A PERCEPTUAL STUDY OF REDUCED VOWELS IN CLEAR AND CASUAL SPEECH

Moon S-J[1], Lindblom B[2] and Lame G[3]
[1]Department of English, College of Liberal Arts, AJOU University, Suwon 442-749, Korea
[2]Department of Linguistics, Stockholm University, Stockholm S-10691, Sweden
[3]Eloquent Technology Inc, 24 Highgate Circle, Ithaca, NY 14580, USA

ABSTRACT

Data are presented on the perception of vowel formant patterns in [w_l] syllables. Perceptual judgements depended on extent of formant transitions, vowel duration and speaking style indicating that listeners do expect undershoot in syllables of this type and that they expect less of it in clear than casual speech. Recognition scores were highest for the most extensive formant movements, particularly in the clear speech condition.

PROBLEM

In a recent study, English vowels in [w_l] syllables were found to be longer, have faster formant transitions and show more peripheral formant patterns in clear speech than in casual style[1]. These changes had the effect of reducing context and duration dependent "formant undershoot" thus shifting formant values closer to ideal "context-free" values. Such decrease in context-dependence makes intuitive sense perceptually, since the phenomena of undershoot tend to reduce intervocalic contrast and therefore create potential difficulties for the listener. However, firm conclusions can only be drawn given data from native listeners. The following experiment was done to shed some light on the perceptual function of the observed formant variations.

EXPERIMENTAL PROCEDURES

The English vowels /i/, /ɪ/ and /ɛ/ were synthesized and embedded in one casually and one clearly spoken version of "Wheelingham", a possible place name. The reason for choosing that context was that, previously[1], undershoot effects had been found to be particularly marked in trisyllabic words. This hybrid synthesis was implemented using KLSYN88 and other software on one of the Vaxstations in the University of Texas Phonetics Laboratory.

The vowel formant patterns (F1, F2, F3) of the synthetic stimulus portions were derived from stylized values based on average data of five speakers[1] plotted in Figure 1, comprising formant values sampled at the frequency maximum of F2 in the [w_l] context. A relatively higher position on the chart implies a larger upward movement of F2 from its low starting point in [w], and a correspondingly larger downward shift back to the low F2 of [l]. Similarly, a relatively higher F1 value means a bigger frequency excursion relative to its location in [w] and [l]. The F3 values are not shown, but were varied according to a similar, but more limited, excursion pattern.

The following aspects were also defined in terms of averages calculated from the earlier speech sample[1] and were obtained for both styles by pooling across tokens, vowels and speakers: timing of formant frequency maxima, contours of F0 and overall amplitude.

To generate the transitions to and from the vowel pattern, smooth cosine functions were used. Great care was taken to ensure continuity at segment

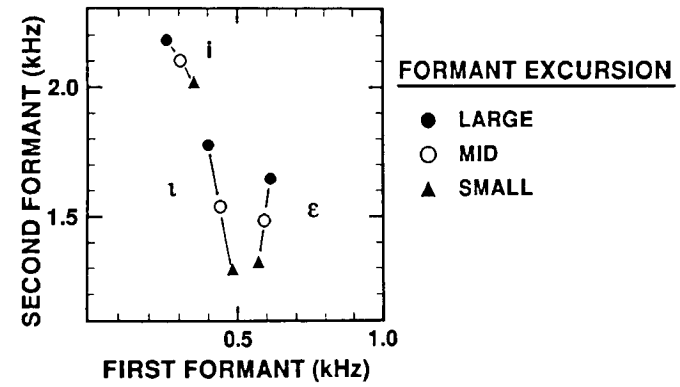

Figure 1. Formant values at approximately midpoints of synthetic vowels. Smooth cosine functions were used as transitions between these patterns and the adjacent [w] and [l] "loci". The entire vowel segments were spliced into a clear and a casual variant of the word ['wɪlɪŋhæm].

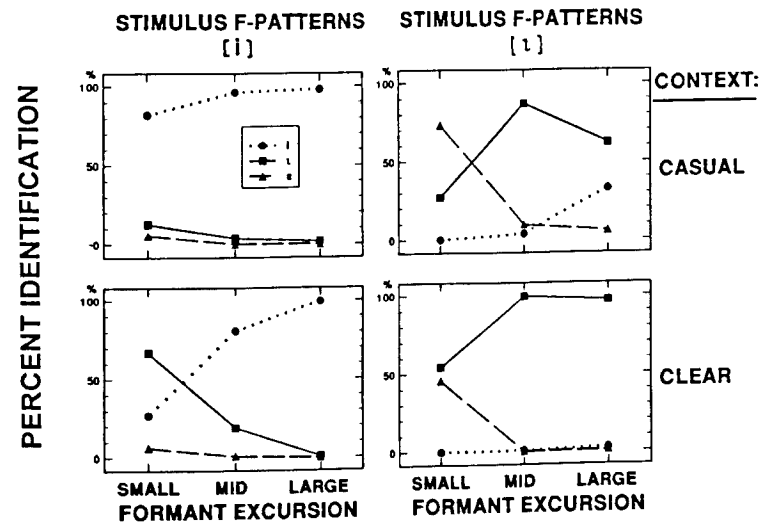

Figure 2. Identification results for stimuli derived from /i/ measurements in the left column and for /ɪ/ in the right. The vertical axis percent identification. Along the x-axes extent of formant transitions. Solid dots, squares and triangles refer to /i/, /ɪ/ and /ɛ/ responses respectively.

boundaries in all parameters including overall amplitude, F0 and voice source characteristics (open quotient and spectral tilt). The synthesized vowel was spliced into the natural speech using cosine tapering at the edges.

A total of 72 stimuli were produced combining 2 contexts (clear and casual), 3 vowels (/i/, /ɪ/, /ɛ/), 3 degrees of formant excursion (or undershoot) and 4 vowel durations. Each vowel had all combinations of three central formant patterns (small, medium and large undershoot) and four durations, 136, 177, 212 and 266 msec chosen to reflect the observed effects of the clear-casual and the open-close conditions.

Five listeners participated. They were graduate students selected as a homogeneous group of American English speakers without marked dialectal features. Each subject participated in three experimental sessions in which they were asked to idetify the stimulus as "Wheelingham", "Willingham" or "Wellingham". The test tapes contained a total of 576 stimuli distributed over the three sessions in randomized sets of 216+216+144 stimuli. There were eight repetitions of every stimulus which means that forty responses per stimulus were collected. The percentages to be reported were calculated with n=40.

RESULTS

Our findings indicate that listener judgements varied as a function of all the variables investigated. Vowel duration, context (casual or clear style of the natural frame), extent of formant transitions and the intended vowel category were all seen to have an effect on the responses.

If "correct identification" is defined in terms of the vowel categories from which the stimuli were derived, we find that all patterns intended as /ɛ/ were identified at a level near 100%. On the other hand, errors were numerous in the labeling of the /i/ and /ɪ/ stimuli. Figure 2 presents the identification results for /i/ in the left column and for /ɪ/ in the right. The vertical axis in all four panels is percent identification. Along the x-axes extent of formants is shown. Here "small", "mid" and "large" excursion corresponds to the triangle, open circle and filled circle respectively of Figure 1. A small formant movement implies a large "undershoot" effect, a large movement means small "undershoot". Data points pertain to percent identification averaged across all four durations. Solid dots, squares and triangles refer to /i/, /ɪ/ and /ɛ/ responses respectively. The two panels of the upper row show how the stimulus F-pattern was classified in the context of the casual frame. Those of the bottom row give the results for the same stimulus pattern in the clear frame.

First note that, for large formant excursions, the recognition score for /i/ is high in both styles. For /ɪ/, identification is near perfect in clear speech but less accurate in the casual context. Note that comparing the upper and lower panels we examine the effect of the style of the frame. Evidently, since for any given vertical position, the top and bottom panels refer to identical synthetic vowel segments, the non-identity of the panels indicates that context has an effect on the responses.

As for small formant excursions, we note that /i/ in casual style is identified resonably well, but in clear speech its score is drastically reduced. Instead /ɪ/ responses are favored. Figure 1 suggest how this effect could be accounted for. It implies that, in the clear context, the small excursion for /i/ (=large undershoot) is interpreted as an instance of /ɪ/ with a large formant movement (=small undershoot). Note the proximity

of the /i/-triangle and the solid /ɪ/-dot of Figure 1. Thus listeners did not expect /i/ to exhibit that much undershoot in clear speech.

That interpretation also clarifies the results for /ɪ/ in the casual context where /i/ responses are seen to increase at the expense of /ɪ/. That the large formant excursions of /ɪ/ get associated with /i/ to some extent would seem to suggest that listeners did not expect an /ɪ/ to show that little undershoot in the casual context.

Looking at the /ɪ/ with a small excursion we find that it is identified as an /ɛ/, especially in the casual context. Figure 1 shows that the /ɪ/ and /ɛ/ stimuli with small excursions are very similar. F1 of this /ɪ/ variant is high.

The strongest duration effects are seen in the responses to the /ɪ/ stimuli. For the /i/ and /ɛ/ stimuli they are more or less absent. As the duration of /ɪ/ is increased, /ɛ/ responses are more and more favored. This pattern is particularly evident in the case of the /ɪ/ variant with small formant excursion. As indicated in Figure 1 this stimulus is very close to the /ɛ/ with the least extensive formant transitions. These findings are compatible with the expectation that, being a more open vowel, /ɛ/ should also be longer. Apart from this /ɪ/-/ɛ/ interaction, duration effects are small indicating that, for this experiment, they were not strong enough to overrule the apparently more important information on formant pattern and the style of the frame.

SUMMARY

The undershoot phenomena evident in previously reported acoustic analyses[1] show up also in the present investigation. Error patterns clearly suggest that listeners expect these effects in [w_l] syllables of the present type. The expectation is that there will be more undershoot in casual than in clear style.

As formants move further and further away from [w] towards an underlying hypothetical "target pattern", recognition scores improve *provided that* the contextual information is consistent. In other words, formant transitions are not judged in absolute terms. Their extent is determined relative to the context. Accordingly, a given pattern can be judged as extensive in casual style but as reduced in clear.

Finally, we note that the vowels /i/ and /ɪ/ when clearly spoken were least confused in the clearly spoken context.

CONCLUSION

Do the present results support the suggestion that the patterns of undershoot observed in clear and casual [w_l] syllables[1] are to be explained as instances of "undershoot compensation" and "facilitation of listener's task"? The present results appear compatible with such a claim, since identification was highest for the most extensive formant movements, particularly in the clear speech condition.

REFERENCES

[1] Moon S-J and Lindblom B (1994): "Interaction between duration, context and speaking style in English stressed vowels", *J Acoust Soc Am* 96(1):40-55.

ACKNOWLEDGEMENTS

The present research was supported by four sources: (i) A grant from the Advanced Research Program of the *Texas Board of Coordination*; (ii) grant No. BNS-9011894 from the *National Science Foundation*; (iii) *Projet Sciences Européen* (ERB4002PL910339); and (iv) by project F 770/93 sponsored by HSFR, *Humanistiska Samhällsvetenskapliga Forskningsrådet* of Sweden.