The description of the production data given above assigns primary importance to closure voicing. For this reason the first consideration in the perceptual experiments will be whether prevoicing alone acts as a cue for voicedness. The second set of experiments will involve a more detailed examination of the perception of VOT, in both synthetic and natural speech.

Some pilot work was carried out with Providence Poles to see how good a match between production and perception could be obtained (Mikoś, Keating and Moslin, 1978). Results indicated that Polish listeners show quite different category boundaries for synthetic VOT continua made up of different ranges of VOT values. Because of the strong range effects, the category boundaries were not well matched to the production categories. It has generally been assumed that the perception of linguistic dimensions such as voicing does not vary according to the range or number of stimuli presented in a task, and that this stability is a characteristic of speech (as opposed to non-speech) perception. Sawusch and Pisoni (1974) showed that the frequency of presentation of any one stimulus did not affect speech category boundaries (but cf. Simon and Studdert-Kennedy, 1978), and it was thought that range would work like frequency. More recently, Darwin and Brady (1978) did obtain range effects in the perception of VOT in English. Also, in Ch. 1 the possibility of range effects in various past studies of VOT perception was mentioned. Therefore, one goal of the present study was to replicate and explore the range effects obtained for VOT in Polish found in pilot work, and to extend the investigation to English. One explanation for the range effects found in pilot work was possible English-language interference. Thus an important

aspect of the present study is replication with monolingual speakers of Polish. Another possible interpretation for the range effects was that the synthetic stimuli used were entirely inappropriate for Polish, and so the results did not reflect subjects' perception of natural speech. Therefore natural speech stimuli were included in the experiments to be described here. A second goal was to identify a set of stimuli for which perceptual categories would correspond to production categories.

2.2.2. Methodology

2.2.2.1. Stimuli and Test Procedure

The stimuli used in this study were i) those synthesized at Haskins Labs by Abramson and Lisker (1970), ii) similar stimuli made at Haskins, and iii) computer-edited tokens of natural speech. i) The original Abramson and Lisker VOT continuum ranges from -150 to +150 msec VOT. In this study two subsets of this continuum were used as test continua. One varied from -100 to +20 msec VOT, and the other from -100 to +50 msec VOT, both in 10-msec steps. Stimuli represent an apical stop followed by the vowel [a]. Voicing lead is present as low-frequency harmonics of the buzz-source. Voicing lag involves F1 cutback and excitement of F2 and F3 by the noise source. There are three formants with onsets of about 300, 1600, and 3400 Hz. F1 and F2 transitions reach steady-state values of 800 and 1250 Hz in 55 msec; F3 reaches a steady-state value of 2500 Hz in 30 msec. There is a weak burst at 3500 Hz which is coincident with the transition onsets. Fundamental frequency begins at 114 Hz and falls to 70 Hz at the end of the stimulus. Stimuli without prevoicing are 440 msec long; prevoicing is added to this

value. Spectrograms and waveform displays of representative stimuli are not equal to the VOT values that would be assigned to them according to the criteria used in measuring production values. This may be due to tape-stretching rather than to different criteria, but in either case an exact match of production and perception cannot be expected. Otherwise, there are some features of the stimuli that are inconsistent or unnatural, as follows: Only the burst of the 0-msec VOT stimulus resembles those shown for natural Polish tokens; that burst, however, is not voiced, as would be expected for a VOT value of 0 msec. In the other stimuli the bursts are unnoticeable. Particularly disturbing is the interval of silence--or low-amplitude voicing--in the lag portion of the +30-msec stimulus. However, the onset of prevoicing and the gradual amplitude increase for the formants is appropriate.

Ten tokens of each stimulus were randomized and recorded (each continuum separately) with an ISI of 4 sec. There was an IBI of 10 sec. and each block contained a complete randomization of the test continuum--16 tokens per block for the -100 to +50 msec continuum test, and 13 tokens per block for the -100 to +20 msec continuum test.[10]

ii) The other synthetic stimuli form a continuum from -20 to +80 msec VOT. They are similar to the ones already described, except that the F1 transition is only 50 msec in duration, and the total (non-prevoiced) duration is only 415 msec. The onset frequencies for the three formants are about 300, 1700, and 3000 Hz. The 5-msec bursts are clearer and more like natural speech in this continuum; note especially the voiced-through burst in the prevoiced stimulus shown. The long lag stimulus appears to have

voicing in the early part of the lag; voice onset in this series is less gradual than in the Abramson and Lisker ones. Waveform displays and spectrograms of representative stimuli are shown in Fig. 2-11.

This synthetic continuum was constructed for use with English-speaking aphasics (Blumstein et al., 1977). The test tape used contained one complete randomization of many tokens of each stimulus, rather than the block-by-block randomization described above. For this study, the entire tape was not used, but instead listeners heard enough of the tape so that there were at least 10 tokens of each stimulus. Thus uneven numbers of tokens of each stimulus were heard, with the following distribution:

VOT	# of tokens heard
-20	10
-10	15
0	10
10	16
20	14
30	10
40	15
50	13
60	13
70	13
80	17

Thus there was a total of 146 tokens presented to the subjects.

The three synthetic-stimuli tests were recorded onto a single high-quality cassette.

The Abramson and Lisker nonsense syllables "da" and "ta" both happen to be real Polish words, and listeners were encouraged to hear them as such. The forced-choice labeling involved underlining the response on an answer sheet which had the two words typed next to each item number.

Data was also obtained in Providence from six English-speaking

control subjects for these three continua. Subjects were run one or two at a time. They wrote t or d on answer sheets.

iii) Two additional sets of stimuli were made; they were constructed by manipulating natural speech. In one set, all prevoicing was removed from tokens of data and dur spoken by MG and from dur and dama spoken by JP. The original tokens with [t] and [d], plus the edited tokens, were presented to Polish listeners for identification to determine the importance of voicing lead per se. Sample waveform displays are given in Fig. 2-12 and 2-13.

The second set consists of two VOT continua, which were made by splicing waveform segments taken from natural tokens. The tokens were tur and dur spoken by JP; these particular tokens were chosen because their bursts were of nearly equal durations and because the tur token had a fairly long VOT. The technique used in making these stimuli was first devised by Terry Halwes at Haskins Labs and was used by Ganong (1978). That study used only stimuli with voiceless bursts; in the present study two continua were made, one with a voiced, and one with a voiceless,burst. Each continuum had five stimuli with VOT values ranging from 12 to 49 msec in approximately 10-msec steps.

The first step of stimulus construction was to isolate the [t]-burst (7.2 msec), the [d]-burst (7.3 msec), and the aspiration from tur (42 msec). This aspiration was then recombined with each of the two bursts (total 49.2 and 49.3 msec). The second step was to note the zero-crossings of the pitch periods in dur, up to 50 msec. With a fundamental frequency of about 100 Hz, the pitch periods were about 10 msec, with a low-energy quasi-period from 7.3 to 11.7 msec in the dur token. The burst+pitch periods were

removed cumulatively and replaced with portions of the burst+aspiration segments of the same duration. For example, the burst and quasi-pitch-period was removed and replaced by the first 11.7 msec of each burst+aspiration segment. Then the burst, quasi-period, and first full pitch period were removed and replaced with the first 20.2 msec of each burst+aspiration segment, and so on for 29.5, 39.1, and 49.1 msec VOT stimuli. Fig. 2-14 shows samples of these stimuli.

There were several experiments with edited natural speech run at the same time, and all the stimuli were recorded onto two high-quality cassettes. The entire set of tests was always given in a single order; the VOT test was the fourth of six. All test stimuli requiring the same forced-choice responses were recorded together, and so the VOT stimuli are combined with others taken from JP's tur and dur. The stimuli from which prevoicing was removed were also combined with other stimuli, and are found in four separate tests.

Twenty-four subjects were run in Poland at the Wrocław Polytechnical Institute of Telecommunications and Acoustics, under the supervision of Dr. Wojciech Majewski. Subjects were paid for their participation. The data for each was obtained in a single session, using headphones. Subjects were run in one of four task-order groups, shown in Table 5.

2.2.2.2. Analysis

The number of [d]-responses to each stimulus for each subject was computed. For the four VOT continua, category boundaries were also computed for each subject. Boundaries were computed by probit analysis (Finney, 1971), using the IBM Scientific Subroutine

TABLE 5

Task orders used in presentation of VOT continua to 24 listeners in Wrocław, Poland. Endpoints (in msec VOT) of each continuum are given. "Natural" refers to the entire set of six tests with edited natural speech, including two VOT continua, 0/+50 and +10/+50.

1)	-20/+80	-100/+50	-100/+20	natural
2)	-100/+20	-100/+50	-20/+80	natural
3)	natural	-20/+80	-100/+50	-100/+20
4)	natural	-100/+20	-100/+50	-20/+80

"PROBT" implemented on the Phonetics Lab's PDP-11/34. This program accepts as input a value along an acoustic dimension, the number of tokens assigned to one of the response categories (here, [d]), and the number of tokens presented to the listener at that value, for each value along the test continuum for each subject. The number of responses is converted to percentages and then to z-scores. The probit analysis determines the parameters of the best-fitting straight line for the z-scores and then outputs the 50% crossover value and the slope of the fitted line, in the units of the acoustic dimension (here, msec VOT).

Subsequent statistical tests were done on the boundary values obtained in this way, unless stated otherwise.

2.2.3. Results with synthetic stimuli

The Polish listeners heard the three synthetic continua in two different orders. Overall, performance was very good. However, four subjects had erratic labeling functions on the -100/+20 msec continuum, with effectively three response categories, and so their results have been excluded from the analysis. Representative labeling functions for the three continua are shown in Fig. 2-15 through 2-17. The -100/+50 msec continuum was designed to correspond to Polish production values, and the -20/+80 msec continuum was designed to correspond to English production values. The -100/+20 msec continuum was designed to include only an idealized "short lag" range of VOT values.

Category boundaries were computed for each continuum for each subject as described in the preceding section. On the -100/+20 msec continuum, four subjects had only one category, [d], and so boundaries could not be computed. However, because these subjects did have

two categories on the other test continua, it was decided not to eliminate their results. Instead it was assumed that their boundaries fell beyond the +20 msec endpoint. That is, these listeners were less influenced by the stimulus range than were other listeners. To represent this fact in the data, these four were arbitrarily assigned boundaries as follows. For the one subject whose other two boundaries were greater than +20 msec, those two boundaries were averaged; for the other three subjects, a value of +21 msec was assigned. Table 6 gives the group data on the boundary values. For the -100/+20 msec continuum, the mean is given both with and without the four assigned values.

Subjects heard the three continua in one of two orders, corresponding to increasing number of lag-value stimuli (-100/+20,-100/+50, -20/+80) or decreasing number of lag-value stimuli (reverse order). The means for each range, divided according to the task order groups, is given in Table 7. A three-way analysis of variance was performed for the factors of Range X Task order X Subjects (nested under task order, repeated measures for range). The results are summarized in Table 8.[11] The analysis of variance indicates that while task order did not significantly affect subjects' category boundaries, both range and the interaction of range with task order did contribute to the variance found in the data. The range effect can be described by saying that, overall, the fewer lag stimuli were present in a continuum, the lower were the subjects' boundaries on that continuum. The interaction with task order can be described by saying that subjects who heard the -100/+20 msec continuum first had higher boundaries (heard more [d]'s) for this continuum than would otherwise be expected; they had higher

TABLE 6

For each range of VOT in synthetic continua, number of subjects for whom boundaries could be computed (N), the highest and lowest boundaries for the group (Range), the mean and standard deviation (SD) of the subjects' boundaries, in msec VOT. The two sets of values given for the -100/+20 msec continuum differ in whether estimated boundaries are included.

Continuum	N	Boundary Range	Mean	SD
-20/+80	20	+10 / +30	+20.4	5.7
-100/+50	20	-20 / +29	+ 5.4	11.5
-100/+20	16	-34 / +19	- .9	13.3
	20	-34 / +27	+ 3.7	

TABLE 7

Mean boundary values in msec VOT for 20 listeners on three synthetic continua differing in range of VOT, presented in two task orders. The overall means are the same as those in Table 6. The task orders are given as the positive endpoint of each continuum used.

Continuum	Mean Boundaries: Task Order 80-50-20	Task Order 20-50-80	Overall
-20/+80	+23.7	+17.7	+20.4
-100/+50	+ 7.7	+ 3.5	+ 5.4
-100/+20	- 5.2	+11.0	+ 3.7
Mean	+ 8.7	+10.76	+ 9.8

TABLE 8

Results of the analysis of variance of boundary values for the factors Range (-20/+80, -100/+50, -100/+20 msec VOT) and Task Order (80-50-20 or 20-50-80)

Source	df	MS	F	p
Task Order	1	62.643		
			<1	-
error	18	193.760		
Range	2	1822.406	21.21	.001
error	36	85.910		
Task Order X Range	2	749.341	8.72	.001
error	36	85.910		

boundaries for this continuum than they did for the -100/+50 msec continuum. In other words, they were less susceptible to range effects on this continuum than were those who heard it last.

To avoid the problem of having to assign arbitrary values in certain cases, whose exact numerical values then affect the group means, the same data was re-analyzed using a less-sensitive non-parametric test on medians. The median for the entire sample was calculated, and the number of boundary values for each continuum falling above and below this median was determined. In this way the exact values of the assigned boundaries have no effect on the results; only their position relative to the other boundary value matters, and it is precisely this information that we want to include in the analysis. Also, using this test allows unequal numbers of values per cell to be tested, so the data for all 24 subjects can be used for the -20/+80 and -100/+50 msec continua.

The median for the entire set of boundary values was 12.5 msec. The number of values falling above and below 12.5 for each continuum is given in Table 9. A Chi-square test of these totals was performed. The χ^2 of 26.63 with df = 2 was significant at the .001 level. Thus we can conclude that the range effects obtained are statistically significant, and are not a result of estimating boundary values for some subjects. Polish listeners show different boundary values depending on the range of stimuli included in the test continuum.

Six American control subjects were also run on these three continua, counterbalanced across the two task orders used in Wroclaw. A seventh subject was eliminated because even on the -100/+50 msec continuum he used all [d]-responses, and so no

TABLE 9

Data used for Chi-square analysis, showing number of boundary values above and below group median of 12.5 msec VOT, for each continuum used.

	Continuum Range			
# Boundaries	-20/+80	-100/+50	-100/+20	Total
Below Median	2	19	13	34
Above Median	22	5	7	34
Total	24	24	20	68

boundary comparisons could be made for him. All the American listeners had only one category on the -100/+20 msec continuum. The results for the six subjects on the remaining two continua are given in Table 10. The same analysis using Range X Task order X Subjects used on the Wrocław data was performed on these data. The -100/+20 msec continuum was assigned boundary values as above: since all subjects' other boundaries were above +20 msec, they were averaged to give an estimate of the third boundary. The analysis of variance resulted in F-ratios that were all less than 1, so no summary table is given. This analysis indicates that the variance in the data is not accounted for by the factors of Task Order or Range. In other words, the American listeners, unlike the Polish listeners, did not show range effects for the three synthetic VOT continua. Their boundaries are nearly identical for the -20/+80 and -100/+50 msec continua, and on a continuum where their preferred boundary region was not represented, they remained consistent in their categorizations.

2.2.4. Results with natural stimuli

2.2.4.1. Prevoicing

As was described earlier, the words dama and dur spoken by JP and the words data and dur spoken by MG were used in making stimuli from which prevoicing had been removed. The object was to see if prevoicing in itself is crucial to a [d]-judgement. Both the edited and unedited [d]-tokens were presented to listeners. Results for the edited tokens were mixed, in that responses to JP's tokens were different from responses to MG's tokens. Listeners judged MG's edited tokens to be [t]'s, while they judged JP's tokens to be

TABLE 10

Results for American control subjects. For each range of VOT in the synthetic continua, the highest and lowest boundaries of the group (Range), the mean and standard deviation (SD) of the subjects' boundaries, in msec VOT. Table 6 gives similar data for the Polish listeners.

Continuum	Boundary Range	Mean	SD
-20/+80	+29 / +46	37.2	6.9
-100/+50	+28 / +44	37.3	6.4
-100/+20	all beyond the +20 msec endpoint		

[d]'s. The mean responses of the 24 subjects are illustrated in Fig. 2-18. A repeated-measures analysis of variance was performed on the number of [d]-responses (out of 10) for the three factors Word nested in Speaker X Condition (prevoiced vs. edited). The results are summarized in Table 11. All factors and interactions were significant.

2.2.4.2. VOT continua

The two VOT continua made by computer-editing had ranges from +12 to +49 msec in approximately 10-msec steps. In addition, the [d]-burst continuum can be extended back to 0 msec VOT by including the [d]-token from which prevoicing has been removed. The original token has 94 msec prevoicing, and voicing continues through the burst so its VOT value after editing would be 0 msec. There is no equivalent [t]-burst stimulus. As can be seen in Fig. 2-18, the edited token produced an average of 80% [d]-responses, and therefore it is not surprising that few subjects reached 100% [d]-responses for the short-lag VOT values in this continuum. Nevertheless, most subjects had two categories on both continua, and boundaries were computed. The effect of [t]- vs. [d]-burst is not considered in this chapter; there was, however, a significant effect which is discussed in Ch. 3. Fig. 2-19 shows averaged labeling functions for 24 subjects.

As with the synthetic continua, some subjects had only one category (here, always [t]), and so arbitrary values were assigned at 1 msec beyond the endpoint, or -1 msec VOT. Therefore two mean boundary values were calculated for the combined sample of [t]-burst and [d]-burst values. The first takes only the boundary values that could be determined by Probit Analysis, and excludes the assigned values; this mean, for 36 boundary values, is 20 msec VOT.

TABLE 11

Results of the analysis of variance of the number of [d]- responses for the factors Speaker (JP or MG), Word nested in Speaker (dur and dama for JP, dur and data for MG), and Condition (original [d]-token or stimulus with prevoicing removed).

Source	df	MS	F	p
Speaker	1	444.081	206.69	.01
error	23	2.149		
Word in Speaker	2	16.677	9.01	.01
error	46	1.851		
Condition	1	858.519	206.81	.01
error	23	4.151		
Speaker X Cond.	1	444.082	206.69	.01
error	23	2.149		
Cond. X Word in Speaker	2	16.677	9.01	.01
error	46	1.851		

The second mean includes the assigned values; this mean, for 48 boundary values, is 17 msec VOT.

The boundaries for this natural continuum can be compared with those obtained on the three synthetic continua. Of course, no direct conclusions about range effects can be drawn since the stimuli are so different in structure, but it can be seen how that structure plus VOT range affects subjects' boundaries. The mean boundary of 20 or 17 msec is closest to that obtained with the -20/+80 msec continuum, and is in fact probably about what would be expected if there <u>were</u> range effects at work, and if both endpoints of a continuum contribute to those effects. However, there is less variation in the boundary values for the natural continua than there is for the synthetic continua, as shown by the low standard deviations of 3.92 and 4.75 for the natural continua, vs. 5.7, 11.5 and 13.3 msec VOT for the synthetic continua.

2.2.5. Discussion

2.2.5.1. Range and task order effects

The 24 listeners in Wrocław heard three synthetic VOT continua in which VOT ranged from -20 to +80, -100 to +50, and -100 to +20 msec, in two task orders corresponding to <u>increasing</u> and <u>decreasing</u> lag values. There was a significant effect of Range on the boundary values obtained for each continuum, in that with more lag stimuli in the continuum, listeners have higher boundaries. There was also a significant Range X Task Order interaction, in that listeners who heard the -100/+20 msec continuum <u>first</u> had higher boundaries than they did on the -100/+50 msec continuum, or than the listeners who heard that continuum <u>last</u>. To interpret this effect with confidence,

we would want data on boundaries for each range continuum presented in various positions and task orders. However, it seems possible that the boundary for the first continuum heard is the least susceptible to range effects. Also, it seems possible that the -100/+20 msec continuum is inherently less prone to range effects than the -20/+80 msec continuum, since the former corresponds better to the range of VOT values found in natural Polish speech. Subjects heard either the -100/+20 msec continuum first, and the -20/+80 msec continuum last, or vice versa. Hearing the -100/+20 msec continuum first, without having heard the others, may lessen the range effects because responses are unbiased. On the other hand, hearing the -20/+80 msec continuum first may confuse listeners immediately, and contribute to increased range effects for continua heard subsequently. That is, each continuum's own range may be one factor in range effects for that continuum, but another factor may be the ranges of previous continua.

In the only systematic study of range effects, Darwin and Brady (1978) varied VOT range within blocks and also varied the order in which the different blocks were presented. They found a stronger range effect for the VOT range included in a single block, and a weaker effect for the overall range of previous blocks to which subjects had been exposed. Their range effects were greater when the ranges extended into the voiced (short lag) end of the continuum: the more voiced were the previously heard sounds, the more likely was a "voiceless" response to a particular stimulus. In the Polish experiment, listeners who heard the -100/+20 msec continuum after they heard the heavily-prevoiced -100/+50 msec continuum heard many more stimuli as voiceless than did listeners who heard

the -100/+20 msec continuum first. However, it should be noted that Darwin and Brady used a much smaller overall range of stimuli (+5/+55), about half of which covers boundary stimuli for English, and in each block used only a 20 msec range of VOT values. Thus their results may not be at all comparable to the results for these experiments in Polish.

Evidence that this is so comes from the fact that the methodology used here not only induced much larger range effects for Polish than Darwin and Brady's did for English, but failed to induce those effects for English. The six American control subjects showed remarkably consistent boundaries on the synthetic continua. The difference between the Poles' and the Americans' perception of these continua suggests that VOT is not as relevant a perceptual dimension for the voicing contrast in Polish as it is in English.

2.2.5.2. Prevoicing

The 24 Wrocław listeners identified tokens of words beginning with [d] from which prevoicing had been removed. Listeners judged speaker MG's tokens to contain [t]'s, while JP's tokens were judged to contain [d]'s. From this result it can be inferred that JP's tokens contained some other cue(s) to voicedness, while in MG's tokens such cues were weak or absent. Sample waveforms of the bursts of two of the pairs were illustrated in Fig. 2-12 and 2-13. Although MG's [d]-burst looks quite different from JP's, both are voiced, and so burst voicing by itself will not explain the difference in responses. In general, the burst intensity and degree of aspiration is greater for MG's [d]-bursts compared to JP's, and in fact her [d]-bursts overall resemble her [t]-bursts. However, with perceptual data available for four tokens only, it is impossible to

determine the relevant acoustic parameters.

MG is not the only speaker whose [d]-bursts often resemble her [t]-bursts, although she is the only one whose tokens were used for perceptual tests. A few of the Wrocław subjects also show this tendency. Fig. 2-20 shows a minimal pair, spoken by one speaker, with extremely similar bursts. Typical features are that both bursts are voiceless, and that the [t]-burst is longer in duration than the [d]-burst.

Prevoicing by itself is clearly a sufficient,and possibly an overriding, cue for voicedness. In the apparent absence of other cues for voicedness, MG's original tokens were uniformly perceived as containing [d]'s.

When prevoicing is removed, other cues to voicing assume some importance, and their relative strengths may determine the percept. Prevoicing is not always a necessary cue for voicedness, since JP's edited stops are still perceived as [d]'s. However, some speakers, such as MG, may in their production provide only one cue for voicedness, prevoicing, and in their speech prevoicing is a necessary cue for voicedness. In natural speech this strategy is perfectly successful. What remains unclear is whether listeners can discriminate between the tokens with only one cue to voicedness and the tokens with more than one.

2.2.6. Conclusions about VOT perception

In this section I have shown that Polish listeners can respond systematically to stimuli with a variety of VOT values, using both natural and synthetic speech stimuli. In most cases listeners show sharp labeling boundaries along the test continua. However, the exact position of these boundaries is subject to strong effects

from the range of VOT values included in each continuum. The natural-speech continua may also have been subject to range effects, but the range of these stimuli was not varied. Still, the natural continua, with no prevoiced stimuli, gave boundary values different from the synthetic continua, which included substantial numbers of stimuli with prevoicing. The boundaries for the natural continua were like those for the synthetic continuum which contained few prevoiced stimuli. To the extent that the natural-speech continua show range effects, then the range effects cannot be attributed entirely to poor-quality, hence ambiguous, synthetic stimuli. However, the range effects are intrinsic to the listeners' perception and not to the stimuli, or even to VOT, since the American control subjects did not show any range effects on these continua.

The result indicates that the VOT dimension does not have the perceptual stability in Polish that it has in English. Polish listeners are able to use VOT as a perceptual cue to voicing contrasts, but their categorizations are easily influenced by extraneous test circumstances.

I have also shown that prevoicing is a sufficient and sometimes a necessary cue for voicedness. Presumably other cues may also be present in the signal, which can,for some speakers, maintain the voicing contrast in the absence of prevoicing. It is this influence of "other" cues that will be addressed in the next chapter, where the role of bursts will be compared to that of VOT.

2.3. General Discussion

2.3.1. Relation of Production and Perception

The strong range effects found in the Polish perception tests

resulted in a variety of VOT boundaries for each subject. When the results of the production and perception studies are compared in an attempt to align the voicing categories, this variety of boundaries make it inevitable that at least some of these boundaries will not correspond to the production categories. Of interest, then, is the question of which VOT continua gave boundary values which are aligned with the production categories.

It is not sufficient to compare an individual subject's own production and perception results in considering this question. Rather, it seems more relevant to ask whether an individual's perceptual boundaries align with the group production categories, since we are interested in group communication. That is, our standards for category matching should be conservative, since a Pole with a high VOT perceptual boundary must nonetheless be able to understand a Pole with relatively short-lag [t]'s. From the group production data, it would seem that the optimal category boundary would be between -20 and +10 msec VOT.

The perception data may be summarized as follows: the -100/+50 msec and -100/+20 msec continua had similar overall mean boundaries; the mean for these two continua is 4.5 msec VOT. The -20/+80 msec continuum had a mean boundary of 20.4 msec VOT. The two natural continua had a mean boundary of 17 msec. The boundary for a more extended range of natural stimuli would perhaps be lower.

It is quite possible that the range of VOT values used in a test continuum is more important than the naturalness of the stimuli in providing a good match between the production and perception categories. The natural speech continua had appropriate values of other acoustic parameters for Polish apical stops, but did not

include lead values in their VOT ranges. These continua resulted in boundary values that fell between the voiced and voiceless production categories only slightly better than did the inappropriate -20/+80 msec synthetic continuum. However, the natural speech continuum with [t]-bursts, the most realistic combination, gave systematically lower boundaries than did the continuum with [d]-bursts, and therefore provides a slightly better match with the production data. The best match, however, is provided by the two continua with the appropriate balance of lead and lag VOT values for Polish. This is true despite the fact that structurally these stimuli do not much resemble natural Polish speech, especially in the bursts.

These continua, the -100/+50 and -100/+20 msec ones, give some negative and low positive boundary values that are particularly good matches for the production data. Consider the distribution of production values shown in Fig. 2-6. Of the 100 boundary values that were computed for labeling functions from Wrocław, 32% fall between the production categories (between -20 and +10 msec), while 68% fall within the [t]-category. Most of the mismatches come from the three continua with few prevoiced stimuli (-20/+80 msec, natural speech continua), and virtually all of the matches come from the two continua with many prevoiced stimuli (-100/+50, -100/+20 msec). In other studies where a mismatch between production and perception has been found (e.g. Lisker and Abramson, 1970), the problem has also been that perceptual boundaries have VOT values that are too high -- too far along into the lag region -- for the production categories. In the past, this mismatch has been attributed to acoustic inappropriateness of the stimuli, but in the present study this explanation seems to be inadequate. The Abramson and Lisker

synthetic stimuli, used in the most appropriate ranges of VOT, provide a better match with the production data than do the natural stimuli, although it must be noted that the "natural" stimuli have themselves been altered in somewhat unnatural ways.

Samples of individual data are shown in Fig. 2-23. An individual's own boundaries are more successful matches for his own production data than they are for the group data. It is surprising that more listeners do not place their perceptual boundaries derived from the minimal pair stimuli in the break between about -30 and 0 msec VOT for minimal pair production. The explanation may lie in the production values for the running speech condition, which occupy this low lead region of the continuum. If perception of minimal pairs corresponded to production of minimal pairs, then substantial realignment of the perceptual categories would then be required for the running speech condition.

The lack of a clear match between production and perception categories, especially with natural speech stimuli, again argues against the primacy of the VOT dimension in Polish voicing contrasts. It may be that Polish listeners are able to relate VOT differences in a general way to their voicing categories, but when asked to use these differences to distinguish two separate categories, the VOT information is not entirely sufficient, and listeners resort to strategies that include taking the overall VOT range into account. The American listeners were able to make voicing judgements without being influenced by the range of stimuli present. Their perception of VOT is stable in a way that the Poles' is not.

The Polish listeners may be using low-level psychoacoustic

information in making their judgements. A natural psychoacoustic boundary is thought to fall at about +17 to +20 msec temporal separation (Miller et al, 1976; Pisoni, 1977 -- see Ch. 1), which corresponds to the Polish labeling performance.

The Polish listeners showed two types of responses. When the range of VOT values in a continuum corresponded more or less to natural production values, they showed boundary values between the production categories -- generally negative and low-positive values. When the range of VOT values in a continuum did not correspond to natural production values, they showed boundary values corresponding to a natural psychoacoustic one. Rather than make linguistically-relevant voicing judgements, they responded according to their ability to distinguish the onset of the stimulus from the separate onset of voicing.

It is not the case that the Polish listeners are simply tapping an inherent distinction that corresponds to one that American listeners already use. The American listeners in this set of experiments showed higher boundary values, corresponding to a different psychoacoustic boundary, than did the Poles: up to about +30 msec VOT for the Poles, but from about +30 to +45 msec for the Americans. None of the continua produced range effects that gave English-like boundaries for the Polish listeners. It would appear then that the Americans responded to the stimuli consistently across VOT ranges according to a language-specific voicing distinction. The Polish listeners did not always do so.

2.3.2. Summary and Conclusions

The production data reported here indicated that the Polish apical stop voicing contrast is essentially one of voicing before

the burst vs. no voicing before the burst. The burst itself is ambiguous, since it may be voiced or voiceless for a voiced stop. However, results of the perceptual identification experiments where prevoicing was removed from tokens of [d] show that the contrast is not so simple. Other cues to voicedness may optionally be present in the signal. Prevoicing (or closure voicing) seems to be a primary cue, but other secondary cues remain to be identified. Speakers seem to vary in their use of the secondary cues in production.

In a perception experiment using VOT continua, subjects' performance depends to a large extent on such experimental factors as range of VOT values included in the current test, as well as in the previously-heard tests. If the range of VOT values in the continuum includes little or no prevoicing (that is, the range is inappropriate for Polish), then the psychoacoustic categories will assert themselves over the expected Polish ones. Thus the match between production and percpetion categories is highly dependent on these range effects.

Zlatin (1974), on the other hand, found an excellent match between production and perception categories for VOT in English. The present study found that American listeners are not subject to the kind of range effects found for Poles. Clearly, then, VOT is not a stable dimension perceptually for Polish listeners, in the way it is for Americans.

In terms of describing production, VOT is a satisfactory dimension, but it provides more detail than is absolutely necessary to describe the relatively simple contrast in Polish. An alternative description was proposed that uses voicing during the closure interval as the defining trait for voicedness. Tokens with any

voicing during closure are [d]'s, and others are [t]'s. However, both descriptions are inadequate in light of the perception data. Other cues to voicing must be sought, during or after the release burst, and it is this requirement that the next chapter addresses.

FOOTNOTES

[1]Michael Jacek Mikoś, native Polish linguist and collaborator in parts of this project.

[2]The medial contrast is analyzed in Ch. 4.

[3]No information about recording equipment used was sent back with the cassette.

[4]The other two pairs containing medial contrasts are analyzed in Ch. 4.

[5]Discrete Fourier Transform displays of both burst and vowel portions of these sentences shows that there is little information present above 2 kHz.

[6]As the vocal folds are adducted, a few pulses may be required to establish a regular vibration pattern. Lisker and Abramson (1967) claim that such irregular pulses have no perceptual significance, but the question remains problematical.

[7]In some cases this exact point was difficult to determine, either because the onset of voicing was very gradual, or because burst frication obscured the onset of voicing.

[8]In such cases it is also possible to make a second VOT measurement from the burst to the second onset of voicing. This is also possible in cases where the prevoicing does not die down completely but shows a decrease in amplitude followed by a voiceless burst. Voicing then resumes immediately after the burst. Measures of the duration of the frication that occurs with the burst have been shown to correlate with voicing distinctions in English (Klatt, 1975; Zue, 1976), but such measurements could not be made with sufficient reliability to be used in this study.

[9]Such a difference might indicate different speeds of consonantal release due to different degrees of muscle force.

[10]An AX discrimination tape was also constructed for the -100 to +50 msec continuum, but pilot work indicated that Polish listeners called most pairs the same. This in itself is suggestive about the perception of these stimuli, but this line of inquiry was not pursued. Only identification data was collected in these experiments.

[11]The same analysis of variance was run using computations of the four "missing values" on the -100/+20 msec continuum, thereby allowing these subjects' data for the other continua to be included in the analysis. The means changed very little with these additions, and the results of the analysis remained the same.

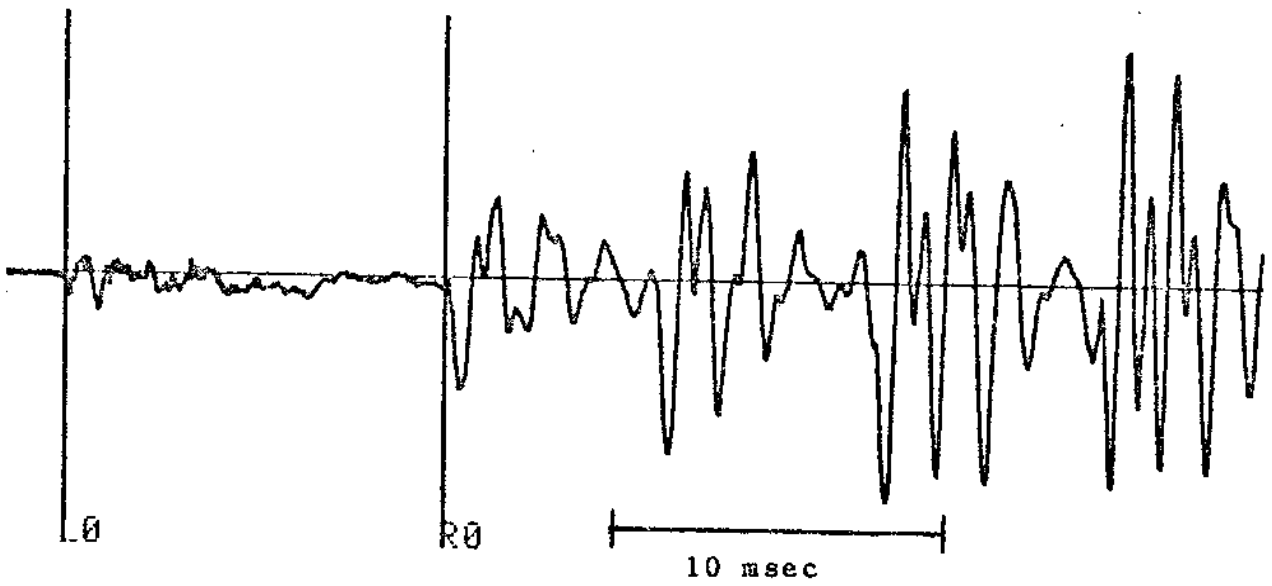

Fig. 2-1a -- Waveform showing landmark points in measuring positive VOT. The left cursor (LØ) is set at the burst, and the right cursor (RØ) is set at voice onset (zero-crossing before first negative peak). The VOT is 11.6 msec. The token is tama spoken by Wrocław subject #4.

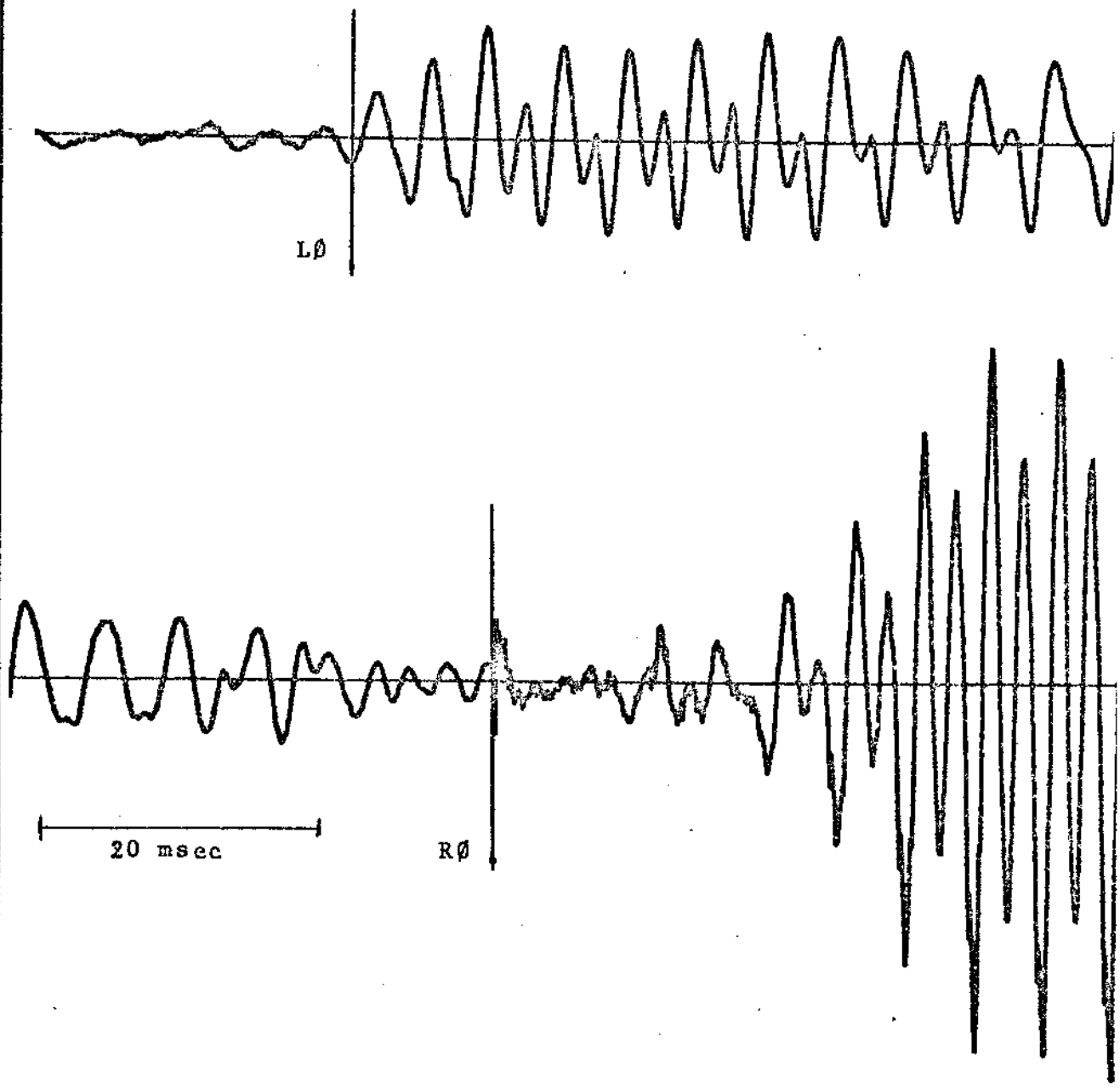

Fig. 2-1b -- Waveform showing landmark points in measuring negative VOT. The left cursor (LØ) is set at the onset of prevoicing, and the right cursor is set at the burst. The token is dym spoken by MG. It is continuous across the two lines on the page.

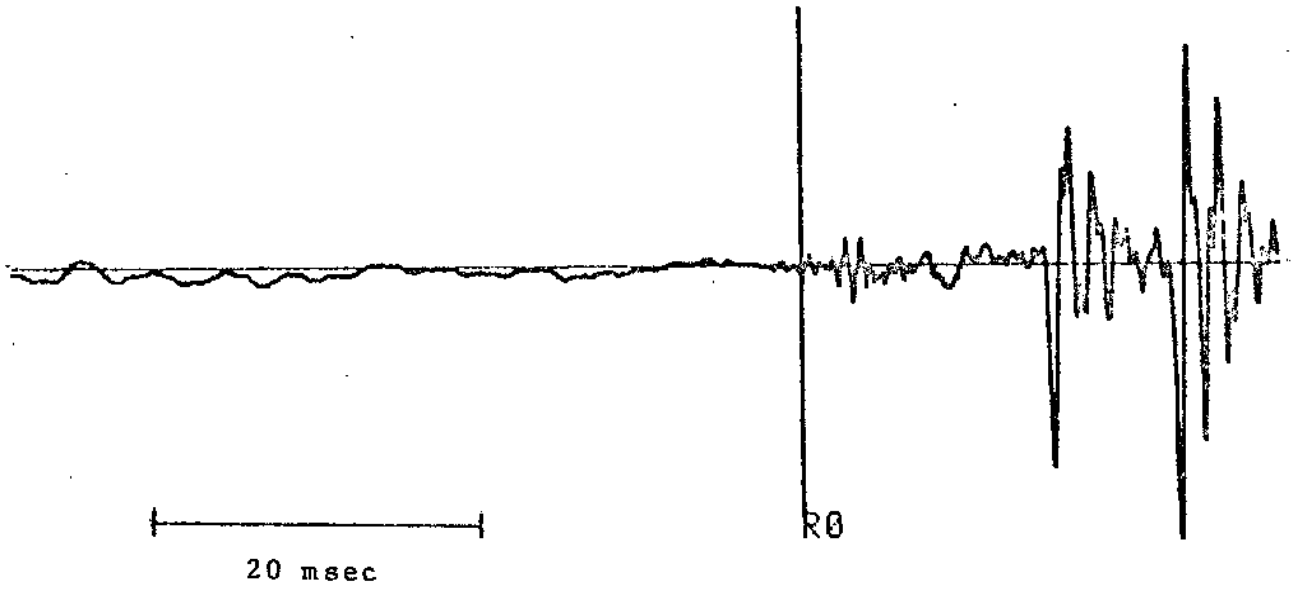

Fig. 2-2 -- Waveform showing prevoicing dying down before burst in voiced token, dama spoken by Wrocław subject #3. The cursor is set at the burst.

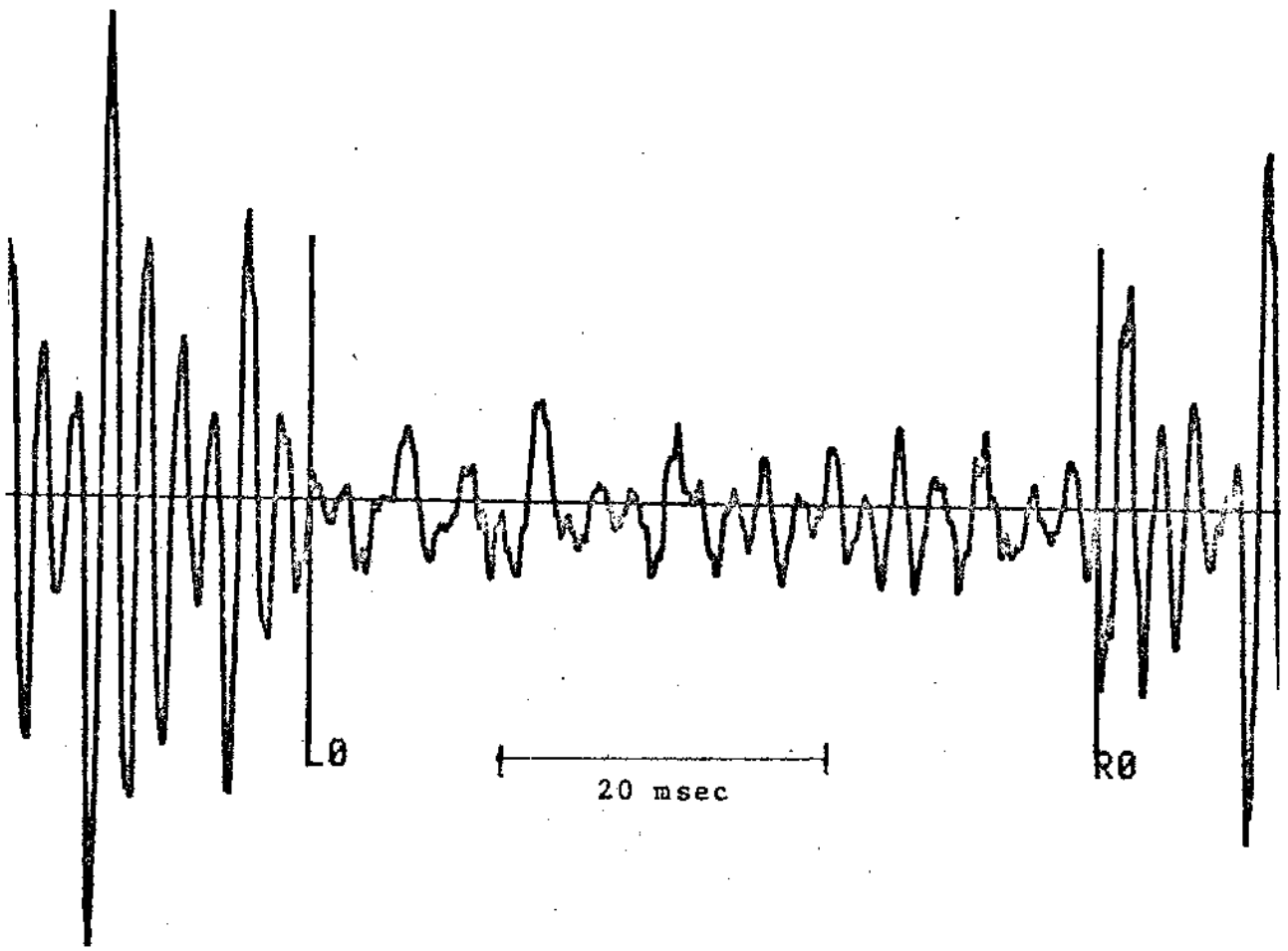

Fig. 2-3a -- Waveform showing estimate of closure duration in [d]-token with no visible burst. The left cursor (LØ) is set at the onset of closure, and the right cursor (RØ) is set at the amplitude increase for the following vowel. The measured estimate is 48.6 msec, rounded down to 45 msec. The token is do (unstressed) in a sentence read by JP.

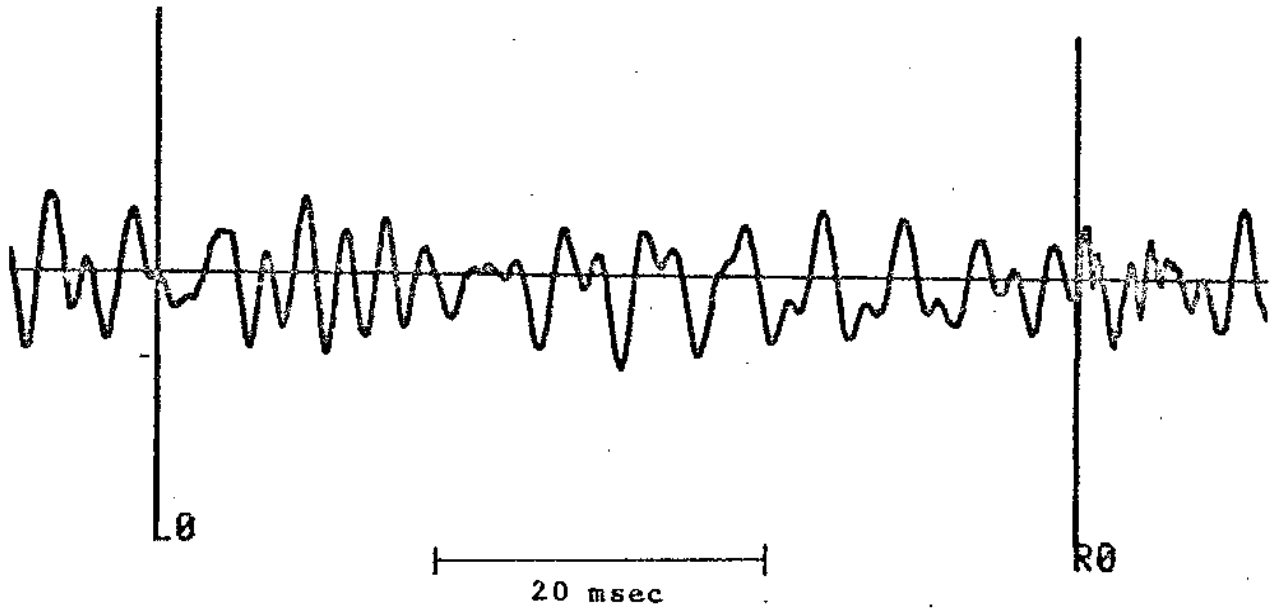

Fig. 2-3b -- Waveform showing measurement of closure duration with burst visible. The oscilloscope display, unlike the hard-copy shown here, makes burst frication and aspiration more distinctive. The measured value from closure onset to burst is 56.6 msec. The token is <u>domu</u> spoken in a sentence by MG.

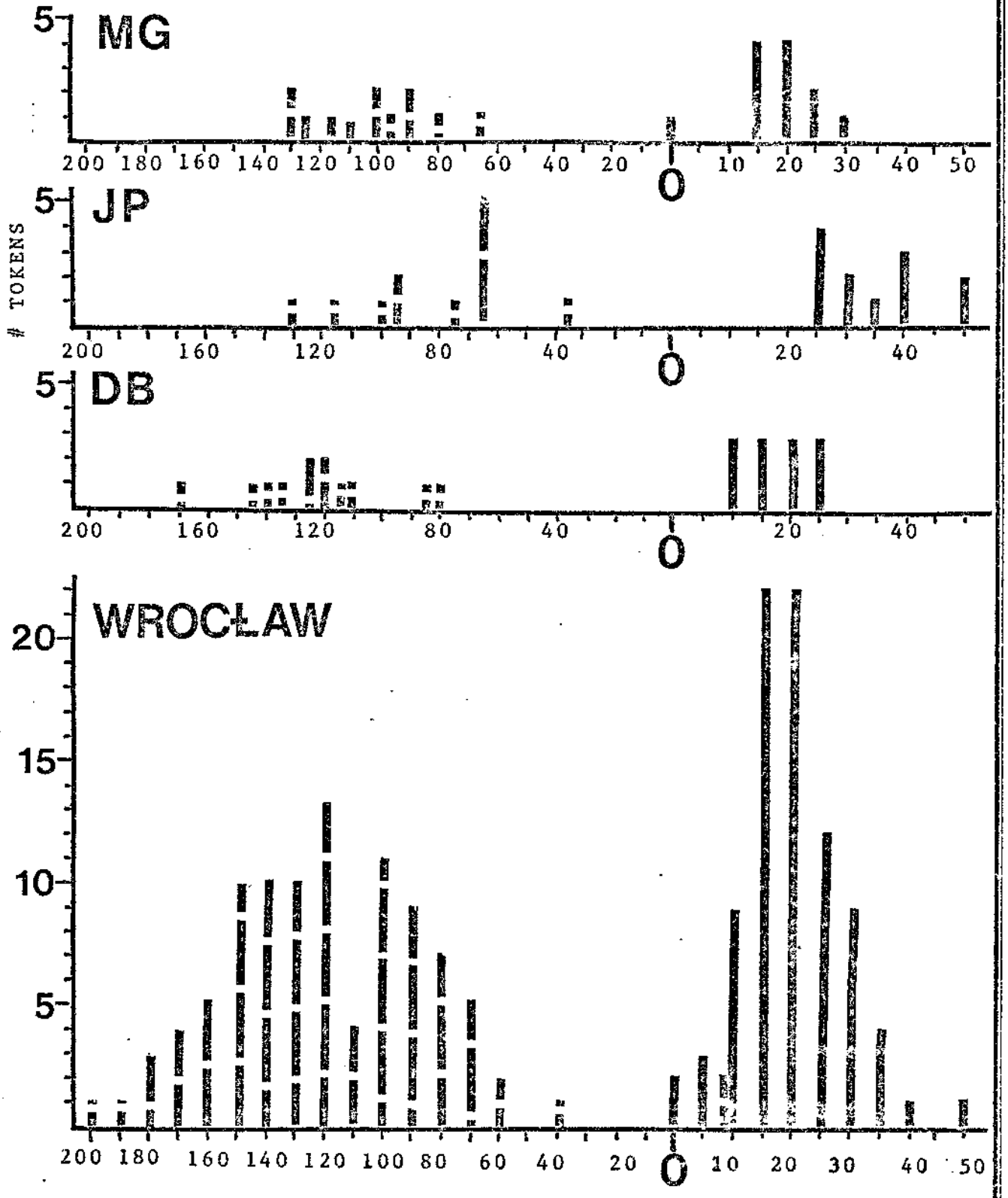

Fig. 2-4 -- Distributions of VOT values for readings of minimal pairs by 3 individuals and by 24 speakers in Wrocław. Note that the scale of positive values is expanded here.

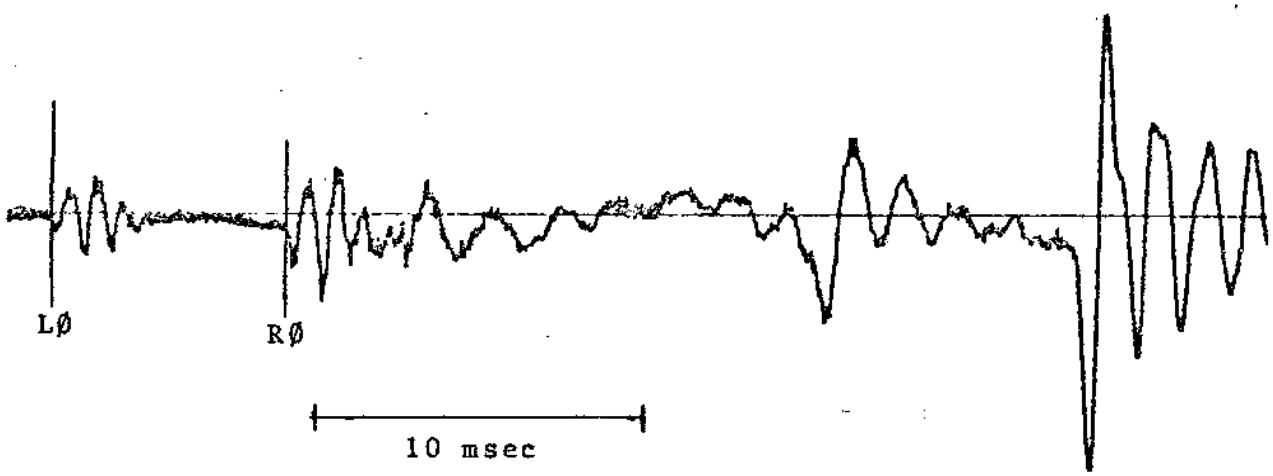

Fig. 2-5 -- Waveform showing double burst. Each cursor is set at one burst. The token is ta spoken by Wrocław subject #2.

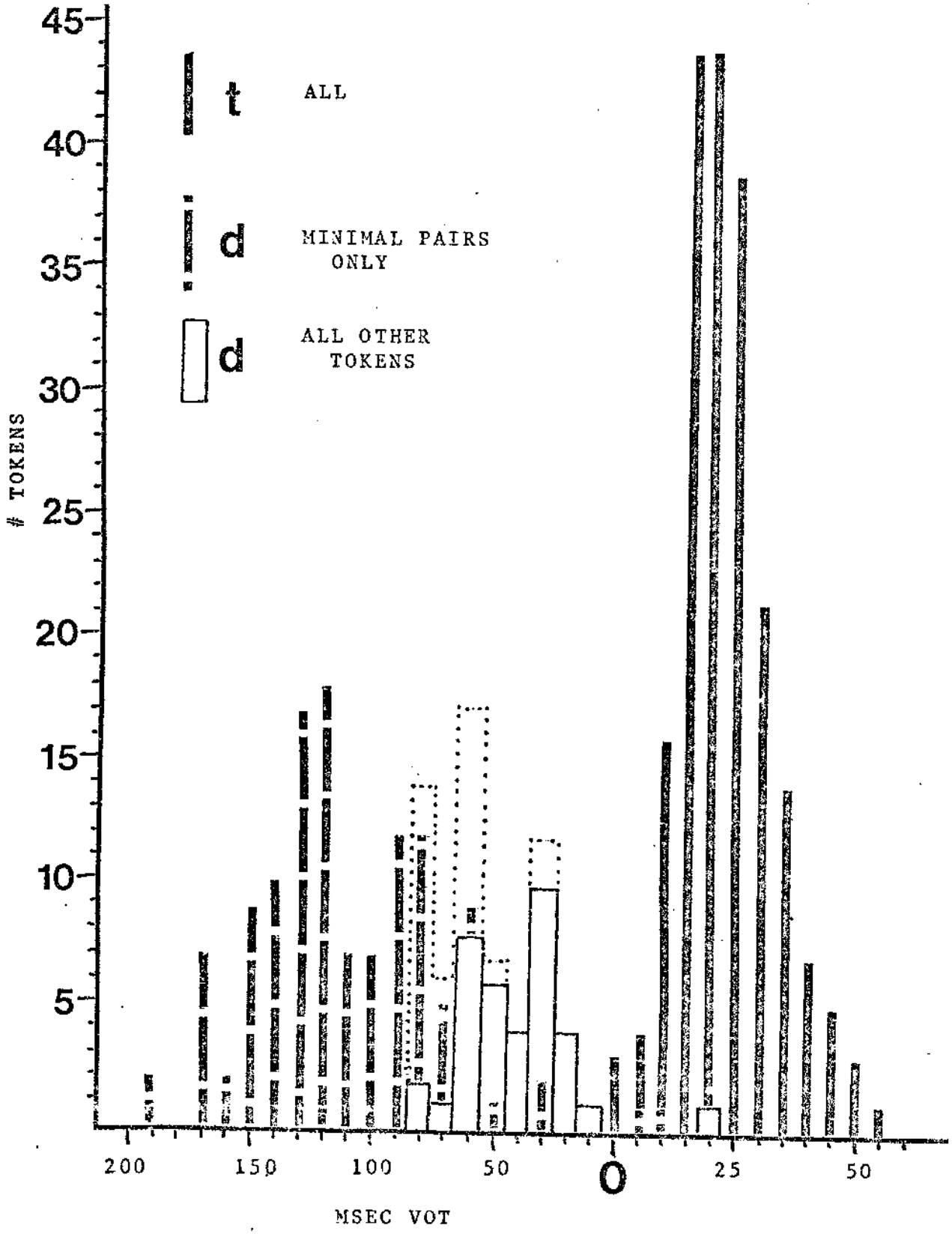

Fig. 2-6 -- Distribution of all measured VOT values. Note the expanded scale for positive values. The dotted distribution represents the combined function for all [d]-tokens.

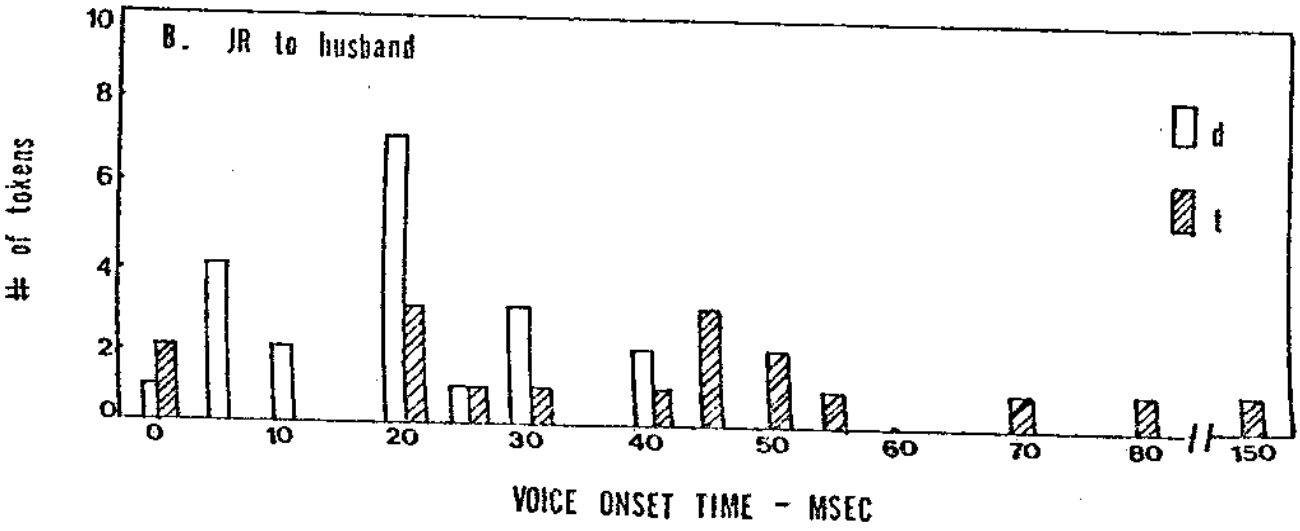

Fig. 2-7 -- Distribution of VOT values from English conversation, taken from Moslin (1978). Note the high degree of overlap between [t] and [d], typical of casual running speech in English.

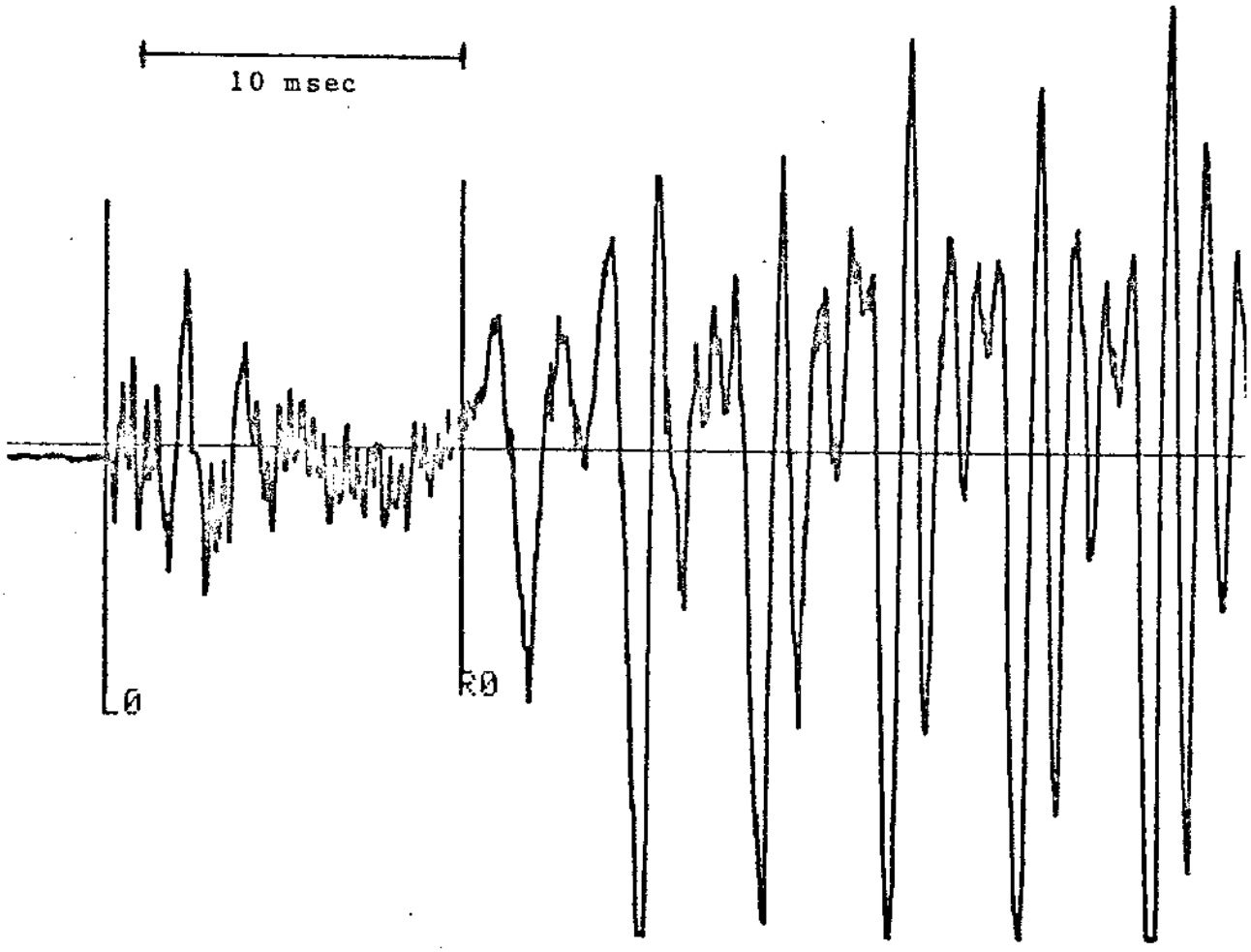

Fig. 2-8a -- Waveform showing aspirated [t]-burst. The VOT value is 11.3 msec. The token is tama read by Wrocław subject #13.

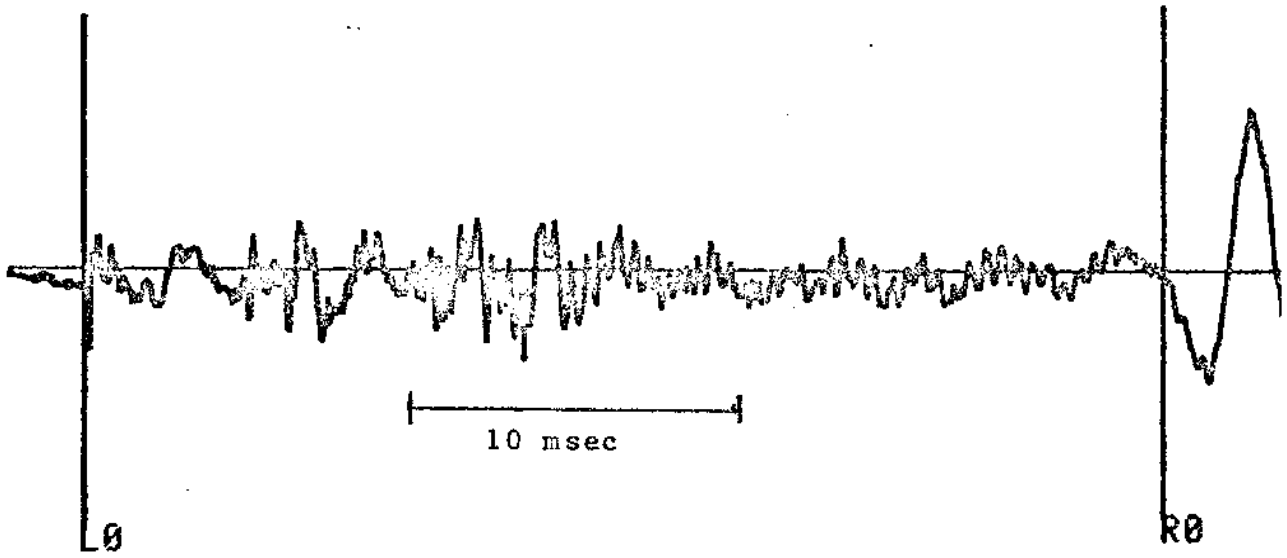

Fig. 2-8b -- Waveform showing long-lag VOT token with a VOT value of 33.4 msec. The lag interval is filled with aspiration. The token is tym read by MG (see Fig. 2-1b for dym by this speaker).

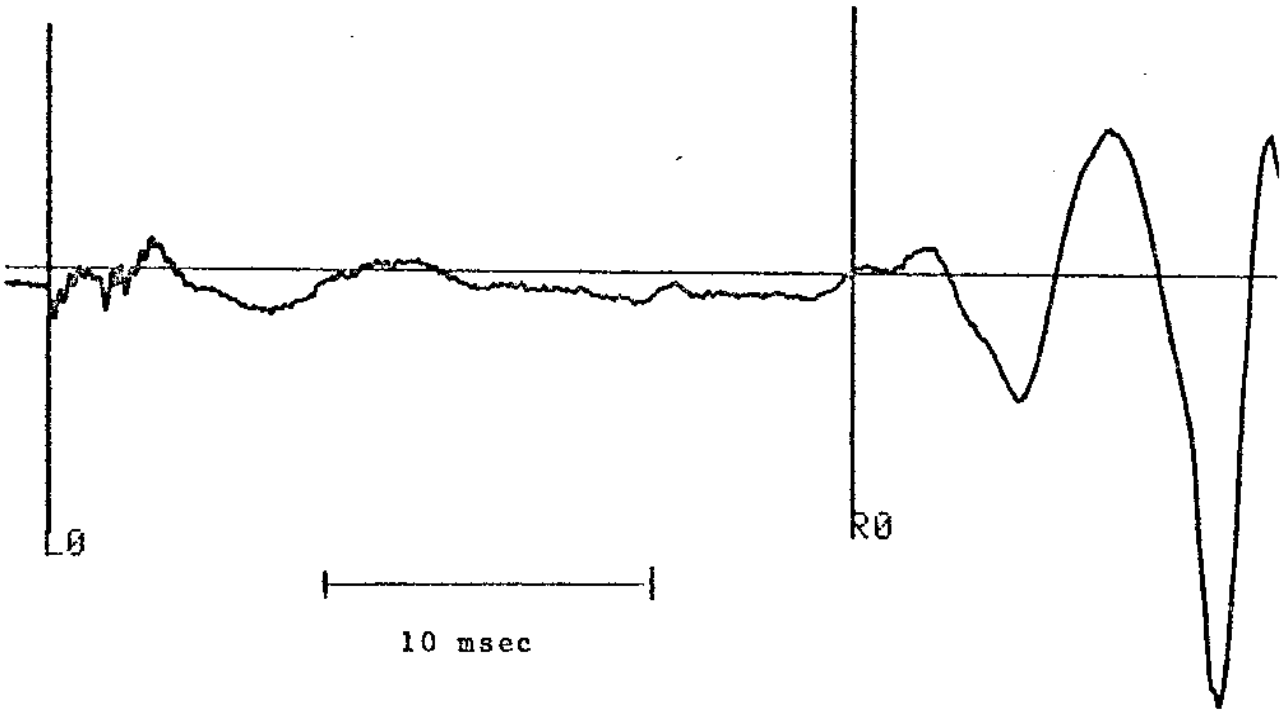

Fig. 2-8c -- Waveform showing moderate-lag VOT token with little apparent aspiration. The VOT value is 24.7 msec. The token is tur read by Wrocław subject #25.

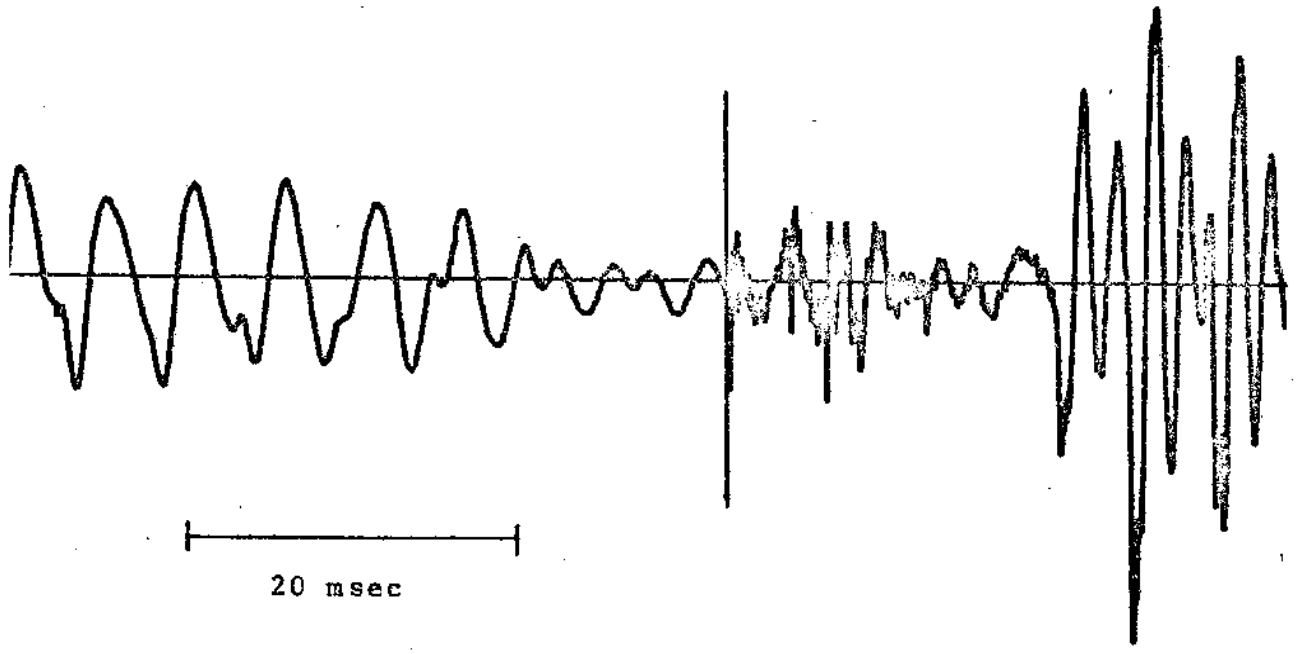

Fig. 2-9a -- Waveform showing voiceless [d]-burst following amplitude decrease in prevoicing. The cursor is set at the burst, which is fricated. The token is <u>domy</u> read by MG.

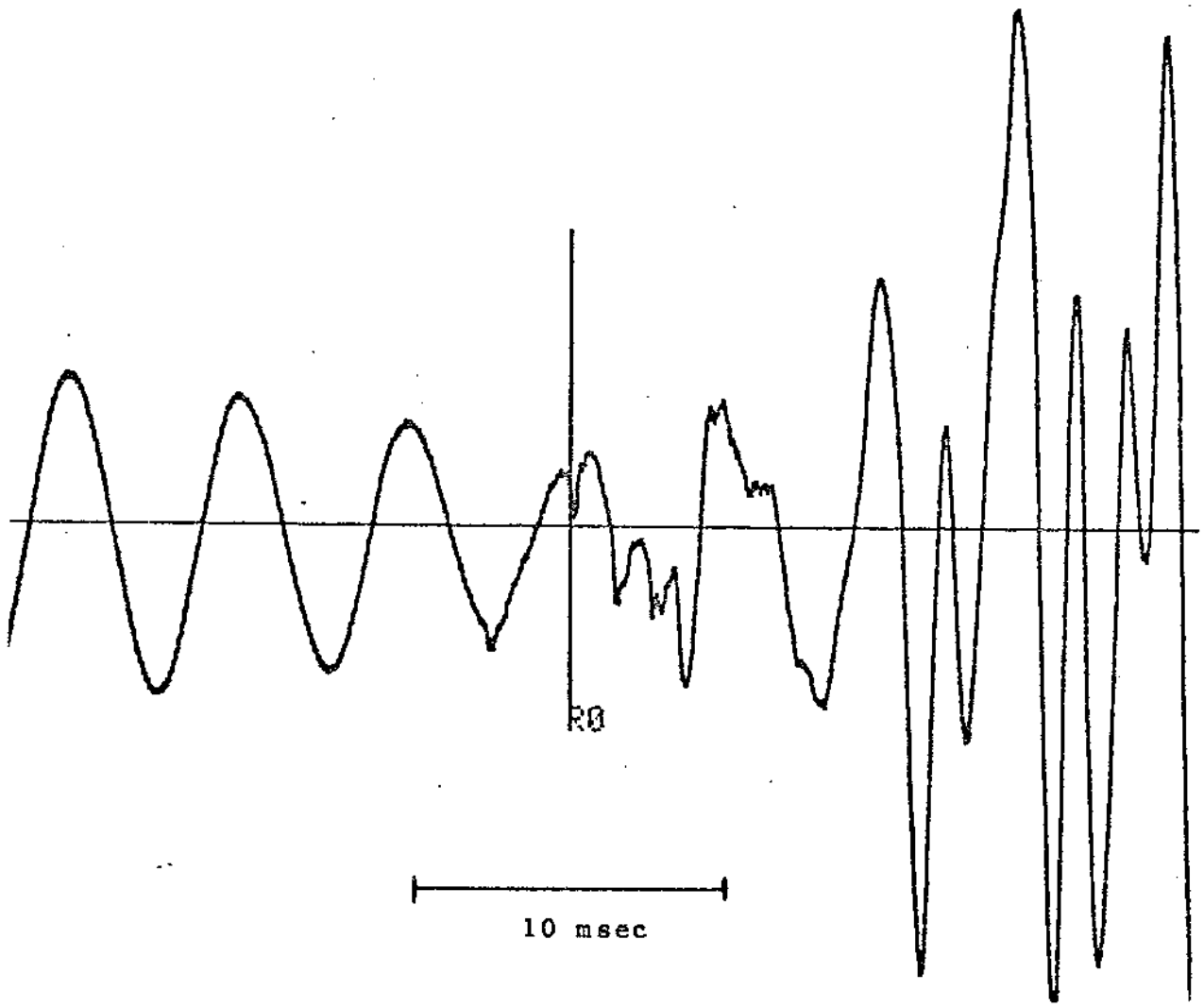

Fig. 2-9b -- Waveform showing [d]-burst superposed on pitch period resembling prevoicing. The cursor is set at the burst. The token is <u>data</u> read by Wrocław subject #22.

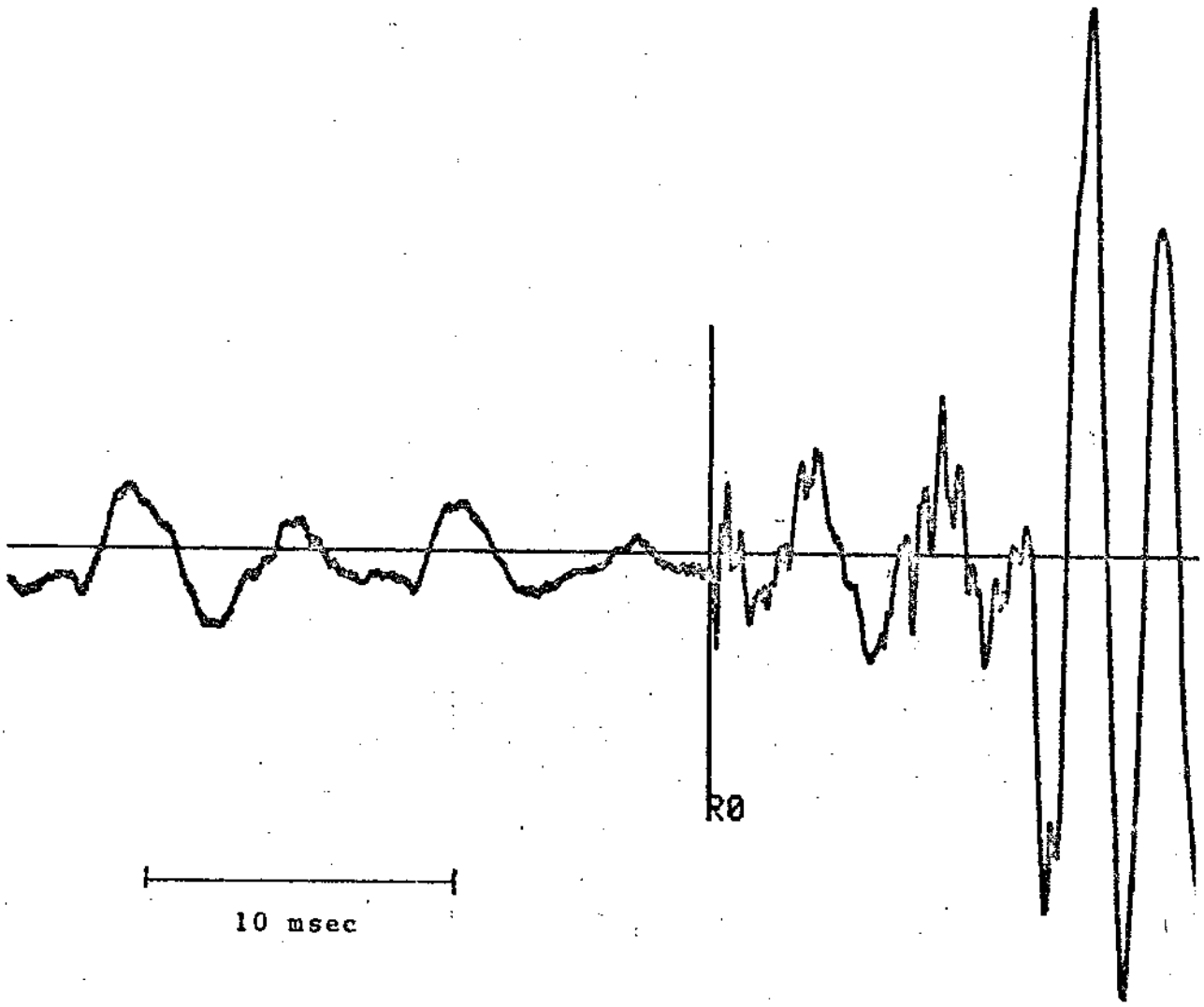

Fig. 2-9c -- Waveform showing typical "voiced" [d]-burst, with voicing pulses within the burst. The cursor is set at the burst. The token is do read by JP.

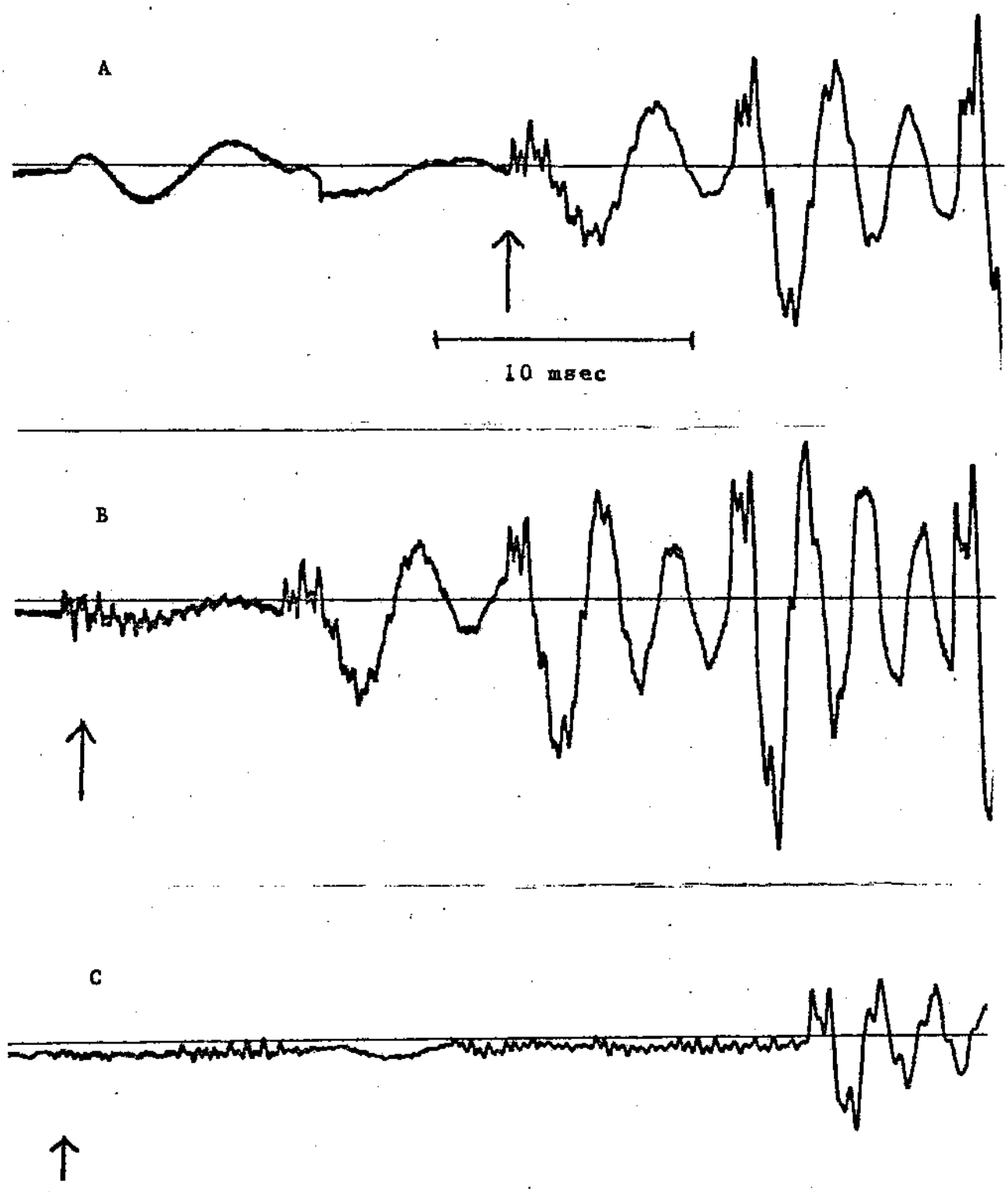

Fig. 2-10a,b,c -- Waveforms showing Abramson and Lisker synthetic VOT stimuli with VOT values of -10, 0, and +30 msec. The arrows indicate the approximate locations of the bursts; note that the bottom display has no clear burst.

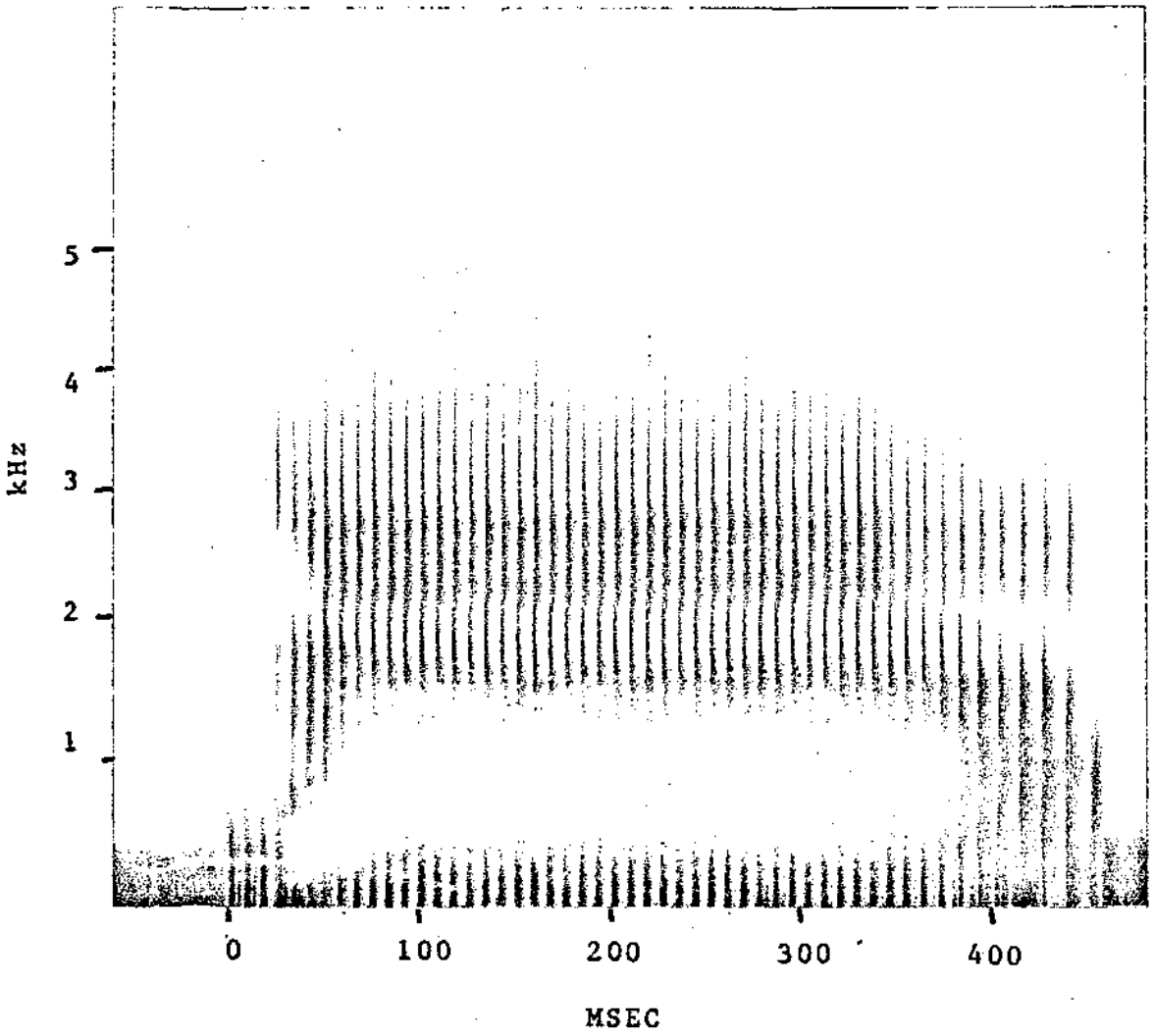

Fig. 2-10d -- Spectrogram of Abramson and Lisker synthetic VOT stimulus, with a VOT value of -20 msec.

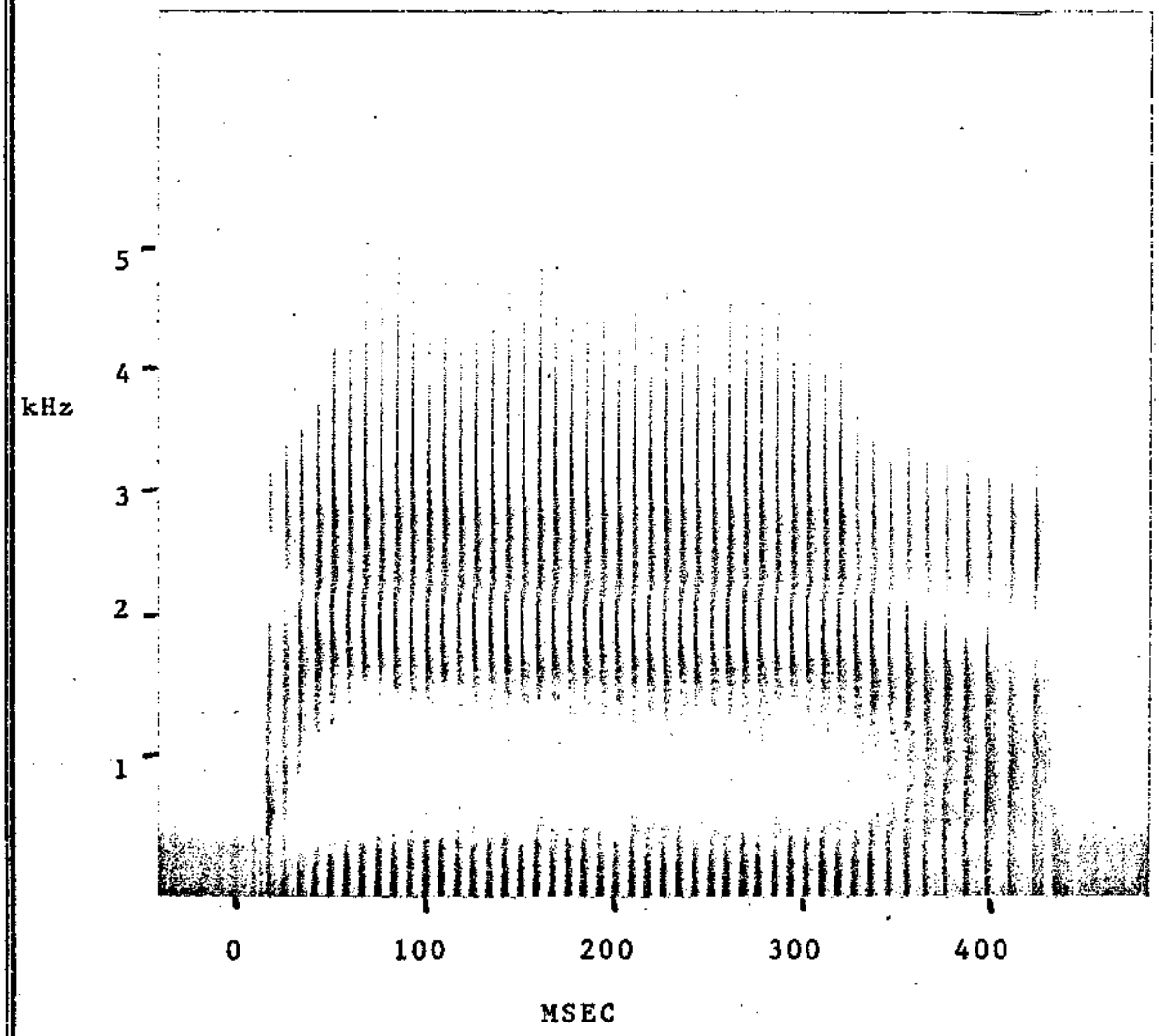

Fig. 2-10e -- Spectrogram of Abramson and Lisker synthetic VOT stimulus with a VOT value of +20 msec.

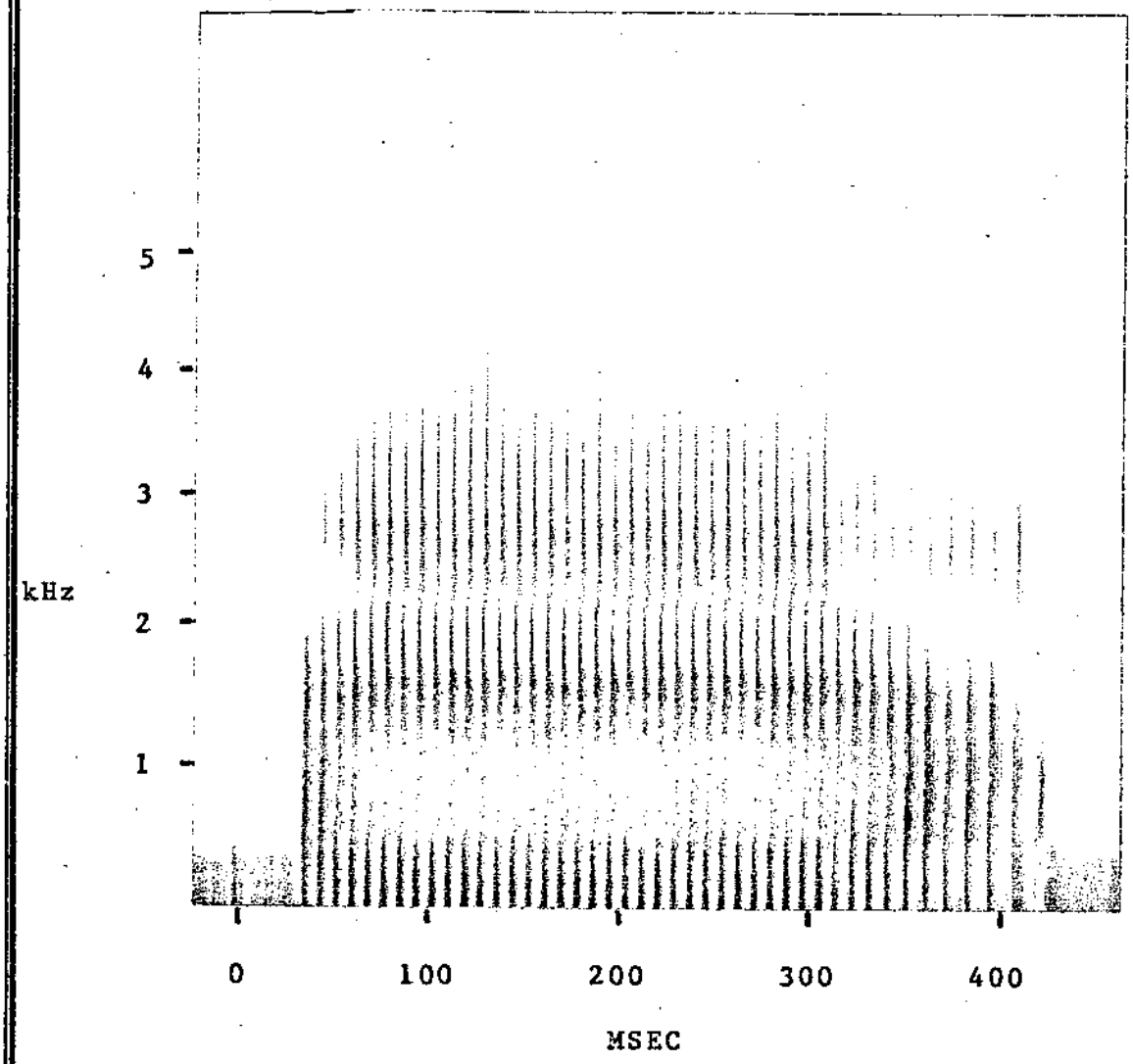

Fig.2-10f -- Spectrogram showing Abramson and Lisker synthetic VOT stimulus, with a VOT value of +50 msec.

Fig. 2-11a

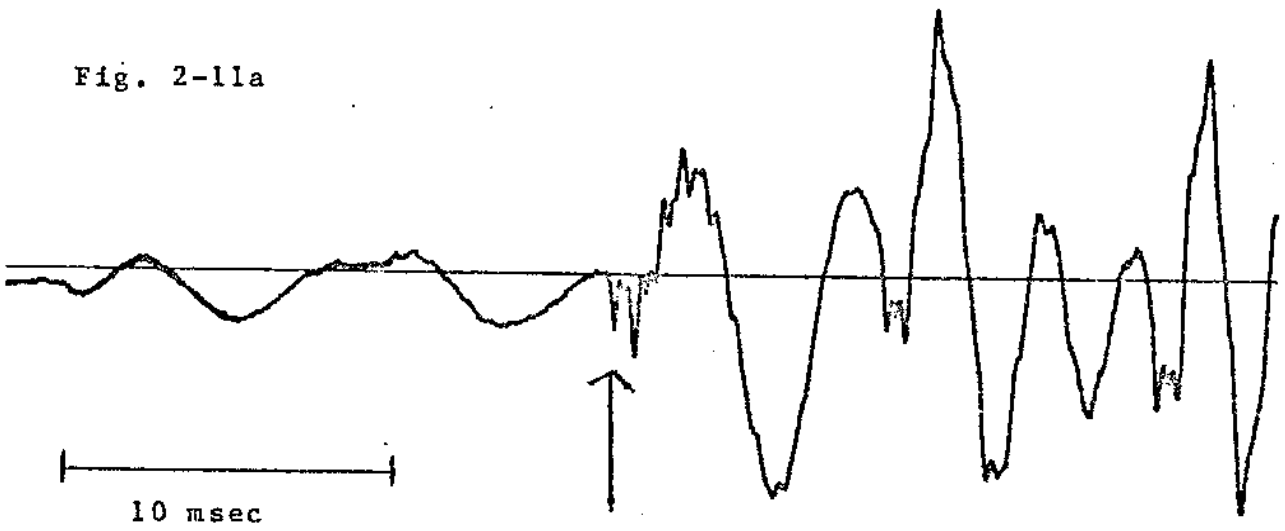

Fig. 2-11b

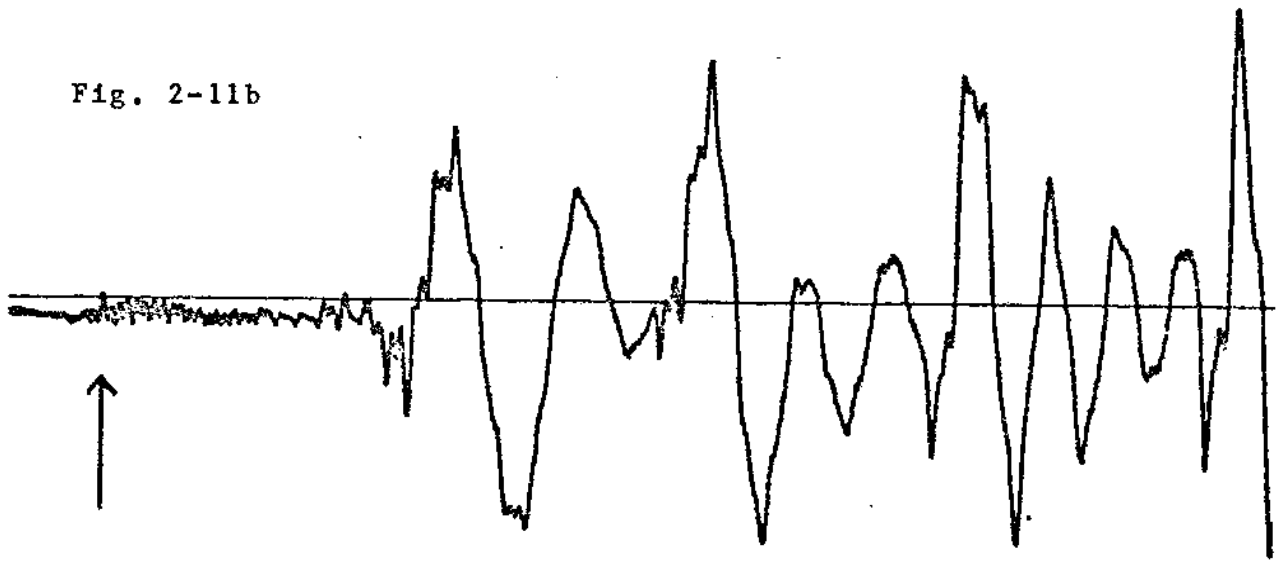

Fig. 2-11a -- Waveform showing onset of Haskins VOT stimulus with nominal VOT value of -10 msec. The arrows in both stimuli indicate where the bursts are located.

Fig. 2-11b -- Waveform showing Haskins VOT stimulus with nominal VOT value of +10 msec.

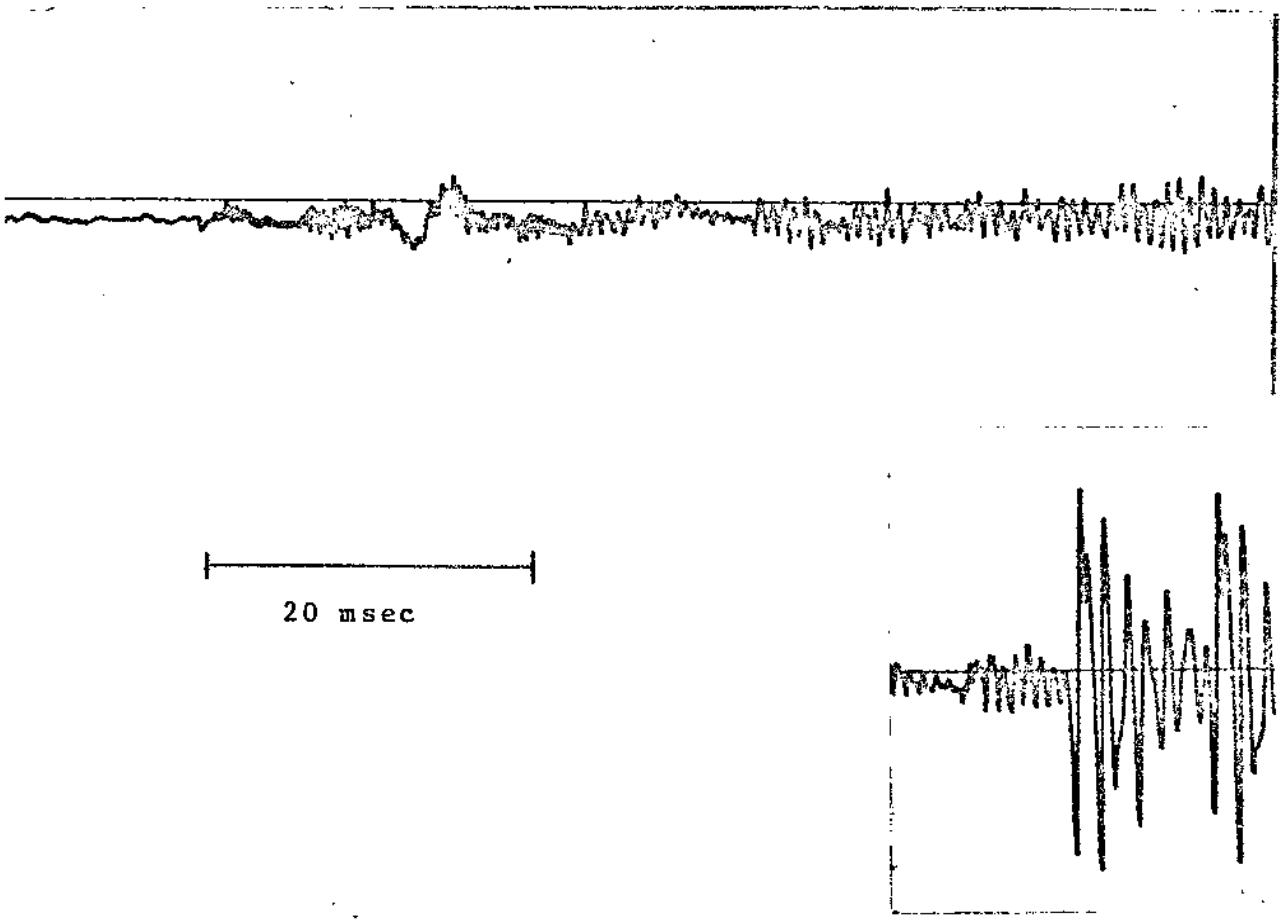

Fig. 2-11c -- Waveform showing Haskins VOT stimulus with nominal value of +80 msec VOT. The display is continuous across the two lines here.

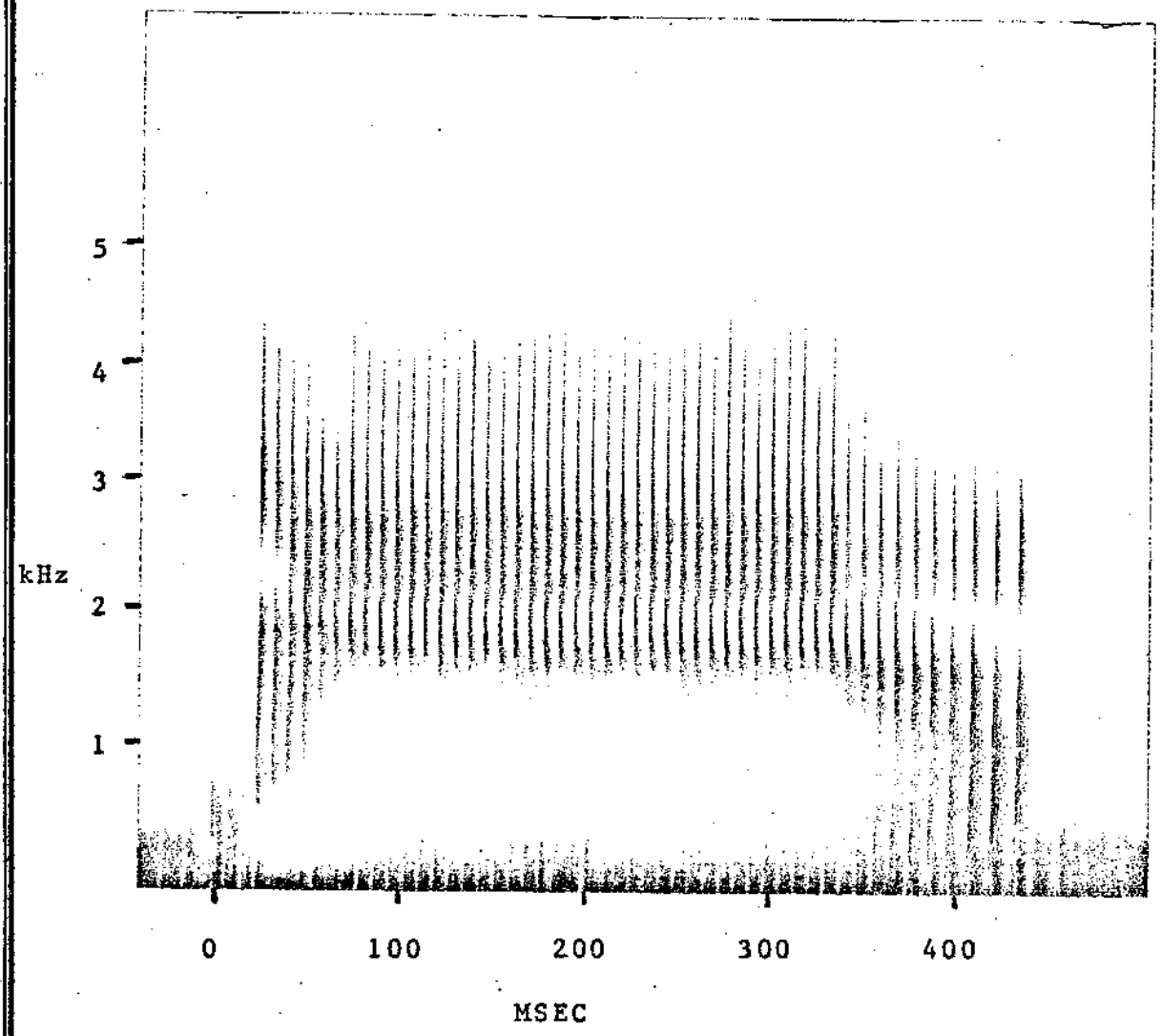

Fig. 2-11d -- Spectrogram showing Haskins synthetic VOT stimulus similar to Abramson and Lisker stimuli, with a VOT value of -20 msec.

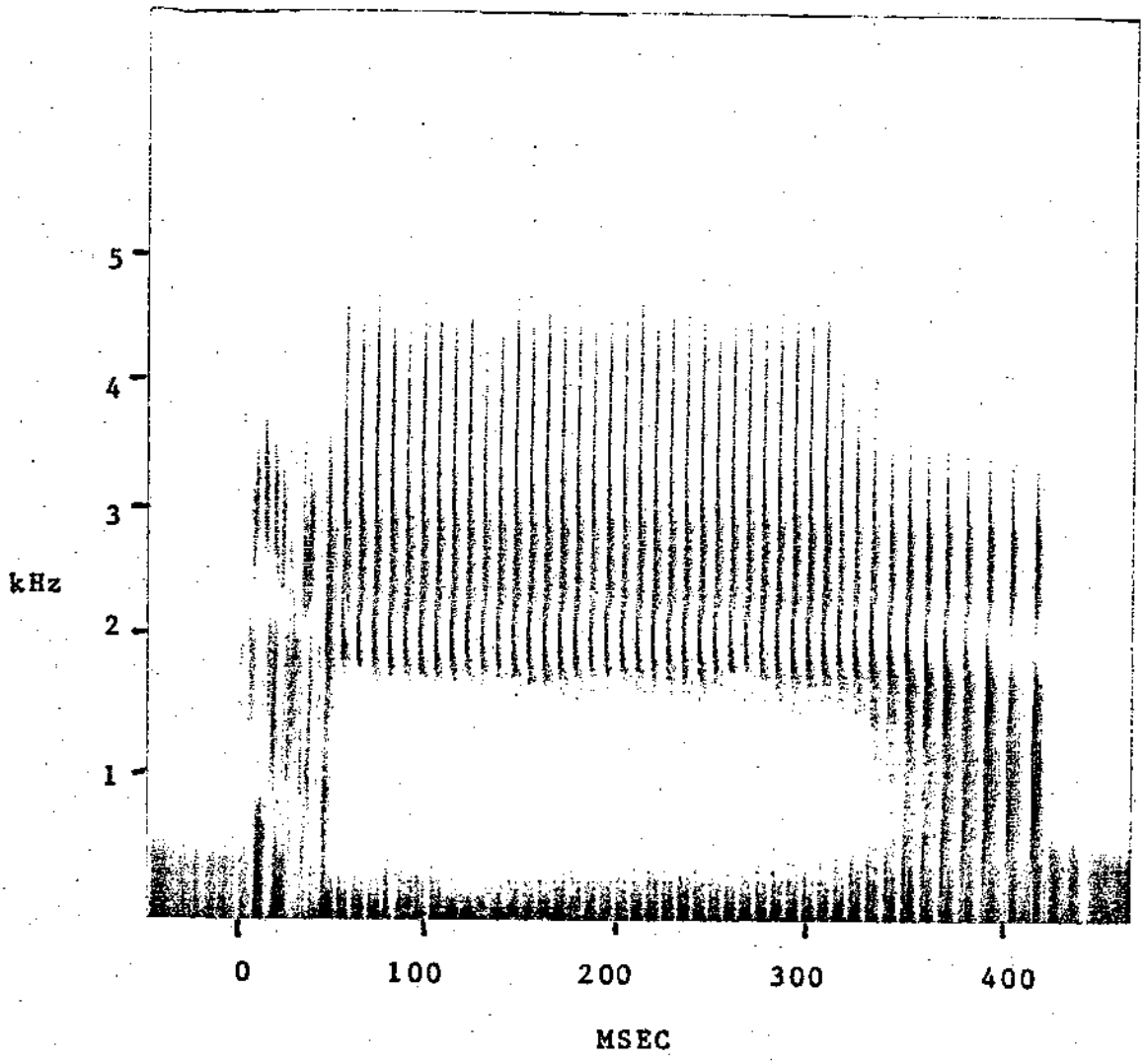

Fig. 2-11e -- Spectrogram showing Haskins synthetic VOT stimulus similar to Abramson and Lisker stimuli, with a VOT value of +50 msec.

Fig. 2-12a

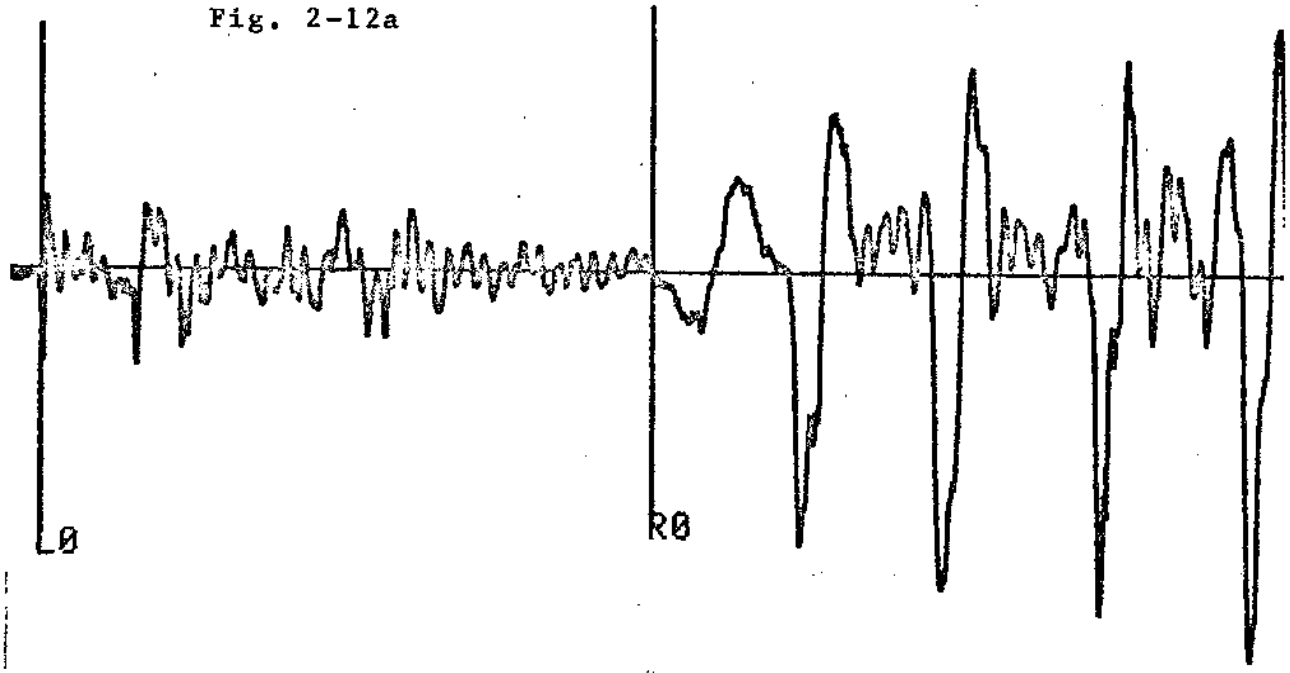

Fig. 2-12b

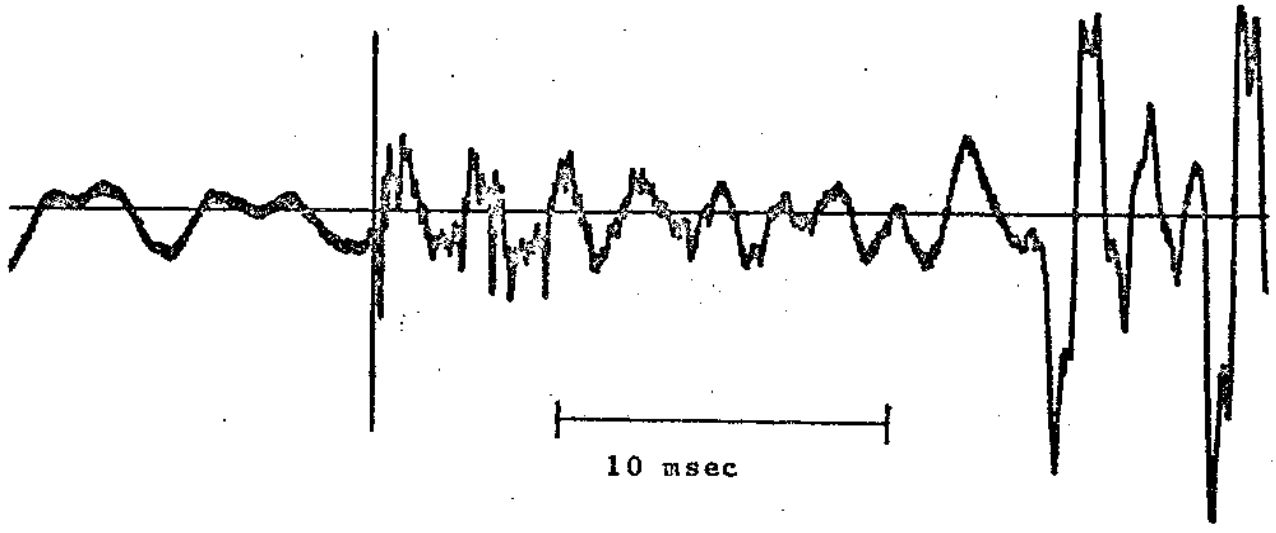

Fig. 2-12a -- Waveform showing token of tata read by MG. The VOT is 18.8 msec as measured between the cursors.

Fig. 2-12b -- Waveform showing token of data read by MG. The cursor is set at the burst. With the prevoicing removed, this token is identified by listeners as [t].

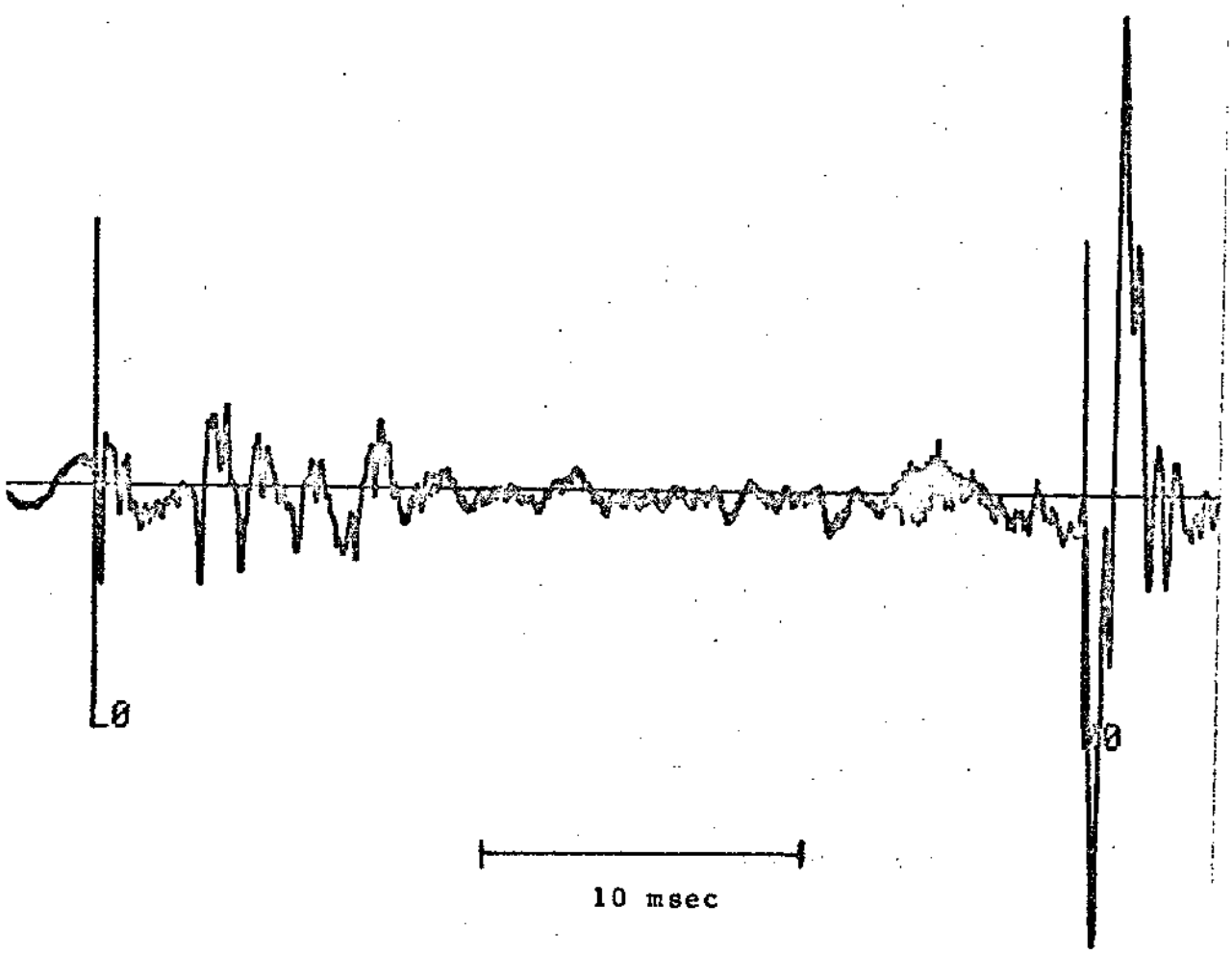

Fig. 2-13a -- Waveform showing token of tama read by JP. The VOT value measured from the cursors is 31.7 msec.

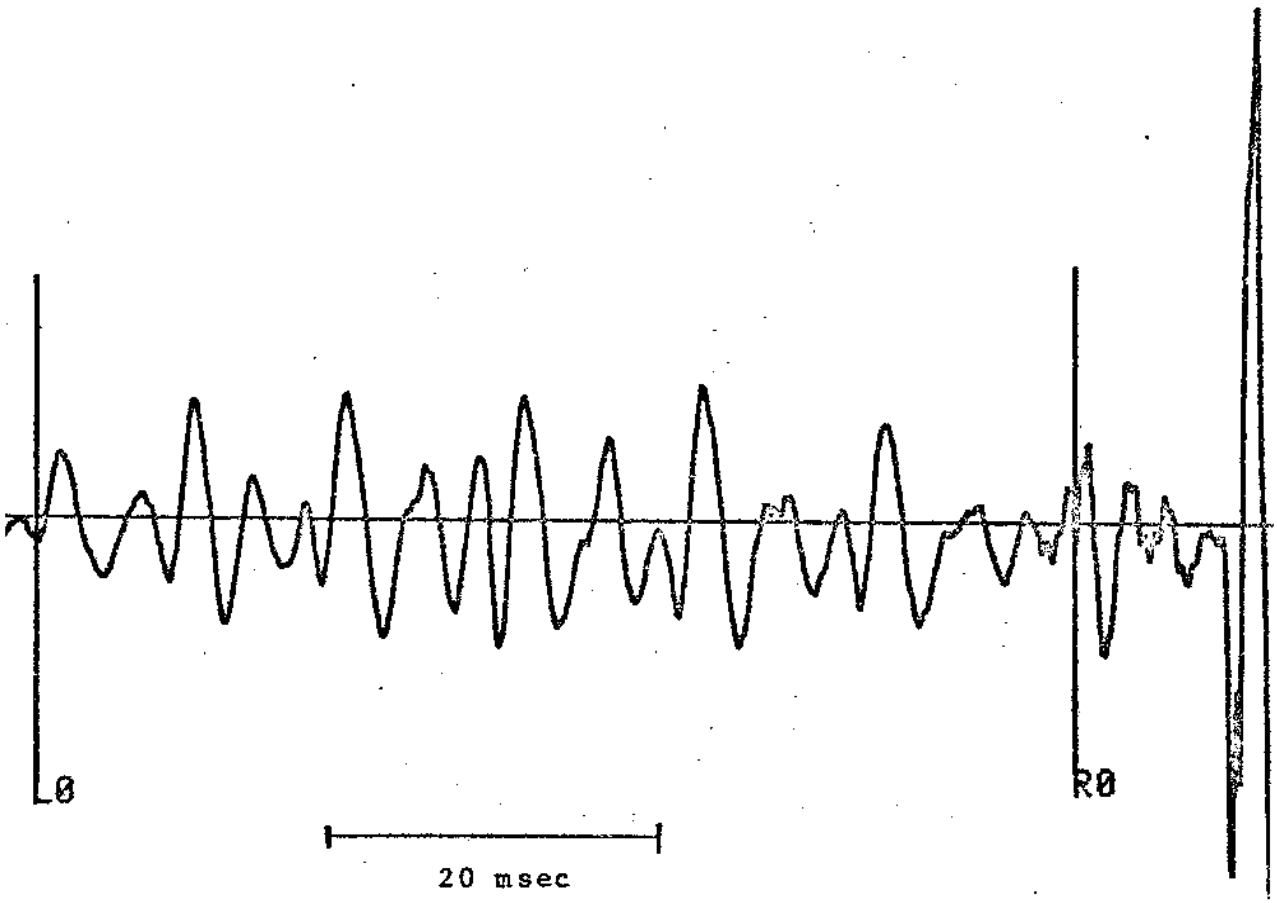

Fig. 2-13b -- Waveform showing token of dama read by JP. The VOT value measured from the cursors is -64.2 msec. Note that the amplitude increase for the following vowel begins immediately after the burst (at RØ). This token, with prevoicing removed, is identified by listeners as [d].

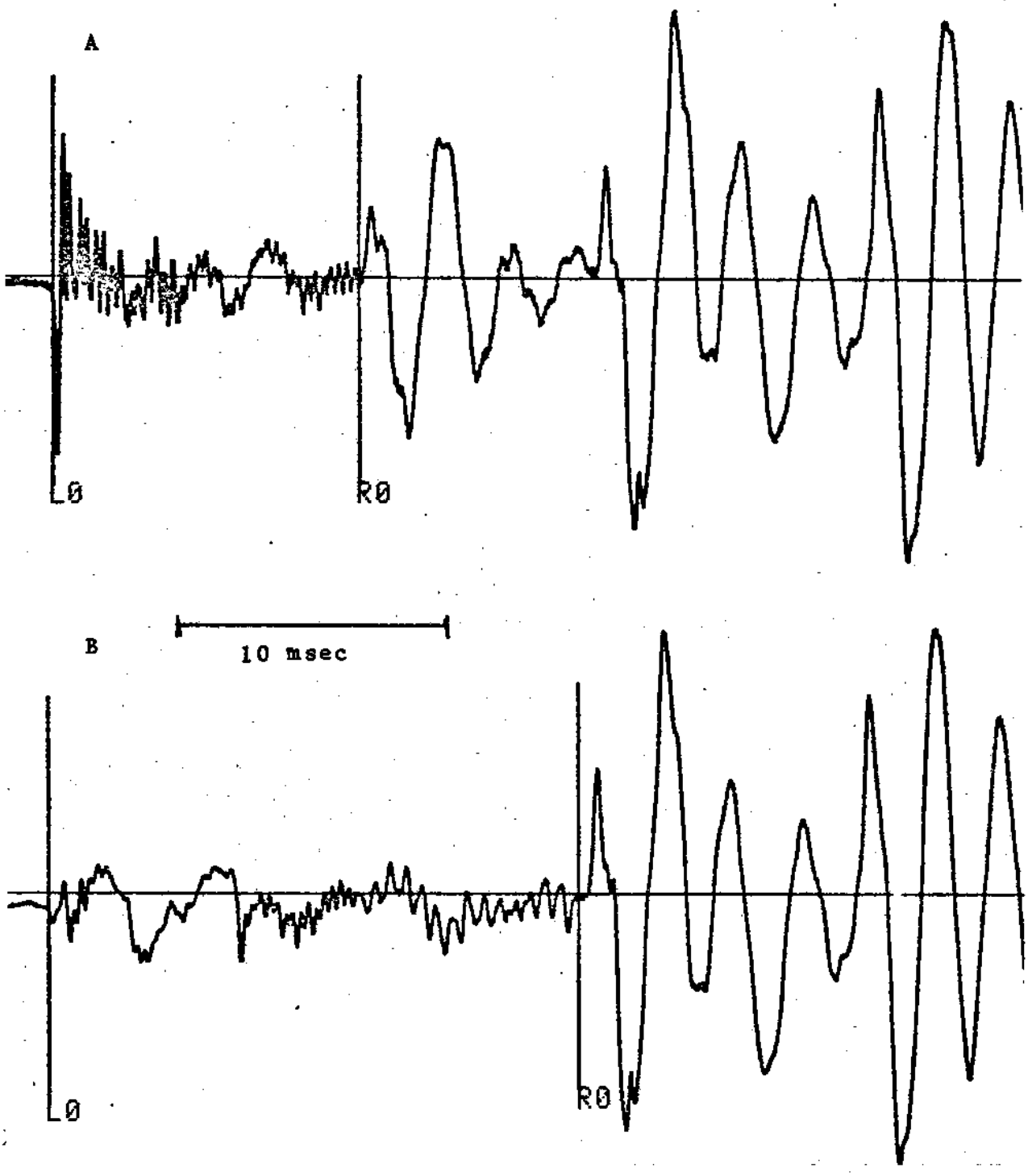

Fig. 2-14a -- Waveform showing natural-edited VOT stimulus with [t]-burst and VOT value of 11.7 msec.

Fig. 2-14b -- Waveform showing natural-edited VOT stimulus with [d]-burst and "VOT" value of 20.2 msec.

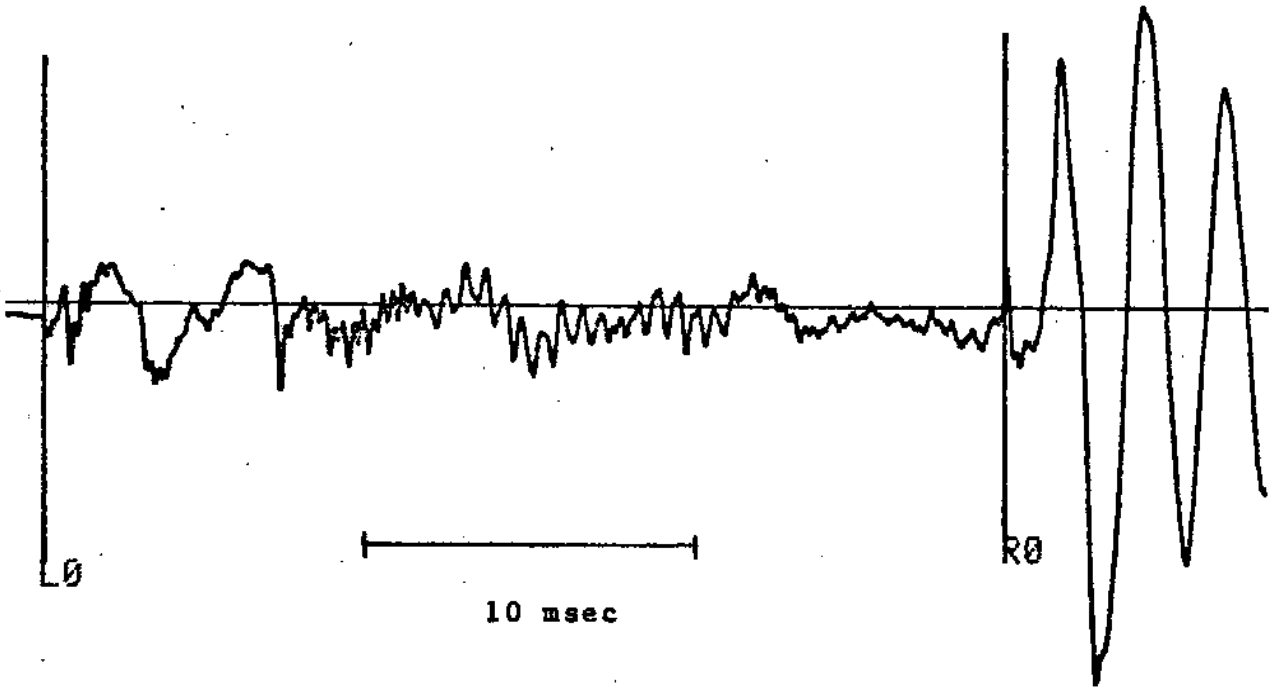

Fig. 2-14c -- Waveform showing natural-edited VOT stimulus with [d]-burst and VOT value of 29.5 msec.

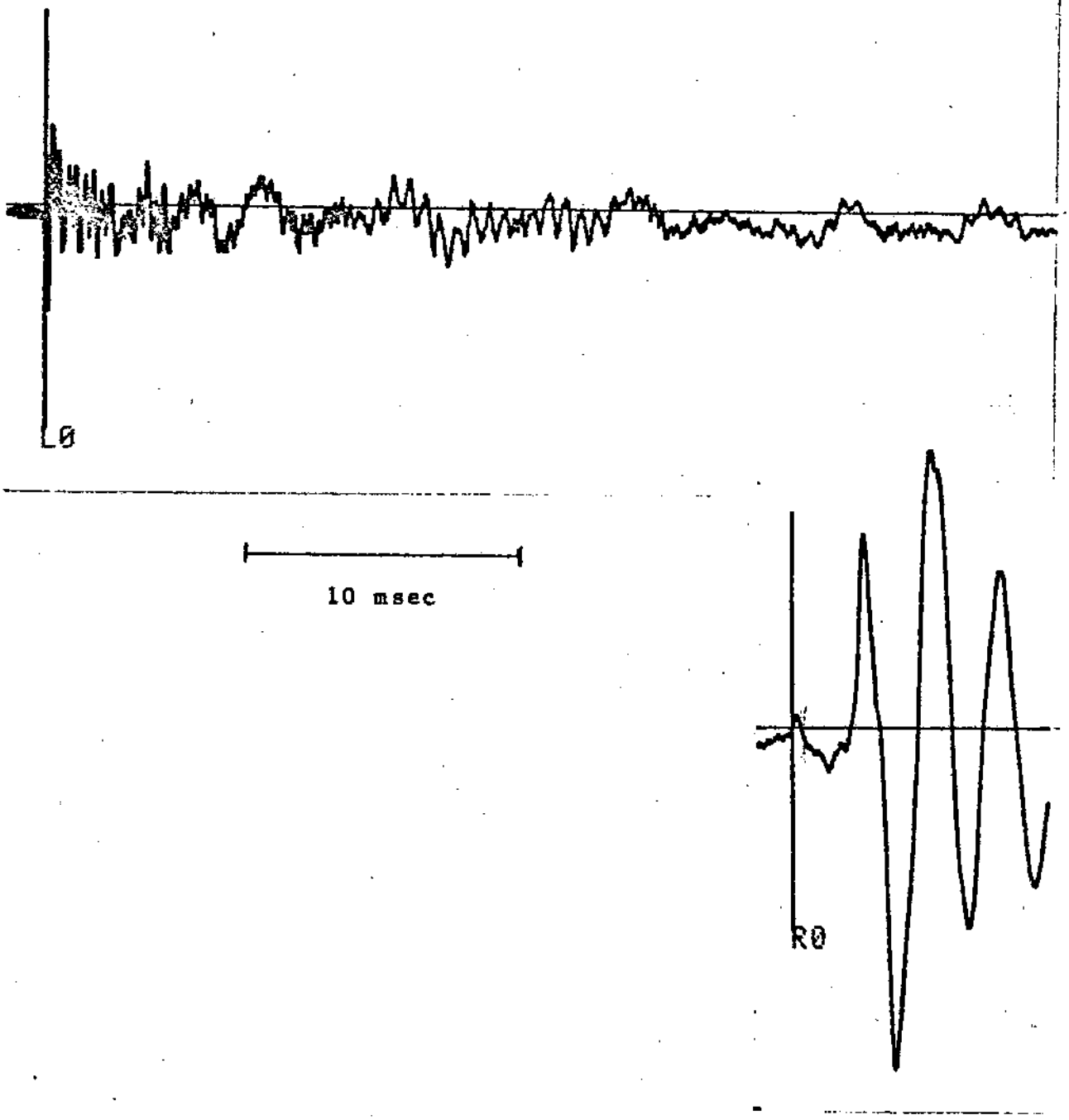

Fig. 2-14d -- Waveform showing natural-edited VOT stimulus with [t]-burst and VOT value of 39.1 msec. This display is continuous across the two lines.

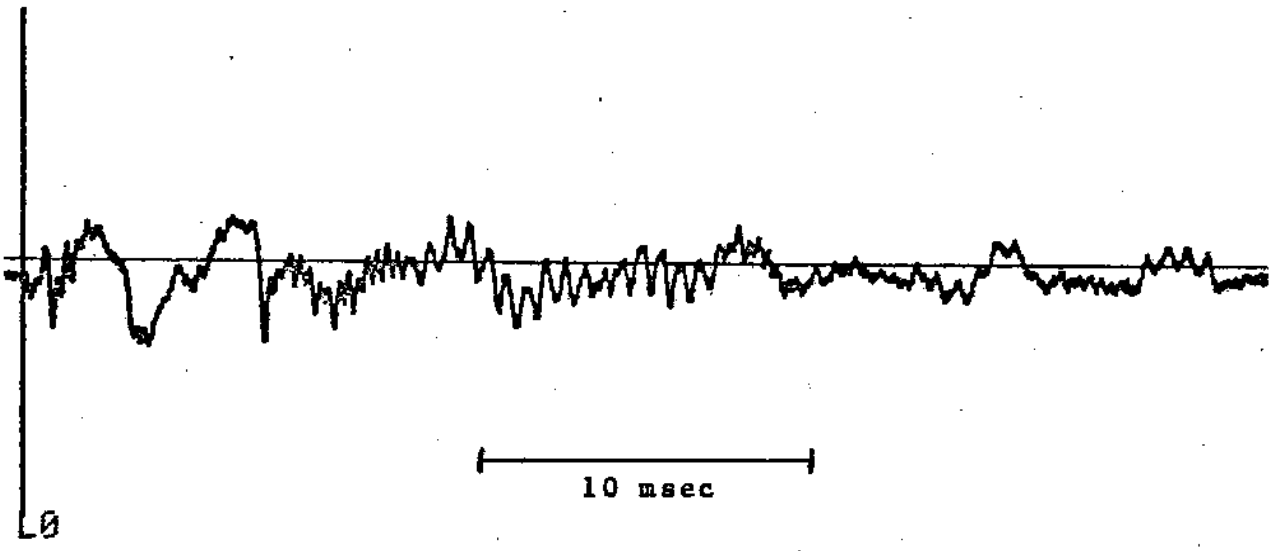

Fig. 2-14e -- Waveform showing natural-edited VOT stimulus with [d]-burst and VOT value of 49.1 msec. The display does not include the entire lag interval.

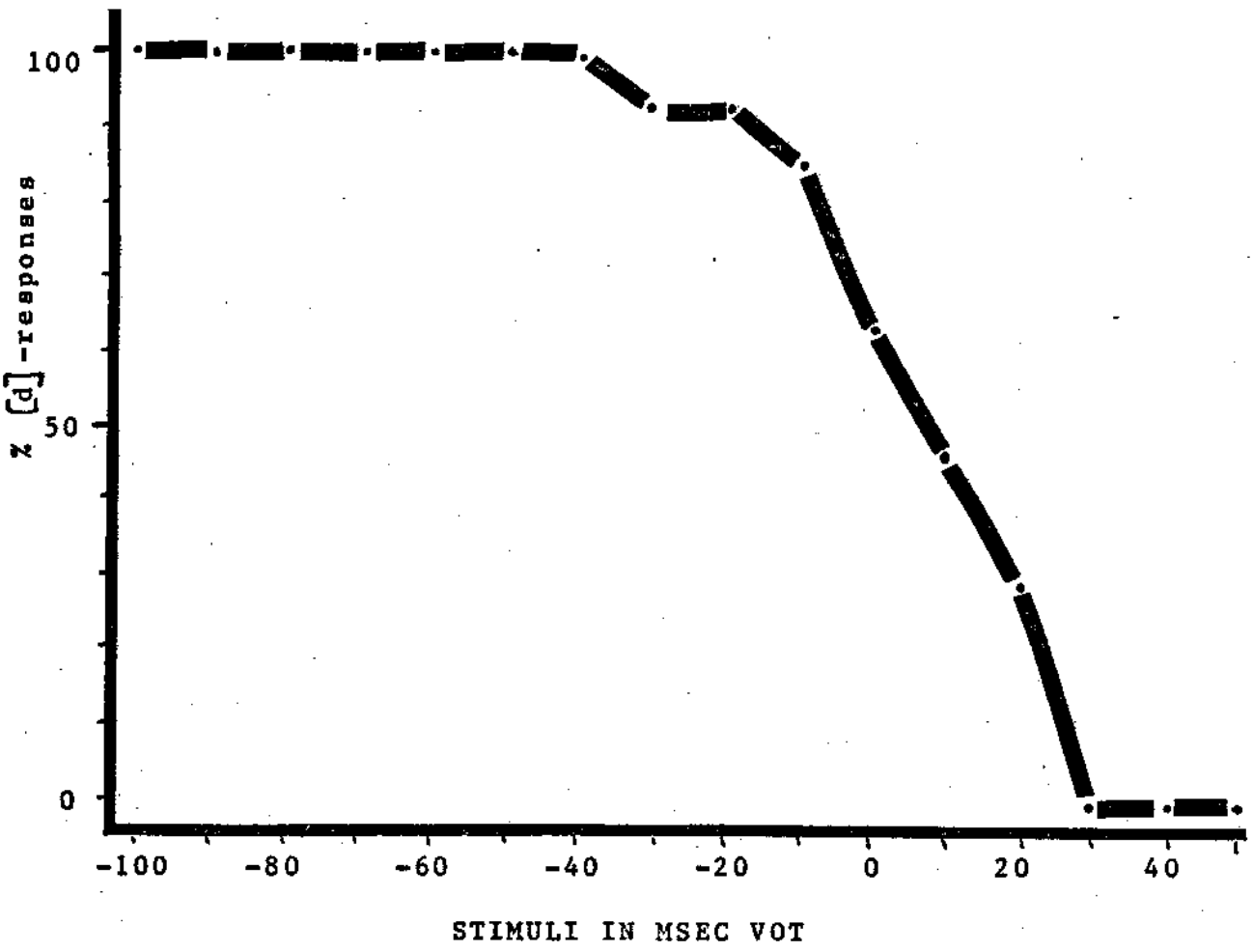

Fig. 2-15 -- Mean identification function for 24 listeners in Wrocław for the synthetic continuum with VOT values from -100 to +50 msec. Collapsing over all listeners causes the function to appear less steep than the individual functions actually are.

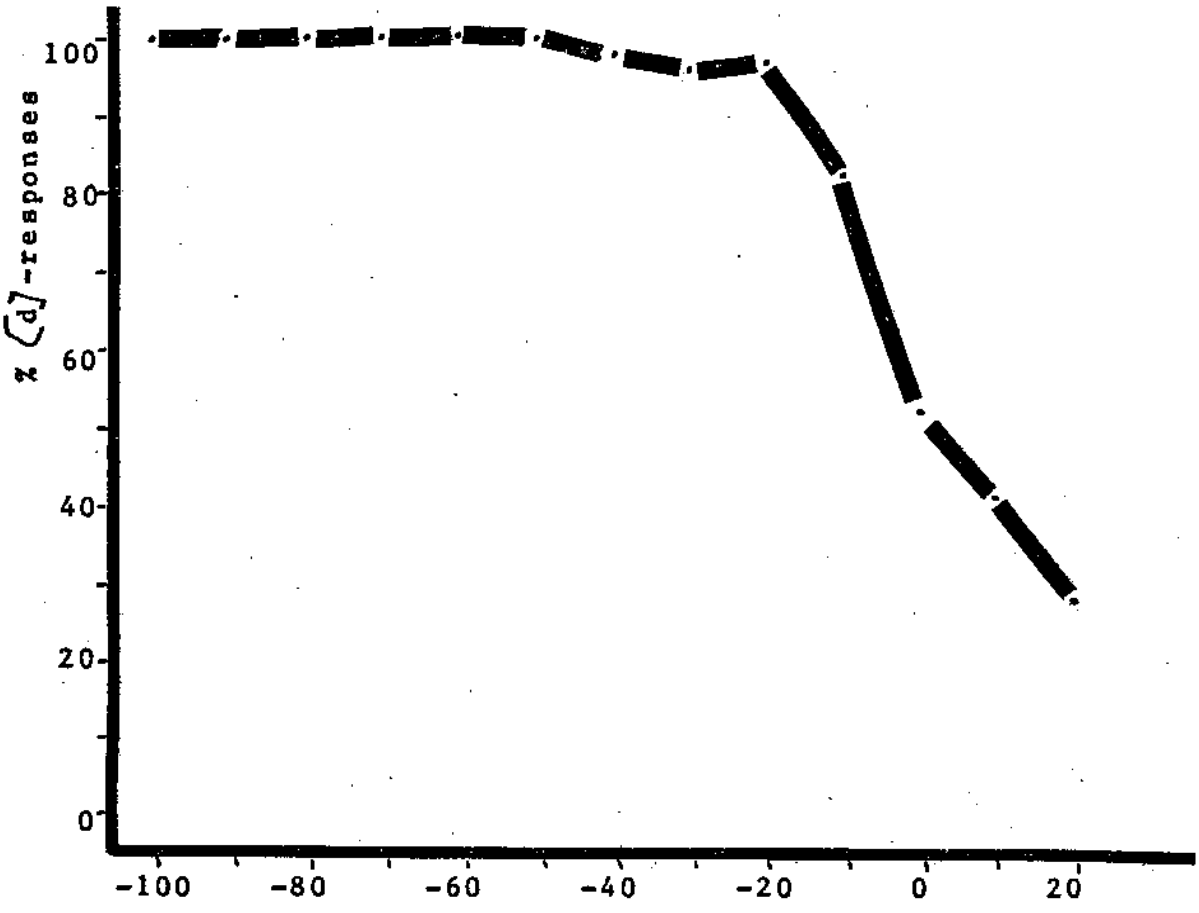

Fig. 2-16 -- Mean identification function for 20 listeners in Wrocław for the synthetic continuum with VOT values from -100 to +20 msec. Four listeners who had three labeling categories were excluded. Collapsing over many listeners causes the function to appear less steep than the individual functions actually are.