IMPROVING SYNTHETIC SPEECH QUALITY BY SYSTEMATIC EVALUATION *

Louis C.W. Pols

ABSTRACT

In the joint Dutch research program for developing a high-quality text-to-speech synthesis system, much emphasis is put on systematic speech quality evaluation. This is not just done to produce performance figures, but even more so to support the developers of the various linguistic and acoustic synthesis modules by indicating to them ways for improvement. This approach compares favourably with most other projects in which no diagnostic testing is done at all, or only once in the final phase in order to produce (incomparable) performance figures which do not lead to further improvements. The joint project is sponsored by SPIN (Dutch National Program for the Advancement of Information Technology).

1. INTRODUCTION

A complete text-to-speech synthesis-by-rule system consists of many different components originating from such diverse areas as text processing, language processing, and signal processing. By improving the performance of single components one hopes to improve the performance of the total system. However, more often than not, different experts in the various fields develop single components and leave the remaining problems to others. For instance the acoustic front end presupposes a correct phonetic input, whereas grapheme-to-phoneme conversion can easily introduce errors here. The intonation module requires correct stress markers, whereas rules to define the position and the character of those markers are not yet fully developed. The morphological decomposition requires error-free word sequences, and the text expander requires knowledge about how to interpret the text. Should, for instance, the digit sequence 14.18 be pronounced as a number, as a money value, or as a time indicator?

Whenever performance figures are given at all, they mainly represent the results of one final test. Such results specify in an absolute or relative way (in comparison to some reference system, such as LPC-resynthesized utterances) the achieved quality of the system, whereas a further diagnostic analysis of the results seldom leads to subsequent modifications of the system.

Especially in the recently started joint Dutch research program for developing a high-quality text-to-speech synthesis system (Pols, 1988), we hope to be able to follow a different line. In an initial evaluation the speech quality at the start of the project is specified (van Bezooijen and Pols, 1987; van Bezooijen, 1988). During the run time of the project, subjective tests will be performed regularly to evaluate the progress, but

* Written version of a paper presented at the Symposium on Advanced Man-Machine Interface through Spoken Language, Makaha, Hawaii, 20-22 November 1988.

even more so to derive information about how to proceed. At the completion of the project a final test will be performed to measure the improvement and to compare, if possible, the results with similar systems in other languages.
A somewhat similar approach is followed in the ESPRIT project SPIN (Speech Interface at Office Workstation) in as far as rule synthesis for French and Italian is concerned (Pols et al., 1987; van Son et al., 1988). In ESPRIT project SAM (Multi - lingual speech input/output assessment, methodology, and standardization) (SAM-partners, 1988) the methodologies for evaluating speech recognizers and speech synthesizers will be further developed and, whenever possible, standardized.
Of course one must realize that, so far, most tests for evaluating the speech quality of text-to-speech synthesizers operate at the segmental level only. We will review those segmental tests, but we will also indicate how tests at the supra-segmental level are going to be developed for sentence intelligibility, global speech quality judgment, and prosodic evaluation.
Another level of evaluation and testing of course involves linguistic processing where results on paper generally suffice to indicate the performance, such as text preprocessing, syntactic analysis, or morphological decomposition. However, even here an acoustic realization and a listening test are sometimes required, for instance to find out whether an incorrect segmentation of a word in morphological components will nevertheless result in a correct pronunciation.

2. SHORT OVERVIEW OF TEST METHODS FOR SYNTHESIS EVALUATION

2.1 Purpose

Once the purpose of a test is identified it will also be easier to choose the appropriate speech material and the method of evaluation. I would like to distinguish four different purposes for developing speech quality tests:

- global testing
- diagnostic testing
- objective testing
- application-oriented testing

Global testing is mainly executed to describe and compare system characteristics in general terms, whether or not in comparison with a reference system or a competing system. The frequently used Mean Opinion Score (MOS) in telecommunication (Goodman and Nash, 1984) is the ultimate example of this, but also a preference judgment by paired comparison, or a magnitude estimation on an 'acceptability' or a 'naturalness' scale are examples belonging to this category.
Diagnostic testing is performed with specific aims in mind and requires a careful choice of the test material. An intelligibility test at the segmental level requires an approach totally different from an acceptability judgment about a number of different algorithms to generate prosodic contours in long sentences.
Objective testing, implying the use of physical means without using listener judgments, is presently virtually non-existent in speech synthesis evaluation. However, in evaluating the performance of analog and digital speech communication channels this approach is quite common. I just have to refer to the Articulation Index (AI) (Kryter, 1962), the Speech Transmission Index (STI) (Steeneken and Houtgast, 1980), or the signal-to-noise measure used in coding evaluation. The ESPRIT-SAM project intends to start research in this area of objective synthesis testing.
Application-specific testing aims at a different line of performance evaluation. Most laboratory conditions will then have to be abandoned; it frequently implies the use of task-specific test material, naive, untrained listeners, probably noisy, reverberant

listening conditions, perhaps interaction in a dialog-type application with or without automatic speech recognition, etc. Examples so far are scarce (Hampshire et al., 1982).

2.2 Test method

From telecommunication testing, psycholinguistics, psychoacoustics, speech audiometry, speech perception, language acquisition, and probably other areas, we have good knowledge about a great variety of subjective test methods. I give a brief overview here only.

Segmental intelligibility method. This method involves the phonemic level but is generally measured at the word level by using simple syllable or word forms of the type CV, CVC, or VCV (V=vowel, C=consonant). Well-known examples of this method are the Modified Rhyme Test (MRT) (House et al., 1965), the Diagnostic Rhyme Test (DRT) (Voiers, 1977), and the use of Phonetically-Balanced (PB) CVC words. There are many aspects of these tests that require careful consideration, e.g.:

- the use of words in isolation, or in a (fixed or variable) carrier phrase
- the use of closed (MRT 6 alternatives, DRT 2 alternatives) or open response sets
- the word type (e.g. CV, CVC, or VCCV)
- the use of meaningful or nonsense words
- equal phoneme probability or phonetically balanced
- language dependence, especially relevant for rhyme tests.

Supra-segmental intelligibility requires more complex test stimuli, such as multi-syllabic words, or sentences. Emphasis is put on such aspects as word stress, word duration, syllable structure, sentence accent, and intonation. From speech audiometry, sets of carefully designed short sentences in various languages are available. However, for synthesis evaluation these sentences are less appropriate because the set is fixed and sentences are easily remembered, whereas also the grammatical structure is too simple and with insufficient variation. In the next paragraph we will discuss some alternative structures for this sentence material.

Paired comparison allows for a direct judgment of pairs of stimuli that only differ in one specific attribute, such as duration rules or intonation contours.

Magnitude estimation involves the judgment of stimuli according to one or more attributes along, say a seven-point scale. Semantic scaling theory can be applied to process this type of data.

Psycholinguistic tests are also used sometimes to evaluate the quality of synthetic speech. Some examples are

- word recall in fixed or free order
- lexical decision (word vs. non-word; sentence vs. non-sentence)
- word monitoring
- phoneme monitoring
- word gating

Speech interference tests are tests in which word or sentence intelligibility is measured against a level of masking noise (Nakatani and Dukes, 1973). Sorin (1982/83) calls this the Equivalent Signal-to-Noise Ratio method (ESNR) in her study about the contribution of pitch contours to the identification of resynthesized sentences. Other similar approaches are the speech reception threshold SRT (Plomp and Mimpen, 1979), and the monosyllabic adaptive interference test MASIT (Eggen, 1988).

Subjective ratings and questionnaires can be used to evaluate the 'linguistic' and 'psychological' aspects of speech understanding: can the sentences be reproduced, how large is the memory load, can one listen to synthetic speech for extended periods of time, can one reproduce the gist of a story, and what about the surface properties, can one comprehend the prose, do children have more difficulty with synthetic speech?

Dave Pisoni and his co-workers at Indiana University certainly have most experience with this type of testing, although it is still in its infancy.

2.3 Test material

From the short overview of test methods given above it will be clear that these various tests use a great variety of speech material, ranging from syllables and words to sentences and paragraphs. Above (under 'segmental intelligibility') we have given already some characteristics of word material, here we will concentrate on sentence material.
Phonetically-balanced short, simple, and meaningful sentences have been developed for English (Egan, 1948), French (Combescure, 1981), and Dutch (Plomp and Mimpen, 1979). The English ones became known as the Harvard Psychoacoustic Sentences (Example: Cook the corn in a large pot of water). In order to lower the predictability and in order to make them more difficult to remember with repeated presentation, Nye and Gaitenby (1974) developed syntactically correct, but semantically anomalous sentences of the type 'The late voice knew the table'. These sentences were called the Haskins sentences.
Pisoni and colleagues have used both types of sentences repeatedly to measure the word recognition in sentence context for various synthesis systems. For an overview, see Pols (1987).
For sentence verification tasks, 3- and 6-word sentences have been used, such as 'Mud is dirty' and 'Birds fly south for the winter', representing true sentences, and 'Rockets move slowly' and 'Beer is a popular contact sport', representing false sentences. Both a true-false reaction and a transcription were required from the subjects (Manous et al., 1985).
Various partners in the ESPRIT-SAM project have recently started a renewed discussion on the structure of sentence test material for synthesis evaluation. The idea of anomalous sentences is attractive since:
- it is a far more natural task than nonsense word identification, although it is of course no real language communication either;
- it hopefully allows for controlled predictability;
- it allows for controlled complexity, for instance in terms of number of words per sentence, number of syllables per word, word frequency, grammatical structure, etc.;
- it creates a very large and always different reservoir of sentences by starting from a (fixed) vocabulary from which words are randomly selected to create specific grammatical structures;
- it might be possible to develop really comparable sentences in different languages, at least in terms of word type and grammatical structure.

Presently, within SAM, we consider five different grammatical structures, instead of just one as in the Haskins sentences. Each grammatical structure will also require a different intonation contour, so, also in that respect we can run a more thorough test. Since presently none of the rule synthesizers is able to use semantic knowledge, it does not matter that the sentences are meaningless. Because of memory overload for the listener we probably will have to limit the number of words per sentence to seven.
One must keep in mind the purpose of the sentence material discussed here: evaluating word intelligibility in sentence context. So it would not be very appropriate to start studying phoneme confusions from the misidentified words. On the other hand the sentence intelligibility for real meaningful sentences will be higher than for these anomalous sentences because of semantic and pragmatic knowledge that normally can be effectively used by the listener.

It is interesting to realize that once this sentence material will be fully developed, it probably means that this test method is ahead of rule synthesizer development itself, since I do not know yet of any text-to-speech synthesis system able to extract from text, and able to generate, a number of different and appropriate prosodic realizations. This situation is contrary to that for speech recognition, where already connected word and continuous speech recognizers are available, at least as laboratory prototypes, whereas no evaluation methods at that level are available yet.

3. SOME EXAMPLES OF SYSTEMATIC EVALUATION

3.1 Segmental intelligibility

None of the presently available rule synthesizers, whether they are diphone-based or allophone-based, have such a good segmental quality that one could further neglect this level and concentrate completely on higher level processing. All present systems will gain speech quality by improving segmental intelligibility. This was true for every system that we evaluated so far:
- the dyadic rule synthesizer (Olive, 1980). By systematic evaluation and subsequent improvement of a great number of CV and VC dyads, both the initial (58.2%) and final consonant (73.5%) intelligibility could be raised to 83% (Pols and Olive, 1983).
- the phoneme intelligibility scores for various diphone-based synthesis systems in several different languages (French, Dutch, Italian) all show room for further improvement (Pols et al., 1987; van Bezooijen and Pols, 1987; van Son et al., 1988).
The absolute scores (see Table 1) are not really important since these strongly depend upon the exact experimental conditions (such as word structure (CVC vs. VCCV and CVVC), and presentation rate), but also the listeners, specific characteristics of the synthesizer (such as prediction order, window size, and bandwidth) and the complexity of the language. But as long as the intelligibility scores for rule synthesis are quite a bit lower than those for the same words resynthesized, one knows that further progress can be made. More specifically, one of the synthesizers required improvement of r-diphones, whereas for certain consonant clusters it might be better to use tri-phones or quadro-phones.

Table 1. Percentage correct phoneme and word intelligibility scores (averaged over 8 subjects) for VCCV and CVVC words, for PCM speech, LPC-resynthesized speech and Italian rule-synthesized speech.

	V	C	C	V	VCCV
PCM-coded speech	89.1	86.9	94.0	79.7	57.8
LPC-15 resynthesized speech	90.2	79.4	91.2	79.9	52.5
Italian rule synthesis	89.5	68.3	78.1	87.8	45.0
	C	V	V	C	CVVC
PCM-coded speec	94.3	90.4	84.1	88.2	63.2
LPC-15 resynthesized speech	88.4	90.8	85.3	87.5	60.2
Italian rule synthesis	74.4	86.5	84.9	76.4	44.4

- in an interactive process the segmental intelligibility of the Dutch allophone-based system will be improved step by step. The initial intelligibility was unacceptably low (van Bezooijen and Pols, 1987), but by modifying the rules and by running small specific tests, for instance for medial plosives only, the system will gradually improve.

Considering the overall consonant error rates reported for DECtalk (13.2 and 17.5 for Paul and Betty, respectively), while using the modified rhyme test with an open response set (Logan et al., 1985), I am almost certain that even this system would benefit substantially from further improvements at the segmental level.

Both for a French (van Son and Pols, 1988) and for two Dutch systems (van Bezooijen, 1988), the intelligibility of consonant clusters was measured recently. Because of the great flexibility of the Dutch language to combine words, the number of medial clusters is almost unlimited, so the test was restricted here to initial and final clusters. However, for French medial (within-word) clusters were taken into account. See Table 2 for some overall results.These data still have to be studied in more detail in order to specify in which way the necessary improvements can be made most effectively.

Table 2. Some overall intelligibility results for initial, medial, and final consonant clusters for French. Scores are percentages correct averaged over 8 subjects.

	Nclusters	PCM	LPC-resynth.	rule-synth.
initial clusters	72	92.0	86.7	62.5
medial clusters	70	85.8	84.5	76.6
final clusters	48	98.2	96.3	70.7

3.2 Supra-segmental intelligibility

Relatively few results have so far been achieved with this level of speech quality evaluation. Greene et al. (1984) used the Harvard and Haskins sentences to evaluate DECtalk (two voices: Paul and Betty). The same did Manous et al. (1984) for Speech Plus Prose- 2000 prototype. In 1980, Pisoni and Hunnicutt had already done this for MITalk-79. Very recently Hazard and Grice (1988) ran a pilot test with newly developed English sentences with the same grammatical structure as the Haskins sentences: 'The ADJ NOUN1 VERB the NOUN2'. Table 3 summarizes the results of these various studies. It will be possible to do more interesting tests as soon as sentences with several different grammatical structures become available; these will require different prosodic characteristics and will introduce more variation for the listeners.

3.3 Quality judgment of intonational aspects in speech

In a recent attempt to improve substantially the prosodic characteristics of rule-synthesized speech, Terken systematically studied natural speech and came up with better rules for intonation. These were evaluated by listening experiments with rule-synthesized diphone speech (Collier and Terken, 1987).

Table 3. Percentage correct word intelligibility scores for natural and synthetic speech using 'Haskins-type' sentences.

	natural	synthetic	type of synthesizer
Nye and Gaitenby (1974)	95	78	Haskins lab. system
Pisoni and Hunnicutt (1980)	97.7	78.7	MITalk-79
Greene et al. (1984)	97.7	86.8/75.1	Paul/Betty DECtalk
Manous et al. (1984)	97.7	64.0	Speech Plus
Hazan and Grice (1988)	98.1	76.6	JSRU synth.-by- rule

For French, and meanwhile for several other languages as well, a set of 20 sentences has been created. These sentences, in principle, should allow for testing several text-to-speech modules such as phonetic rules, diphone concatenation, and prosodic processing (SAM Extension phase report, 1988). The corpus contains simple as well as complex sentences, with words of various complexity in terms of length, stress, affix structure, morphological structure, phoneme realization, etc.

3.4 Quality judgment of prosodic analyses from text

Kager and Quené (1987) are developing an algorithm that, directly from Dutch text, derives pause locations and can indicate which words should get sentence accent. A first performance check was done by comparison with actual realizations of a specific speaker. However, a better check would be to run listening experiments on acceptability in order to study perceptual tolerance. These experiments are presently in preparation.

4. CONCLUSIONS

Although phoneme and word intelligibility of most rule synthesis systems is not yet good enough, there is a growing need for intelligibility and acceptability tests at the sentence level. The use of unpredictable, anomalous, short and rather simple, sentences seems to be a good choice at the intelligibility level. Grammatically more complex and longer sentences are generally required for naturalness and acceptability judgments. Only multilingual standardization will allow for comparison of performance figures.

ACKNOWLEDGEMENTS

In the SAM and SPIN project we work together with TNO Institute for Perception in Soesterberg and get support from ESPRIT. The program on 'Analysis and synthesis of speech' is supported by the Dutch SPIN.

REFERENCES

Allen, J., Hunnicutt, M.S. & Klatt, D.H. (1987). From text to speech. The MITalk system, Cambridge Univ. Press, 216 pag.

Bezooijen, R. van (1988). "Evaluation of the quality of consonant clusters in two synthesis systems for Dutch", Proc. SPEECH '88, 7th FASE Symp., Edinburgh, Book 2, 445-452.

Bezooijen, R. van & Pols, L.C.W. (1987). "Evaluation of two synthesis-by-rule systems for Dutch", Proc. Europ. Conf. Speech Techn., Edinburgh, Vol. 1, 183-186.

Collier, R. & Terken, J. (1987). "Intonation by rule in text-to-speech applications", Proc. Europ. Conf. Speech Techn., Edinburgh, Vol. 2, 165-168.

Combescure, P. (1981). "20 listes de dix phrases phonétiquement équilibrées", Revue d'Acoustique 56, 34-38.

Egan, J.P. (1948). "Articulation testing methods", Laryngoscope 58, 955- 991.

Eggen, B. (1988). "Evaluation of speech communication quality with a Monosyllabic Adaptive Speech Interference Test", to be published in Speech Comm.

Greene, B.G., Manous, L.M. & Pisoni, D.B. (1984). "Perceptual evaluation of DECtalk: A final report on Version 1.8", Research on Speech Perc., Progress Report 10, Indiana Univ., 77-127.

Goodman, D.J. & Nash, R.D. (1984). "Subjective quality of the same speech transmission conditions in seven different countries", IEEE Trans. Comm. 30, 642-654.

Hampshire, B., Ruden, J., Carlson, R. & Granström, B. (1982). "Evaluation of centrally produced and distributed synthetic speech", STL- QPSR 2-3, 18-23.

Hazan, V. & Grice, M. (1988). "Intelligibility tests for the assessment of synthetic speech using semantically anomalous sentences", SAM-report, University College London.

House, A.S., Williams, C.E., Hecker, M.H.L. & Kryter, K.D. (1965). "Articulation testing methods: Consonantal differentiation with a closed response set", J. Acoust. Soc. Amer. 37, 158-166.

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

Kryter, K.D. (1962). "Methods for calculation and use of the articulation index", J. Acoust. Soc. Amer. 34, 1689-1697.

Logan, J.S., Pisoni, D.B. & Greene, B.G. (1985). "Measuring the segmental intelligibility of synthetic speech: Results from eight text-to-speech systems", Research on Speech Perc., Progress Report 11, Indiana Univ., 3-31.

Manous, L.M., Greene, B.G. & Pisoni, D.B. (1984). "Evaluation of Prose - The Speech Plus text-to-speech system. I. Phoneme intelligibility and word recognition in meaningful sentences", Speech Research Lab. Technical Note 84-04, Indiana Univ.

Manous, L.M., Pisoni, D.B., Dedina, M.J. & Nusbaum, H.C. (1985). "Comprehension of natural and synthetic speech using a sentence verification task", Research on Speech Perc., Progress Report 11, Indiana Univ., 33-57.

Nakatani, L.H. & Dukes, K.D. (1973). "A sensitive test of speech communication quality", J. Acoust. Soc. Amer. 53, 1083-1092.

Nye, P.W. & Gaitenby, J.H. (1974). "The intelligibility of synthetic mono-syllabic words in short syntactically normal sentences", Haskins Labs. SR-37/38, 169-190.

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

Pisoni, D.B. (1982). "Perception of speech: The human listener as a cognitive interface", Speech Techn., Vol. 1, Nr. 2, 10-23.
Pisoni, D.B. & Hunnicutt, S. (1980). "Perceptual evaluation of MITalk: The MIT unrestricted text-to-speech system", Proc. IEEE-ICASSP80, 572-575.
Plomp, R. & Mimpen, A.M. (1979). "Improving the reliability of testing the speech reception threshold for sentences", Audiology 8, 43-52.
Pols, L.C.W. (1987). "Quality evaluation of text-to-speech synthesis systems", Deliverable of ESPRIT-project 1541 SAM, also IFA-report nr. 94, 31 pag.
Pols, L.C.W. (1988). "Joint Dutch research program for developing a high-quality text-to-speech synthesis system", written version of paper at ASA/ASJ Meeting, Honolulu, Nov. 1988. See also this IFA Proceedings.
Pols, L.C.W., Lefèvre, J.-P., Boxelaar, G.W. & Son, N. van (1987). "Word intelligibility of a rule synthesis system for French", Proc. Europ. Conf. Speech Techn., Edinburgh, Vol. 1, 179-182.
Pols, L.C.W. & Olive, J.P. (1983). "Intelligibility of consonants in CVC utterances produced by dyadic rule synthesis", Speech Comm. 2, 3-13.
Pratt, L.R. (1987). "Quantifying the performance of text-to-speech synthesizers", Speech Techn., Vol. 3, Nr. 4, 54-64.
SAM-partners (1988). "Multilingual speech assessment methods (SAM)", Proc. SPEECH '88, 7th FASE Symp., Edinburgh, Book 1, 137-143.
Sorin, C. (1982/1983). "Evaluation de la contribution de F0 à l'intelligibilité", Recherches Acoustiques, CNET, Vol. 7, 141-155.
Son, N. van, Pols, L.C.W., Sandri, S. & Salza, P.L. (1988). "First quality evaluation of a diphone-based speech synthesis system for Italian", Proc. SPEECH '88, 7th FASE Symp., Edinburgh, Book 2, 429- 436.
Steeneken, H.J.M. (1982). "Ontwikkeling en toetsing van een nederlandstalige diagnostische rijmtest voor het testen van spraakkommunikatiekanalen", TNO Inst. for Perception, report IZF 1982-13, 30 pag.
Steeneken, H.J.M. & Houtgast, T. (1980). "A physical method for measuring speech-transmission quality", J. Acoust. Soc. Amer. 67, 318- 326.
Voiers, W.D. (1977). "Diagnostic evaluation of speech intelligibility", In: M. Hawley (Ed.), Speech intelligibility and speaker recognition, Dowden, Hutchinson and Ross, Stroudsburg, 374-387.