ASSIMILATION OF VOICE AND PERCEPTION OF VOICING: EFFECTS OF PHONETIC CONTEXT

I.H. Slis & R.J.H. van den Berg

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

0. Abstract

The results of an earlier experiment contained indications that the degree of voicing in the phonetic context affected the perception of voicing in Dutch two-obstruent sequences. This was confirmed in a separate perception experiment. The articulatory/acoustic measurements obtained in a production experiment refute an explanation in terms of a perception mechanism in which regularities in speech production are embodied. The phonetic context effect appears to be a purely perceptual phenomenon.

1. Introduction

Up to only a few decades ago, assimilation of voice in Dutch two-obstruent sequences was investigated by linguists who scored instances of assimilation by ear, often after a single presentation of the utterance. The results of these investigations varied considerably, leading to a great variation of opinions upon the subject [12]. In this contribution we will try to show that one of the possible causes for this lack of agreement may be found in the phonetic context of the two-obstruent sequences.
According to Crystal's [7] definition, assimilation is '*the influence of one sound segment upon the articulation of another so that the two sounds become more alike, or identical*'. In line with this definition, we too consider assimilation to be an essentially articulatory phenomenon. If in a two-obstruent sequence assimilation of voice takes place, both consonants will be produced with the same vocal fold setting: vibrating or non-vibrating. This point of view was the basis for a number of articulatory/acoustic measurements relating to vocal fold behaviour during the production of the obstruent sequence [12].
The consonantal sequence in which assimilation of voice has taken place may be perceived as a sequence of two voiced (or two voiceless) consonants. However, one may perceive two consonants as having the same voicing status in spite of the fact that assimilation of voice did not take place [14]. In that case it is obvious that voicing cues (i.e. acoustic cues to the voicing status of the consonants in question) other than the auditory result of the presence or absence of vocal fold vibration are used by the listener. At the Institute of Phonetics in Nijmegen (IFN) the effect of voicing cues on the perception of voicing in two-obstruent sequences is being investigated in a series of experiments. As was the case with single voiced and voiceless obstruents in Dutch [13], a number of cues were found to affect the perception of voicing in such C_1C_2 sequences [1,2,3].
One of the factors that may affect the perception of voicing in C_1C_2 sequences appears to be the degree of voicing in the consonants in the context. Indications to that effect were found in an earlier study [4] set up to investigate the most suitable type of stimuli for a large series of experiments on the perception of voicing in C_1C_2 sequences. In this paper we will briefly discuss this study (section 2). The results gave rise to an experiment specifically designed to investigate such phonetic context effects on the perception of voicing in C_1C_2 sequences [5]. This experiment is presented in section 3 of this paper. In section 4, several hypotheses will be forwarded that may explain the results obtained. In order to be able to choose between the hypotheses a production experiment was run, which is discussed in section 5.
All experiments employed heterorganic two-obstruent sequences (C_1C_2) to avoid problems arising from the use of (homorganic) geminates. Because of restrictions inherent in Dutch [6] the sequences consisted of a phonologically voiceless obstruent (C_1) followed by a phonologically voiced one (C_2).

2. Investigation of optimal stimulus form

In this first experiment [4] we investigated the perception of voicing in two-obstruent sequences that were part of two successive syllables ($C_i VC_1$-C_2VC_f). One of the aims of this study was to investigate whether the linguistic status of the stimuli would affect the perception of voicing in such sequences. To this purpose the C_1C_2 sequences were embedded in three types of linguistic context:
(a) a word pair that was part of a meaningful sentence;
(b) the same word pair in isolation;
(c) an utterance made up of two meaningless syllables; these nonwords were obtained by changing the initial consonant (C_i) of the first word and the final consonant (C_f) of the second word of the same pairs as used in conditions (a) and (b).

All stimuli were generated by means of a speech-synthesis-by-rules system available at the Institute of Phonetics Nijmegen [11]. Eighteen subjects participated, who identified the consonants and indicated what sequence they had heard in a forced choice task with four response alternatives: voiced-voiced, notation (++); voiceless-voiced, notation (-+); voiceless-voiceless, notation (--); and voiced-voiceless, notation (+-). This last sequence is irregular in Dutch according to the generally accepted phonological rules, but it was nevertheless included, because the subjects felt the need for this response category.

No differences in the perception of voicing in C_1C_2 sequences were observed between the sentence and word pair conditions. However, a significant (χ^2= 12.30, df=6, p<.01) difference was found between word pairs and nonwords.
Three possible explanations for this difference offer themselves:

1) A lexical explanation: the listener is inclined to interpret the perceived sounds so that they make up an existing word. We may expect, therefore, that the responses show a bias towards the perception of meaningful words, and consequently towards the perception of a voiceless consonant followed by a voiced one, which yields a string of unaltered words. In those cases where a nonword can be changed into a word by a shift in the voicing status of one of the members of the C_1C_2 sequence, the meaningful word is expected to prevail. A shift away from the voiceless-voiced responses can then be expected. Nothing of the kind is observed; on the contrary, nonwords show more voiceless-voiced responses than word pairs.
2) A phonological explanation: the listener's perception is subject to his knowledge of phonological rules, particularly in a 'language-mode' of listening. Therefore, we expect that the subjects will perceive more 'regular' sequences in word pairs than in nonwords. So, we expect a higher number of voiced-voiced responses in word pairs with obstruent-stop sequences, and a higher number of voiceless-voiceless responses in word pairs with obstruent-fricative sequences [6]. The results are to the contrary: we observed more irregular sequences (voiced-voiceless) in words than in nonwords.
3) A phonetic explanation: a change in the sound structure of the context might have affected perception. Since the linguistic and phonological explanations did not adequately predict the observed response patterns, we were left with the phonetic explanation. The nonwords were derived from the word pairs by altering the initial (C_i) and the final (C_f) consonant. Therefore, the only phonetic difference between the word pairs and the nonwords was in the C_i and the C_f. So, if a difference in the phonetic context affects the perception of voicing in the C_1C_2 sequences, it is obvious that the alterations in C_i and C_f must be the cause for the perceptual differences observed.

An analysis of the results showed that in those cases where C_i and/or C_f was changed into a voiceless consonant, the number of responses containing voiceless C_1's and C_2's increased. A more detailed analysis suggested that changes in the voicing status of C_i were related to changes in C_1 responses, and changes in the voicing status of C_f to changes in C_2 responses. Since this phonetic context effect was not expected, we had not controlled for the voicing status of the phonetic context when generating the stimulus material. In order to investigate the effect more systematically, a new experiment, specifically designed for this purpose, was carried out.

3. The effect of voiced/voiceless contexts on the perception of voicing in C_1C_2 sequences

The effect of voicing in the phonetic context on the perception of voicing in two-obstruent sequences was investigated in synthetically generated nonwords of the type $C_i VC_1C_2VC_f$. Both syllables were stressed, the first by a rise, the second by a fall in the fundamental frequency contour [8]. The vowel in both syllables was an /ɑ/. The phonetic context, the independent variable, was formed by C_i and C_f. Both could be either an /s/ (voiceless context), or an /n/ (voiced context). The C_1C_2 sequences used were all possible heterorganic combinations of labial and dental obstruents, viz. /pd, tb, fd, sb, pz, tv, fz, sv/. On the basis of the results of previous experiments the synthesis parameters were chosen so as to yield stimuli that were ambiguous with respect to the perceptual voicing status of C_1 and C_2. This implies that the stop-stop sequences had a closure interval of 125 ms, and the other sequences one of 140 ms. The stimuli were synthesized without periodicity during the closure interval. Procedure and response categories were as described above.

Table 1: Frequencies of perceived voicing in C_1C_2 sequences as a function of the voicing status of the context (in %).

context	(++)	(-+)	(--)	(+-)
n--n	33.6	44.8	10.3	11.4
s--n	19.5	61.6	12.2	6.7
s--s	14.5	50.3	27.2	8.0
n--s	26.3	35.2	22.0	16.6

The results (see Tables 1 and 2) showed a highly significant effect of voicing status of the phonetic context. With a voiced C_i, viz. /n/, a significantly (χ^2=107.77, df=1, p<.001) higher number of voiced C_1 percepts was observed than with a voiceless C_i, viz. /s/. With a voiced C_f (/n/) significantly (χ^2=86.87, df=1, p<.001) more voiced C_2's were perceived than with a voiceless C_f (/s/).
The voicing status of C_i was found to have no significant effect on the perception of C_2, nor did the voicing status of C_f affect C_1 perception. Therefore, it would seem that effects of voicing in the context are restricted to the syllable. However, it remains possible that such effects can occur over longer temporal distances, and thus across syllable boundaries.

Table 2: Frequencies of C_1 and C_2 responses as a function of the voicing status of the initial and final context (in %).

context	C_1=(+)	C_1=(-)	C_2=(+)	C_2=(-)
n--	43.8	56.2	69.9	30.1
s--	24.4	75.6	73.0	27.0
--n	35.5	64.5	79.8	20.2
--s	32.7	67.3	63.1	36.9

4. Discussion

In this section we will discuss four different hypotheses that may explain the results obtained. The first two are based on the assumption that the perceptual mechanism uses its awareness of regularities in speech production. The other two hypotheses are purely perceptual in nature.

A1) Perceptual compensation of coarticulatory differences

Let us assume that a difference in the degree of voicing in the context leads to a different production of the C_1C_2 sequence. In that case, there is a ground for a mechanism like perceptual compensation. According to this mechanism listeners perform perceptual corrections for differences in the production of natural speech that arise from contextual influences [10]. The result of these corrections is that no differences are perceived. In synthetic speech stimuli in which these articulatory/acoustic differences are absent, the same compensation mechanism will lead to perceptual differences. When we apply this to voicing in the context, we come to the following argument.
In order to explain our present results, we would have to assume that in natural speech C_1 (or C_2) is

produced with stronger 'voicing' if C_i (or C_f) is voiceless as compared to the condition where C_i (or C_f) is voiced. This may be seen as a kind of emphasized articulatory contrast. The listener's compensation for this articulatory contrast will lead to the perception of the same sequence in all contexts. If, however, the stimulus is ambiguous, as was the case in the present experiment, the same mechanism will result in the perception of a more voiceless sequence in a voiceless context, and a more voiced sequence in a voiced context.

A2) Perceptual expectation of coarticulatory effects

The listener may be inclined to perceive the things he expects, in other words he may be the victim of selective perception. In order to explain our results we must assume that the listener expects to hear a voiceless C_1 (or C_2) in combination with a voiceless C_i (or C_f). This expectation must be based on facts in natural speech. Therefore, we have to assume that voiceless consonants in the context lead to devoicing of (some of) the nearby consonants. From a coarticulatory viewpoint this is a plausible position.

The hypotheses A1 and A2 are mutually exclusive, since they assume opposite effects in production. So, articulatory measurements must enable us to make a choice between the two, or, in case no differences are found, refute them both.

B1) A perceptual-phonological explanation

In this purely perceptual hypothesis we assume that a sequence of speech sounds is recognized in terms of a sequence of 'bundles of phonological features' to which phoneme labels are attached. Context effects as the one found in Experiment 2 may occur when a correctly identified feature of C_i or (C_f) is erroneously attributed to C_1 or (C_2). If this type of erroneous attributions in fact occur at the phonological level, it is likely that the acoustic duration of the intervening phoneme (the vowel) is of no consequence. In that case no effect of (intervening) vowel length is expected. A second factor that can be expected to induce this type of attribution errors is the resemblance between the two 'phonological feature bundles'. In a way similar to the processes involved in producing slips of the tongue, we may expect an increase in the number of attribution errors if the two phonemes (context and target) have more features in common.

B2) A perceptual-phonetic explanation

In this hypothesis we assume that the error occurs at a more peripheral level, viz. that of acoustic cue integration. During this stage the acoustic cues are held in a preperceptual auditory storage (PAS) [9]. The time span of PAS is about 200-250 ms. So, if the temporal distance between context phoneme and target phoneme is less than this time span, the cues for the two phonemes are simultaneously present in PAS. In such a situation misattributions of cues may occur, resulting in a cue being erroneously taken as a voicing cue to the wrong sound segment. In this way effects of voicing in the phonetic context on the perception of voicing in C_1C_2 sequences may be explained. Assuming that the 'strength' of cues plays a role, we expect that such errors are likely to be more frequent with an increase in 'cue strength' (voicedness or voicelessness) of the context phonemes. So, the frequency of erroneous cue attribution may be expected to be dependent on the degree of voicedness or voicelessness. The notion 'degree of voicing' may, for example, be operationalized as the position of C_i (or C_f) on a 'voicing scale' depending on e.g. VOT (or VTT). Furthermore, a greater temporal proximity of the context phoneme and target phoneme may also promote misattributions of voicing cues. So, the (phonological) duration of the vowels intervening the context phoneme and the target phoneme, that is the V's in a $C_iVC_1C_2VC_f$ sequence, is expected to interact with the effect of voicing in the phonetic context on the perception of voicing in C_1C_2 sequences.

The hypotheses B1 and B2 make different predictions with respect to the effect of vowel duration and with respect to a possible effect of the position of C_i (or C_f) on the voicing scale.

In order to decide which of the four hypotheses outlined above prevails, two more experiments need to be carried out. The first one, a production experiment, will enable us to find out whether the context affects articulation of the C_1C_2 sequence. If articulatory effects are found, we may decide on hypothesis A1 (perceptual compensation) if a voiceless context (C_i and/or C_f) leads to more voicing in C_1C_2, or on hypothesis A2 (perceptual expectation) if it leads to less voicing in C_1C_2.
If no context effects are found in articulation there is no ground to maintain hypotheses A1 and A2, and a purely perceptual explanation would seem appropriate. The crucial experiment for the choice between hypotheses B1 and B2 would be one in which the degree of voicing in the context and the duration of the intervening vowels are varied. If perceptual errors result from an erroneous attribution of already recognized phonological features, neither gradations of voicing in the context, nor the durations of the vowel phonemes are expected to affect the perception of voicing in the C_1C_2 sequence. If, on the other hand, the errors are located in the cue integration stage, we expect to find effects of gradation of voicing and of vowel length.
The production experiment will be discussed in the next section. The perception experiment has as yet not been carried out.

5. The effect of voicing in the context on the production of C_1C_2 sequences

The production experiment did in fact consist of three parts, referred to as part (a), (b), and (c), respectively. In each of the three, five male speakers participated, who were asked to read the stimulus materials. The acoustic signal was recorded via a microphone, vocal fold activity by means of an electrolaryngograph. Both these signals were registered on photographic paper with a UV-recorder (SE oscillograph 6008). In the oscillograms we related the moment of voice termination (VTT) to that of oral closure, and the moment of voice onset (VOT) to that of oral release. According to criteria derived from single voiced and voiceless consonants, the voicing status of C_1 and C_2 was assessed, and thus whether assimilation of voice had occurred or not. For a detailed description of the procedure and criteria, see [12].

In part (a) the stimuli were the same as in the previous experiment, embedded in a short carrier phrase, viz. 'doe die $C_iVC_1C_2VC_f$ om'. Employing this type of stimuli resulted in a very low frequency of produced assimilation in obstruent-stop sequences (all obstruent-fricative sequences were progressively assimilated). This was probably due to the fact that the speakers were aware of the central role of the nonword (the only element to vary in the sentences) and may therefore have been inclined to pronounce it with great care. So, we had five other speakers read two additional series (parts (b) and (c)) consisting of meaningful sentences in which the C_1C_2 sequence was part of two adjacent words (C_iVC_1-C_2VC_f). In these sentences C_i and C_f were either voiceless (single consonants or consonant clusters) or voiced (single nasals). The C_1C_2 sequences used in these two series were all heterorganic obstruent-stop sequences. In part (b) the C_1C_2 sequence followed a stressed syllable, in part (c) it preceded a stressed syllable. On the text sheet syllables that had to be stressed were underlined. The speakers were instructed to read the sentences as spontaneously as possible.

Table 3: Frequencies of assimilation of voice as a function of voicing in the context (in %).

context	(++)	(-+)	(--)
+...+	31.7	41.7	26.7
-...+	30.0	50.0	20.0
-...-	31.7	38.3	30.0
+...-	26.7	33.3	40.0

In contrast with part (a) assimilation of voice, either regressively (i.e. two voiced consonants) or progressively (i.e. two voiceless consonants) occurred rather frequently. In line with earlier measurements [14] stress on the syllable preceding the C_1C_2 sequence (part (b)) favoured progressive assimilation, and stress on the following syllable (part (c)) favoured regressive assimilation. However, in none of the three parts of this experiment did we observe a significant effect of voicing in the context on the production of the C_1C_2 sequences, that is on assimilation of voice in those sequences. For this reason, and because the number of speakers was rather low, we pooled the obstruent-stop data from the three parts of the experiment. These pooled data are given in Tables 3 and 4. As may be clear from the figures no significant context effect was found.

Table 4: Frequencies of produced voiced and voiceless C_1 and C_2 as a function of voicing in initial and final context (in %).

context	C_1=(+)	C_1=(-)	C_2=(+)	C_2=(-)
+...	29.2	70.8	66.7	33.3
-...	30.8	69.2	75.0	25.0
...+	30.8	69.2	76.7	23.3
...-	29.2	70.8	65.0	35.0

6. Conclusion

Since we did not find any effects of voicing in the phonetic context (C_i and/or C_f) on the production of the two-consonant sequence C_1C_2, we conclude that these results refute the first two hypotheses, viz. perceptual compensation (A1) and perceptual expectation (A2) of articulatory differences. Thus we are left with the two purely perceptual hypotheses (B1 and B2). The question of whether we have to look for an explanation in terms of an erroneous attribution of a phonological feature (that is to the wrong phoneme), or whether the error occurs at the cue integration stage, cannot be settled by the present data. To address this issue an experiment needs to be run in which the degree of voicedness/voicelessness in the context (C_i and/or C_f) is systematically varied by choosing C_i and C_f from a continuum. Besides, by varying the time interval between context and target phoneme, we may be able to assess whether the domain over which phonetic context effects do take place is determined in durational terms or in terms of number of phonemes, and thus whether the effect originates in PAS, or from misattributions on a higher, phonological level.

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, "The perception of voicing in Dutch two-obstruent sequences", *this issue*, 1987.
[4] R.van den Berg, I.Slis, "Perception of assimilation of voice as a function of segmental duration and linguistic context", *Phonetica 42(1)*, 25-38, 1985.
[5] 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.
[6] G.Booij, *"Generatieve fonologie van het Nederlands"*, Het Spectrum, Utrecht/Antwerpen, 1981.
[7] A.Crystal, *"A first dictionary of linguistics and phonetics"*, Andre Deutsch, London, 1980.
[8] A.van Katwijk, *"Accentuation in Dutch: an experimental linguistic study"*, van Gorcum, Amsterdam/Assen, 1974.
[9] D.Massaro, *"Experimental Psychology and Information Processes"*, Rand McNally, Chicago, 1975.
[10] B.Repp, "Perceptual integration and differentiation of spectral cues for intervocalic stop consonants", *Perception and Psychophysics 24(5)*, 471-485, 1978.
[11] I.Slis, "Some remarks on speech synthesis by rule", *Proceedings Institute of Phonetics, University of Nijmegen 2*, 83-99, 1978.
[12] 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 and L.Shockey (eds.) *In honour of Ilse Lehiste/Ilse Lehiste Pühentusteots*, Foris Publications, Dordrecht/Cinnaminson, 225-240, 1986.
[13] I.Slis, *"The voiced-voiceless distinction and assimilation of voice in Dutch"*, Unpublished Doctoral Dissertation, University of Nijmegen. SR-54: 127-132, 1985.
[14] 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.