are created, and the appropriateness to context of different prosodic contours is judged [14]. The existing software could easily be adapted to these new demands. Before running more of such *functional* tests, one should not forget to run *form* tests as well, in order to find out whether a certain synthesizer has the knowledge and the capabilities to produce proper prosody (properly synchronized with the (stressed) syllable, dependent on the position of the syllable in the word and of the word in the sentence). SAM acquired some experience in doing this for Italian and English mono- and multi-syllabic words [13].

6. VARIABILITY OVER LISTENERS AND TESTS

So far it has been common practice in synthesis evaluation to use a representative number of subjects to run the listening experiments, after which the average results, plus perhaps some measure of variability, are presented as the experimental outcome. In SAM we wondered whether listener variability might contain systematic effects. Whether individual behavior might be predictable (from psycho-acoustic and/or audiometric tests), whether it is consistent over various tests, whether training is consistent, and whether certain listeners are more sensitive to specific speech deficiencies than others. Several types of sentences and VCVs were used to collect individual intelligibility data, for synthetic and natural speech, in quiet and in noise, for 60 listeners, homogeneous in terms of age, hearing acuity, exposure to synthetic speech and mother tongue. Despite the homogeneity of this listener group, the score variability was substantial, more so for synthetic than for natural speech, and more so at sentence level than at nonsense-word level. The implication is that between-group effects (such as with respect to age, experience/expertise, or hearing impairment) may be hard to show because of the large within-group variability. This is exactly what was found in a

number of SAM pilot studies for English, Italian, and German [20]. In correlating the results on different tests [16] it was shown that performance on VCV material was not related to performance on sentence material, justifying once again that tests at both levels are required, and that *good auditors* not necessarily are also *good comprehenders*. Since there was also a lack of relation between performance on natural speech in noise vs. synthetic speech, this seems to suggest that listeners do not have a general capability to process degraded speech, whatever the type of degradation. However, there was some indication that certain listeners have an overall skill in processing synthetic speech, whatever the type of test material, be it sentences or VCV-words.

7. OBJECTIVE EVALUATION

Although the human listener is the ultimate judge of the quality of the speech produced by a rule synthesizer, acquiring these subjective judgements for each new condition is very laborious and time consuming. There is a strong desire to simplify and automatize these measurements. In analogy to the objective (physical) measures sometimes used in evaluating speech communication channels (such as the articulation index AI or the speech transmission index STI), SAM also studied the possibilities for doing something similar for the objective evaluation of the quality of synthetic speech. However, there are several complicating factors:

- there is no unique, best, reference realization for each word or sentence
- not just the acoustic characteristics of the speech signal are important, but also text pre-processing and linguistic realization determine the final speech quality

There is also a potentially very beneficial point in this approach: it can be analytic and predictive.

So far several overall characteristics of synthetic speech have been compared with those for natural speech, from one or more speakers and languages. These overall characteristics imply F0 and pause distributions [2], long-term average spectra and its variation over time [26], and correlation patterns reflecting the statistics of the dynamics of speech in the frequency and in the time domain [17]. Up to what level these measures actually define the naturalness of speech, is still a widely open point for research. In the final year of the project the research on correlation patterns was intensified [18].

8. DISCUSSION

In the preceding sections it has been mentioned several times that, for specific tests, software had been developed. Actually the suite of four tests (segmental, SUS, overall quality, and prosodic) are built into one package, called SOAP, for Speech Output Assessment Package [19]. Version 4.0 is well documented and contains demonstrations for all the implemented tests, most of the time for several different languages.

Although this certainly is an improvement compared to the situation up to now, the presently presented tests are not formally standardized. If the research and the user community considers them to be good and appropriate, this will happen after all. It is only fair to say that SAM so far has neglected the text preprocessing and linguistic part of text-to-speech. Of course these aspects of a rule synthesizer require evaluative attention too.

First attempts have been done to evaluate synthetic speech quality at paragraph level [9]. One approach is to use open content questions that give an indication of the comprehensibility of a read text. The other is to ask ratings on a number of scales, such as naturalness, fluency, and correctness of pronunciation.

Also the application side so far did not receive sufficient attention [35]. It is worth to study the completion of realistic tasks (such as listening to synthesized e-mail texts over the telephone), which will frequently imply some form of human-machine dialogue. The main task of listening to synthetic speech for a certain purpose could also be extended with (realistic) secondary tasks. Naive users (including children, elderly people, second-language learners) and adverse listening conditions (background noise, reverberation, competing speech, telephone line) add on to real-live conditions. One should also compare performance with other possible modalities, such as presenting the text on a screen, or the (limited) use of canned speech. Finally it is worth to consider the additional benefits of using a stylized visual image of (the lips of) the speaker [6].

Although progress is still relatively slow, we do believe that the development of objective, diagnostic evaluation methods for synthetic speech is a scientific challenge with great potential. Progress will have to rely on proper knowledge in speech signal processing and in the processes of speech production and perception in general [7, 30].

REFERENCES

[1] Bailly, G., Benoit, C. & Sawallis, T.R. (Eds.), *Talking machines: Theories, models, and designs*, Elsevier Science Publishers B.V., 523 pp., 1992.

[2] Barry, W.J., Grice, M., Hazan, V. & Fourcin, A.J., "Excitation distributions for synthesised speech", *Proc. Eurospeech'89*, Paris, vol. 1, pp. 353-356, 1989.

[3] Benoit, C., "An intelligibility test using semantically unpredictable sentences: Towards the quantification of linguistic complexity", *Speech Communication*, vol. 9, pp. 293-304, 1990.

[4] Benoit, C., "Semantically unpredictable sentences (SUS) - theoretical basis for generation", In: Howard-Jones (1992a), App. 5, 1992.

[5] Benoit, C., Erp, A. van, Grice, M., Hazan, V. & Jekosch, U., "Multilingual synthesiser assessment using semantically unpredictable sentences", *Proc. Eurospeech'89*, Paris, vol. 2, pp. 633-636, 1989.

[6] Benoit, C., Lallouache, T, Mohamadi, T. & Abry, C., "A set of French *visemes* for visual speech synthesis", In: Bailly et al. (1992), pp. 485-504.

[7] Benoit, C. & Pols, L.C.W., "On the assessment of synthetic speech", In: Bailly et al. (1992), pp. 435-441.

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

[9] Bezooijen, R. van & Pols, L.C.W., "Performance of text-to-speech conversion for Dutch: A comparative evaluation of allophone and diphone based synthesis at the level of the segment, the word, and the paragraph", *Proc. Eurospeech'91*, Genova, vol. 3, pp. 1117-1120, 1991.

[10] Bladon, A., "Evaluating the prosody of text-to-speech synthesizers", *Proc. Speech Tech'90*, pp. 215-220, 1990.

[11] Carlson, R., Granström, B. & Nord, L., "Segmental evaluation using the Esprit/SAM test procedures and monosyllabic words", In: Bailly et al. (1992), pp. 443-453.

[12] Fourcin, A., "Assessment of synthetic speech", In: Bailly et al. (1992), pp. 431-434.

[13] Grice, M., Vagges, K. & Hirst, D., "Prosodic form tests", *Part of SAM final report*, 12 pp., 1992.

[14] Grice, M., Vagges, K. & Hirst, D., "Prosodic function tests", *Part of SAM final report*, 14 pp., 1992.

[15] Goldstein, M., Lindström & Till, O., "Assessing global performance of speech synthesizers: Context effects when assessing Naturalness of Swedish sentence-pairs generated by 4 systems using 3 different assessment procedures (free number Magnitude Estimation, 5- and 11-point category scales)", *Part of SAM final report*, 19 pp., 1992.

[16] Hazan, V., "Quantification of listener variability", *Part of SAM final report*, 19 pp., 1992.
[17] Houtgast, T. & Verhave, J.A., "A physical approach to speech quality assessment: Correlation patterns in the speech spectrogram", *Proc. Eurospeech'91*, Genova, vol. 1, pp. 285-288, 1991.
[18] Houtgast, T. & Verhave, J.A., "An objective approach to speech quality", *Part of SAM final report*, 17 pp., 1992.
[19] Howard-Jones, P., "SOAP, Speech output assessment package, version 4.0", *SAM report*, 1992a.
[20] Howard-Jones, P., "Specification of listener dimensions", *Part of SAM final report*, 21 pp., 1992b.
[21] Jekosch, U., "The cluster-identification test", *Part of SAM final report*, 13 pp., 1992.
[22] Jekosch, U. & Belhoula, K., "WORDGEN manual", In: Howard-Jones (1992a), App. 9, 23 pp.
[23] Manous, L.M., Pisoni, D.B., Dedina, M.J. & Nusbaum, H.C., "Comprehension of natural and synthetic speech using a sentence verification task", *Indianan Univ. Progress Report*, vol.11, pp. 35-57, 1985.
[24] Nye, P.W. & Gaitenby, J.H., "The intelligibility of synthetic monosyllabic words in short syntactically normal sentences", *Haskins Labs* SR-37/38, pp. 169-190, 1974.
[25] Olive, J.P., "A new algorithm for a concatenative speech synthesis system using an augmented acoustic inventory of speech sounds", *Proc. ESCA Workshop on Speech synthesis*, Autrans, pp. 25-30, 1990.
[26] Pavlovic, C.V., Rossi, M. & Espesser, R., "Perceived spectral energy distributions for Eurom.0 speech and for some synthetic speech", *Proc. XIIth Int. Congr. Phon. Sciences*, Aix-en-Provence, vol. 5, 418-421, 1991.
[27] Pols, L.C.W., "Assessment of text-to-speech synthesis systems:, In: A. Fourcin et al. (Eds.), *Speech input and output assessment. Multilingual methods and standards*, Ellis Horwood Ltd., Chichester, Chapter III, pp. 53-81 & 251-266, 1989.
[28] Pols, L.C.W., "Improving synthetic speech quality by systematic evaluation", *Proc. ESCA Tutorial Day on Speech input/output assessment and speech databases*, Noordwijkerhout, The Netherlands, pp. 3-11, 1989.
[29] Pols, L.C.W., "'Standardized' synthesis evaluation methods", *Proc. Int. Workshop on int. coord. and stand. of speech databases and ass. techniques*, Kobe, pp. 53-60, 1990.
[30] Pols, L.C.W., "Gaining phonetic knowledge whilst improving synthetic speech quality?", *J. of Phonetics*, vol. 19, 139-146, 1991.
[31] Pols, L.C.W., Evaluating the performance of speech input/output systems. A report on the ESPRIT-SAM project", *Proc. DAGA'91*, Bochum, pp. 139-150, 1991.
[32] Pols, L.C.W., "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, pp. 387-416, 1991.
[33] Rossi, M., Espesser, R. & Pavlovic, C.V., "The effects of an internal reference system and cross-modality matching on the subjective rating of speech synthesisers", *Proc. Eurospeech'91*, Genova, vol. 1, pp. 273-276, 1991.
[34] Santen, J.P.H. van, "Perceptual experiments for diagnostic testing of text-to-speech systems", *Computer Speech and Language*, to be published.
[35] Silverman, K., Basson, S. & Levas, S., "Evaluating synthesis performance: Is segmental intelligibility enough?", *Proc. ICSLP'90*, Kobe, pp. 981-984, 1990.
[36] Son, N. van, Pols, L.C.W., Sandri, S. & Salza, P.L., "First quality evaluation of a diphone-based speech synthesis system for Italian", *Proc. Speech'88, 7th FASE Symp.*, Edinburgh, book 2, pp. 429-436, 1988.
[37] Spiegel, M.F., Altom, M.J., Macchi, M.J. & Wallace, K., "Comprehensive assessment of the telephone intelligibility of synthesized and natural speech", *Speech Communication*, vol. 9, pp. 279-291, 1990.
[38] Spiegel, M.F., Macchi, M.J. & Gollhardt, K.D., "Synthesis of names by a demisyllable-based speech synthesizer (Spokesman)", *Proc. Eurospeech'89*, Paris, vol. 1, pp. 117-120, 1989.