EVALUATION OF TWO SYNTHESIS SYSTEMS FOR DUTCH: PHONEMES AND CONSONANT CLUSTERS

Renée van Bezooijen

1.0 INTRODUCTION

In december 1985 a national research program (SPIN) was launched to improve the quality of text-to-speech for Dutch. Presently there are two systems in a more or less advanced stage of development: a diphone based system, developed at the Institute for Perception Research (IPO) in Eindhoven (Elsendoorn, 1984), and an allophone based system, developed at the Institute of Phonetics of the University of Nijmegen (Boves, Buiting, Kerkhoff, & Wester, 1986; Kerkhoff, Loman, & Boves, 1986).
This paper describes the first three tests in a series of tests in which the quality of the two systems 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 described in this paper pertains to the intelligibility of single vowels and consonants; the second to the intelligibility of initial and final consonant clusters; and the third to the overall quality (intelligibility, naturalness, pleasantness) of initial and final consonant clusters. For all three tests information will be provided on the method used and a description will be given of the main results (sections 2.0, 3.0, and 4.0, respectively). In section 5.0 some concluding remarks will be made and plans for the future will be presented.

2.0 PHONEME INTELLIGIBILITY

2.1 Aim

The aim of this test was to diagnostically evaluate the intelligibility of synthesized vowels and consonants in all possible phonetic contexts in which they can occur in Dutch. In this test unstressed vowels were excluded as well as initial and final consonant clusters. The latter were evaluated in the second and third tests (see below).

2.2 Method

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 four different structures were needed to contain all permissible initial and final CV and VC combinations and all permissible medial CC and VV combinations: one monosyllabic type of the form CVC and three bisyllabic types of the forms VCV, VCCV, and CVVC. The - generally meaningless - stimulus words were constructed by randomly combining the CV, VC, CC, and VV sequences. Most two-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 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 (Linear Predictive Coding) analysis. The LPC stimuli were included to serve as a reference.
Sixteen subjects took part in the experiment as listeners. Most of them were university students or research assistants, from various faculties and departments. The subjects had no experience in listening to synthetic speech.
The 16 listeners were randomly divided over four groups of four listeners each. The stimuli were presented in blocks consisting of one of the four word types produced with one of the three systems. The order of the blocks varied per listener group. Within the blocks there was a fixed random order of words. The subjects were requested to identify the individual sounds of the stimuli by successively pressing the orthographically labeled keys on a terminal. Use was made of a test station with an IBM PC XT as central controller and four Tandy model 102 terminals as keyboards for the subjects. The only restriction imposed on the responses was the structure of the different word types and 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 and became accustomed to the task. The total experiment, including training and breaks, took about six hours.

2.3 Results

The responses of the listeners were processed in terms of percentages correct phoneme identification and phoneme confusion matrices. 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 possible listener effects, we first considered the percentages 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 (n=4) had the lowest scores (mean of 29.6%), those with linguistic but no phonetic background (n=8) intermediate scores (mean of 37.5%), and those with both linguistic and phonetic backgrounds (n=4) the highest scores (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 as a function of the type of (re)synthesized speech.
The mean percentages correct words and phonemes per word position are given in table 1, separately for the three systems and the four word types. Clearly, the

allophone scores are lowest, the diphone scores intermediate, and the LPC scores highest (for detailed information on the intelligibility and confusions of the separate vowels and consonants, see Van Bezooijen, 1987).

Table 1. Mean percentages correct vowels, consonants, and words, separately for the four word types and the three systems. In the last column the total numbers of stimulus words are given. All = allophone, dip = diphone.

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

Table 1 furthermore shows that there are considerable differences in the scores as a function of word position. For example, it can be observed that the 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, in Dutch there are no short vowels and voiced obstruents in word final position, which would restrict the number of alternatives the subjects can choose from (they had been told that all stimuli were words that can be pronounced in Dutch). 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 (Pols, Lefevre, Boxelaar, & Van Son, 1987).

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, although it is very small for the LPC. 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 when inspecting the percentages correct for the individual phonemes, especially the consonants (see Van Bezooijen, 1987). Extreme contrasts are initial vs. final (s) (6.7% vs. 67.8%) for the allophone system and final vs. medial (k) (90.0% vs. 33.3%) for the diphone system (for the meaning of the notation, see Appendix A). Obviously, in order to gain a complete overview of the segmental intelligibility of a system, it is necessary to test the phonemes systematically in all possible word positions, which requires 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 makes it possible to obtain detailed information on phoneme intelligibility, 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 at hand. 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 "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.

3.0 THE INTELLIGIBILITY OF CONSONANT CLUSTERS

3.1 Aim

The aim of the second evaluation test was to diagnostically evaluate the intelligibility of initial and final consonant clusters, i.e. groups consisting of two or three initial consonants (e.g. (sp-) as in (spOt) and (spr-) as in (sprut) and groups consisting of two, three, or four final consonants (e.g. (-rt) as in (kOrt), (-rts) as in (kOrts) and (-rtst) as in (kOrtst)). Medial consonant clusters, which always contain a syllable boundary, have been evaluated in the phoneme intelligibility test, in the VCCV-words.

3.2 Method

The stimulus material consisted of 41 initial clusters (35 of two consonants and 6 of three consonants) and 102 final clusters (42 of two, 43 of three, and 17 of four consonants). One could say that this inventory comprises all consonant clusters that one is likely to encounter in a normal Dutch text, except for a few very low-frequent and/or un-Dutch ones. No distinction was made between clusters containing a morpheme boundary and clusters that (can) occur root internally.
All 143 clusters were embedded in nonsense words, the not-to-be-tested end of which was always (t). There were two words per cluster, differing only in the vowel. So, for example, the initial cluster (pr-) occurred in (prOt) and (pryt), and the final cluster (-pst) in (tApst) and (tepst). All stimulus words were synthesized with the allophone and diphone systems. In addition, two reference conditions were included: LPC (Linear Predictive Coding: analysis/resynthesis of naturally spoken speech) and PCM (Pulse Code Modulation: digitized natural speech). The speaker of the LPC and PCM stimuli was the same speaker from whose speech the diphones had been

derived.
There were eight listeners, evenly divided over four groups. All were students and researchers with a linguistic and/or phonetic background. None of them had any experience in listening to synthetic speech. Each group of listeners heard the systems and stimuli in a different order. They identified the clusters in an open response task. Nothing was told about the size of the clusters. Use was made of the same listening facilities as described for the first test. The complete experiment, including instructions, practice (without feedback), and breaks, took 3 hours and 15 minutes.

3.3 Results

3.3.1 Initial clusters

The percentages correct for the initial clusters are given in table 2. The rank order is the same as in the phoneme test: the percentage is lowest for the allophone version, followed by the diphone, LPC, and PCM versions, in that order. Apparently, for digitized natural speech (PCM) virtually 100% correct cluster identification is possible. Analysis/resynthesis with LPC causes a drop of about 20%, and concatenation results in a drop of another 25%. So, in the diphone condition one out of two clusters was identified correctly, and in the allophone condition one out of three. In all four conditions the scores for the three-consonant clusters (CC3) exceed those for the two-consonant clusters (CC2), perhaps as a result of more co-occurrence restrictions within the clusters and a smaller number of tokens (n=6 versus n=35) in the Dutch language.

Table 2. Percentages correct for the initial consonant clusters realized with the four systems. CC2 = two-consonant clusters, CC3 = three-consonant clusters.

	CC2 (n=35)	CC3 (n=6)	Total (n=41)
PCM	95.00	96.88	95.27
LPC	73.21	84.40	74.84
Dip	49.82	54.17	50.46
All	33.40	44.79	35.06

In the PCM condition the scores are uniformly high (>75% correct for all individual clusters). In the other three conditions the scores cover the whole continuum from 100% to 12.5% (LPC) or 100% to 0% correct (diphone and allophone). There does not seem to be any systematic relationship between cluster characteristics (e.g in terms of (combinations of) phonetic classes) and intelligibility. There are a few systematic errors, such as (fr-) -> (xr-) and (sf-) -> (sx-).
The variation in the scores of the eight listeners is smallest for the PCM system, followed by the allophone, diphone, and LPC systems, in that order. Apparently, for the initial clusters the amount of variation in the listener scores is not related to the quality of the stimuli, as one would perhaps expect. None of the correlations between the listener scores for the four systems was significant at the 5% level, so the performance of the listeners seems to be system dependent. This might point to the existence of system specific listener biases. "Local" listener biases, i.e. idiosyncratic patterns in the perception of a particular sound or cluster, occurred frequently as

well.

3.3.2 Final clusters

Whereas the analysis of the responses for the initial clusters was fairly straightforward, the computation of the final cluster scores presented some problems. These had to do with the frequent occurrence of stop intrusions (e.g. (tANt) -> (tANkt)) and stop deletions (e.g. (tOmpt) -> (tOmt) and (tiptst) -> (tipst)) in the listener responses as compared to the original stimulus words (for an extensive discussion of these phenomena, see Van Bezooijen, 1988). Since these intrusions and deletions were also present in the PCM condition, they were interpreted in terms of natural production and perception phenomena, which have nothing to do with the quality of the realization of the stimuli and which should therefore not be counted as errors. In table 3 the percentages correct final cluster identification are given, both on the basis of the original computation, in which "wrong is wrong", and on the basis of the alternative computation, in which responses containing intrusions or deletions are counted correct.

Table 3. Percentages correct for the four versions of the final consonant clusters according to two different calculation methods. In parentheses, the difference between the two methods is given.

	CC2 (n=42)	CC3 (n=43)	CC4 (n=17)	Total (n=102)
Original, "wrong is wrong" method				
PCM	93.75	83.87	56.99	83.46
LPC	84.97	75.00	55.89	75.92
Dip	65.77	56.68	41.54	57.90
All	26.34	14.97	8.46	18.57
Alternative method with intrusions and deletions counted correct				
PCM	95.98(+ 2.23)	92.73(+ 8.86)	85.67(+28.68)	92.89(+ 9.43)
LPC	89.29(+ 4.32)	83.58(+ 8.58)	76.47(+20.58)	84.80(+ 8.88)
Dip	68.16(+ 2.39)	60.90(+ 4.22)	56.25(+14.71)	63.11(+ 5.21)
All	26.34	15.99(+ 1.02)	13.60(+ 5.14)	19.85(+ 1.28)

Table 3 reveals the same rank order in final cluster intelligibility as found for the initial clusters: allophone, diphone, LPC, PCM. However, taking the alternative scores as a point of departure, the absolute percentages correct may be seen to be different: the scores for the final diphone and LPC clusters are higher and those for the final allophone clusters lower. Moreover, it can be observed that the CC2 have been identified best, followed by the CC3 and CC4. Thus, in contrast with the initial clusters, the number of correct final cluster identifications decreases with increasing cluster complexity. It is not clear what these differences are due to.

As for the types of errors, it appears that in the PCM and LPC conditions (x) and (m) have caused relatively many identification problems. These errors occur in the diphone condition as well, supplemented with (f) -> (g) confusions. In the allophone condition the types of errors are very diverse, and there are hardly any clusters that have been identified correctly more than 50% of the time.

Listener performance is rather variable but systematic in the sense that variation tends to increase as the intelligibility of the system decreases. Also, the correlations among the listener scores for "adjacent" conditions (PCM and LPC, LPC and diphone, diphone and allophone) are significant. Both findings contrast with those observed for the initial clusters. On the other hand, here also there are some striking examples of systematic listener biases in the perception of certain sounds and clusters.

4.0 THE OVERALL QUALITY OF CONSONANT CLUSTERS

4.1 Aim

The aim of this test was threefold:

1) To compare the overall quality (naturalness, intelligibility, pleasantness) of PCM, LPC, diphone, and allophone consonant clusters.
2) To assess the relationships among naturalness, intelligibility, and pleasantness judgments.
3) To determine the relationship between global estimates of intelligibility and the results from the detailed identification task employed in the consonant cluster intelligibility test.

4.2 Method

The stimulus material consisted of 26 clusters, divided more or less evenly over the different cluster types (initial and final) and sizes (two, three, and four consonants). The clusters were embedded in common Dutch words, the form of which was always (C)CCVC for the initial clusters and CVCC(C)(C) for the final clusters. In other words, the not-to-be-tested end of the word consisted always of a single consonant.
The clusters (the listenerers were asked to ignore the rest of the stimulus words) were judged on three 10-point scales: unnatural - natural, low intelligibility - high intelligibility, unpleasant - pleasant. The procedure in the listening task was as follows: First, the four versions of each word (one per system) were presented consecutively with a 1 sec interval. In this way the listeners got an idea of the range covered by the four systems for the stimulus cluster at hand. Then the same series was presented three times with an interstimulus interval of 2 sec during which each stimulus version was given a score on each of the three scales. After the procedure had been completed for one cluster, there was a pause of 4 sec before the series for the next cluster were presented. The order of the different systems was varied randomly per stimulus. Moreover, there were two stimulus orders, one with first the initial and then the final clusters (i/f), and one the other way around (f/i). Within the two types of clusters, the two-, three-, and four-consonant clusters were mixed at random.

Eight listeners participated, comparable to but not identical with the subjects used in the consonant cluster intelligibility test. The listeners were evenly divided over two groups. One group got the one stimulus order (i/f), the other group got the other (f/i). The listeners noted their scores in a booklet containing as many pages as there are clusters. At the top of each page the word was given, in normal Dutch orthography, with the target cluster underlined. This way the listeners knew which word was intended. Underneath the stimulus word there were three blocks (one for each scale) of each four rows (one for each stimulus version). This way the listeners could easily compare the scores given on any one scale for the four versions of each stimulus word. The complete experiment, including instructions and practice (without feedback), took 50 minutes.

4.3 Results

Since it is rather awkward to work with the individual listener scores, the reliability of the means of the listener scores was assessed, separately for the three scales and the four systems, with the so-called Ru-coefficient (Asendorpf & Wallbott, 1979). The values for the PCM scores could not be computed because of a lack of intra-stimulus and inter-stimulus variance (too many 9- and 10-scores). The other coefficients ranged from .28 to .78. Reliability varied systematically as a function of scale: it was lowest for pleasant, intermediate for natural, and highest for intelligible. Several, non-conflicting explanations may account for this finding. It can be hypothesized that the scale "intelligible" has a much clearer and more objective meaning than the other two scales. It could also be that the individual clusters differ more with respect to intelligibility than with respect to pleasantness and naturalness. And finally, there are indications that some listeners have scored the scale "intelligible" differently from the other two scales in the sense that with intelligibility the scale positions had an absolute meaning, whereas with naturalness and pleasantness the meaning of the scale positions was variable. In the latter case, the allophone and PCM-versions were always assigned the scale extremes, independent of possible variations in quality. Whatever the reasons for the differences in reliability may be, some of the coefficients are so low that the results of the subsequent analyses, in which nevertheless use has been made of the means, should be interpreted with some caution.

A comparison of the mean scores, averaged over stimuli, reveals that for each listener there is a virtually constant rank order of the four systems, which is identical for all three scales: quality decreases going from PCM, to LPC, to diphone, to allophone. In general the scores for intelligibility are higher than those for naturalness and pleasantness. The intelligibility judgments have probably been influenced positively by the fact that the listeners knew which cluster was intended.

In order to assess whether the three scales contain independent information about the quality of the clusters or whether they all measure the same thing, product-moment correlations were computed between the scale scores (averaged over the eight listeners) for the 26 stimuli, separately for the four systems. All coefficients were significant at the 5% level, except those for pleasant/natural and pleasant/intelligible in the PCM-condition, which is probably due to too little variance in the PCM-scores. So, in general, the three scales appear to covary, i.e. those stimuli that receive high respectively low scores on one scale, will tend to receive high respectively low scores on the other scales. The correlations between pleasantness and naturalness are so high that one of the two is certainly redundant. However, the correlations with intelligibility are quite a bit lower. Perhaps the unshared variance is due to the fact that intelligibility is, up to a certain extent, independent of naturalness and pleasantness.

As a last point, I compared the intelligibility differences between the four systems as they emerged from the cluster indentification task in terms of percentages correct with the scores on the intelligibility scale resulting from the present test. The most striking finding is the fact that the difference in intelligibility between the diphone and LPC clusters appears in practice to be much larger than the listeners estimated in the scaling test. This is due both to an overestimation of the diphone intelligibility and an underestimation of the LPC intelligibility. The allophone intelligibility has been overestimated as well in the scaling test.

5.0 CONCLUDING REMARKS

In this contribution, the results and set-up of three evaluation tests were presented, two diagnostic tests, pertaining to the intelligibility of synthesized phonemes and consonant clusters, and one global test, pertaining to the overall quality of synthesized consonant clusters. In the phoneme intelligibility test, there was one reference condition, consisting of LPC analysis/resynthesis. In the consonant test there were two reference conditions: LPC analysis/resynthesis and PCM, i.e. digitized natural speech.

All three tests revealed the same intelligibility and overall quality rankorder, going from allophone synthesis, with the lowest score, via diphone synthesis to LPC resynthesis. As was to be expected, in the two cases in which a PCM version of the stimuli was included, it came out best. Although both synthesis systems appear to be in need of improvement, in the case of the diphone system correction can be limited to a restricted number of phonemes and phoneme combinations, whereas in the case of the allophone system correction is necessary for the majority of the phoneme inventory, in many different contexts. Although improvement should primarily aim at enhancing intelligibility, the aspects of pleasantness and naturalness, which seem to be partially independent of intelligibility, should not be neglected.

A point of general methodological interest is the rather considerable listener variation which manifests itself, in different forms, in all three tests. This is one of the topics we are planning to look further into in the near future.

In view of the size and time-consuming nature of the two diagnostic intelligibility tests, we are also planning to develop a number of diagnostic subtests, each pertaining to a particular phonetic subclass, as well as a quick test for globally evaluating segmental (including consonant cluster) intelligibility. The latter test will probably be used to compare the intelligibility of the two diphone inventories that are presently available at the IPO, from two different speakers.

After these methodologically oriented activities, the systematic evaluation of the two synthesis systems will be taken up again. The next steps will probably pertain to the quality of unstressed syllables and prosody.

APPENDIX A. PHONEME NOTATION

Vowels		Consonants	
bad	A	put	p
bed	E	bad	b
bid	I	tak	t
bod	O	dak	d
put	U	kat	k
baad	a	goal	G
beet	e	fiets	f
biet	i	vat	v
boot	o	sap	s
buurt	y	zat	z
boek	u	sjaal	S
beuk	@	jaquet	Z
		lach	x
bijt	EI	mat	m
buit	UI	nat	n
bout	AU	lang	N
		lat	l
		rat	r
		jat	j
		wat	w
		had	h

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

Asendorpf, J. & Wallbott, H.G. (1979) Masse der Beobachtungsuebereinstimmung: ein systematischer Vergleich. Zeitschrift fuer Sozialpsychologie, vol. 10, pp. 243-252.

Bezooijen, R. van (1987) Evaluation of two synthesis-by-rule systems for Dutch: segmental intelligibility. Report no. 1, Stichting Spraaktechnologie, Utrecht.

Bezooijen, R. van (1988) Evaluation of two synthesis systems for Dutch: intelligibility and overall quality of initial and final consonant clusters. Report no. 3, Stichting Spraaktechnologie, Utrecht.

Boves, L., Buiting, H., Kerkhoff, J., & Wester, J. (1985) Development of flexible and modular systems for automatic speech synthesis on the basis of linguistic and phonetic knowledge. IFN-Proceedings, Nijmegen, vol. 9, pp. 48-53.

Elsendoorn, B.A.G. (1984) Heading for a diphone speech synthesis for Dutch. IPO

Annual Progress Report, vol. 19, pp.32-35.

Kerkhoff, J., Loman, H., & Boves, L. (1986) FONPARS2: A compiler for the implementation of phonetic rules. IFN-Proceedings, Nijmegen, vol. 10, pp. 75-80.

Pols, L.C.W., Lefevre, J-P., Boxelaar, G., & Son, N. van. (1987) Word intelligibility of a rule synthesis system for French. Proceedings of the European Conference on Speech Technology, Edinburgh, vol. 1, pp. 179-183.