Proceedings of The Institute of Acoustics

FORWARD MASKING AND THE PERCEPTION OF STOP CONSONANTS: PSYCHOPHYSICAL AND NEUROPHYSIOLOGICAL EXPERIMENTS.

H. Spenner and J.V. Urbas

Max-Planck-Institute for Biophysical Chemistry, Dept. of Neurobiology,
D-3400 Goettingen, FRG.

INTRODUCTION

Research on the acoustic properties of speech sounds relevant to perception has met significant difficulties particularly concerning stop consonants. While vowels are fairly well described by a number of spectral maxima (formants) and relatively slow amplitude changes, stop consonants are short events showing rapid spectral changes. Furthermore, they cannot be isolated in time because of coarticulation, the transitions from and to the neighbouring speech sounds being important clues for the recognition of stops. In fact, it has not been possible to determine invariant acoustic features for the different stops, that could explain their invariant perception. This has led to the assumption of speech specific (phonetic) perceptual processes, e.g. plosive perception by reference to articulatory processes [1]. Mechanisms involving articulation have been proposed to explain the following phenomenon in the perception of stop consonants: removing the gap or silent interval corresponding to vocal tract closure preceding the release of a stop leads to "suppression" of the acoustically present consonant, e.g. SPIN is then heard as SIN.
Psychophysical experiments [2] have led to the conclusion that the articulatory process of closure of the vocal tract provides an invariant clue to the listener that production of a stop has occurred, while acoustic features do not. An alternative approach to this phenomenon is based on the similarity of the experimental design to the psychophysical paradigm of forward masking, where an auditory stimulus disrupts or weakens the perception of another presented a few milliseconds later. Applied to this phenomenon, the fricative would be the masker and the stop consonant the probe.

A purely psychophysical approach to this hypothesis is difficult because of the complexity of the consonant-vowel (CV) onset. Measuring masked thresholds would require an a priori decision on which components of the CV onset are relevant for stop perception. Including neurophysiological measurements with animals not only avoids this problem, but in addition eliminates any possibility of interpretation of data that relies on speech-specific processes. Masking is a perceptual phenomenon, resulting in elevated thresholds for the probe in the presence of the masker. With equivalent stimulus arrangements in neurophysiological experiments, a reduction of the discharge rate of the recorded neurone or auditory nerve fibre is observed [3,4]. Concerning complex sounds, there is some evidence that central neural responses tend to favour changes in the signal's amplitude and spectral composition [5,6]. Only a few neurophysiological studies used either natural or synthetic speech stimuli [7,8]. Of immediate interest are experiments, where a variety of speech-like stimuli were presented [9,10]: cat auditory nerve fibres showed reduced response rates to a vowel-like sound when preceded by a consonant-like sound. In addition, the response patterns of the fibres during the formant transitions of consonant-like stimuli contained clues about the preceding context.

METHODS

Stimuli. Tape recorded naturally spoken German words (STAHL and SPOTT) were low-pass filtered (7 kHz), A/D-converted (16.7 kHz) and stored on disc. Amplitudes were normalized to give equal peak values for the vowels. The actual stimuli were produced by digital revision. The beginning and end of the intervals between the fricative and the consonant-vowel (CV) onsets were determined by visual inspection of the signal and print-outs of the sampled values. The CV onset was moved to the offset of the fricative (0 ms interval). Stimuli having intervals up to 60 ms with a step-size of 7.5 ms (psychoacoustics) and 15 ms (neurophysiology) were produced by inserting the appropriate number of zero-amplitude values. To include stimuli that are similar to isolated stops (the parts of the signal supposed to be affected by forward masking) the CV syllables were truncated 18 ms after their onset, the cut being smoothed with a cosine window (fall-time 6 ms). For a second set of experiments masker levels were varied by processing the naturally spoken fricatives to have the same peak amplitude as the following vowels (0 dB masker level) and relative masker levels of -10 dB and -20 dB. Finally, the different gaps were inserted. The original levels of fricatives were -24 dB (STAHL) and -16 dB (SPOTT). For the neurophysiological measurements additional stimuli were used: 120 ms intervals were inserted to see whether there was any suppression at longer than psychoacoustically relevant intervals. Isolated vowels [a:} and [o] were produced by removing the 18 ms CV onset, to make sure the neurone recorded from was in fact responding to the consonant part of the stimulus.

Psychoacoustics. Only the word stimuli were used. Stimuli were presented to the subject via headphones (Sennheiser HD 230) at an average level of 75 dB SPL, selected randomly and D/A-converted by computer. Responses were given on-line to the computer. In the first experiment the subject reported whether he heard the stop-consonant (Yes-No task). The second, testing the effect of fricative levels, was carried out as an AXB task, the subject deciding whether the second stimulus (X) sounded more like the first (A) or the third (B) presentation. The 0 ms word stimulus was randomly placed in position A or B, the other being taken by the 60 ms token. For Position X one of the nine gap lengths was randomly selected.

Neurophysiology. Anaesthetised adult cats were placed into a stereotactic frame, paralysed, and artificially respirated. During each experiment the animals were continuously infused with a mixture of 5 mg/kg/h of gallamine triethiodide and 2 mg/kg/h of sodium pentobarbital. End-tidal CO_2 and body temperature were controlled, the electrocardiogram was monitored audiovisually. Recordings of single neurone activity were made in the thalamus (MGB) using carbon fibre microelectrodes. Sound stimuli were delivered either in a free-field environment or via headphones (Sennheiser HD 424). Neurone activity was stored in the form of a peri-stimulus time histogram (PSTH, binwidth 1 ms). Each stimulus was presented 50 times with a pause of about 2 s between presentations. To test a complete set of stimuli required about 1 hr (not counting measurements of the neurone's threshold (tuning) curve). The stimuli were identical to those used in the psychophysical experiments except that in addition to the word stimuli the supplementary stimuli were also used, and that the number of intervals was reduced choosing a minimal step-size of 15 ms.

For comparison with perceptual data it is useful to display neurophysiological data as an average of all recorded cells (population response) rather than to refer to single neurone activity. Response rates were determined by counting the neurone discharges within a time window placed on that part of the PSTH corresponding to the relevant part of the stimulus. Response latencies (9-12 ms) and the width of the window (18 ms) are set after analysis of the neurone's response to the isolated CV onset. Some variability of the cells' responsiveness is to be observed, especially when recording over the considerable time necessary for presentation of complete sets of stimuli. Therefore the neurone's discharge preceding the silent gap of the stimulus, including some spontaneous activity and the response to the masker, was used as a control. A (quite stable) correction factor was derived to calculate corrected discharge rates for the CV onset.

RESULTS

Psychoacoustics. Results for the words STAHL and SPOTT with 6 subjects are shown in Fig. 1 as percentages of stop detection as a function of the silent interval between the fricative and the CV onset. Critical interval lengths (50 % detection) as estimated by regression analysis are about 10 ms for [t] and 20 ms for [p]. Fig. 2 (STAHL) and 3 (SPOTT) show the results for different levels of fricatives with 2 trained subjects: the critical intervals were longer with increasing masker level. Threshold curves for different levels are significantly separated when the interval is 15 ms or less (STAHL) and 22.5 ms or less (SPOTT, $p < 0.05$, Page's trend test). Regression analysis yields critical interval lengths of 8, 14, and 16 ms (STAHL) and of 10, 17, and 23 ms (SPOTT) for fricative levels of -20, -10, and 0 dB, respectively.

Fig. 1

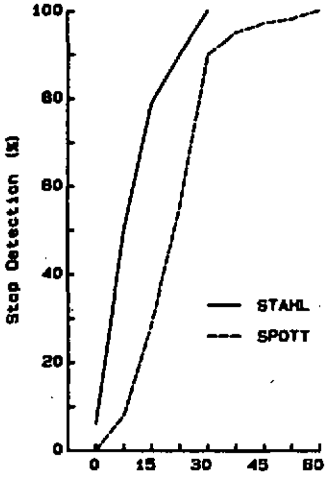


Fig. 2

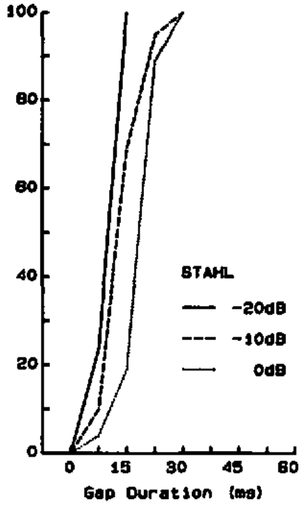
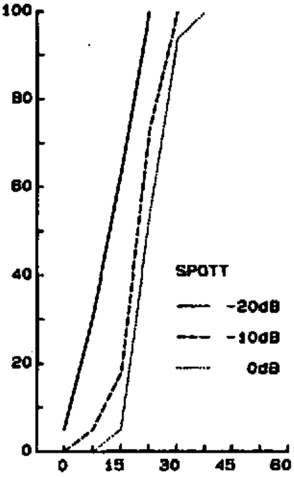
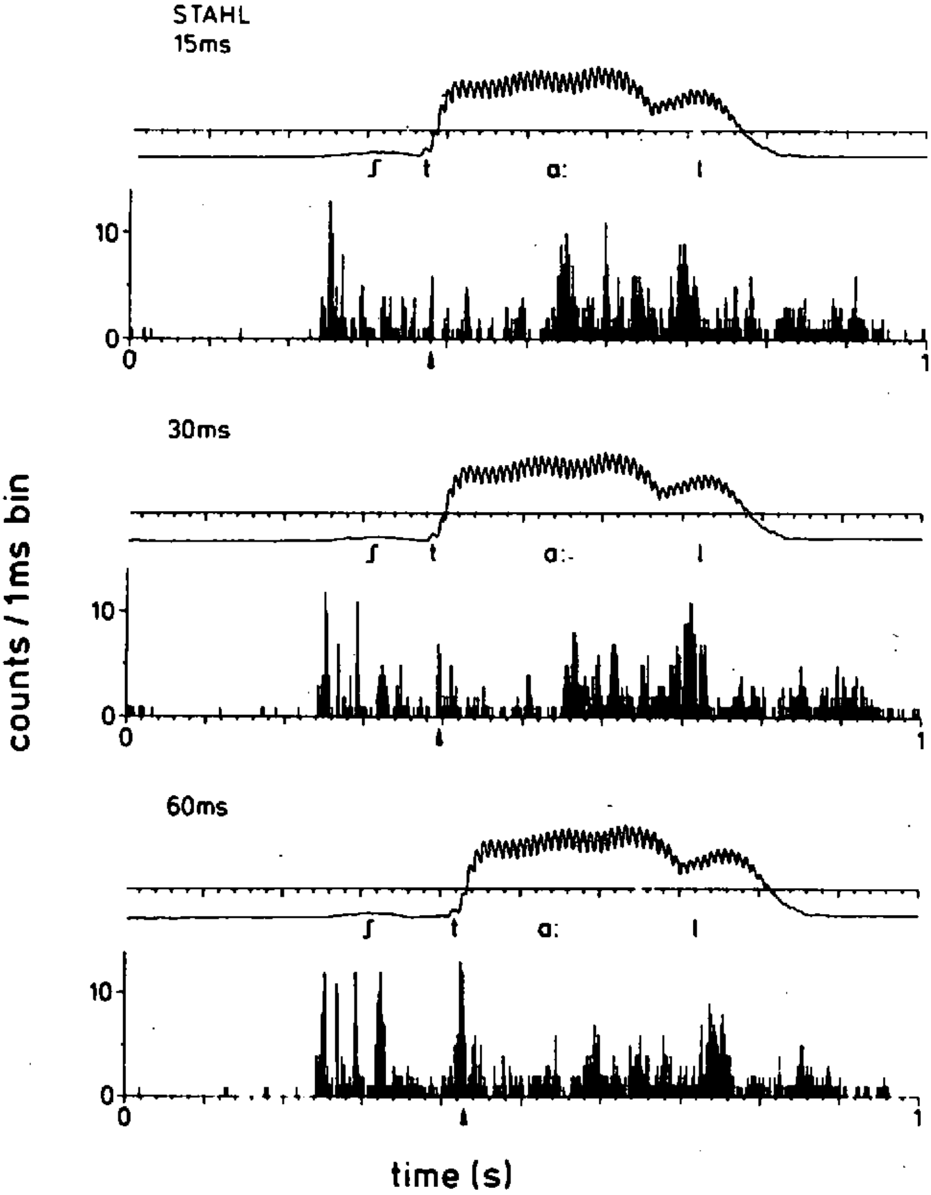


Fig. 3



FORWARD MASKING AND STOP CONSONANTS

Fig. 4



Neurophysiology. For the first experiment recordings were obtained from a total of 34 MGB neurones, some of which responded to one of the CV onsets only, some to both. All neurones also responded to various other parts of the stimuli. A typical example of recordings from one neurone using the word STAHL is shown in Fig. 4 for intervals of 15, 30, and 60 ms. Response rates for the consonant part are increasing with longer intervals. Similar results were found with all cells recorded. Further analyses are based on those data only, that were recorded from cells to which complete stimulus sets were presented. If comparisons between stimuli or sets of stimuli were made, only the data obtained from neurones that responded to all the stimuli involved were included. No significant difference in discharge rate between gaps of 60 and 120 ms has been found when using the word stimuli. Furthermore, rates for interval lengths up to 60 ms are of similar magnitude for both word and CV onset stimuli. (Wilcoxon test: $p > 0.10$). With the CV onset stimuli, however, rates for 120 ms intervals are either significantly greater (SP, $p < 0.05$) or show a trend in this direction (ST, $p < 0.08$). Since the data are from the same cells, the higher rates for the CV onset stimuli at 120 ms vs. 60 ms gap length (SP vs. SPOTT: $p < 0.05$; ST vs. STAHL $p < 0.01$) might be evidence for backward reduction of the discharge evoked by the CV onset stimuli due to the following vowel. To test the forward masking hypothesis, Page's trend test was used. The statistical decision is whether an ordered series of treatments yields similarly ranked results: thus longer silent intervals should lead to significantly greater discharge rates. Data for gaps up to 60 ms are included, for both word (as suggested by statistical evidence) and onset stimuli (resulting in a more conservative analysis). For all four stimulus sets there are clearly significant increases of CV onset-evoked discharge rates with longer gaps ($p < 0.01$). These effects are displayed in Fig. 5 as recovery functions, showing relative average discharge rates with respect to the maximum response. The second experiment - combining variation of fricative levels and the different silent intervals - required extensive recording time with each cell. In addition, fairly high discharge rates evoked by "unmasked" CV onsets were necessary to obtain data for all conditions to be compared. Consequently, only 15 neurones are included in the analysis. As the results are very similar to those described above, it may be assumed, however, that they give a valid description of the neurones' behaviour. Recovery functions for two of the stimulus ensembles (STAHL and ST) are shown in Fig. 6. Again response rates are significantly greater with increasing gap length for all three masker levels (Page's trend test, $p < 0.05$). In addition, involvement of forward masking should lead to a reduction of the response to the probe when the fricative sound pressure level is increased. In fact there is a significant masker level effect for intervals up to 45 ms (STAHL), 60 ms (ST), and 30 ms (SP; Page's trend test, $p < 0.05$ for all sets). Data for SPOTT are not considered because they are based on only 2 neurones.

DISCUSSION

Psychoacoustics. The gap durations found necessary to perceive a stop consonant are within the range of 10-50 ms reported in the literature for [p] and [t]. In both experiments, critical intervals are longer for [p] than for [t]. This finding is in agreement with articulatory conditions, the closure of the vocal tract being shortest for alveolar stops [11]. The results also confirm the general finding, obtained with synthetic stimuli, that the

Fig. 5

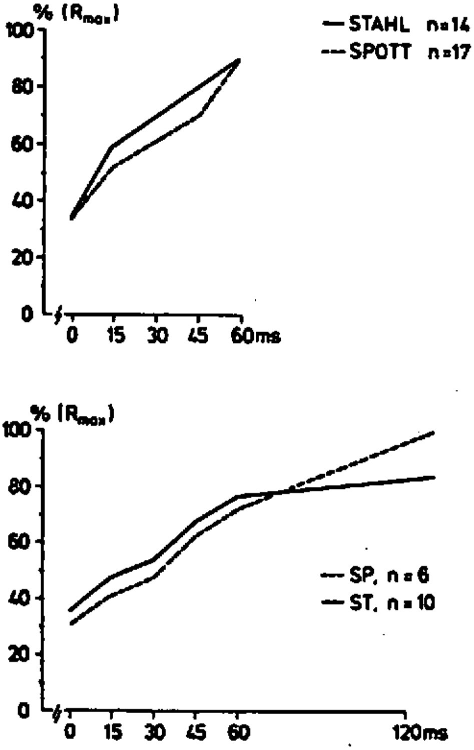
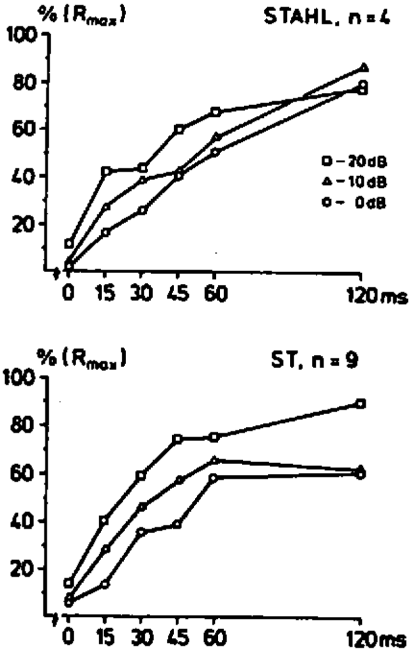

% (Rmax)
100
80
60
40
20
0
0
15
30
45
60ms
— STAHL n=14
--- SPOTT n=17
% (Rmax)
100
80
60
40
20
0
0
15
30
45
60
120ms
--- SP, n=6
— ST, n=10

Fig. 6


% (Rmax)
STAHL, n=4
100
80
60
40
20
0
15
30
45
60
120ms
□ - 20dB
△ - 10dB
○ - 0dB
% (Rmax)
ST, n=9
100
80
60
40
20
0
15
30
45
60
120ms

perception of bilabial stops requires longer preceding intervals than the perception of alveolars [11]. The acoustic properties of the stimuli used in this study do not contradict the hypothesis of forward masking being involved. The fricatives have broad-band spectra, showing maxima above 1 kHz. Therefore, forward masking should be more effective for higher frequency components of the following CV onsets. At lower frequencies, the formants of the [a:], and consequently the transitions for [t], have more energy than their counterparts for [o] and [p]. In addition, the acoustic energy of the fricative is less in STAHL than in SPOTT in precisely those frequency regions where the CV onsets have maxima. Consequently, one would expect less forward masking of the [t], yielding shorter critical intervals - which in fact have been found. The results obtained with different fricative levels directly support a forward masking hypothesis, since there is no evidence that in natural speech the duration of vocal tract closure is dependent on the sound pressure level of the preceding fricative. Similar results are found with "classical" stimuli like noise or tone bursts: at higher masker levels the masked threshold is reached at longer masker-probe delays.

Neurophysiology. As we are comparing the responses of single neurones of anaesthetized cats to results obtained in psychophysical experiments with humans, it should be emphasized that this study is considering only the acoustic properties of the speech sounds and not their phonetic aspects. The underlying assumption is that perception requires a correlate in the spatio-temporal activity of the brain (possibly only the cortex) [12]. We believe that sensory representations are complex patterns of neural activity produced by transduction of adequate external signals (in this case of the speech sounds) by the sensory organs. To yield perception these representations are to reach some "higher perceptual centre" (e.g. the cortex). Although the quantitave relationship of neural activity and perceptual threshold is not yet known, the data presented show that a preceding component of the speech signal can reduce the activity evoked by CV onsets to such an extent that its representation does not reach the "perceptual centre" or is too weak for perception to be probable to occur. The neurophysiological results for neural activation, evoked by CV onsets, have a qualitative similarity to the psychoacoustical results, though the data cannot be compared directly. Furthermore, the neural recovery functions found for the speech stimuli are in good agreement with results of classical forward masking studies in the auditory periphery [3,4]. An initial quick rise of discharge rates is found for short gaps and then, for progressively longer gaps, the rate of recovery slows until no further effects on neural activity are found. While in the auditory nerve the extent of discharge reduction is dependent on the excitation produced by the masker, we sometimes find effects in neurones that do not respond to the fricative. Obviously, this is due to inhibitory processes rather than caused by adaptation. In fact, there is evidence that inhibition occurs in the auditory pathway from the cochlear nucleus onwards [13], and that it persists or may be more marked at higher levels [5].
The results obtained using different fricative levels are further evidence that forward masking processes are involved in the neural representation of speech sounds. Again there is a similarity of psychophysical and neurophysiological results for gaps of 30 ms and less. Comparing these recordings from the thalamus to data obtained in the auditory nerve [3,4], we find the amount and time constant of discharge reduction to be dependent on the masker sound

pressure and not on the level of excitation it produces in the neurone. It may be assumed that auditory nerve fibres respond more vigorously to the fricatives when their levels are increased (provided the sound pressures used are below saturation level). Their activity evokes both excitatory and inhibitory neurones at higher levels. Thus the reduction of response to the CV onsets is probably due to short-term adaptation, occurring at the level of the auditory nerve, and to inhibitory processes at later stages. Both types of discharge reduction decay with time, thus allowing for increasing responses to the CV onsets with longer silent intervals.

With respect to theories of stop-consonant perception, these results support a physiological explanation. If, as our data suggest, the representation of CV onsets in the auditory system is disrupted or seriously weakened unless they are preceded by a silent gap of sufficient length, it is reasonable to assume that forward masking rather than speech-specific processes account for the phenomenon of "stop suppression".

REFERENCES

[1] A.M. Liberman, F.S. Cooper, D.P. Shankweiler, and M. Studdert-Kennedy, 'The perception of the speech code', Psychol. Rev. 74, 431-461, (1967).

[2] M.F. Dorman, L.J. Raphael, and A.M.Liberman, 'Some experiments on the sound of silence in phonetic perception', J.A.S.A. 65, 1518-1532, (1979).

[3] D.M. Harris and P. Dallos, 'Forward masking of auditory nerve fiber responses', J. Neurophysiol. 42, 1083-1107, (1979).

[4] R.L. Smith, 'Short-term adaptation in single auditory nerve fibers: some poststimulatory effects', J. Neurophysiol. 40, 1098-1112, (1977).

[5] O.D. Creutzfeldt, F.C. Hellweg, and C. Schreiner, 'Thalamocortical transformation of responses to complex auditory stimuli', Exp. Brain Res. 39, 87-104, (1980).

[6] I.C. Whitfield and E.F. Evans, 'Responses of auditory cortical neurones to stimuli of changing frequency', J. Neurophysiol. 28, 655-672, (1965).

[7] T. Watanabe and H. Sakai, 'Responses of the cat's collicular neuron to human speech', J.A.S.A. 64, 333-337, (1978).

[8] T. Hashimoto, ' Information processing of speech sounds in the medial geniculate and the inferior colliculus', Proc. Jpn. Acad. 56, 294-299, (1980).

[9] B. Delgutte, 'Representation of speech-like sounds in the discharge patterns of auditory-nerve fibers. J.A.S.A. 68, 843-857, (1980).

[10] B. Delgutte and N.Y.S. Kiang, 'Speech coding in the auditory nerve. IV. Sounds with consonant-like dynamic characteristics. J.A.S.A. 75, 897-907, (1984).

[11] P. Bailey and A.Q. Summerfield, 'Information for speech: observations on the perception of /s/-stop clusters', J. Exp. Psychol. Human Percept. 6, 536-563, (1980).

[12] O.D. Creutzfeldt, 'Cortex cerebri', Springer, Berlin Heidelberg New York, (1983).

[13] E.F. Evans and P.G. Nelson, 'The responses of single neurones in the cochlear nucleus of the cat as a function of their location and the anaesthetic state', Exp. Brain Res. 17, 402-427, (1973).