THE PERCEPTION OF VOICING IN DUTCH TWO-OBSTRUENT SEQUENCES

R.J.H. van den Berg

Institute of Phonetics, University of Nijmegen
P.O. Box 9103, 6500 HD Nijmegen, The Netherlands

0. Abstract

Perceived voicing in Dutch two-obstruent sequences (C_1C_2), tested in synthetic VCCV nonwords, was shown to depend not only on the amount of periodicity present in the sequence (VOT and VTT), but also on the intensity of frication noise, and on the durations of the second consonant and the preceding vowel. The duration of the first consonant and the speed and range of formant transitions showed no significant effect. Furthermore, these parameters appeared to be independent cues to perception.

1. Introduction

Because of obligatory final devoicing in Dutch no voiced obstruents occur word-finally, e.g. *goed* (good): /xud/ -> /xut/. Therefore, no phonological voicing opposition exists word-finally, and words as *bod* (bid) and *bot* (bone) are phonetically equivalent. As a consequence assimilation in two-obstruent sequences (C_1C_2) with respect to the feature 'voice' can only take place in sequences of which the first consonant (C_1) is voiceless and the second (C_2) is voiced. This initial (i.e. before assimilation takes place) voicing status of the obstruents is the one occurring in an environment in which assimilation cannot take place, e.g. /dɪt bukɪs xut/ (this book is good).
Assimilation is essentially an articulatory phenomenon [4]. Therefore, if in a sequence of an (initially) voiceless C_1 and voiced C_2 assimilation with respect to 'voice' did take place, both consonants are produced with the same setting of the articulatory feature 'voice', that is both are produced with either vibrating or non-vibrating vocal folds.
Over the years, assimilation of 'voice' in Dutch, which can take place word-internally in compounds as well as across word boundaries, has received a good deal of attention. So far, the aim of the research has been to discover linguistic (and extralinguistic) regularities in the occurrences of assimilation. Two phonological assimilation rules have been formulated.

(1) If C_2 is a plosive, assimilation is regressive, that is C_1 takes on the voicing status of C_2. The result is a sequence of two voiced consonants, e.g. *wit boek* (white book): /wɪt buk/ -> /wɪd buk/.
(2) If C_2 is a fricative, assimilation is progressive, that is C_2 adapts to C_1. The result is a sequence of two voiceless consonants, e.g. *wit zand* (white sand): /wɪt zɑnt/ -> /wɪt sɑnt/.

These rules were formulated on the basis of data obtained by linguists who listened to utterances, often only to one occurrence as in radio broadcasts or lectures, and noted down cases of assimilation. The decisions about the voicing status of the obstruents were made on the basis of what one heard and they were more often than not made by one perceiver only, the researcher. Moreover, these researchers implicitly assumed that if a voiced (or voiceless) consonant was perceived, a voiced (or voiceless) consonant was produced. However, it is a well-known fact that the perception of voicing is not only affected by the acoustic correlate of presence or absence of vocal fold vibrations, but also by a number of other acoustic cues [6]. In view of this and of the fact that assimilation is an articulatory process, the data obtained by means of this perceptual method can at best be considered as only indirect evidence of assimilation.
A more direct method of establishing whether assimilation did occur would be to measure vocal fold activity during the production of the two-obstruent sequence. Slis [8] took this methodological consequence and performed articulatory/acoustic voice measurements of two-obstruent sequences in which assimilation could occur. The voicing status of the obstruents was established by relating the measurements to those obtained for single voiced and voiceless consonants. In the light of the data thus obtained rule (1) above in particular became contestable.
Slis [9] also made a direct comparison of articulatory/acoustic voice measurements and perceptual voicing judgements of the same natural speech stimuli. From this comparison it appeared that no one-to-one relationship exists between the two types of data: the voicing status assigned on the basis of the presence or absence of vocal fold vibrations was not an adequate predictor of the voicing judgements obtained. However, it is possible that the two consonants did become more alike in some other articulatory feature(s), the acoustic correlates of which may have triggered the perception of two voiced (or two voiceless) consonants.
As stated above, the assimilation rules for Dutch are based on perceptual data. The researchers who formulated them may have been able to distinguish the acoustic correlate of presence or absence of vocal fold vibration from other acoustic cues relevant to the perception of voicing. But these other cues may also have (mis)guided their voicing judgements. Therefore, the study of the relation between acoustic cues and voicing judgements of two-obstruent sequences is of importance for the description of assimilation with respect to the feature 'voice' in Dutch.
The question of what acoustic parameters affect the perception of voicing in two-obstruent sequences was addressed in a series of experiments employing synthetic speech stimuli. Synthetic speech was chosen since it allowed for a ready manipulation of the parameter(s) under investigation and for a strict control on the other parameters. In order not to complicate matters, only one parameter at a time was varied in the earlier experiments, and only after some knowledge was gained about the effects on perception of the various parameters an experiment in which they were covaried was performed. The results of all these experiments are presented below.
Investigated were the effects of voice onset time (VOT), voice termination time (VTT), frication noise intensity, and duration and range of formant transitions [1] as well as duration of C_1 and C_2, and of the preceding vowel [2].

2. Method

The stimuli were generated by a 'speech-synthesis-by-rules' system. In this system a string of phoneme labels and prosodic condition signs is transformed into a string of labels indicating successive segments. Parameter values for each segment are read from a table containing target values and timing data for each parameter (a 'phoneme' representation). These values are adapted for context and prosodic conditions by a set of rules (into an 'allophone' representation). Subsequently, these parameter values for allophone-sized segments are converted into parameter values for segments of pitch period size. These are used as input for the calculation of the synthetic speech signal. The fundamental frequency depended on the intrinsic F_0, stress, and declination, and varied around a mean of about 150 Hz. At the allophone-size level the program allows the user to set parameters at self-chosen values.

2.1. Speech material

To preclude effects due to phonetic context [3] the C_1C_2 sequences were part of VCCV nonwords with a strict control on the vowels. The obstruents included in the research were the labial and alveolar plosives and fricatives. Velar consonants were excluded because in Dutch the phoneme /g/ occurs only in loan words, and the voicing opposition in the velar fricatives is of a doubtful status.
The sequences are described as a voiceless plus voiced consonant, because these were the labels used in the input string for the synthesis system.
Because informal listening showed that synthesizing C_1 plosives with the release burst counteracted the perception of C_1 as voiced, and since I wanted the acoustic signal to be ambiguous with regard to cues that were not under investigation, all C_1 plosives were synthesized without a release burst. For the same reason, the consonantal segments (apart from the stimuli in the VOT and VTT experiments) were synthesized without periodicity but for the first 10 ms of the C_1 segment in which the periodic source amplitude dropped to zero.

2.2. Subjects

In each experiment 12 subjects participated. In the experiment on the durations of C_1 and C_2, and in the covariation experiment the number of subjects was 20. All subjects were university students (ages 19-32) and were paid for their services.

2.3. Procedure

After the stimuli were synthesized they were recorded onto audiotape in random order. Each stimulus was recorded three times in successsion, with a one-second interval between repetitions and an intertrial interval of five seconds, in which the subjects made their response. The subjects' task was to listen to all three repetitions of a stimulus, to identify the consonantal sounds in the sequence, and to indicate in a forced choice task what sequence they had heard. To this purpose the response alternatives were orthographically represented on a score sheet, for example *abda*, *apda*, *apta*, and *abta*, standing for /ɑbdɑ/, /ɑpdɑ/, /ɑptɑ/, and /ɑbtɑ/. The last response category, the voiced-voiceless sequence, which is phonologically inadmissible in Dutch, was not included in the VOT and frication noise intensity experiments. Subjects were tested individually in a sound-treated booth. The stimuli were presented over headphones at a comfortable listening level. Experimental trials were preceded by ten practice trials to allow the subjects to get used to the synthetic speech and to the task. Informal interviews after the tests showed that none of the subjects had experienced difficulties in performing the task, and that all judged the synthetic speech to be of good quality.

2.4. Data analysis

The response categories were labelled (++) for voiced-voiced responses, (-+) for voiceless-voiced, (--) for voiceless-voiceless and (+-) for voiced-voiceless responses. For each stimulus the frequencies of the response categories were assessed. The resulting matrix of frequencies was analyzed according to Goodman's loglinear model [5]. This model was specifically developed for frequency data with more than one independent variable. The statistic employed is a multidimensional chi square.

3. Voice Onset Time (VOT)

The effect of VOT was tested in 16 sequences, all combinations of /p,t,s,f/ + /b,d,z,v/. A uniform VOT continuum was opted for, and therefore the durations of the consonantal segments were set at a constant value of 75 ms each, counting the silent interval of the plosives and the noise portion of the fricatives. The durations of the preceding and following vowel were set at 90 and 170 ms respectively. Five VOT values were employed: -150, -75, -30, 0, and +20 ms. These values are relative to the end of the segment representing C_2, that is the end of the silent interval for plosives and the moment of frication noise offset for fricatives.
The effect of VOT was significant: $\chi^2 = 246.23$, df=8, $p<0.001$. VOT did not interact with sequence type: the same general pattern was observed for all sequences. From the data in Table 1 it is clear that late VOT's (that is with no periodicity present in the consonantal segments) favour (--) responses, as could be expected. With earlier VOT's the responses shift via (-+) towards (++).

Table 1: Response frequencies for five VOT's (in %).

VOT	(++)	(-+)	(--)
-150	55.7	33.9	10.4
-75	39.6	46.4	14.1
-30	21.4	52.6	26.0
0	13.5	29.7	56.8
+20	9.9	26.6	63.5

Comparison of the -30 and 0 ms VOT conditions shows that the 30 ms stretch of periodicity at the end of the C_2 segment is a strong cue to C_2 perception: it raises the number of [+voice] C_2 percepts by 30.8%. But the number of [+voice] C_1 percepts also increases, although to a lesser extent (7.9%). Apparently, the stretch of periodicity is also taken as a voicing cue to C_1. This seems to indicate that cues from rather distant portions of the acoustic signal can be integrated into a perceptual unit. However, the effect of this cue appears to become reduced with increasing distance.

4. Voice Termination Time (VTT)

In testing the effect of VTT the same 16 sequences were used. The durations of the segments were as in the VOT experiment. Five VTT values were employed:

0, 40, 75, 110, and 150 ms. These values are relative to the beginning of the C_1 segment.
As can be seen from Table 2 VTT affects perception significantly: χ^2= 451.16, df=8, p<0.001. Again there was no interaction with sequence type: the same pattern was found for all sequences. The responses were in line with what was expected: short VTT's led to (--) and long VTT's to (++) responses, whereas for the intermediate VTT's (40 and 75 ms) the highest frequencies for (+-) were observed, although the frequencies for (++) and (-+) also increased in comparison with a VTT of 0 ms. This is most likely due to the fact that a voiced-voiceless sequence is phonologically inadmissible in Dutch. Probably, phonological restrictions affected perception [7], and the subjects (having perceived periodicity) resorted to the (++) and (-+) categories.

Table 2: Response frequencies for five VTT's (in %).

VTT	(++)	(-+)	(--)	(+-)
0	3.1	16.7	66.1	14.1
40	6.8	24.5	37.5	31.3
75	28.6	32.3	14.1	25.0
110	55.7	32.3	4.2	7.8
150	62.5	31.8	2.6	3.1

These data, too, show indications that cues from rather distant parts of the acoustic signal are integrated into a perceptual unit. The 40 ms stretch of periodicity at the beginning of the C_1 segment not only raised the number of [+voice] C_1 percepts by 20.9%, but also the number of [+voice] C_2 percepts by 11.5%.

5. Frication Noise Intensity

Again the same 16 sequences were tested. The durations of all segments were controlled by the timing rules of the synthesis system. Since C_1 plosives contained no release burst, the variation of the noise intensity was restricted to C_2 (the noise portion of the fricatives and the release burst of the plosives). The six amplitude values used were chosen so as to cover the voiced-voiceless continuum without exceeding naturalness limits. This resulted in a 3 dB step size for fricatives and 6 dB steps for plosives.

Table 3: Response frequencies for six noise level values (in %). Obstruent-plosives only.

noise	(++)	(-+)	(--)
low	21.9	35.4	42.7
	19.8	36.5	43.8
|	11.5	38.5	50.0
V	6.3	36.5	57.3
	12.5	22.9	64.6
high	6.3	27.1	66.7

The overall effect of noise intensity was significant: χ^2= 30.74, df=10, p<0.01. However, the interaction with sequence type was also significant: noise intensity did show an effect for obstruent-plosives, but not for obstruent-fricatives. The difference is most likely due to the different step sizes involved. The direction of the effect is as was expected: with increasingly higher noise levels more (--) responses were given at the cost of (++) and (-+). As may be clear from the data in Table 3, the effect was rather weak, which may have been due to the fact that other parameters were not set at adequate values. Most likely the segmental durations, controlled by the built-in synthesis rules, were too long and biased the responses to (--) and (-+).

6. Formant transitions

Range and duration of the F1, F2, and F3 transitions into and out of the C_1C_2 sequence were tested in /pd/, /tb/, /fd/, and /sb/. The consonantal durations were set at 60+65 ms for /pd/ and /tb/, and at 70+70 ms for /fd/ and /sb/. Speed and range of the transitions were controlled by setting the moment of transition onset and the time within which the shift takes place. For VC_1 transitions moment of onset was relative to the beginning of the C_1 segment, for C_2V transitions it was relative to the boundary between the silent interval and the release burst.
Only a limited range of values could be used, because in informal listening it appeared that long transitions led to the perception of glides and short transitions to the loss of the principal perceptual cue to the place of articulation. Three types of VC_1 transitions: -40;40, -40;60, and -20;40 (moment of onset and transition time respectively) were combined with four types of C_2V transitions: -10;95, -10;75, 0;75, and 0;55.
No overall effect on perception was observed: χ^2= 31.10, df=33, ns. Also, the interaction between VC_1 and C_2V transitions and the main effect of C_2V transitions were not significant. A small effect of VC_1 transitions (p<0.05) appeared to be due to one sequence only (/pd/), the responses to which showed no coherent pattern.

7. Durations of C_1 and C_2

Since the durations of both C_1 and C_2 were varied it seemed advisable to have a clear acoustic boundary between the two consonants. So, only combinations of a fricative (noise) with a plosive (silent interval) were used: /fd,sb,pz,tv/. The durations of the preceding and following vowel were set at 90 and 160 ms respectively. While limiting the total duration of the sequence to 150 ms, both the durations of C_1 and C_2 were varied in 5 steps of 15 ms starting at 45 ms. This resulted in 15 combinations of durations (C_1 and C_2 respectively): 45-45, 45-60, ... 45-105, 60-45, ... 60-90, ... 105-45. All stimuli were synthesized under two stress conditions: stress on the first or on the second syllable. Stress was synthesized by means of a prominence lending rise and fall of F_0.
The interactions of sequence type and stress pattern with duration were not significant. This signifies that duration had a similar effect for the various sequences and both stress patterns, so for the duration results the data were pooled over these conditions. Since the design was not fully crossed for C_1 and C_2 duration, the effect of C_1 duration could only be tested for the various levels of C_2 duration separately, and the effect of C_2 duration only for the various levels of C_1 duration separately.
For none of the levels of the C_2 duration variable did the factor of C_1 duration affect the frequency distribution of the four response categories. Even if only the responses to C_1 were considered, no significant effect was obtained. C_2 responses, too, were not affected by C_1 duration.
On the other hand, for a C_1 of 45 ms (C_2 ranging from 45 to 105 ms), for a C_1 of 60 ms (C_2: 45-90 ms), and for a C_1 of 75 ms (C_2: 45-75 ms) the effect of C_2 duration was significant. Longer C_2 durations led to more (--) and (+-), and less (++) and (-+) responses. The picture becomes even clearer if C_2 responses only are considered: longer C_2 durations led to more [-voice] C_2 percepts. No effect of C_1 duration on C_1 perception was observed.
So, what the effect of this manipulation seems to be boiling down to is that C_2 duration affects C_2 perception, longer durations giving rise to more [-voice] percepts. This effect is strong enough to affect the frequency distribution of the four response categories, wheras C_1 duration does not have any effect at all.
The significant effect of stress pattern manifested itself in that more (++) and less (--) responses were given if stress was on the first than if it was on the second syllable.

8. Preceding Vowel Duration

This effect was tested in the sequences /pd, tb, fd, sb, pz, tv, fz, sv/. The vowels in the VCCV nonwords were either a phonologically long /a:/ or a phonologically short /ɛ/, to test a possibly differential effect of preceding vowel duration for vowels of different phonological length. The durations of /pd/ and /tb/ were 60+65 ms, of the other sequences 70+70 ms. The following vowel had a duration of 160 ms. Stress was either on the first or on the second syllable (stress-1 and stress-2 respectively), realized by a prominence lending rise and fall. To avoid a clash between stress and duration longer preceding vowel durations were used in the stress-1 condition (80, 120, and 180 ms), than in the stress-2 condition (55, 80, and 120 ms).
Vowel type and sequence type did not interact with vowel duration, so the data were pooled over these conditions. Preceding vowel duration significantly (χ^2= 21.25, df=6, p<0.001) affected the response distribution in the stress-1 condition, but not in the stress-2 condition. Under stress-1 longer durations led to more (++) and (+-), and to less (-+) and (--) responses (see Table 4).

Table 4: Response frequencies for three preceding vowel durations (in %).

stress-1	(++)	(-+)	(--)	(+-)	C_1=(+)	C_2=(+)
80	10.4	31.8	38.0	19.8	30.2	42.2
120	17.7	30.2	31.3	20.8	38.5	47.9
180	26.6	22.4	27.6	23.4	50.0	49.0
stress-2	**(++)**	**(-+)**	**(--)**	**(+-)**	**C_1=(+)**	**C_2=(+)**
55	12.5	27.6	45.3	14.6	27.1	40.1
80	10.4	28.6	42.2	18.8	29.2	39.1
120	15.6	29.2	31.3	24.0	39.6	44.8

From a comparison of the 80 and 120 ms conditions under stress-1 with the same conditions under stress-2, it appeared that the non-significant effect for stress-2 is not due to the difference in stress, but rather to the smaller absolute range in durations. However, if only C_1 responses are considered, the effect of vowel duration is significant for stress-1 (χ^2= 15.82, df=2, p<0.001) as well as for stress-2 (χ^2= 7.76, df=2, p<0.05), although the effect is still larger for stress-1. Perception of C_2 was not affected significantly in both stress conditions.
So, preceding vowel duration affects C_1 perception, with longer durations leading to more [+voice] C_1 percepts. In stress-1 this effect is large enough to influence the distribution of the four response categories.
The interaction vowel type x duration was not significant. However, since vowel duration mainly affected C_1 responses, the interaction was also tested with C_1 responses as the dependent variable. For stress-1 the interaction was significant: χ^2= 6.78, df=2, p<0.05. With increasing vowel duration the number of [+voice] percepts increased more rapidly for /ɛ/ than for /a:/, so it seems that an /ɛ/ is perceived as longer than an /a:/ of the same duration. From this it may be inferred that an internal representation of a vowel's intrinsic duration might play a role in its perceived duration.

9. Covariation of parameters

To study possible interactions between parameters an experiment was run in which those parameters that showed a significant effect were covaried. Four C_1C_2 sequences were used: /fd,sb,pz,tv/. The duration of the C_1 segment was 50 ms, that of the second vowel 160 ms. The six parameters that were varied were VOT (in three steps of 0, -25, and -50 ms), VTT (in two steps of 0 and 40 ms), noise intensity of the fricatives (two levels with a difference of approximately 10 dB), duration of C_2 (in three steps of 50, 75, and 100 ms), duration of the preceding vowel (in two steps of 80 and 120 ms), and stress pattern (stress on the first or on the second syllable).
Some parameters interacted with sequence type due to the fact that for some sequences the effect of some parameters was more powerful, rather than to a totally different response pattern. Therefore, the data were analyzed for all four sequences separately. It appeared that for each sequence all six factors had a highly significant effect (p<0.001) on perception and that the response patterns were in line with the earlier findings. The few significant interactions that were obtained seemed to be incidental, since, if an interaction was observed, it was found for one out of four sequences only.

10. Conclusion

These results show that perception of voicing in Dutch two-obstruent sequences does not depend solely on the presence/absence of periodicity. Furthermore, it seems justified to conclude that the factors that affect the perception of voicing in two-obstruent sequences (viz. VOT, VTT, frication noise intensity, stress pattern, C_2 duration, and preceding vowel duration) do so independently.

References

[1] R.van den Berg, "The effect of varying voice and noise parameters on the perception of voicing in Dutch two-obstruent sequences", *Speech Communication 5(4)*, 355-367, 1986.
[2] R.van den Berg, "Effects of duration on the perception of voicing in Dutch two-obstruent sequences", *J.Phonetics*, subm.
[3] R.van den Berg, I.Slis, "Phonetic context effects in the perception of voicing in C_1C_2 sequences", *J.Phonetics 15(1)*, 39-46, 1987.
[4] A.Crystal, "*A first dictionary of linguistics and phonetics*", Andre Deutsch, London, 1980.
[5] L.Goodman, "The analysis of multidimensional contingency tables when some variables are posterior to others: a modified path analysis approach", *Biometrika 60*: 179-192, 1973.
[6] L.Lisker, "Rapid vs Rabid: a catalogue of acoustic features that may cue the distinction", *Haskins Lab.Stat.Rep.Speech Res. SR-54*: 127-132, 1978.
[7] D.Massaro, M.Cohen, "Phonological context in speech perception", *Perception and Psychophysics 34(4)*: 338-348, 1983.
[8] I.Slis, "Assimilatie van stem in het Nederlands", *Glot 5*: 235-261, 1982. Also as: I.Slis, "Rules for assimilation of voice in Dutch", in: R.Channon & L.Shockey (eds) "*In honour of Ilse Lehiste/Ilse Lehiste Pühentusteots*", Foris Publications, Dordrecht, 225-240, 1986.
[9] I.Slis, "Assimilation of voice in Dutch as a function of stress, word boundaries, and sex of speaker and listener", *J.Phonetics 14*: 311-326, 1986.