# A JOINT DUTCH RESEARCH PROGRAM FOR DEVELOPING A HIGH-QUALITY TEXT-TO-SPEECH SYNTHESIS SYSTEM *

Louis C.W. Pols

ABSTRACT

Since November 1985 a Dutch national research program is on its way aiming at the realization of a laboratory prototype system for high-quality unlimited text-to-speech synthesis-by-rule for the Dutch language. Six research laboratories work together in this program supported by SPIN (Dutch National Program for the Advancement of Information Technology). About 15 subtasks have been defined ranging from defining the optimal structure of spoken text, morphological parsing and grapheme-to-phoneme conversion, to studying speaker characteristics, intonation contours, and spectro-temporal speech parameters as a function of sentence accent, speaking rate, and local context. Both allophones and diphones are used as basic units. Formal methods for evaluating the segmental and supra-segmental speech quality are being developed and applied during the various phases of realization. In the final two years of the project an optimal system architecture will be developed in order to integrate the single components and in order to achieve an operational implementation.

## 1. INTRODUCTION

Despite the availability of several commercial text-to-speech synthesizers, mainly for American English, there is still a strong need for more fundamental knowledge to improve the speech quality of these and other systems.
Even MITalk (Allen et al., 1987), or its commercialized versions, which probably can be considered as the presently most advanced system with the highest quality, still has many drawbacks (Klatt, 1987) not to talk about other simpler systems. These shortcomings are reflected at many different levels, from text input to spoken output:
- it is unknown which text structures are best to be spoken
- complex grammatical structures are not recognized
- semantic knowledge is not used
- word stress and sentence accent are not well represented
- variable speaking rates and appropriate pauses are not realized
- the system generally speaks with only one (male) voice
- various speaking styles and emotional states are not considered
- appropriate intonation contours are desired
- other prosodic aspects need more attention

* Written version of invited paper presented at special session on 'Speech Processing in Human-Machine Interaction - An International View', at the ASA/ASJ Conference in Honolulu, Hawaii, 14-18 November 1988. This paper was given on behalf of the SPIN Steering Committee.

- synthetic speech is generally overarticulated, whereas unstressed syllables are lacking
- segmental characteristics are far from optimal
- systems should have a more natural speech quality, allowing for extended listening, also in noise and over the telephone
- at least in Europe a universal approach for many different languages would be desirable.

For these and many other reasons several national and international research programs for improving speech synthesis have recently been initiated.
Such a program was also suggested in the Netherlands by a working group on Language and Speech Technology in 1984. Fortunately somewhat later also the possibilities to finance such a program became available through a governmental stimulation program in Information Technology. Quickly a specific research program on 'Analysis and Synthesis of Speech' was written and approved.
The program started in November 1985 with a total budget of 6 million Dutch guilders (1 US$ is about 2 Dfl). Originally five, later six, institutes participated. The program is coordinated by A. Cohen (Utrecht). A Steering Committee, with a representative from each of the institutes, is responsible for the content of the program and for the progress of activities:
- Institute of Phonetic Sciences, University of Amsterdam (L.C.W. Pols)
- Institute of Phonetics, Utrecht (M.P.R. van den Broecke, later S.G. Nooteboom)
- Institute of Phonetics, Catholic University Nijmegen (L. Boves)
- Phonetics Laboratory, Leyden University (V. van Heuven)
- Institute for Perception Research IPO, Eindhoven (S.G. Nooteboom, later R. Collier)
- PTT Dr. Neher Laboratories, Leidschendam (B. van Heugten).

A Program Committee, made up of experts and representatives of industry, has an advisory and monitoring function.
Part of the budget was used to extend the existing computer facilities at the various institutes in such a way that optimal cooperation and data exchange would be possible. For historical and practical reasons we normalized on VAX systems under VMS. This also permits straightforward use of the computer network SURFNET. The remaining budget is used to appoint researchers on specific projects, which will be discussed in more detail below.

## 2. AIMS OF THE PROJECT

The overall general aims are fourfold:
- Integrating existing expertise and developing it still further among the participating research groups with respect to analysis and synthesis of speech;
- Acquiring more insight into the fundamental knowledge necessary for future application in a number of specific research areas;
- Building software and hardware systems for analyzing and (re)synthesizing speech;
- Transmission of knowledge jointly acquired for distribution to industry.

More specifically the following parallel and partly competing research projects have been initiated (the code is very global and just indicates a category: L for linguistic, P for prosodic, A for acoustic, R for realization, and E for evaluation; also the names of the researchers involved are given):
Project L1 - Optimal characteristics for texts to be spoken (Irena Petric)
Project L2 - From text to speech via a lexicon (Jo Lammens)

Project L3 - Morphological decomposition by using pattern matching (Jeroen Reizevoort)
Project L4 - Morphological parsing using a morpheme lexicon (Josée Heemskerk)
Project L5 - Automatic prosodic analysis without using a lexicon (René Kager and Hugo Quené)
Project P1 - Intonation (Jacques Terken)
Project P2 - Rate changes (Wieke Eefting)
Project P3 - Effects of speaking rate on spectro-temporal characteristics of speech (Rob van Son)
Project A1 - Physical correlates of perceived speaker identity and speaker characteristics (Berry Eggen)
Project A2 - Pole-zero analysis (Johan de Veth)
Project A3 - Rules for allophone synthesis by analyzing diphone speech (Louis ten Bosch)
Project A4 - Rules for allophone synthesis (Henk Loman)
Project A5 - Stressed and unstressed diphones (Rob Drullman)
Project R1 - Stand-alone text-to-speech system (René Deliège)
Project R2 - Software development and exchange (Daan Broeder)
Project R3 - Technical support (Philippe Alain)
Project R4 - Implementation (vacancy)
Project E1 - Evaluation of the quality of synthetic speech (Renée van Bezooijen).

Although we consider this, according to Dutch standards, a very big cooperative project, it is equally clear that this effort in itself will not suffice to solve the full problem of high quality text-to- speech synthesis for Dutch. It is therefore mandatory that whatever will be developed in this program, as individual or integrated components, will remain available for further research. This is also important in view of the integration of this work with other related activities at national and international level. Although every single project had at its start a well-defined research program with steps to go and goals to achieve, and although project leaders, researchers, and discussion teams were carefully chosen, the projects so far were executed in a rather isolated way. This will have to change in the final two years of the project in order to allow for a real integration of the various components into an operational prototype system.

## 3. POSSIBLE STRUCTURE FOR THE INTEGRATED PROTOTYPE

### 3.1 Modules

Although the discussion about the structure that will be chosen is not at all finished yet, it may nevertheless be useful to indicate in which direction we are thinking. On the one hand it shows what we consider to be a realistic and advanced system, on the other hand it shows where the present projects fit, what should be extended, integrated, or perhaps stopped, and what is still missing.
In our discussions about a useful structure for the total system we are certainly influenced by the DELTA system (Hertz et al., 1985). We presently consider some 15 modules (M) which will (probably) be put in a strict serial order. The data structure will be parallel with synchronized data streams. The output of each module should be synchronized with the information already present in the data stream. Below we describe the modules as they have so far been identified. The order in which they are described is probably also the order in which they should operate, although changes are still possible.

M1. TEXT-SCAN
This first module will use the sequence of characters as its input. It specifies which sequence of characters forms a word, and also what type of word (lexical, space, punctuation, number, abbreviation, acronym), and which sequence of words forms a sentence. Perhaps, also paragraphs can be identified. This module is more or less available.

M2. TEXT-ANAL
This text analysis module is a somewhat speculative one since it should determine various text characteristics such as whether this text is a database, or instructive, or informative in content, whether the voice should be that of a man, a woman, or a child, whether the rate should be slow, normal, or fast, and whether the declamation should be lifely, business-like, or for children. This module is presently non-existent, although project L1 should make a contribution to this area.

M3. EXPANDER
This expander module specifies the words in their full graphemic form. It handles such things as capital letters, numbers, abbreviations. At several research institutes in the Netherlands such a module is available.

M4. MORF-G
Here the morphological structure of each lexical word is specified in order to guide the grapheme-to-phoneme conversion and the word stress assignment. The characteristics per morpheme are also assigned, such as prefix, stem, or suffix, native or foreign origin, but also whether the suffix carries word stress, attracts it, or is neutral in that respect. The grammatical word class is indicated (noun, adjective, verb).
Presently three different approaches are followed in the Netherlands, one uses rules and phonotactic constraints to detect morpheme boundaries (Berendsen and Don, 1987; Kerkhoff et al., 1984; Daelemans, 1988), the other is pattern-based (Project L3), whereas the third is lexicon-based (Project L4). The last one is attractive for several reasons, such as the possible completeness of a morpheme set (presently 14,264 morphemes) with information about pronunciation and grammatical category (important since also for Dutch the righthand member of a morphologically complex word determines word category). Unfortunately, the segmentation (including spelling rules to recognize transformed morphemes) will frequently lead to several alternative morphemic representations of an input word. Through CELEX (1985) we have access to a testbed of 130,000 morphematized words of which 3,876 were randomly selected for a first test. For 91% of these words a correct analysis was generated, for 75% the correct analysis was also the most likely candidate, for 30% it was the only candidate, on average 3.6 alternatives per word were found. For only about half of the errors the incorrect alternative will actually lead to a wrong pronunciation. Further progress is foreseen by improving the morphological filters and by including probabilities. A subset of high-frequency words can also be added to the lexicon (Baart and Heemskerk, 1988). The performance of this system should be compared with the other two approaches. The rules in a rule-based system can become very complex and mutually dependent, whereas the word category is not provided. The pattern-matching approach is based on Liang (1983), who developed it originally for word hyphenation. He felt that a lexicon could never be complete whereas rules would never be perfect and always language dependent. The idea is to test with every word in a text lexicon how well the present pattern match performs. Whenever there is an error, store as many extra symbols as necessary to achieve again a correctly recognized pattern. One has to try to find an optimum for keeping the best patterns. This approach in principle is very fast and flexible. By applying lexical morphology, not just morpheme boundaries but

also boundary markers are used. From a test lexicon of 123,093 morphematized words, 64,096 words with native affixes were selected. The program generated for this set 9,932 different patterns, describing 84,6% of the morpheme boundaries.

M5. SYNT
Unfortunately, this syntactic module is right now a rather speculative one within this project. It should define the syntactic structure of the sentence in terms of such categories as subject and predicate. Baart (1987) has done some work in this direction. What is available, however, is a Markov parser developed in ESPRIT project 291/860 (Linguistic analyses of European Languages) (Vittorelli, 1987).

M6. PROS
The prosodic structure of each sentence in terms of intonation domains (separated by 'heavy boundaries' frequently realized by pauses) and phi domains (roughly identical to syntactic word groups or phrases) is derived in this module. Project L5 is totally devoted to this problem and has already achieved some very interesting results (Kager and Quené, 1987). It derives the prosodic structure directly from text without full syntactic parsing. From a lexicon of about 400 function words it is decided whether a word is a function word; all other words are labeled as content words. Grammatical word categories, like verbs and adjectives, are only specified in as far as they are relevant for prosodic analysis including accent marking.
In a first evaluation of this algorithm, naturally realized pauses and accents in a read-aloud text (706 sentences, 12,129 words) were compared with rule-derived pause locations (at boundaries of intonation domains) and words indicated as getting sentence accent. From all actually realized heavy boundary positions (1,570) the program correctly indicated 82%, from all others (9,853) 3% was wrongly marked. From all (4,458) accentuated words 83% was correctly predicted, from all other words (9,853) 7.4% was wrongly marked as having sentence accent. Of course one specific natural realization cannot be the only check, a listening test will follow to study the perceptual tolerance.

M7. INTON-PATT
This presently non-existent module should specify which intonation pattern (in terms of the IPO intonation grammar: 't Hart and Collier, 1975)) should be attributed to which intonation domain.

M8. TEMPO
This module should specify the speaking rate at which an intonation domain should be produced. Again this module is far from being realized, but project P2 should contribute to this (Eefting, 1988), whereas also project P3 will give more insight in the ways these rate changes should be realized in a spectro-temporal way. A detailed analysis of all vowel segments in one page of text spoken by our 'norm' speaker already showed that this highly trained speaker did not show any additional reduction from normal to fast speech, at least in terms of the formant position of one 'stationary' position in each vowel segment (van Son and Pols, 1988). There was of course very much variation in vowel realization depending on local context, but very little variation in 'stationary' vowel realization for the same segment in normal- or fast-rate speech. Next, the dynamic formant contours will be studied in more detail.

M9. SYL-G
Whether this module, which defines the syllable structure at the graphemic level, is necessary at all, is not yet clear.

M10. GRAFON
For the grapheme-to-phoneme conversion there are several options of existing programs for Dutch.

M11. STRESS
Once the phoneme transcription is available, this module should define the syllable structure of each word and add to that the word internal stress. In several existing grapheme-to-phoneme conversion programs this part is included. Langeweg (1988) also developed a separate module.

M12. FONOLOG
This again may be part of GRAFON, it controls several phonological processes which necessarily have to come after STRESS.

M13. ALLOFON
Depending on the type of synthesizer, this module defines the synthesis elements.

M14. INTON-ATOM
This module specifies which actual intonation contours should be used on the basis of all preceding information. Project P1 is working towards that goal (Collier and Terken, 1987). So far the intonation of longer passages of speech sounds rather boring, probably because of the application of always the same standard recipe (fixed sequence, size, and timing of movements with a fixed declination). Variation in size of pitch movements, declination resets, and variation in type of hat patterns all appeared to contribute to a more lively intonation. More study will be made of the declination parameters in relation to convergence domains, both in isolated sentences and for sentences in context. Finally micro-intonation at the segmental level will be studied.

M15. DURATION
This module specifies the actual spectral information per frame on the basis of the earlier defined (allophonic) synthesis elements. Also the segmental duration is defined here. Some work in this area has been done at IPO, mainly along the line of adapting the Klatt rules for American English to British English and to German.

The control modules for the acoustic front-end for this synthesizer are not yet specified in much detail. This has to do with the fact that so far two possible approaches have been left open. One is diphone-based, the other is allophone-based.

3.2 Diphone-based approach
Presently there are already two Dutch diphone sets for two different male speakers. Soon a third set will be compiled which this time will not just contain diphones taken from stressed syllables in nonsense utterances, but also a (sub)set of unstressed diphones (project A5). Furthermore these utterances will have been spoken by our 'norm' speaker, from whom also the normative intonation contours have been derived. There is some reluctance to try to imitate one specific (professional) speaker. However, we feel that it is better to have a good imitation than to have a marginal overall quality. By the time more insight is derived in the specific speaker characteristics (project A1), it should not be too difficult to go from one voice to the other, although that probably will require a much more detailed source-filter description than the present all-pole LPC approach. This is one of the reasons that in Project A2 the ARMA- based pole-zero analysis is evaluated. The presently implemented robust pole-zero analysis gives consistent results for vowel sounds (van Golstein Brouwers et al., 1988).

3.3 Allophone-based approach
This second approach (allophone-based) is potentially more powerful: the better the rules, the higher the quality. However, as everyone knows, there are no good and efficient analysis procedures which easily hand you these ultimate rules. In project A4 one is working hard to improve the present set of rules. In project A3 an intermediate approach is followed by systematically studying spectro-temporal characteristics in (clusters of) diphones. There is some hope that we will end up with a system in which we can combine the good basic quality of diphone speech with the flexibility of allophone speech.

In the preceding scheme no separate module LEXICAL ACCESS was included. Various modules of course require lexical access. However it might be better to split up the various functions of the lexicon and integrate each separate part in the appropriate modules.

3.4 Speech quality evaluation
Contrary to most other projects, we try to do an ongoing evaluation of the speech quality of the developed systems, both at the start of the project as well as during its development and at completion (project E1). So far an initial evaluation at the segmental level has been completed (van Bezooijen and Pols, 1987), whereas also the intelligibility of consonant clusters has been systematically evaluated (van Bezooijen, 1988). Because of similarity of approach these results can actually be compared with results for other diphone- based systems for French (Pols et al., 1987) and Italian (van Son et al., 1988), as evaluated in ESPRIT project SPIN (Speech interface at office workstation). Meanwhile other diagnostic tests have been performed to compare the two different Dutch diphone sets, and to measure the progress in allophone rule development for a subset of plosives. Other tests will be executed soon to evaluate the PROS module, and also a test about intelligibility and acceptability at sentence level will soon be performed. It is intended to use for this last test syntactically correct but semantically anomalous sentences of say 7 words each as suggested in the ESPRIT project SAM (Multilingual speech input-output assessment, methodology, and standardization). These words will be high-frequency mono-syllabic words with which, randomly, an unlimited number of sentences can be generated. All these sentences will be constructed according to about five predefined specific grammatical structures.

## 4. CONCLUSIONS

This joint Dutch research program is now about three years on its way. Where before most research was done in isolation, now there is good cooperation between most researchers involved. Comparable computer facilities allow for intensive data and software exchange. Regular meetings of research workers force people to consider their own progress in relation to the originally set goals. However, the more progress is made, the better we also realize how much still has to be done. This is especially true with respect to the integration of the various components into a complete system. One approach will be to add gradually components to the existing IPO diphone synthesis system DS (van Rijnsoever, 1988) whenever they become available. In this way their performance can be tested against and eventually compared with other components. However, DS was never designed to include these additional modules, which means that somewhat ad-hoc solutions will have to be chosen every now and then. Another approach would be to develop a complete new structure from scratch. In a way this is

more attractive, although presently we do not yet have a complete overview of the 'ideal' structure. Such a realization would also require additional manpower and computer resources that are presently not available, although the program budget still leaves some space for it. In the near future these matters will have to be solved. We will certainly continue to report about this interesting cooperative project in the open literature. It is also our firm intent to produce the final version at the end of the project as a real research tool, with which further research will be possible.

## ACKNOWLEDGEMENTS

The Dutch joint research program on 'Analysis and synthesis of speech' is supported by SPIN.

## REFERENCES

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

Baart, J.(1987). Focus, syntax, and accent placement, Doctoral dissertation, Leyden University.

Baart, J.L.G. & Heemskerk, J.S. (1988). "The problem of ambiguity in morphological analysis for a Dutch text-to-speech system", Proc. SPEECH '88, 7th FASE Symp., Edinburgh, Book 3, 959-965.

Berendsen, E, & Don, J. (1987). "Morphology and stress in a rule-based grapheme-to-phoneme conversion system for Dutch", Proc. Europ. Conf. Speech Techn., Edinburgh, Vol. 1, 239-242.

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

CELEX-report (1985). Proposal for creating a national, multilingual, lexical database, University of Nijmegen.

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

Daelemans, W.M.P. (1988). "Grafon: A grapheme-to-phoneme conversion system for Dutch", Coling 1988, Budapest, 133-138.

Golstein Brouwers, W.G. van, Veth, J.M. de & Boves, L. (1988). "Robust ARMA analysis for estimation of vocal tract parameters", Proc. SPEECH '88, 7th FASE Symp., Edinburgh, Book 1, 305-312.

't Hart, J. & Collier, R. (1975). "Integrating different levels of intonation analysis", J. Phon. 3, 235-255.

Hertz, S.R., Kadin, J. & Karplus, K. (1985). "The DELTA rule development system for speech synthesis from text", Proc. IEEE, 73, 1589- 1601.

Kager, R. & Quené, H. (1987). "Deriving prosodic sentence structure without exhaustive syntactic analysis", Proc. Europ. Conf. Speech Techn., Edinburgh, Vol. 1, 243-246.

Kerkhof, J., Wester, J. & Boves, L. (1984). "A compiler for implementing the linguistic phase of a text-to-speech conversion system", Linguistics in the Netherlands, 111-118.

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

Lammens, J.M.G. (1987). "A lexicon-based grapheme-to-phoneme conversion system", Proc. Europ. Conf. Speech Techn., Edinburgh, Vol. 1, 281-284.

Langeweg, S.J. (1988). The stress system of Dutch, Doctoral dissertation, Leyden University.

Liang, F.M. (1983). Word hy-phen-a-tion by computer, Diss., Stanford.

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.

Rijnsoever, P. van (1988). From text to speech: User manual for Diphone Speech program DS. IPO Manual nr. 88.

Son, R.J.J.H. van & Pols, L.C.W. (1988). "Differences in formant values of Dutch vowels due to speaking rate", Proc. SPEECH '88, 7th FASE Symp., Edinburgh, Book 2, 429-436.

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.

Vittorelli, V. (1987). "Linguistic analysis of the European languages", ESPRIT '87, Achievements and Impact, Part 2, 1358-1365.