International Symposium on Prosody, 18 September 1994, Yokohama, Japan

ACOUSTIC STUDY OF CLEAR SPEECH: A TEST OF THE CONTRASTIVE HYPOTHESIS

John J. Ohala

University of California, Berkeley
E-mail: ohala@cogsci.berkeley.edu

I. INTRODUCTION.

Structural linguistics teaches that one of the principal properties of speech sounds is being different from each other: the essence of a phoneme is that it is not any other phoneme. In addition, many structuralist theories of sound change are based on notions of preservation of contrast between phonemes. From this one might expect that it should be possible to observe speakers' efforts at maintaining contrasts in speech. One situation where this would be expected is in repeated speech, i.e., where a speaker repeats a word after receiving feedback that it has been misapprehended as another similar word. This paper reports results of an analysis of such repeated speech samples in order to evaluate the 'contrastive hypothesis'.

II. METHOD

A. The Speech Sample

All of the data analyzed came from speech samples obtained in a laboratory setting: the recordings were made in the IAC chamber in Department of Linguistics, University of Alberta using studio-quality recording equipment. The speakers were 10 paid volunteers, five males and five females, recruited from among the non-linguistic student population at University of Alberta.

After being ushered into the recording booth and placed before a CRT, a microphone, and a mouse, a subject (S) was instructed that he/she was part of a two-person team in an experiment in which we were evaluating how various channel characteristics affect the intelligibility of speech. When prompted by the CRT S was to speak the word displayed on the CRT. S was told that there was another subject in a room in another wing of the building who would be listening to the speech and attempting the figure out what word had been spoken and indicating their responses by pressing a button which would cause the chosen word or, if they had no idea what the word was, a "?" to appear on the CRT to the right of the prompt word. We would be adding noise and distortion of various sorts to the channel in a way that would be undetectable to S. Thus errors in recognition might occur. If the

word recognized matched the word just spoken, S was to indicate that it was correct by using the mouse to click on a box so labeled on the CRT. But if the listener's response did not match the word just spoken, S would be given a chance to signal with the mouse that the response was an error and to repeat the target word. Only one repetition was allowed. S was instructed to do his/her best to convey the words and was told that there would be a premium pay rate for a high rate of accurate transmission. In fact, there was no other listener; the "responses", both correct and incorrect, that appeared on the CRT to the right of the prompted word (after a suitable delay to simulate human response time) were generated by the same computer that issued the prompts. All subjects received the 'premium' pay.

Twelve phonetically similar words -- words that the speakers might plausibly believe would be highly confusable --were prompted and recorded; see Table 1.

Table 1. Words studied in the experiment.

Orthographic representation	Phonetic transcription
bait	beit
bayed	beid
bed	bɛd
bet	bɛt
bid	bɪd
bit	bɪt
paid	peid
pat	pæt
ped	pɛd
pet	pɛt
pid	pɪd
pit	pɪt

These words were chosen to be relatively close sounding in order to maximize the need for clear speech (when a misperception was cued) and to highlight certain well-documented phonetic features which differentiate them. For example, Voice Onset Time (VOT) differentiates the words starting with /b-/ vs. /p-/; voicing during stop closure and the duration of the preceding vowel differentiates those ending with /-d/ vs. /-t/; duration and formant frequencies differentiate those with different vowels, /ɪ/ vs. /ei/ vs. /ɛ/ vs. /æ/; within this set of vowels, /ei/ is heavily diphthongized, whereas the remainder are typically considered monophthongs (or at least have a much more subtle degree of diphthongization).

The schedule of prompts and pseudo-responses presented by the computer to S had the following characteristics.

1. The order of the words prompted (Table 1) were randomized differently for each subject (this was a quasi-randomization due to the constraint given below in (3)).

2. The total number of prompts were between 445 and 455.

3. The initial twelve prompts consisted of the twelve target words presented once each.

4. All the pseudo-responses to the initial 12 prompted words were correct. These 12 utterances constituted the "original" pronunciation of those words which was compared with the pronunciation given later in the session.

5. After the initial twelve trials, approximately 1/3 of the pseudo-responses were incorrect, i.e., were intentionally made not to match the prompted word. Such incorrect pseudo-responses triggered S's repetition of the target word, presumably with a pronunciation that would be clearer. The interval between such incorrect pseudo-responses varied, randomly, between 0 and 12 trials.

6. The incorrect pseudo-responses were any of the other 11 words from Table 1 or a non-specific 'misperception' which appeared as '?' on the CRT. Each of these 12 misperceptions (the 11 words other than the target word + 1 ('?') were given once each to each of the 12 target words yielding a potential maximum of 12 x 12 = 144 repetitions in response to incorrect pseudo-responses.

7. The inter-trial interval was controlled by the S's rate of clicking his/her indication of correct/incorrect response. In general, however, the whole session took about 50 minutes to an hour.

It should be noted here that many Ss found the task extremely frustrating and tedious. Muttered comments are evident on the tape ('this is not working', 'not again?' (in reaction to an erroneous CRT response), 'thank God' (when the 'end of experiment message appeared), etc.). Nevertheless, there is no indication that Ss failed to perform as expected or that they figured out that there was no human listener at the other end of the channel.

B. Analysis

Using high quality playback equipment, the recordings of the Ss' speech were digitized and analyzed using the CSRE analysis system. The sampling rate was 10 kHz, yielding a usable audio frequency range of just under 5 kHz, which was sufficient to include all acoustic details of interest.

The digitized signals were then displayed on a CRT in two time-aligned formats: waveform and spectrogram. Using cursors, the following points were marked by hand: the onset of the stop burst, the onset of voicing measured with respect to the stop burst (a negative value if voicing preceded the stop release and a positive value if voicing followed), the offset of the vowel. Given these initial settings the following time points, where formant frequencies were to be measured, were done semi-automatically: 25%, 50% and 75% of the full duration of the vowel. In the case of the voiced stops the five measurement points, 0%, 25%, 50%, 75%, and 100% were simply the stop release and the vowel offset and 3 equidistant points in between. In the case of the voiceless stops, the 0% mark was the onset of voicing after the stop release and the remaining points 25%, 50%, 75%, and 100% of the interval up to the offset of the vowel. In other words, the vowel formant frequencies were measured for the fully voiced portions of any given word; in the case of the words starting

with the voiceless stops, this portion started (typically) well after the stop release and was coterminous with voice onset. The advantage of this demarcation of the vowel in the case of those following the voiceless aspirated stops is that it was assured that valid measures of the formants could be obtained whereas it would have been extremely problematic to measure formants during the period of voiceless aspiration. The disadvantage is that this implied some degree of non-comparability between those vowels following the voiceless and voiced stops, i.e., the vowels in *paid* and *bayed* might not be readily compared in spite of their being the same phoneme since the time alignment might be different.

Formant frequencies were measured semi-automatically in that automatic formant tracking (based on LPC spectra) was done. If the operator judged that the formant tracks so derived were reasonable (i.e., that they could not have been better located by the operator), then those values were automatically "logged" into an ASCII text file for further statistical analysis.

The following measures were obtained for each of the words analyzed:

Table 2. Values measured for each utterance analyzed.

- SR: Initial stop release
- VO Moment of voice onset
- Voff Vowel offset (= onset of closure of final stop)
- F1, F2, F3 (formants 1, 2, and 3) at each of 5 measurement points in the vowel (as defined above)

Measures Derived from the above

- Voice Onset Time (VOT) = SR - VO
- Vowel Duration (VD) = Voff - SR

Missing Data

Occasionally, a speaker either failed to utter anything given the CRT prompt, failed to repeat a "misperceived" word, or uttered the wrong word. A few tokens were lost due to recording errors. Finally, a few tokens had acoustic properties which made them unmeasurable; this happened most often in the cases of extracting F2 and F3. In all, 3.1% of the data points were lost for the above reasons. This still left sufficient data from which to extract useful generalizations.

III. RESULTS

A. Duration

In addition to the three conditions of speaking (original, Rep. 1, Rep. 2), vowel duration varies due to the duration intrinsic to a give vowel [1]. High vowels tend to be shorter than low vowels, other things being equal, and "tense" or diphthongized vowels are longer than

monophthongs. In addition, vowels are shorter before final voiceless consonants (/t/ vs. /d/ in this sample). The pre-vocalic consonants in this sample, /b/ and /p/ could affect vowel duration indirectly: the aspiration (positive VOT) of a pre-vocalic /p/ obscures the precise boundary of the vowel onset. Some of the aspiration has clear vowel-like resonances (except for F1) but it is not known how to apportion the duration of the aspiration between that due to the consonantal release and that due to the vowel [4].

Table 3 gives the mean durations (and standard deviation (s.d.)) of the 12 target words in the three conditions, original, repetition 1, and repetition 2. It will be recalled that 'original' is the condition where speakers were unaware of the high rate at which their speech would be misperceived. For this study, this condition constitutes the control or neutral condition. In the 'repetition 1' condition, the speakers had already experienced the high rate (c. 33%) of misperception, although that particular repetition of the word was not given as a reaction to a misperception. 'Repetition 2' was the condition where the word was uttered immediately after a misperception and should constitute therefore the clearest style of pronunciation.

Table 3. Mean durations (and *standard deviations*) of vowels in the 12 test words. The population of these distributions was 10 in the case of the original version but varied from 100 to 119 in Rep. 1 and Rep. 2 (out of a possible maximum of 120, the lesser numbers being due to missing values).

WORD	ORIGINAL	REP. 1	REP. 2
bayed	303.2 (*89.0*)	289.4 (*62.9*)	309.1 (*61.5*)
paid	236.2 (*65.7*)	260.4 (*58.9*)	268.8 (*63.0*)
bed	230.2 (*63.7*)	227.8 (*57.4*)	227.6 (*51.2*)
bid	186.1 (*49.6*)	183.0 (*40.0*)	191.8 (*43.2*)
ped	176.7 (*41.8*)	183.4 (*85.6*)	178.1 (*29.1*)
bait	176.0 (*21.0*)	178.1 (*26.1*)	184.5 (*25.7*)
pat	172.7 (*19.5*)	171.2 (*28.7*)	180.9 (*29.1*)
pid	148.5 (*32.1*)	153.0 (*30.9*)	158.1 (*30.9*)
bet	143.6 (*13.9*)	144.2 (*22.3*)	149.9 (*22.6*)
bit	131.7 (*27.1*)	123.4 (*19.0*)	129.3 (*19.9*)
pet	115.8 (*18.2*)	123.0 (*18.2*)	129.5 (*24.1*)
pit	96.78 (*24.9*)	106.5 (*17.3*)	110.8 (*18.3*)

From this data it can be seen that:

(a) the range of durations increases from the Orig. to Rep. 1 and Rep. 2, however, this

doesn't necessarily mean that the s.d. increases. In fact the s.d. increases from the orig. to Rep. 2 in only 4 cases, and from Rep. 1 to Rep. 2 in 7 cases and the increase is often quite small. When shifting to clearer speech the tolerance of vowel durations may become smaller.

(b) The mean duration increases in the majority of cases from Orig. to Rep. 1 and Rep. 1 to Rep. 2. Specifically, there is an increase in duration in 7 of the 12 cases from Orig. to Rep. 1 and 10 cases in Orig. to Rep. 2 and in Rep. 1 to Rep. 2. In many cases, however, the magnitude of the duration increase is quite small.

In sum, there is a tendency for clear speech to have longer syllable durations than less clear speech and for the "tails" of the distribution to be long although there is not a strong tendency for the variability to increase.

B. Contrastive Duration

The above analysis does not address the question raised in the introduction of whether speakers alter duration of a word in a way to make it contrast with the word in the pseudo-response. That is, if the target word *bayed*, with the longest intrinsic duration, received the pseudo-response *pit*, with the shortest intrinsic duration, would the speaker, when repeating the target (in Rep. 2), make the duration of *bayed* even longer than normal? More generally, for all pairs of words W_i, W_j, would it be the case that the duration of W_i given the pseudo-response W_j (symbolized $W_i|W_j$), be altered -- increased or decreased -- in proportion to the normal difference between W_i and W_j. Using $W_i|0$ to symbolize the original duration of a word, i.e., W_i given '0' response, we formulate the hypothesis as in (1).

(1)
$$(W_i|W_j) - (W_i|0) \propto (W_i|0) - (W_j|0)$$

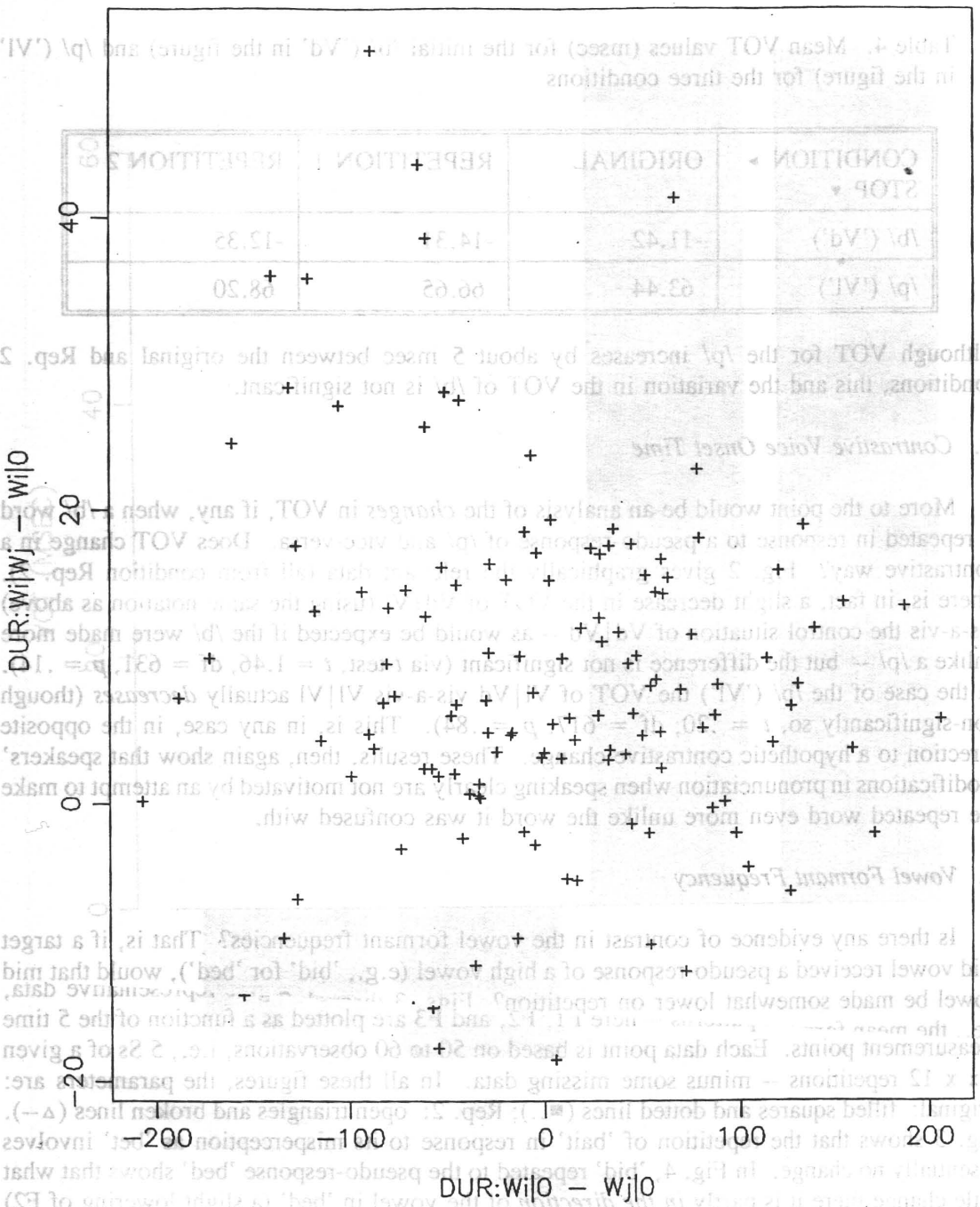
We might relax the 'proportional to' part of the hypothesis and merely expect that the relation would be monotonic. The test of this hypothesis is given in Fig. 1, where it is the mean durations pooled across all 10 subjects which were used for any given W_i duration.

Obviously, the hypothesis is completely unsupported by these data. There is no significant correlation between the two parameters. Apparently speakers do not vary duration in a way to contrast it or make it more different from the original. Assuming it is replicated, this is an important finding which has considerable theoretical significance. It suggests that, once established, the pronunciation norms of a language which exploit a number of distinctive features, among them duration, are not so elastic -- at least the norms are not relaxed in order to make speech clearer. This could be good news for the task of automatic speech recognition.

C. Voice Onset Time

English uses two voice onset times (VOT) to differentiate words, one which has a positive VOT ($\approx$ 60 msec) for /p/ and another which has a value closer to zero for /b/ [2]. The latter may sometimes show negative VOT, i.e., voicing starting before the stop release. VOT

Fig. 1 Test of Contrastive Vowel Duration



is known to vary with the degree of stress [3]. Since, as we might imagine, clearer speech tends to be more stressed, we should see some increase in the degree of separation of the VOT for the /p/ and /b/ in the present data. Table 4 gives the values of VOT for the /b/ ('Vd') and /p/ ('Vl') in the three conditions.

Table 4. Mean VOT values (msec) for the initial /b/ ('Vd' in the figure) and /p/ ('Vl' in the figure) for the three conditions

CONDITION ▸ STOP ▾	ORIGINAL	REPETITION 1	REPETITION 2
/b/ ('Vd')	-11.42	-14.31	-12.35
/p/ ('Vl')	63.44	66.65	68.20

Although VOT for the /p/ increases by about 5 msec between the original and Rep. 2 conditions, this and the variation in the VOT of /b/ is not significant.

D. Contrastive Voice Onset Time

More to the point would be an analysis of the *changes* in VOT, if any, when a /b/ word is repeated in response to a pseudo-response of /p/ and vice-versa. Does VOT change in a contrastive way? Fig. 2 gives graphically the relevant data (all from condition Rep. 2). There is, in fact, a slight decrease in the VOT of Vd|Vl (using the same notation as above) vis-a-vis the control situation of Vd|Vd -- as would be expected if the /b/ were made more unlike a /p/ -- but the difference is not significant (via t-test, $t = 1.46$, df $= 631$, $p = .14$). In the case of the /p/ ('Vl') the VOT of Vl|Vd vis-a-vis Vl|Vl actually *decreases* (though non-significantly so, $t = .20$, df $= 617$, $p = .84$). This is, in any case, in the opposite direction to a hypothetic contrastive change. These results, then, again show that speakers' modifications in pronunciation when speaking clearly are not motivated by an attempt to make the repeated word even more unlike the word it was confused with.

E. Vowel Formant Frequency

Is there any evidence of contrast in the vowel formant frequencies? That is, if a target mid vowel received a pseudo-response of a high vowel (e.g., 'bid' for 'bed'), would that mid vowel be made somewhat lower on repetition? Figs. 3 through 7 give representative data, i.e., the mean formant patterns -- here F1, F2, and F3 are plotted as a function of the 5 time measurement points. Each data point is based on 50 to 60 observations, i.e., 5 Ss of a given sex x 12 repetitions -- minus some missing data. In all these figures, the parameters are: original: filled squares and dotted lines (■..); Rep. 2: open triangles and broken lines (▵--). Fig. 3 shows that the repetition of 'bait' in response to its misperception as 'bet' involves essentially no change. In Fig. 4, 'bid' repeated to the pseudo-response 'bed' shows that what little change there it is partly *in the direction* of the vowel in 'bed' (a slight lowering of F2) although there is also a slight lowering of F1 (which is *away* from the vowel in 'bed'. In Fig. 5, 'bed' repeated to the pseudo-response 'bid' shows that what little change there is is

Fig. 2 Contrastive Voice Onset Time (Rep. 2)

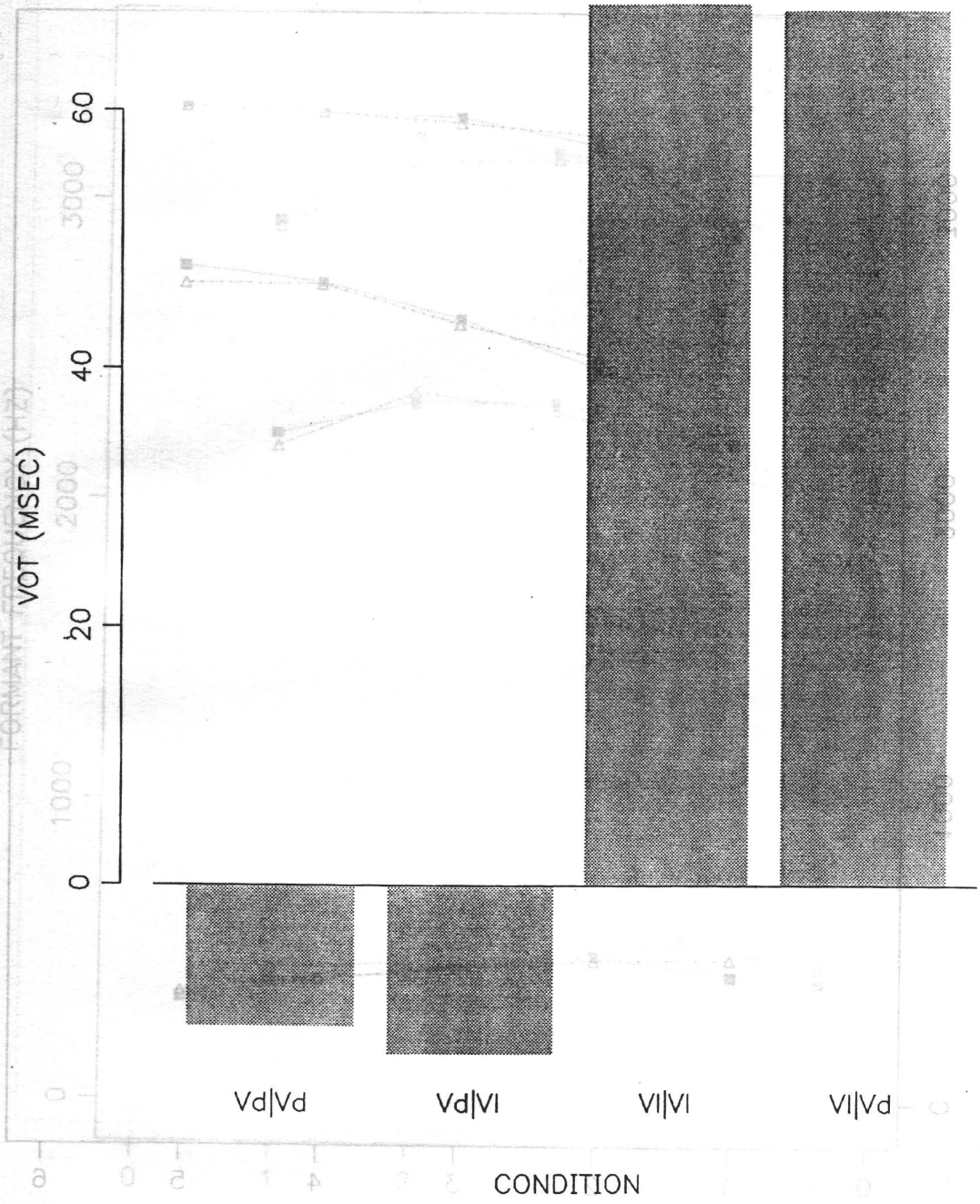


Fig. 3

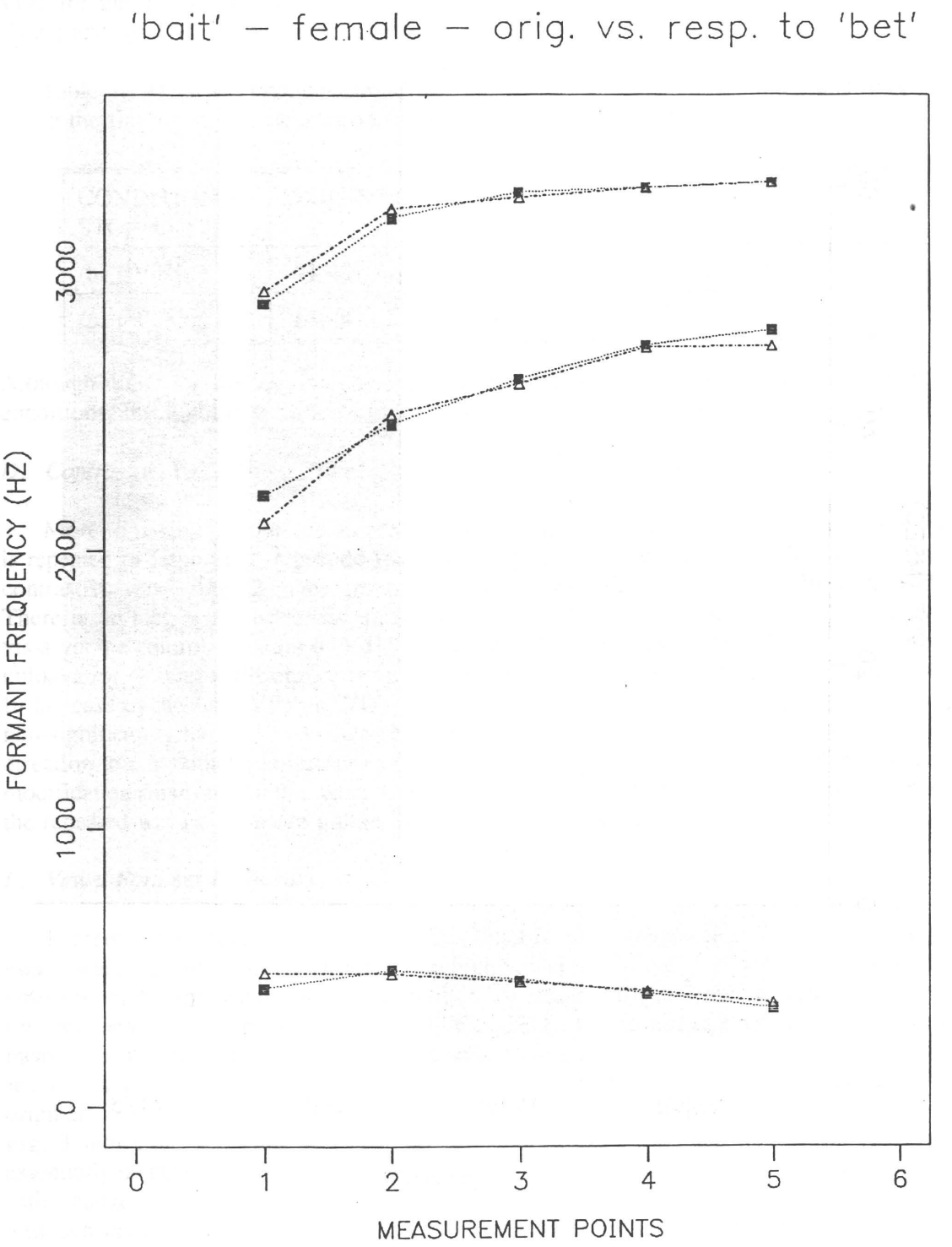

'bait' – female – orig. vs. resp. to 'bet'
FORMANT FREQUENCY (HZ)
0
1000
2000
3000
0
1
2
3
4
5
6
MEASUREMENT POINTS

Fig. 4

'bid' – female – orig. vs. resp. to 'bed'

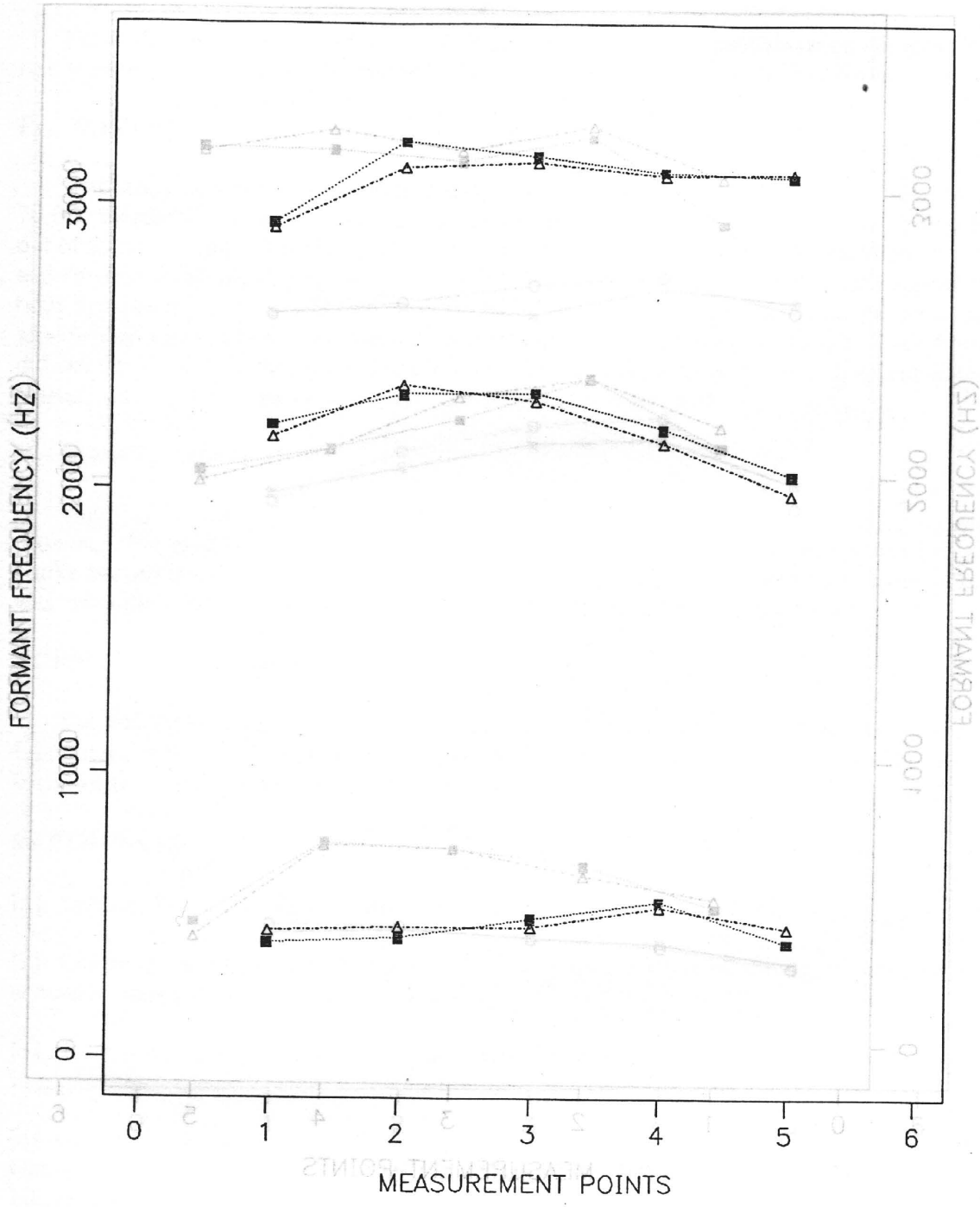


Fig. 5

'bed' – female – orig. vs. resp. to 'bid'

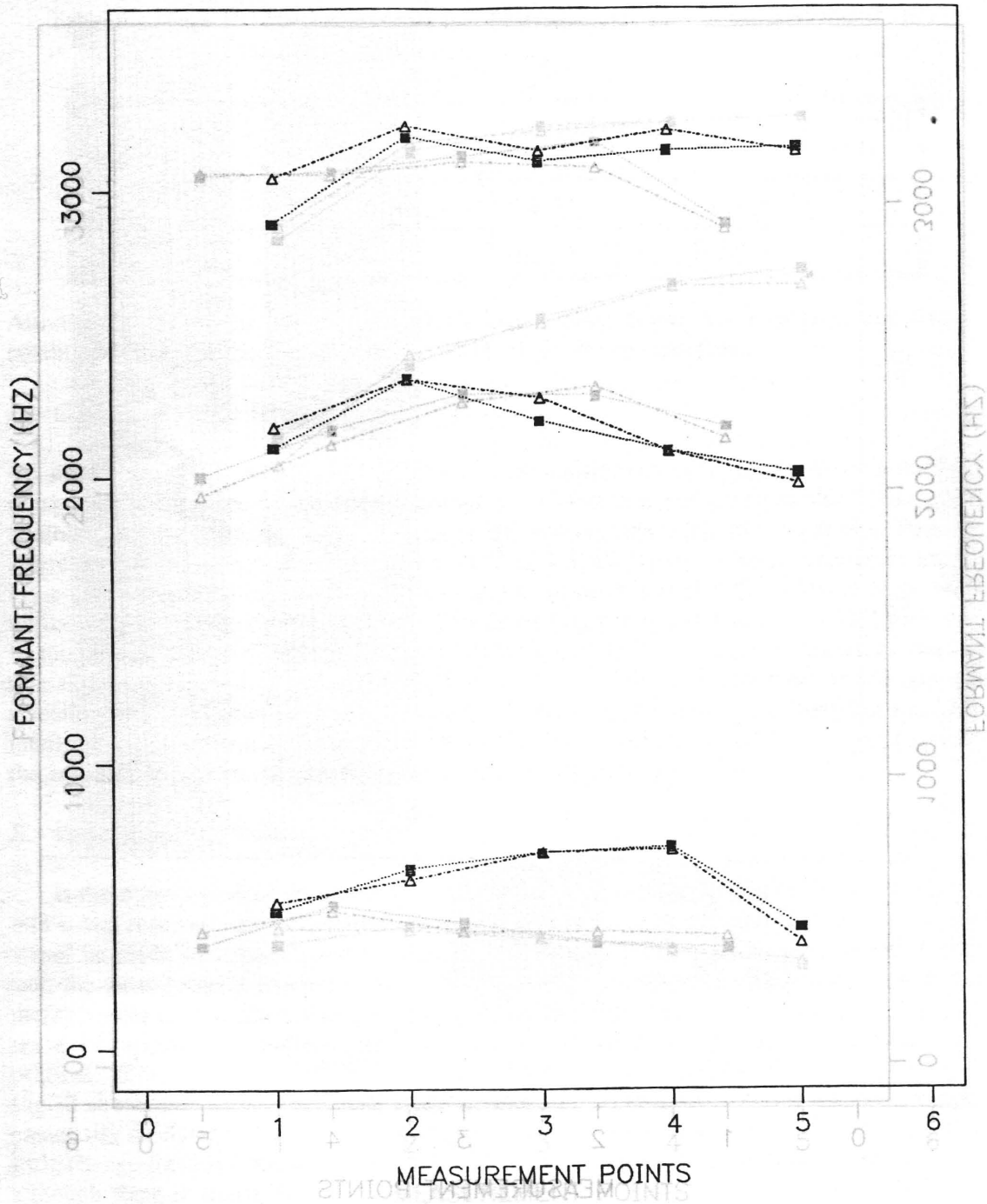


Fig. 6

'paid' — male — orig. vs. resp. to '?'

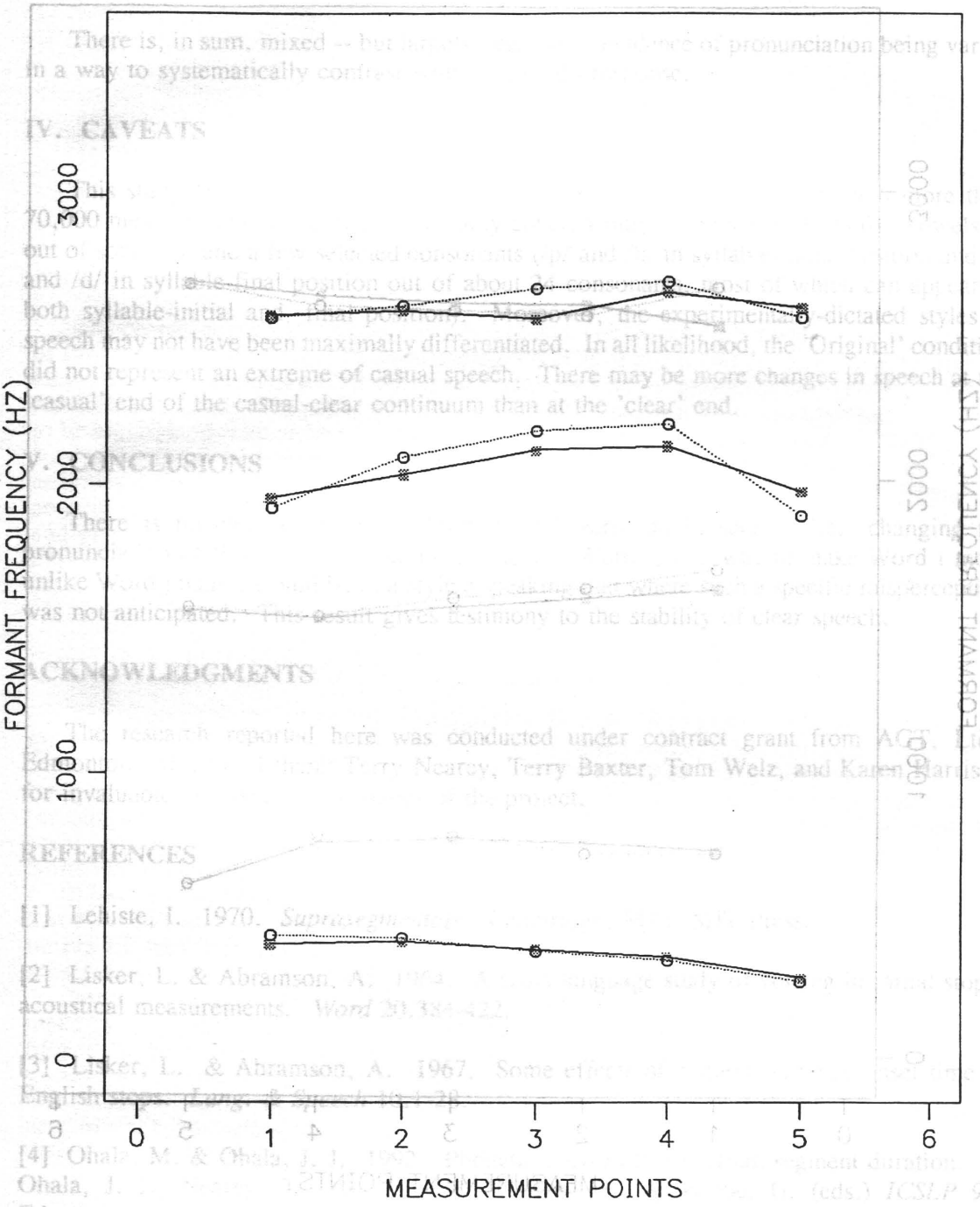
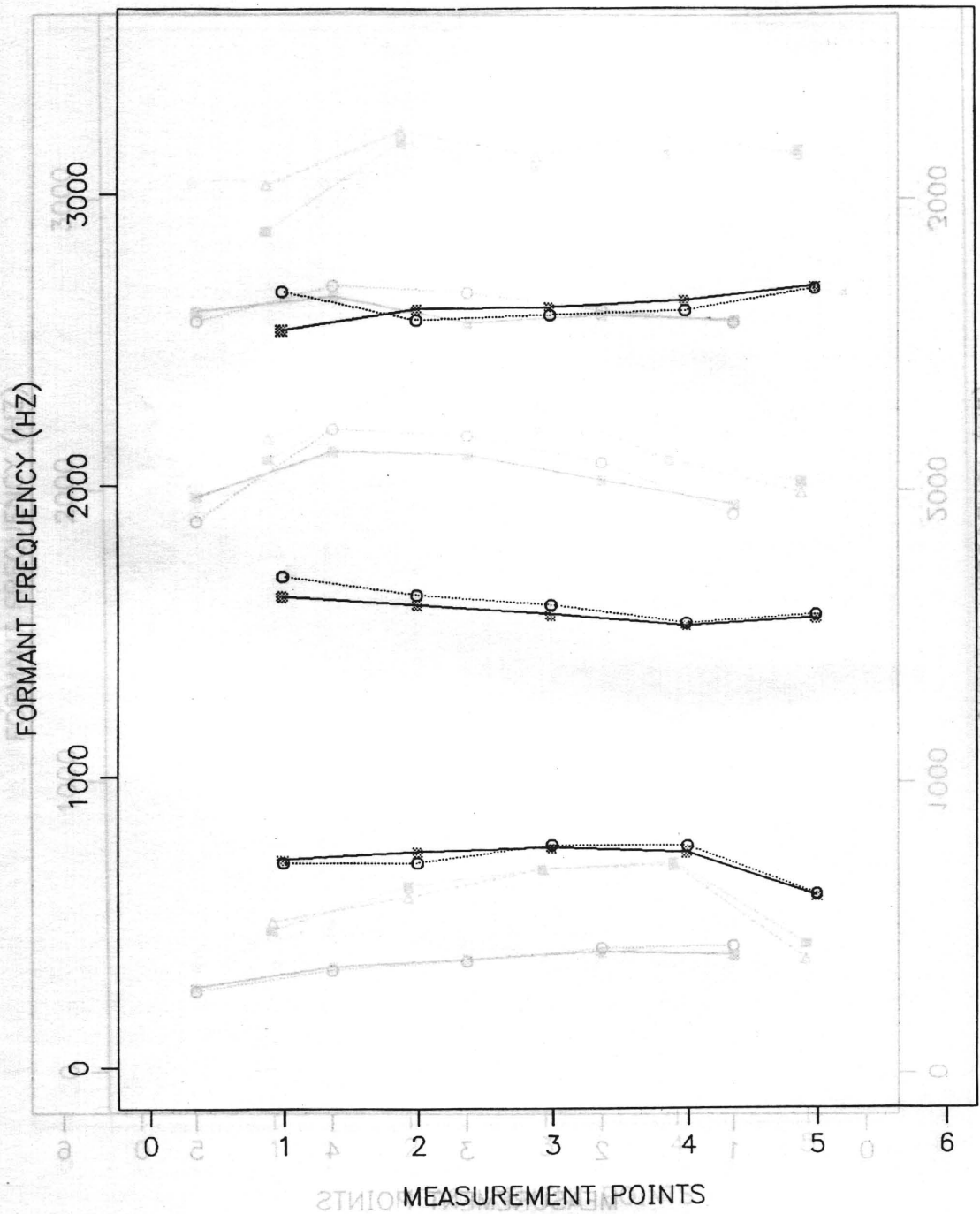
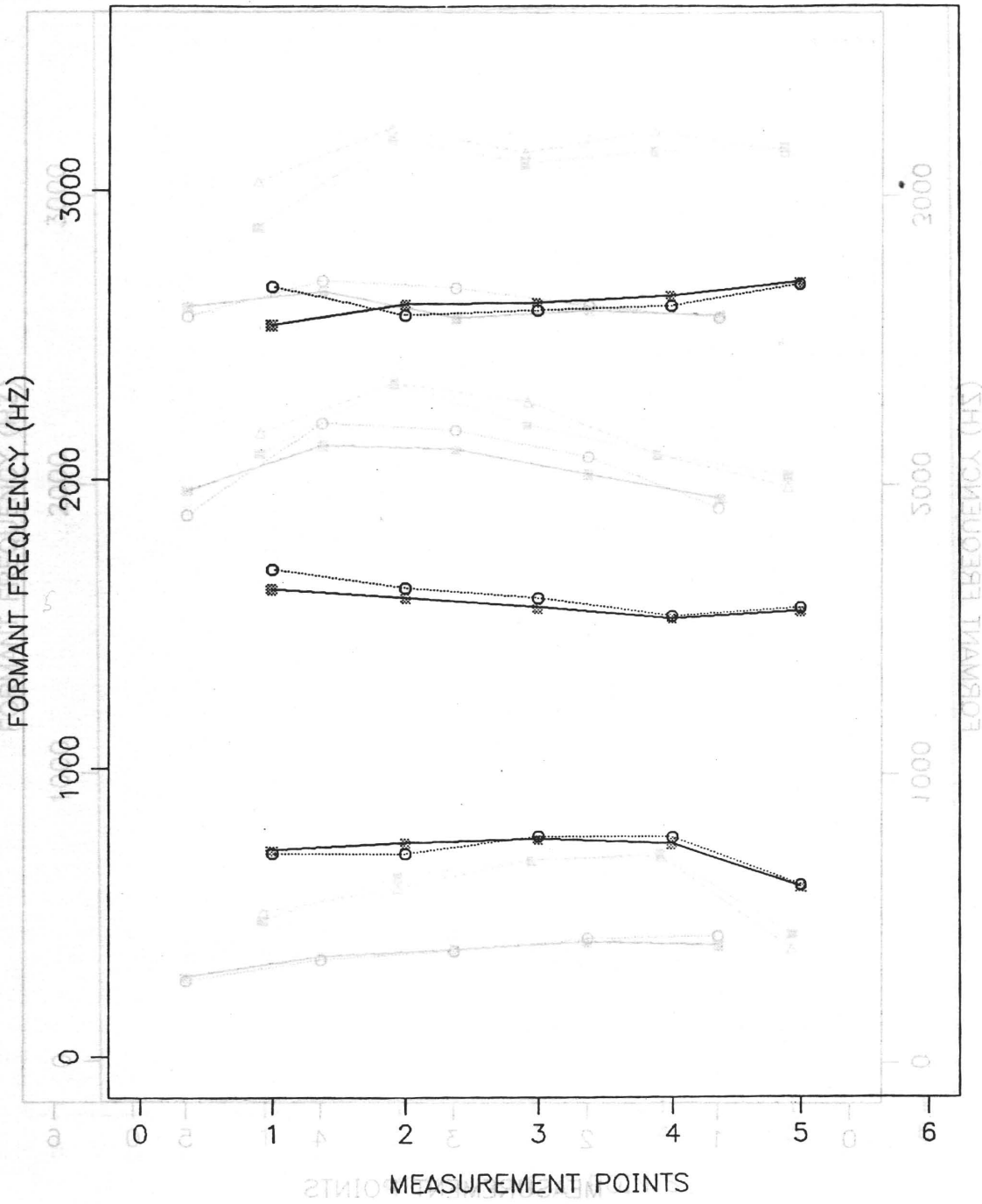


Fig. 7

'pat' – male – orig. vs. resp. to 'pet'



'pat' — male — orig. vs. resp. to 'pet'



partly *in the direction* of the vowel in 'bid' (a slight rise in F2). In Fig. 6, 'paid' in response to the non-specific misperception cued by '?' shows that F2 (and F3) goes a bit higher on repetition -- this would imply that its diphthongal character is being emphasized. In Fig. 7, 'pat' repeated in response to 'pet' shows a slight increase in F2 and this is *in the direction* of the vowel in 'pet'.

There is, in sum, mixed -- but largely negative -- evidence of pronunciation being varied in a way to systematically contrast with the pseudo-response.

IV. CAVEATS

This study had many limitations dictated by time and resources. Although more than 70,000 measurements were made, the study covered only a few selected English vowels (4 out of some 16) and a few selected consonants (/p/ and /b/ in syllable-initial position and /t/ and /d/ in syllable-final position out of about 24 consonants, most of which can appear in both syllable-initial and -final position). Moreover, the experimentally-dictated styles of speech may not have been maximally differentiated. In all likelihood, the 'Original' condition did not represent an extreme of casual speech. There may be more changes in speech at the 'casual' end of the casual-clear continuum than at the 'clear' end.

V. CONCLUSIONS

There is no clear evidence of "contrastive" variation in speech, i.e., changing the pronunciation of Word i, given the misperception Word j, in a way to make Word i more unlike Word j than it would be in a style a speaking that where such a specific misperception was not anticipated. This result gives testimony to the stability of clear speech.

ACKNOWLEDGMENTS

The research reported here was conducted under contract grant from AGT, Ltd., Edmonton, Alberta. I thank Terry Nearey, Terry Baxter, Tom Welz, and Karen Harrison for invaluable assistance in all stages of the project.

REFERENCES

[1] Lehiste, I. 1970. *Suprasegmentals*. Cambridge, MA: MIT Press.

[2] Lisker, L. & Abramson, A. 1964. A cross-language study of voicing in initial stops: acoustical measurements. *Word* 20.384-422.

[3] Lisker, L. & Abramson, A. 1967. Some effects of context on voice onset time in English stops. *Lang. & Speech* 10.1-28.

[4] Ohala, M. & Ohala, J. J. 1992. Phonetic universals and Hindi segment duration. In Ohala, J. J., Nearey, T., Derwing, B., Hodge, M., & Wiebe, G. (eds.) *ICSLP 92*. Edmonton: University of Alberta. 831-834.