



EVALUATION OF TWO SYNTHESIS-BY-RULE SYSTEMS FOR DUTCH

R. van Bezooijen, L.C.W. Pols ¹

ABSTRACT

The intelligibility of two synthesis-by-rule systems for Dutch was assessed by means of a comprehensive test consisting of 768 CVC, VCV, VCCV, and CVVC stimuli that included all phoneme combinations permitted in Dutch. The test was administered to 16 subjects. Use was made of an open response task. It appeared that phonetic and/or linguistic knowledge had a strong effect upon the listeners' intelligibility scores. Moreover, phoneme intelligibility proved to vary considerably as a function of position within the word. And finally, the types of confusions were found to differ systematically for the two systems.

INTRODUCTION

The aim of a recently started national research program (SPIN) is to improve the quality of text-to-speech synthesis-by-rule for Dutch. Presently there are two such systems: a diphone based system, developed at the Institute for Perception Research (IPO) in Eindhoven (ref 1), and an allophone based system, developed at the Institute of Phonetics of the University of Nijmegen (ref 2). This paper describes a first test in a series in which the quality of the two systems, with respect to intelligibility, naturalness, and perceptual encoding, will be evaluated at successively more complex levels, going from words to phrases, sentences, and texts. The evaluation will be carried out twice: once at the beginning of the SPIN program, in order to obtain detailed diagnostic information as to which aspects of the synthesis output should be improved, and once at the end of the program, in order to assess whether improvement has been effective. In addition, our research has an important methodological component in that insight will be gained into different evaluation methods. The first test aimed at establishing the intelligibility of all phonemes and diphones in all phonetic contexts. Unstressed vowels were excluded as well as initial and final consonant clusters. In this paper emphasis will be on methodological matters concerned with the construction of the test, effects of listener characteristics, variations in the intelligibility scores as a function of phoneme word position, and systematic patterns in the phoneme confusions.

OUTLINE OF STUDY AND METHODS

The 15 Dutch vowels (including 3 diphthongs) and 22 consonants were combined to all possible CV, VC, CC, and VV sequences (C=consonant, V=vowel). Only those combinations were retained that are permissible in Dutch. Examples of phonotactic constraints are the absence of short vowels in syllable and word final position and the absence of voiced obstruents in word final position.

It appeared that words of 4 different structures were needed to contain all permissible initial and final CV and VC combinations and all permissible medial CC and VV combinations: 1 monosyllabic type of the form CVC and 3 bisyllabic types of the forms VCV, VCCV, and CVVC. The - generally meaningless - stimulus words were constructed by randomly combining the CV,

¹ Institute of Phonetic Sciences, University of Amsterdam, Herengracht 338, 1016 CG Amsterdam, The Netherlands

VC, CC, and VV sequences. Most phoneme combinations occurred between 1 and 4 times and most phonemes between 50 and 100 times. The test material comprehended a total of 307 CVC words, 173 VCV words, 267 VCCV words, and 21 CVVC words.

All stimulus words were synthesized both with the allophone and with the diphone system. In addition, a subset of 90 CVC, 90 VCV, 90 VCCV, and all 21 CVVC words were spoken onto tape by the same speaker from whose speech the diphones had been derived, and resynthesized using a tenth order LPC analysis. The LPC stimuli were included to serve as a reference. The interstimulus interval was 3 sec for the CVC and VCV words and 4 sec for the VCCV and CVVC words.

Sixteen subjects (10 males, 6 females) took part in the experiment as listeners. The average age was 27 years, ranging from 24 to 34. Most of them were university students or research assistants, from various faculties and departments: Dutch, foreign languages, medicine, physics, etc. The subjects met three conditions: (1) no experience in listening to synthetic speech, (2) no hearing problems, and (3) typing experience. The subjects were paid for their participation.

The 16 listeners were randomly divided over 4 groups of 4 listeners each. The stimuli were presented in blocks consisting of 1 of the 4 word types produced with 1 of the 3 systems. The order of the blocks varied per listener group. Within the blocks there was a fixed random order of words. The experiment was divided into three parts with breaks of about half an hour in between. The subjects were requested to identify the individual sounds of the stimuli by successively pressing the orthographically labeled keys on a terminal. The only restriction imposed on the responses was the structure of the different word types and the information that the stimuli should be pronounceable in Dutch.

The test was preceded by a training session of about 90 minutes in which the subjects practiced in giving unambiguous responses. The training started with 160 naturally spoken stimuli of the same type as the experimental stimuli. Feedback was given on notational errors. Then the response task was practiced with 60 allophone, 60 diphone, and 60 LPC stimuli, without feedback. This part of the training was an exact copy of the test itself. The total experiment, including the training and the breaks, took about 6 hours.

In the experiment use was made of a test station with an IBM PC XT as central controller and four Tandy model 102 terminals as keyboards for four subjects at a time (ref 3). The test stimuli were prerecorded on audio-tape and played to the subjects with a Studer tape recorder type A 80/R via headphones type Sennheiser HD 414SL. The subjects were seated in a sound-insulated studio, each at a separate table. Their responses were stored in result files. The experimenter was in an adjacent room, handling the tape recorder and the IBM.

RESULTS

The responses were processed in terms of percentages correct phoneme identification and phoneme confusion matrices. Matrices containing the number of mistakes per phoneme combination were computed as well, but have not yet been analyzed and will therefore not be referred to. The computations were based on an automatic comparison of the files in which the correct responses were stored and the files with the responses of the listeners. Before this comparison was made, the listener responses were lined up in case of a missing character or a missing word (the latter happened in only 0.2% of the cases), and obvious typing errors were corrected (e.g. ";pat" was changed into "pat"). The correction was done by hand. Since we were interested in listener effects, we first considered the per-

centages correct words, averaged over systems and word types, separately for each listener. The overall listener scores ranged from 15.9% to 52.0%. From the variation in the scores a very clear pattern emerged in the sense that the subjects with no linguistic or phonetic background had the lowest scores (n = 4, mean of 29.6%), those with linguistic but no phonetic background intermediate scores (n = 8, mean of 37.5%), and those with both linguistic and phonetic backgrounds the highest scores ((n = 4, mean of 48.5%). The rank order of the subjects was very stable (product-moment correlations exceeding .85) across the different word types (except for the very limited CVVC set), but much less so across the different systems (r's of around .65). Apparently, a subject's performance may vary rather considerably as a function of the type of (re)synthesized speech.

The overall percentages correct words and phonemes per word position are given in Table 1, separately for the 3 systems and the 4 word types. Clearly, the allophone scores are lowest, the diphone scores intermediate, and the LPC scores highest. However, in the context of the present paper we do not think it useful to give a detailed account of the quality of the systems tested; the diagnostic data are mostly for internal use.

Table 1. Percentages correct identification, averaged over 16 listeners and separated for 3 systems and 4 word types, for phoneme position and for words. In the last column the total number of stimuli is given for each word type and each system.

	C	V	C		Word	N of stim.
Allophone	44.0	76.4	55.2		21.2	307
Diphone	68.8	79.3	79.3		46.2	307
LPC	81.6	92.4	85.4		65.9	90
	V	C	V		Word	N of stim.
Allophone	74.7	28.3	79.0		20.2	173
Diphone	73.6	60.7	90.9		43.6	173
LPC	87.9	79.6	96.5		69.2	90
	V	C	C	V	Word	N of stim.
Allophone	70.8	35.0	31.5	80.7	10.1	267
Diphone	74.3	59.8	60.7	92.8	31.5	267
LPC	91.9	76.6	81.5	96.3	58.3	90
	C	V	V	C	Word	N of stim.
Allophone	33.3	65.2	64.3	47.6	6.8	21
Diphone	67.0	90.8	62.5	80.4	3.3	21
LPC	74.7	87.5	77.7	84.5	47.0	21

Of more general interest are the considerable differences in the mean scores as a function of word position. For example, it can be observed that the mean scores for the initial phonemes are consistently lower than those for the final phonemes. This holds for both consonants and vowels. However, it is not clear to what extent this difference results from phonotactic constraints in Dutch or reflects true differences in intelligibility. As stated above, there are no short vowels and voiced obstruents in word final position, which would restrict the number of alternatives

the subjects can choose from. Although this was not explicitly mentioned to the subjects, they *had* been told that the stimuli were words that can be pronounced in Dutch, so it cannot be precluded that they were aware of the restricted response possibilities. Of course, the fewer response categories, the higher the chance of guessing the right response, which would explain the better performance on final phonemes. That initial phonemes are not always identified at a better rate than final phonemes appears from the evaluation of a French synthesis system (ref 4).

On the other hand, the same explanation cannot account for the fact that the lowest consonantal intelligibility is found for medial C in the VCV words, since there are no restrictions on the consonants occurring in that position. This tendency manifests itself for all three systems, but for the LPC stimuli it is almost negligible. So, the relatively bad quality of medial C would seem to be a characteristic of the allophone and diphone systems rather than of Dutch or human speech in general. Apparently intervocalic consonants are harder to synthesize adequately than initial or final consonants.

The effect of position-dependent differences on the intelligibility scores is even clearer for the individual phonemes, especially the consonants. Extreme contrasts are initial vs. final /s/ (6.7% vs. 67.8%) for the allophone system; final vs. medial /k/ (90.0% vs. 33.3%) for the diphone system; and initial vs. final /w/ (100.0% vs. 50.0%) for the LPC system. Obviously, in order to gain a complete overview of segmental intelligibility it is necessary to test the phonemes systematically in all possible word positions, requiring the use of different types of stimulus words. With the use of just CVC words, as is general practice, only partial and possibly misleading information is obtained.

Just like the comprehensiveness of the test material yields detailed phoneme intelligibility scores, the open response task yields precise information on the confusions. Together, these data provide indications as to which phonemes should be improved in what respects. In fact, by categorizing the confusions, it was possible to detect some very systematic confusion tendencies, particular to the two synthesis systems. It thus appeared that with respect to place of articulation the allophone stimuli showed a strong tendency to be perceived as too "fronted", whereas the diphone stimuli showed a - somewhat weak - tendency to be perceived as too much "backed". Moreover, whereas the number of voice confusions was fairly similar for the two systems, the allophone stimuli usually sounded too voiced, whereas the diphone stimuli sounded too unvoiced.

A more detailed account of the evaluation results can be found in a forthcoming IFA SPIN report (ref 5).

ACKNOWLEDGMENT

This research was supported by the Foundation for Speech Technology, which is funded by the Dutch National Program for the Advancement of Information Technology (SPIN).

REFERENCES

1. B A G Elsendoorn, IPO Annual Progress Report, Eindhoven 19, 32-35 (1984)
2. L Boves, IFN Proceedings, Nijmegen 10, 18-22 (1986)
3. L C W Pols, G W Boxelaar. Proc. IEEE-ICASSP86, 901-904 (1986)
4. L C W Pols, J P Lefevre, G W Boxelaar, N van Son, This Proceedings
5. R van Bezooijen, L C W Pols, IFA SPIN Report (forthcoming)