CLINICAL LINGUISTICS & PHONETICS, 1989, VOL. 3, NO. 2, 151–161

Speech cue enhancement in intervocalic stops

WALTER F. SENDLMEIER

Max-Planck-Institut für Psycholinguistik, Nijmegen (NL) and Institut für Kommunikationsforschung und Phonetik, Bonn (FRG)

(*Received 8 August 1988; accepted 6 December 1988*)

Abstract

The place of articulation feature for stop consonants is subject to many errors in speech processing by hearing-impaired listeners. Attempts to improve the recognition of initial and final stop consonants by lowering the level of the first formant or—with a different approach—by narrowing the formant bandwidth of the first five formants only very partially led to satisfactory results. Intervocalic stops were used in the present investigation, on the one hand because the spectral information is represented twice (in the VC as well as in the CV-transition) and, on the other hand, because the closure duration offers additional information to the listener. The modification of bandwidth led to no noticeable improvement in the /b, d, g/ discrimination. A change of the closure durations affected the identification of the stops, especially the /b/. The modification of the transitions of the second and third formants optimized the recognition rates for /b/ and /g/.

Introduction

Besides improving speech intelligibility for hearing-impaired listeners with the help of frequency-dependent linear amplification systems—the assistance of which, however, declines with increases in environmental noise—one could think of a selective enhancement of certain speech cues which help the hearing-impaired listener in phoneme discrimination (Pickett, Revoile and Danaher, 1983). If the acoustic characteristics responsible for an improvement in discrimination and identification of speech sounds can be identified by certain methods of signal processing, the practical value of such research can be seen in serving to advance diagnosis, auditory training and further developments of hearing aids. In the latter case, the potential of experimental results on intelligibility improvement by methods of signal processing is a function of how well new programmable hearing-aids can achieve the respective signal modifications.

In the beginning of the 1970s, Pickett and his colleagues at Gallaudet began to investigate the transition detection discrimination of hearing impaired subjects with synthetic stimuli (e.g. Martin, Pickett and Colten, 1972; Danaher, Osberger and Pickett, 1973). They tried to enhance this discrimination by means of signal modification. One important result of these experiments with two formant stimuli is the finding that the presence of the first formant causes a reduction in F2-transition detection acuity. This masking by F1 which affects the higher second formant was greater for hearing-impaired than for normal subjects. In case the first formant also contained a transition, this led to an even stronger upward masking effect of F1. A reduction of F1 by 10–15 dB caused a clear reduction of the masking effect.

0269-9206/89 $3.00

Hannley and Dorman (1983) also systematically investigated the effect of a reduction of the F1-intensity in relation to the F2-intensity in two-formant stimuli. They used CV-syllables with the vowel /a/ so that the centre frequencies of F1 and F2, in the steady state vowel, were quite close to allow effects of broadened auditory filters to emerge (i.e. predominantly upward spread of masking). The level of F1 with respect to the level of F2 was reduced by 0, 6, 12, 18 dB respectively. The greatest effect was found for /g/ when F1 was reduced by more than 6 dB, in which case the hearing impaired showed the same discrimination rate as the normal subjects. This result can be seen as evidence for an upward masking effect in identifying phonetic categories like stops. Hannley and Dorman, however, conclude that this effect only occurs for strongly reduced, i.e. two-formant, synthesized stimuli in which F1 and F2 begin simultaneously and in which only the transition of F2 contains information on the place of articulation. Dorman, Lindholm and Hannley (1985) investigated the upward masking effect of F1 for more complex, five formant stimuli synthesized with the parallel version of the Klatt-synthesizer. The stimuli were presented to hearing impaired subjects with different sound pressure levels with and without F1 respectively. The experiment did not indicate that removing F1 from the signal improved the identification of /b, d, g/. The authors conclude that the detection of the spectral temporal information by subjects with slight to moderate hearing impairment sufficed to suppress a potential upward masking.

Summerfield, Foster and Tyler (1985) chose a different approach in their study of speech cue enhancement in stop consonants. They attempted to investigate the impact of formant bandwidth on the discrimination of initial and final stops in synthetic one-syllable stops. They found that a broadening of the bandwidth led to reduced discrimination both for normal listeners and for listeners with cochlear impairment. However, no general improvement was found for both groups by narrowing formant bandwidths. There was only a slight tendency, in stimuli in which bandwidths were 1/2ed, for the normal listeners to show some improvement in discriminating final stops. For hearing impaired subjects, no improvement of stop discrimination could be found. The question remains, however, whether these results obtained with whispered one-syllable words can be generalized to similar manipulations with natural and LPC-resynthesized voiced speech stimuli.

Furthermore the question remains as to what effects might be produced by the combination of narrowing formant bandwidths and the manipulations shown by Pickett and co-workers to reduce the masking of F2 by F1, i.e. modifying formants in different ways with respect to the bandwidths and thus effecting different changes in the relative amplitude of the first three formants (for a more detailed discussion on the relationship between frequency selectivity and the perception of speech see Rosen and Fourcin, 1986).

Besides the published suggestions for an enhancement of the discrimination of stops by hearing-impaired subjects, one can conceive further possibilities. For example, so far an attempt has not been made to modify the onset and end frequencies of stop–vowel-transitions and vowel–stop-transitions by means of resynthesis procedures in such a way that the place of articulation cues of the stops are more prominent than in the original natural stimuli. *Prominence* here is not used in the sense of syllabic compression where the low-intensity consonantal cues are boosted relative to the more intense adjacent vowels, as for example in Gordon-Salant's (1986) work on the recognition of natural and time/intensity altered CVs by subjects with normal hearing; rather, 'prominence' is understood in the sense of a

selective enhancement of certain linguistically–phonetically relevant speech cues, without changing the overall level of the speech signal.

Turning to intervocalic stops, closure duration is a further parameter with a strong effect on the perception process. Closure duration cannot only have a direct effect on the perception of the place of articulation (Port 1981; Repp 1984), but also an indirect effect on the processing of other features like the transitions from the stops to the adjacent vowels. Tarter, Kat, Samuel and Repp (1983) showed in their study that the processing of intervocalic stops results from a complex interaction of a number of features, especially transitions and closure duration. There are several reasons for working with intervocalic stops: in running speech the majority of consonants occur in intervocalic position. Here the detrimental effects of forward and backward masking are maximized due to the fact that the cues are brief, time-varying, and of low intensity. In addition, the cues are located in the mid- to high-frequency region of the speech spectrum which poses problems for impaired listeners who often have reduced sensitivity and impaired spectro–temporal resolution in this frequency region. Furthermore these impaired subjects may be required to detect the cues close to threshold where the effects of forward and backward masking of stop cues by adjacent vowels can be greater than normal.

The above considerations led to the question whether it is possible to obtain an enhancement in the discrimination of intervocalic stops by hearing impaired subjects if the following parameters are modified: (1) transitions with respect to the start and end frequencies (which means variation of the steepness of the gradients of the transitions); (2) the bandwidth of certain formants in the VC and CV transitions implying the respective variation of the formant amplitudes (the first formant is treated differently from the second and third formants); and (3) the closure durations of the stop consonants.

Experimental details

A necessary condition for generalizing the results of speech perception experiments is that the speech stimuli sound as natural as possible. Unfortunately, this condition is not met by a number of experiments which were carried out during the last 30–40 years on the processing of acoustic cues of speech stimuli. The present perception experiment started from natural speech stimuli, more precisely from the German words *loben* (praise), *Loden* (coarse, woollen cloth), *logen* (lied). These natural stimuli were modified by means of LPC analysis and resynthesis. Both the unmodified and modified stimuli had a very natural sound after resynthesis.

The stimuli had no symmetrical VC and CV transitions. This was due to the fact that the vowels on either side of the consonant are different in both quality and degree of stress, and also, of course, to their naturalness. In most investigations on the dominance of either the CV, or the VC transitions in the perception of intervocalic stops, stimuli which were symmetrical with respect to the transitions were synthetically constructed in order to eliminate any influence other than that of the position (see e.g. Grunke and Pisoni, 1982). Since this study did not concern the question of the dominance of the two transitions of intervocalic stops, symmetry was regarded as less important than naturalness.

The factors investigated with respect to the recognizability of intervocalic stops were: (1) the bandwidth and thus the relative amplitude of the first three formants of the VCV transitions (normal vs. modified); (2) the start and end frequencies of the VC

and CV transitions (normal vs. modified); (3) the closure duration (short vs. normal vs. lengthened); and (4) the place of articulation (b vs. d vs. g). Thus, this study was carried out as a $2 \times 2 \times 3 \times 3$ factorial design.

Stimuli

In order to produce the test stimuli, the three words *loben*, *Loden*, and *logen*, spoken by a male speaker, were recorded on tape in the recording studio of the Phonetic Institute of Nijmegen University (using a Sennheiser MKH 415 T microphone, a Studer 80 R 089 taperecorder, and a LGR 30B tape, with a tape speed of 19 cm/sec). After A/D-conversion at 10 kHz sampling frequency (low pass filtering at 5 kHz, 24 dB/oct butterworth filter), the stimuli were analyzed with an LPC-based computer program for speech research, written by L. Vogten (1983). This program analyzes the digitized speech signal in terms of the parameters of a source-filter model for speech production. In such a model, the acoustic properties of the process of speech production are described as a variable filter which is excited by a variable source signal. The filter parameters describe the spectral pattern of a certain segment as a product of five second-order filters. Pairs of such filter parameters constitute spectral resonances which can be interpreted in terms of formant frequencies and bandwidths (Markel and Gray, 1976). Important for this experiment is the fact that, on the parameter level, every single parameter can be interactively modified by a graphic display and can then, after resynthesis (and D/A conversion including low pass filtering, see above), be studied as to its perceptual effect.

(1) The bandwidths of the stimuli were only modified for the first three formants within a time interval of 50 ms before closure to 50 ms after release. The bandwidth of F1 was broadened by the factor 3, thus causing a respective lowering of the amplitude (since there is an approximately inverse relationship between bandwidth and amplitude; for further details on the precise relationship between both parameters for recursive second order filters see Hess, 1989). The bandwidths of the second and third formants were narrowed by multiplying the original bandwidths with the factor 0·3; thus causing an increase of the amplitudes of these formants. By these different modifications of the bandwidths of the first, second and third formants respectively, aspects of the different approaches of Pickett and his colleagues, aiming at a lowering of the upward masking, and of Summerfield *et al.* aiming at increasing the acuity of certain spectral cues were combined.

(2) A modification of the start and end frequencies of the VC and CV transitions were only made for F2 and F3, as shown schematically in Figs. 1a–1c. The time interval of these modifications is 50 ms per transition. The amount and direction of change of the transitions of F2 and F3 in the test stimuli were determined by pilot studies with three normal listeners who were asked to identify different versions of transition modifications as to the place of articulation category with the stimuli presented close to perception threshold.

(3) The closure durations of the original natural stops varied only slightly. They were 78 ms for /b/, 74 ms for /d/ and 82 ms for /g/ (see Figs. 2a–2c). The clearly longer closure durations for the labial vs. dental and velar stop consonants which are reported in the literature (e.g. Lehiste, 1970) were not found in these stimuli spoken in isolation. Starting from the closure durations of the natural stimuli, two further versions of each stimulus were produced, i.e. a short version obtained by a 42 ms shortening and a long version obtained by a 60 ms lengthening of the original stimulus.

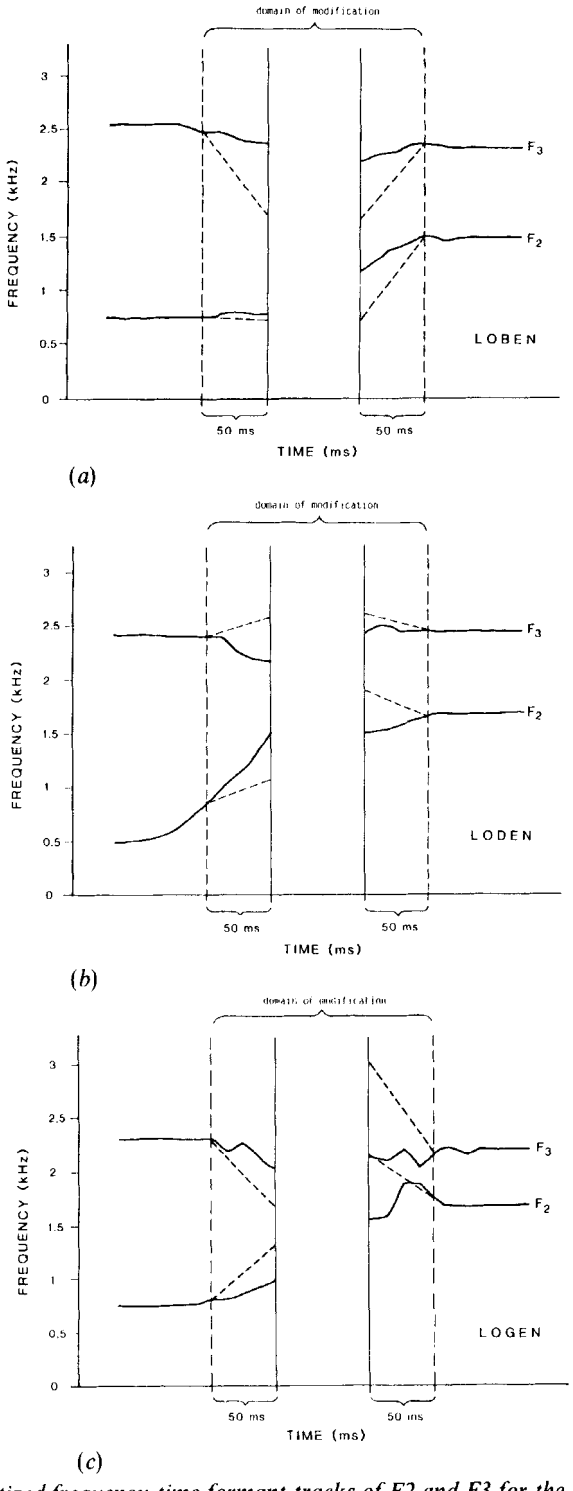
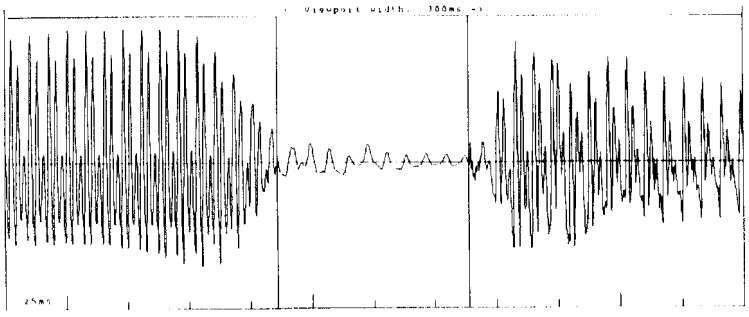
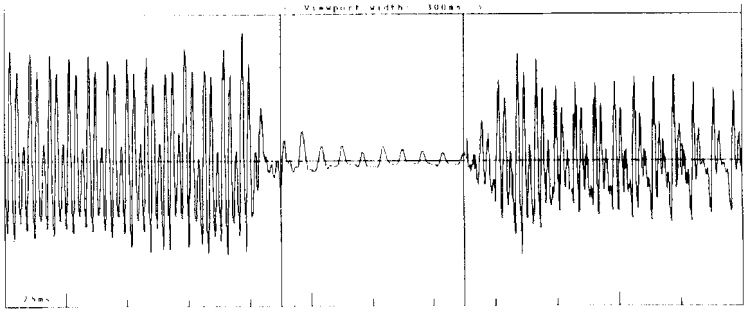


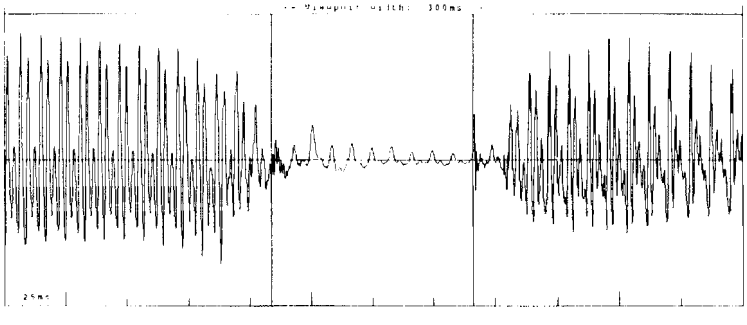
Figure 1. *Schematized frequency-time formant tracks of F2 and F3 for the three stimuli* loben, loden, logen, *each track starting in the middle of the pre-stop vowel and ending in the middle of the post-stop vowel. The VC transitions are considered as starting 50 ms before closure and the CV transitions are considered as ending 50 ms after release. The solid lines indicate the transition tracks of the unmodified stimuli; the dotted lines indicate the transitions of the modified versions.*



(*a*)



(*b*)



(*c*)

Figure 2. *Sections of the waveforms of the three stimulus words* loben, loden, logen *including the closures of the intervocalic stop consonants. The original closure durations of the three intervocalic stops are almost the same.*

(4) The variable place of articulation was given by the well-articulated pronunciation of the voiced stops /b/, /d/ and /g/ in the three words *loben*, *Loden*, and *logen*.

Subjects

A group of 10 subjects with hearing impairments of cochlear origin was tested (average pure tone thresholds at 0·25, 0·5, 1, 2 and 4 kHz were between 25–65 dB, with an increase of hearing loss in the higher frequencies; bone conduction thresholds within 10 dB of air conduction thresholds). The subjects' average age was 58·7 years (range 41–71 years). The audiometric data for the better ear of the impaired listeners are given in Table 1. All subjects were native speakers of German.

Table 1. *The audiometric data (hearing loss dB) for the better ear of the impaired listeners.*

Subject	Age (years)	Sex (m/f)	kHz				
			0·25	0·50	1·0	2·0	4·0
1	54	m	35	40	40	50	60
2	41	m	30	30	35	45	65
3	63	f	40	35	40	45	45
4	58	m	30	30	35	50	60
5	68	f	25	25	35	45	55
6	58	f	35	35	40	40	45
7	66	f	30	25	30	45	55
8	59	m	25	30	30	40	60
9	49	f	35	35	35	40	55
10	71	f	30	25	35	40	50

Procedure

The 36 different stimuli were recorded on tape nine times in a randomized sequence. The stimuli were presented mono-aurally to the listeners' better ear through a TDH-39 headphone at 60 dB SPL (approximately 10–25 dB SL). Listeners were tested individually in a sound attenuating chamber. Before stimulus presentation, the subjects were told that they would hear examples of *loben*, *Loden*, *logen*; they were instructed to identify each stimulus and to make responses by circling their choices on a response sheet.

Results

The recognition rate was derived by counting the number of correct identifications for each stimulus. An analysis of variance was computed over subjects and additionally over replications. The following description of the results is based primarily on the analysis over subjects which in general led to lower F-values than the analysis over replications, due to the fact that the variance between subjects was larger than between replications. If not stated otherwise, the F-values refer to the ANOVA over subjects.

The main effect with the highest F-value (F (2, 18) = 56·270; $p < 0·001$) was

obtained for the place of articulation feature. Both /b/ and /d/ were identified equally well with a rate of about 70%, whereas /g/ was identified to a much lesser degree, i.e. 45%. Differences between identification rates of /b/ and /g/, and /d/ and /g/ respectively, are highly significant. The difference between /b/ and /d/ versus /g/ may be explained, at least partly, by the greater hearing loss of most subjects in the higher frequency range, which affects /g/ more than the /b/ and /d/.

For the duration feature, a significant main effect was also found (F (2, 18) = 11·73; $p < 0{\cdot}001$). Besides the clearly higher identification rate of the stimuli with a lengthened closure duration versus the shortened stimuli, there was also a significant effect between the shortened and the normal stimuli in favour of the latter (post hoc Tukey test, $a = 0{\cdot}05$). The third significant main effect (F (1, 9) = 8·313; $p < 0{\cdot}05$) was found for the transition feature. Modified stimuli were identified to a higher degree than the unmodified stimuli.

The main effect concerning the modification of the bandwidth and thus the modification of the relative amplitudes of the first three formants was not significant (F (1, 9) = 3·467; $p = 0{\cdot}095$); however, the effect was strong enough to justify speaking of a tendency in favour of a higher identification rate of modified versus normal stimuli.

The interaction between place of articulation and duration was highly significant (F (4, 36) = 8·961; $p < 0{\cdot}001$). For /b/, lengthening as well as shortening led to a significantly different perception from the normal version (Post hoc Tukey, $a = 0{\cdot}05$); lengthened and shortened versions differed at the one per cent level (see Fig. 3). This result is in accordance with the fact that, on the average, intervocalic realizations of /b/ have a longer closure duration than both velar and alveolar voiced stops (Lehiste, 1970), which makes listeners likely to perceive stops with a long closure duration as being bilabials (cf. Repp, 1984). For /d/ and /g/, the opposite effect could have been expected: an improvement of the identification by shortening and a decrease by lengthening the stimuli. However, no such effect was found, although there was a tendency for a better identification of /d/ after shortening, so that in comparison to the shortened /b/ it proved to be significantly higher (even distinctly at the 1% level). Surprisingly, there was even a decrease in identification for the shortened /g/.

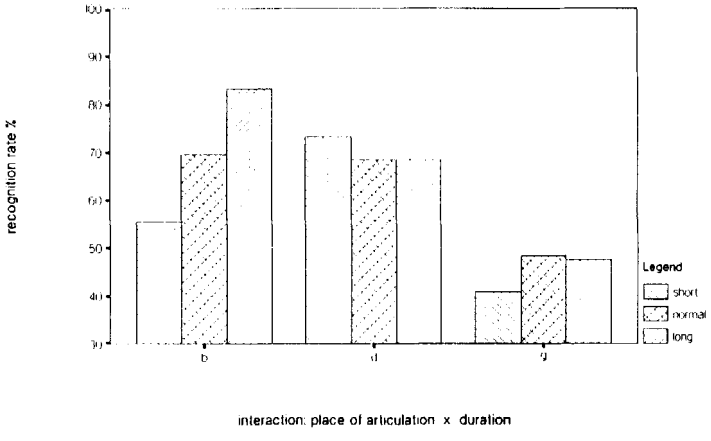


Figure 3. *Effect of the interaction between the factors place of articulation and duration. Within the categories only for /b/ significant differences could be found at the 0·1% and 5% level. For a more detailed description see text.*

The differences in the interaction between the place of articulation and the transition features were just below the significance level ($F(2, 18) = 3{\cdot}369$; $p = 0{\cdot}057$). When analyzed for replications, however, the interaction proved to be significant ($F(2, 16) = 5{\cdot}326$; $p < 0{\cdot}05$). For /b/ and /g/, the transition modified stimuli were identified more easily than the unmodified stimuli (Post hoc Tukey, $a = 0{\cdot}05$). For /d/, there was a slight tendency towards an easier identification of the unmodified stimuli (see Fig. 4). These results may be explained by the fact that /b/ and /g/ are the extremes on the place of articulation continuum, which—in case of an enhancement of the transition gradients—leads to a perception of these features as more prominent. However, /d/, takes a middle position on the continuum, so that a modification of the second and third transitions necessarily leads to a shift to one of the adjacent categories.

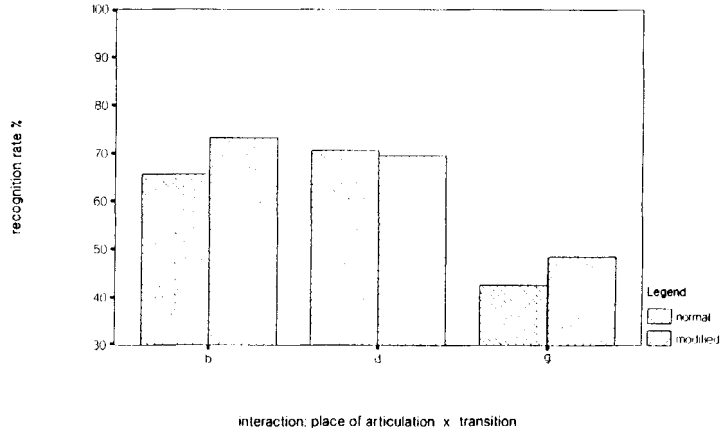


Figure 4. *Effect of the interaction between the factors place of articulation and transition. In the* ANOVA *over subjects the F value just failed to show a significant effect. In the analysis over replications within the /b/ and the /g/ categories a significant difference at the 5% level (post hoc Tukey) between the modified and the unmodified stimuli was found (see Figs 1a–1c).*

Discussion

The place of articulation feature for stop consonants is subject to many errors in speech processing by hearing-impaired listeners. This is why efforts to enhance the distinctive cues of these contrasts seem especially valuable. Attempts to improve the recognition of initial and final stop consonants by lowering the level of the first formant or—with a different approach—by narrowing the formant bandwidth of several formants in the same direction, has only very partially led to satisfactory results. Intervocalic stops were used in the present investigation, on the one hand because the spectral information is represented twice (in the VC as well as in the CV transition) and, on the other hand, because the closure duration offers additional temporal information to the listener.

The modification of bandwidth, more precisely: the broadening of the first formant accompanied by a simultaneous narrowing of the F2 and F3 bandwidths, led to no noticeable improvement in the /b, d, g/ discrimination. Ignoring the effected

change of the amplitudes, this result confirms the limitation of improvement that can be achieved by narrowing formant bandwidths (Horst, 1987).

Depending on the place of articulation a change of the closure durations affected the identification of the stops. Especially for /b/ a clear improvement was obtained by lengthening the closure duration, and a clear decrease in correct identifications was found by shortening it; for /d/ and /g/ there was no clear tendency towards a better recognition after shortening the closure duration as one might have expected from the literature on closure durations. This matches the results of Picheny, Durlach and Braida (1985) who found that more intelligible, clearly spoken sentences are twice as long as conversationally spoken sentences; the difference in duration between the two speech modes is not only accomplished by adding pauses but also by increasing durations of individual speech sounds. This effect might—at least partially—be explained by a reduction of forward and backward masking.

The modification of the transitions of the second and third formants optimized the recognition rate for /b/ and /g/ as the extremes of the transition continuum. This means that an increase in the steepness of the gradients enhances the place of articulation specific information. For /d/ this does not apply due to its middle position on the continuum of the stimuli used.

From these results initial implications for a speech cue enhancement of voiced intervocalic stop consonants for hearing impaired subjects can be inferred. However, up to now the relevance of these results is limited to stops investigated within their special vowel context. Further experiments with a variety of vowel contexts and different speakers are needed to verify and further specify the presented results. After these results have been tested by further investigations such that general rules can be deduced and schematically used, they may be practically applied, for example, in the construction of programmable hearing aids or in designing speech training procedures.

Acknowledgements

I should like to thank Quentin Summerfield, Bruno Repp and Nigel Hewlett for their valuable comments on an earlier version of the manuscript, and Jack Ryalls for helpful suggestions for improving the text.

References

DANAHER, E. M., OSBERGER, M. J. and PICKETT, J. M. (1973) Discrimination of formant frequency transitions in synthetic vowels. *Journal of Speech and Hearing Research*, **16,** 439–451.

DORMAN, M. F., LINDHOLM, J. M. and HANNLEY, M. T. (1985) Influence of the first formant on the recognition of voiced stop consonants by hearing-impaired listeners. *Journal of Speech and Hearing Research*, **28,** 377–380.

GORDON-SALANT, S. (1986) Recognition of natural and time/intensity altered CVs by young and elderly subjects with normal hearing. *Journal of the Acoustical Society of America*, **80,** 1599–1607.

GRUNKE, M. E. and PISONI, D. B. (1982) Some experiments on perceptual learning of mirror-image acoustic patterns. *Perception and Psychophysics*, **31,** 210–218.

HANNLEY, M. T. and DORMAN, M. F. (1983) Susceptibility to intraspeech masking in listeners with sensorineural hearing loss. *Journal of the Acoustical Society of America*, **74,** 40–51.

HESS, W. (1988) *Digitale Filter*. Teubner: Stuttgart.

HORST, J. W. (1987) Frequency discrimination of complex signals, frequency selectivity, and speech perception in hearing-impaired subjects. *Journal of the Acoustical Society of America*, **82,** 874–885.

LEHISTE, I. (1970) *Suprasegmentals*. Cambridge, Mass: MIT Press.

MARKEL, J. D. and GRAY, A. H. (1976) *Linear Prediction of Speech*. Berlin: Springer.

MARTIN, E. S., PICKETT, J. M. and COLTEN, S. (1972) Discrimination of vowel formant transition by listeners with severe sensorineural hearing loss. In G. Fant (Ed), *Speech Communication Ability and Profound Deafness*. Washington, D.C.: Bell Association, pp. 119–133.

PICHENY, M. A., DURLACH, N. I. and BRAIDA, L. D. (1985) Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech. *Journal of Speech and Hearing Research*, **28,** 96–103.

PICKETT, J. M., REVOILE, S. G. and DANAHER, E. M. (1983) Speech cue measures of impaired hearing. In J. V. Tobias and E. D. Schubert (Eds), *Hearing Research and Theory*. New York: Academic Press, pp. 57–92.

PORT, R. (1981) Linguistic timing factors in combination. *Journal of the Acoustical Society of America*, **69,** 262–274.

REPP, B. (1984) Closure duration and release burst amplitude on the perception of stop consonant place of articulation. *Language and Speech*, **27,** 245–254.

ROSEN, S. and FOURCIN, A. (1986) Frequency selectivity and the perception or speech. In B. Moore (Ed), *Frequency Selectivity and Hearing*. New York: Academic Press.

SUMMERFIELD, Q., FOSTER, J. and TYLER, R. (1985) Influences of formant bandwidth and auditory frequency selectivity on identification of place of articulation in stop consonants. *Speech Communication*, **4,** 213–229.

TARTTER, V., KAT, D., SAMUEL, A. and REPP, B. (1983) Perception of intervocalic stop consonants: The contribution of closure duration and formant transitions. *Journal of the Acoustical Society of America*, **74,** 715–725.

VOGTEN, L. L. (1983) *Analyse, zuinige codering en resynthese van spraakgeluid* (Unpublished PhD thesis, TH Eindhoven).