PARTICIPATION IN (INTER)NATIONAL PROJECTS ON SPEECH TECHNOLOGY EVALUATION

Louis C.W. Pols

1. INTRODUCTION

Compared to the amount of money and manpower spent on the development of speech technology systems, there is relatively little attention for assessing the performance of those systems. However, such evaluations could be very useful:
- for comparing different systems already on the market,
- as an objective measure to define the improvement of systems under development,
- for diagnostic purposes,
- and for specifying the sensitivity to various (external) parameters, such as background noise, vocabulary size, or speaker variability.

Field tests under actual conditions frequently show deficiencies or reduced performance compared to laboratory tests under controlled conditions. Our institute recently became involved in three speech technology projects in which our participation is mainly focussed on the evaluation part. I will briefly introduce the three projects here and will each time describe in somewhat more detail our (planned) contribution with respect to evaluating the speech quality of the various systems.

2. ESPRIT PROJECT 64 "SPEECH INTERFACE AT OFFICE WORK-STATION" (SPIN)

This five-year project started in July 1984 and has industrial partners in three European countries (France, Germany, and Italy) and university partners in two other countries (Greece and Holland). The Dutch participation is a joint venture between the Institute of Phonetic Sciences of the University of Amsterdam and the TNO Institute for Perception in Soesterberg. As far as the speech interface is concerned, algorithmic and hardware development will take place at the following levels:
- Speech coding. Various coding algorithms are studied for a speech store-and-forward application. By now it is quite sure that a multi-pulse linear predictive coder at 9.6 kb/s will be implemented.
- Speech recognition. Research will concentrate on speaker-independent isolated-word recognition with an extension to connected words.
- Speaker verification. Admissible users of a workstation will have to be verified on the basis of a few spoken words, and then accepted, whereas impostors should be rejected.
- Speech synthesis. For French, Italian, and Greek, text-to-speech synthesis-by-rule systems will be developed based upon diphone dictionaries, prosodic rules, a unified rule compiler, and linguistic processing.
- Hardware implementation. This will involve the development of a general speech interface connected with an existing workstation, as well as

dedicated hardware with a modular architecture for coding, synthesis, recognition, and verification. For coding also VLSI will be considered.
- Ergonomy. Two applications of the speech interface together with other interfaces in an office workstation are envisaged: document preparation and telephone management.
- Evaluation. Only the speech quality of the coders and the rule-synthesizers will be evaluated, no resources were available for evaluating speech input performance as well.

For this, as well as for all other ESPRIT projects, half-yearly interim reports are produced, from which the technical parts are publicly available. With respect to evaluation methods a state-of-the-art report was written (Boxelaar and Pols, 1985), as well as a report about a preliminary evaluation of the word intelligibility and the speaker identifiability of five medium-band coders (Boxelaar, this volume). For a general outline of the topics involved in comparative evaluation of the speech quality of speech coders and text-to-speech synthesizers, I refer to Pols and Boxelaar (1986). In that paper one can also find a short description of a portable, stand-alone, multi-subject intelligibility-testing device which was developed in order to ease the execution of word intelligibility tests with up to four subjects at a time and in which various levels of data processing are efficiently programmed.
The word intelligibility measurements for the various speech coders were performed by using phonetically-balanced lists with 50 nonsense words each, of the type consonant-vowel-consonant, spoken by three male and three female speakers. As a reference system, an analog telephone-bandwidth speech channel was used. None of the coder software prototypes showed as good a performance as the reference system.
Speaker identifiability was tested with increasingly longer speech fragments from isolated words up to two sentences, spoken by the same six speakers and processed by the various coders. These speakers belonged to a laboratory population of 130, and were known to the 12 listeners who had to identify the speakers. Listener performance was not very stable and for a few speaker-coder combinations, the speaker was not recognized at all, sometimes not even when the final (longest) fragment was presented again in its natural form. This area of voice identification is an almost unexplored area of research (Schmidt-Nielsen and Stern, 1985).

The rule-synthesis systems for the three specific languages (French, Italian, and Greek) are still under development and cannot yet be tested as complete text-to-speech systems. However, as a first step a systematic evaluation of all diphones in the diphone dictionaries will be done. For the full set of 1250 French diphones this has been done recently by using CVVC- and VCCV-type words, the responses from the eight native French listeners are presently being processed (Pols, Lefèvre, and Boxelaar, in preparation). Subsequent tests will involve suprasegmental aspects of the rule-synthesis systems, like duration and intonation at word and sentence level.

3. SPIN PROJECT "ANALYSIS AND SYNTHESIS OF SPEECH"

This five-year research program started in December 1985 and is funded by the Dutch SPIN organization, a joint effort from three governmental departments to promote strategic research in information technology. The project is carried out jointly by four phonetic institutes at the universities of Am-

sterdam, Leiden, Nijmegen, and Utrecht as well as by the Institute for Perception Research (IPO) in Eindhoven. The major aims of this program are:

- Integrating existing expertise and extending new expertise among partners, with respect to analysis and synthesis of speech.
- Requiring more insight into fundamental knowledge necessary to achieve a laboratory prototype of a fully automatic text-to-speech conversion system for Dutch of high speech quality.
- Building software and hardware systems for analyzing and (re)synthesizing speech. As a first step all five laboratories have, or have acquired (partly through this program), comparable hardware (VAX or microVAX under VMS).
- Transmission of jointly-acquired knowledge for distribution to industry.

Research topics will include analysis and (re)synthesis methods (e.g. LPC, ARMA, temporal decomposition, multipulse excitation, speaker characteristics), grapheme-to-phoneme conversion (e.g. rules vs. lexicon, syllable vs. morpheme boundaries, relevance of morphological/syntactic analysis), basic units (allophones vs. diphones), rule compiler, intonation and prosody, and text composition. Beyond those substantial attention will be devoted to a diagnostic, comparative evaluation of the various phases of development of the rule synthesizers. Especially at the suprasegmental level, new evaluation methods will have to be developed.
Recently a state-of-the-art intelligibility evaluation has been executed (at the word level) for the two available systems, one diphone-based (Elsendoorn, 1984), one allophone-based (Boves et al., 1986) in the form they presently are. In order to test all diphones in an efficient way, the same word forms (CVVC and VCCV) as mentioned before have been used, however, extended with CVC and VCV words. These additional words make it possible to test, for instance, also CV diphones with short vowels, which cannot occur in open syllables in Dutch. The PC-based system for efficiently performing these word intelligibility tests, also has been used for this evaluation (van Bezooijen and Pols, in preparation).
Subsequent tests will include words with consonantal clusters, and multisyllabic words in order to evaluate stressed and unstressed syllables. After that, the prosodic rules will be diagnostically evaluated with short sentences. The synthesized sentences will be compared with natural utterances, as well as with various deviated forms in between the natural and the rule-synthesized sentence. The comparison could be a preference judgment or a scaled quality judgment using various terms.

4. ESPRIT PROJECT "MULTILINGUAL SPEECH INPUT-OUTPUT ASSESSMENT, METHODOLOGY, AND STANDARDIZATION"

This is a project proposal which still has to be approved by ESPRIT, although a (preliminary) start in the form of a so-called "definition phase" will take place early 1987. This three-year program is building upon several existing national research programs on, or involving, speech technology assessment. As such programs can be mentioned the French GRECO program on speech data bases (Carré et al., 1984), the British ALVEY program on speech input technology performance (Holmes, 1985), or the Dutch SPIN program mentioned earlier. This new Esprit project intends to develop and provide databases, speech workstations, protocols, and methodologies which will enable speech synthesizers and speech recognizers to be assessed on a European basis (multilingual methodology). Presently the partners come

from the UK, France, Italy, Denmark, and Holland. German and Greek participation is being considered. The Dutch partner will be the TNO Institute for Perception, Soesterberg in collaboration with the Institute of Phonetic Sciences - Amsterdam, whereas the PTT speech research laboratory at the Dr. Neher Laboratory has also shown a vivid interest. The development and promulgation, on a European-wide basis, of universally accepted standards for the assessment of present and future generations of speech technology equipment is the major aim. One can specify three components:

- The evaluation and development of materials using, for example, corpora of word-, sentence-, and continuous-speech type.
- The application of those materials together with quantitative subjective test methods to assess speech engineering products and systems.
- The support of fundamental research into the nature of speech production, perception, and processing, so that future generation speech technology devices can be both developed and assessed.

At various places in the world speech data bases are being compiled. However, one does not recognize much cooperation in those efforts, the aims are all different, the amount of segmentation and labelling varies (from not-at-all to very-detailed), the storage media differ widely (from analog and digital storage to PCM/VCR technology), and the accessibility and exchangeability is most of the time a problem.
Baker et al. (1983) give an overview of data bases as of 1983. Pallett (1985; 1986) from the American National Bureau of Standards gives suggestions for standardized recording and testing.
The NATO RSG10 speech data base (Bridle et al., 1983; Pols, 1982) seems to be the only multilingual one, but is limited to isolated and connected digits.
The French GRECO speech data base is rather extensive and is supposed to be a basis for both speech technology assessment as well as for phonetic research, however, so far the material is unlabelled and not yet accessible by a data base management system. Storage on optical disc or CD techniques are being considered.
The new DARPA speech data base is probably the most advanced in several aspects. It involves a number of sentences spoken by 600 speakers from 7 major regional dialects. All these sentences will be phonetically transcribed and provided with time segmentation markers by using the Spire facility at MIT (Zue et al., 1986).
Several military research establishments are also gathering large data bases, mainly with specific military requirements, such as choice of specific vocabulary, use of oxygen mask or g-forces, or high noise levels.
Japanese speech research laboratories have worked together to compile a speech data base of 15 digits, 35 digit sequences, 63 function words, 110 monosyllables (complete set for Japanese), 110 city names, and a few sentences spoken by 75 male and 75 female speakers (Itahashi, 1986).

For diagnostic evaluation of speech recognizers and speech synthesizers it is very well possible that use can be made of speech material available for measuring the intelligibility of speech communication channels (Goodman and Nash, 1982) or of speech material used in speech audiometry (Kapteyn and Smoorenburg, 1985). Logan et al. (1985) used for instance the modified rhyme test (MRT), both with six response alternatives (House et al., 1965) as well as with an open response version, to measure the initial- and final-consonant intelligibility for eight different text-to-speech systems including one male and one female voice of DECtalk (Gutcho, 1985), the English

version of the Swedish multilingual Invofox (Carlson et al., 1982), and some other rather simple systems like Votrax Type'n'Talk and Street Electronics Echo.
Furthermore it is certainly worth considering the concept of objective methods, as used for testing the intelligibility of communication channels, like the articulation index (AI) (Kryter, 1962) or the speech transmission index (STI) (Steeneken and Houtgast, 1980).
Apart from using speech data bases, there are certainly various other (more basic) approaches to evaluate speech recognizers, like comparison with a reference system (Chollet and Gagnoulet, 1982), or using the human equivalent noise reference (HENR) (Moore, 1977), the vocabulary or phoneme confusion matrix, the effective vocabulary capability (EVC) (Taylor, 1981), or the relative information loss (RIL) (Woodard and Nelson, 1982). These and other approaches will certainly be considered in this new ESPRIT project.

5. REFERENCES

Baker, J.M., Pallett, D.S. & Bridle, J.S. (1983), "Speech recognition performance assessments and available data bases", Proc. IEEE-ICASSP83, 527-530.

Bezooijen, R. van & Pols, L.C.W. (in preparation), "Evaluation at the word level of two text-to-speech synthesis systems for Dutch", Paper to be presented at the 11th Int. Congress of Phonetic Sciences, Tallinn, Aug. 1987.

Boxelaar, G.W. (this volume).

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, 21 pages.

Boves, L., Buiting, H., Kerkhoff, J. & Wester, J. (1986), "Automatic text-to-speech conversion and vice versa", IFN-Proceedings 10, 18-19.

Bridle, J.S. et al. (1983), "Connected word recognition for use in military systems", AC/243 (Panel III/RSG10) Project One Report, 43 pages.

Carlson, R., Granström, B. & Hunnicutt, S. (1982), "A multi-language text-to-speech module", Proc. IEEE-ICASSP82, 1604-1607.

Carré, R., Descout, R., Eskenazi, M., Mariani, J. & Rossi, M. (1984), "The French language database: Defining, planning, and recording a large database", Proc. IEEE-ICASSP84, 42.10.1-4.

Chollet, G. & Gagnoulet, B.P. (1982), "On the evaluation of speech recognizers using a reference system", Proc. IEEE-ICAASSP82, 2026-2029.

Doddington, G.R. & Schalk, T.B. (1981), "Speech recognition: turning theory to practice", IEEE Spectrum 18, 26-32.

Elsendoorn, B. (1984), "Heading for a diphone speech synthesis for Dutch", IPO Annual Progress Report 19, 32-35.

Goodman, D.J. & Nash, R.D. (1982), "Subjective quality of the same speech transmission conditions in seven different countries", IEEE Trans. COM 30, 642-654.

Gutcho, L. (1985), "DECtalk - A year later", Speech Techn. 3, 98-102.

Holmes, J.N. (1985), "Speech technology in the U.K. Alvey Program", Speech Techn. 3, 44-47.

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.

Itahashi, S. (1986), "A Japanese language speech database", Proc. IEEE-

ICASSP86, 321-324.
Kryter, K.D. (1962), "Methods for the 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 Perception Progress Report 11, 3-31.
Moore, R.K. (1977), "Evaluating speech recognizers", IEEE Trans. ASSP 2, 178-183.
Pallett, D.S. (1985), "Performance assessment of automatic speech recognizers", J. Res. Nat'l Bureau of Standards 90, 371-387.
Pallett, D.S. (1986), "A PCM/VCR speech database exchange format", Proc. IEEE-ICASSP86, 317-320.
Pols, L.C.W. (1982), "How humans perform on a connected-digits data base", Proc. IEEE-ICASSP82, 867-870.
Pols, L.C.W. & Boxelaar, G.W. (1986), "Comparative evaluation of the speech quality of speech coders and text-to-speech synthesizers", Proc. IEEE-ICASSP86, 901-904.
Pols, L.C.W., Lefèvre, J.-P. & Boxelaar, G.W. (in preparation), "Intelligibility of words produced by a rule synthesis system for French", Paper to be presented at European Conf. on Speech Techn., Edinburgh, Sept. 1987.
Schmidt-Nielsen, A. & Stern, K.R. (1985), "Identification of known voices as a function of familiarity and narrow-band coding", J. Acoust. Soc. Amer. 77, 658-663.
Steeneken, H.J.M. & Houtgast, T. (1980), "A physical method for measuring speech-transmission quality", J. Acoust. Soc. Amer. 67, 318-326.
Taylor, M.M. (1981), "Issues in the evaluation of speech recognition systems", DCIEM draft paper.
Woodard, J.P. & Nelson, J.T. (1982), "An information theoretic measure of speech recognition performance", In: Proc. Workshop on standardization for speech I/O technology, Ed. D.S. Pallett.
Zue, V.W., Cyphers, D.S., Kassel, R.H., Kaufman, D.H., Leung, H.C., Randolp M., Seneff, S., Unverferth, J.F. III & Wilson, T. (1986), "The development of the MIT Lisp-machine based speech research workstation", Proc. IEEE-ICASSP86, 329-332.