Implications of rapid spectral changes on the categorization of tonal patterns in speech perception

David House

ABSTRACT

In order to study what kinds of categorization features can be used in the perception of tonal contours in speech, and to determine how these features interact with spectral changes, four perception experiments were carried out. The stimuli were comprised of synthetic /a/ vowels with different tonal patterns, and listeners were presented with an ABX-test configuration. In the first experiment, steady-state vowels were used. In the three other experiments, a gap in the tonal pattern consisting of an intensity drop preceded and followed by formant transitions was introduced in different places in the vowel. The results of the tests suggest that tonal movement in vowels can be categorized in terms of pitch movement such as rise-fall and fall-rise or in terms of pitch levels. Listeners who categorized in terms of pitch levels demonstrated tendencies which can be interpreted as heightened attention to pitch frequency immediately following spectral changes. These findings can have implications for production and perception models of tone and intonation.

1. INTRODUCTION

In studying the functions and tasks of the auditory system in speech perception, at least two levels of frequency dependent analysis can be distinguished. A first order frequency analysis is carried out based on the mechanical properties of the basalar membrane and characteristic frequencies and temporal

responses of auditory-nerve fibers. This first-order analysis can be seen as providing the raw materials for a second-order analysis of pitch and timbre (Plomp, 1976; Green, 1976; Young and Sachs, 1979; Gelfand, 1981).

Current psychoacoustic research and physiological modelling experiments are generally in agreement that this second-level analysis, which resolves spectral information and pitch, involves some degree of central processing (Plomp, 1976; Delgutte, 1982). Although many models of pitch perception presuppose a spectral resolution of the lower harmonics from which pitch can then be derived (Wightman, 1973; Goldstein, 1973; Terhardt, 1974) it is still unclear as to what degree of spectral resolution is necessary for pitch perception and exactly what type of information (spatial or temporal) is used (Sachs, Young, and Miller, 1982; Delgutte, 1982). A further question involves the interaction between spectral cues and pitch cues in speech perception.

The present study is concerned with two issues related to fundamental frequency perception and spectral analysis in speech. The first issue is an attempt to define short-term memory categorization features in speech-like pitch movements. The second issue is an attempt to study how such categorization features of pitch might interact with spectral changes during pitch movement.

Two candidates for pitch movement categorization features can be proposed. One would be a continuous pattern storage involving categories such as rise-fall and fall-rise. The other possibility would be the storage of discrete pitch frequencies at given time intervals with movement then being interpolated after the pitch analysis.

Where interaction with spectral changes is concerned, categorization features based on continuous pattern storage might be sensitive to disturbances by such rapid spectral

changes as formant transitions. On the other hand, spectral changes associated with consonants could provide perceptual boundaries which, when related to linguistic structure, might facilitate discrete pitch frequency resolution.

2. METHOD

A. Stimuli design and synthesis

Fundamental frequency cues in speech such as syllable stress, word accents, word tones, etc. are tightly related to lexical items and are therefore, as categories, acquired and processed centrally as are other features having linguistic functions. To attempt to separate lexical function from intonation contours, question-statement categories have often been used (Hadding-Koch and Studdert-Kennedy, 1964; Fourcin, 1980). In the present study, however, an ABX test design was used to force listeners to create categories instead of using presupposed linguistic categories. The basic categories expected to be created by the listeners were rise-fall and fall-rise.

A Klatt software synthesizer and a VAX digital computer were used to synthesize a Swedish /a/ vowel with formant frequencies of 600, 925, 2540, and 3320 Hz. (Klatt, 1980; Fant, 1973). Vowel duration was 300 msec. including 30 msec. intensity onset and offset. Fundamental frequency was systematically varied to create 18 different stimuli (Fig. 1). The Fo contour for stimulus A (also stimulus 1), designed to elicit rise-fall categories, rose from 120 Hz to a turning point of 180 Hz and then fell to an end point of 100 Hz. The Fo contour for stimulus B (stimulus 12), designed to elicit fall-rise categories, began at 120 Hz falling to 80 Hz and then rising to 160 Hz. The difference in end-point frequency was designed to test the effect of end-point variation on the rise-fall, fall-rise categories, i.e. movement pattern recognition versus discrete frequency analysis. The 18 stimuli were constructed by

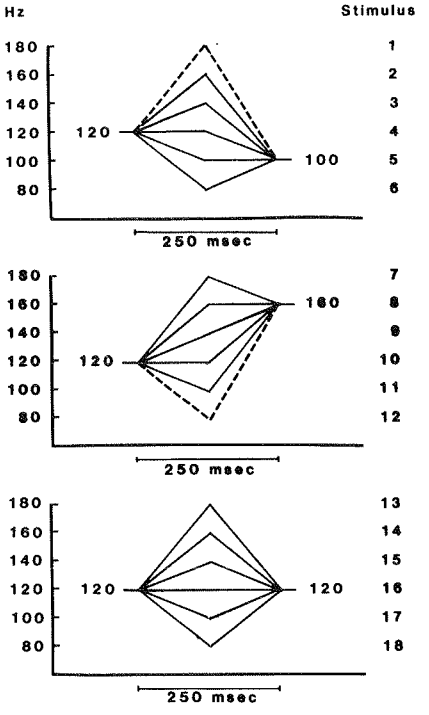

Figure 1. Stylized Fo contours for the first set of stimuli, steady-state vowel. The dashed lines (stimuli 1 and 12) represent stimuli A and B of the ABX configuration.

systematically varying the turning point in steps of 20 Hz from 80 Hz to 180 Hz with three different end-point configurations: 100 Hz, 160 Hz and 120 Hz.

Varying the Fo parameter in the Klatt synthesis program, as is often the case in natural speech, causes concomitant intensity variations. An informal listening test using LPC versions of selected stimuli with corrected intensity parameters demonstrated that the difference between corrected stimuli and uncorrected stimuli was marginal and in most cases could not be perceived at all. In view of the marginal perceptual nature of the intensity variation when compared to Fo variation, the intensity variations in the stimuli used in the experiments were not corrected. This is not to say, however, that intensity, especially in conjunction with Fo variation, is not perceptually relevant, but simply that intensity variation due to Fo changes in the context of these experiments is viewed as a concomitant feature of Fo.

An ABX-type test was constructed with a 1.3-second pause between stimulus A and B and a 2.2-second pause between stimulus B and X. Between stimulus X and the following A stimulus there was a 4.6-second pause. The X stimuli were randomized by a computer program and the test was divided up into blocks of 10 stimuli each. The test consisted of 100 ABX stimuli, the first block of 10 being a buffer block to acquaint the listeners with the test. Each of the 18 different stimuli occurred five times in the remaining 90 stimuli. All 18 stimuli were presented once before any stimulus was repeated. The same test configuration was also used for the second, third and fourth versions of the test. The stimuli were recorded on tape using a Revox PR99 tape recorder. A pause of about 10 seconds was included between each block. The total duration of the test was about 15 minutes.

To test the effects of rapid spectral changes on the categorization of the tonal patterns, four different versions of the test were produced creating four experiments. The first version consisted of the steady-state /a/ vowel as described above. In the second version, a gap, consisting of an intensity drop preceded and followed by formant transitions, was

introduced in the tonal pattern (Fig. 2). The first set of formant transitions extended from 60 msec. into the vowel to 75 msec. The second set extended from 135 msec. to 150 msec. into the vowel. The intensity drop began 80 msec. into the vowel and ended at 130 msec. This created /abaa/-like stimuli. Otherwise the tonal patterns were the same as in the first version (Fig. 3). In the third version, the gap was placed in the last half of the vowel; the transitions beginning and ending at 150 - 165 msec. and 225 - 240 msec. respectively, and the intensity drop extending from 170 to 225 msec. (Fig. 4). This created /aaba/-like stimuli. In both second and third versions the actual turning point frequency was present. In the fourth version, however, the gap was placed in the middle of the vowel; transitions beginning and ending at 105 - 120 Hz and 180 - 195 Hz respectively, with the intensity drop extending from 120 to 180 Hz. (Fig. 5). These stimuli were /aba/-like.

These manipulations were done to test which parts of the tonal pattern are most susceptible to possible disturbances caused by rapid spectral changes. All other elements of the test configuration were held constant throughout the four different versions.

B. Subjects

Five members of the secretarial staff at the Department of Linguistics and Phonetics, Lund University and 31 first-term students in phonetics participated in the experiments. All of the subjects had normal hearing and were native speakers of Swedish. To avoid fatigue effects, each of the four tests was administered on a separate occasion.

The five staff members participated in all four tests while each individual student participated in only one of the four tests. The total number of listeners for each test was 11 for test 1, 11 for test 2, 10 for test 3, and 14 for test 4.

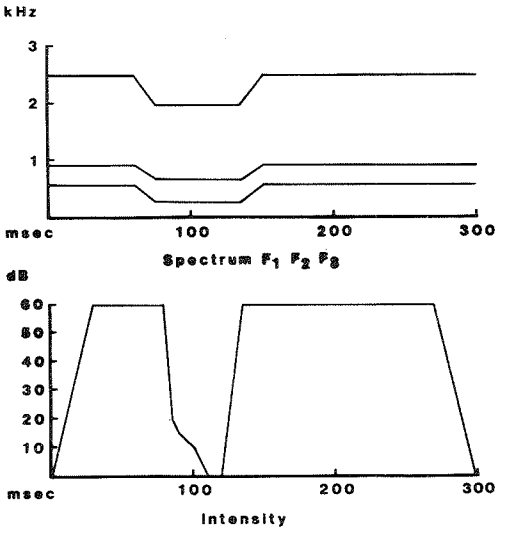

Figure 2. Spectral and intensity manipulations in the second version of the stimuli (abaa-like stimuli).

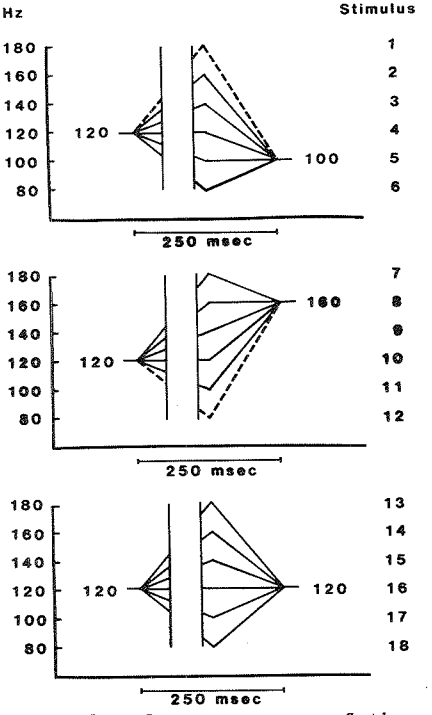

Figure 3. Stylized Fo contours of the second version of the test (abaa-like stimuli). The dashed lines (stimuli 1 and 12) represent stimuli A and B of the ABX configuration. The gaps represent spectral and intensity manipulations.

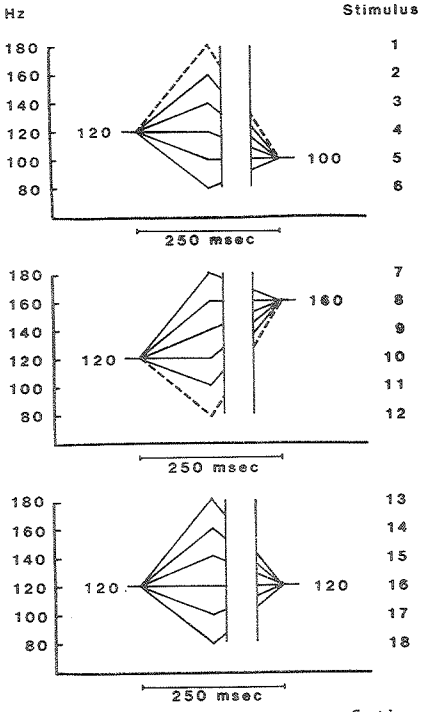

Figure 4. Stylized Fo contours of the third version of the test (aaba-like stimuli). The dashed lines (stimuli 1 and 12) represent stimuli A and B of the ABX configuration. The gaps represent spectral and intensity manipulations.

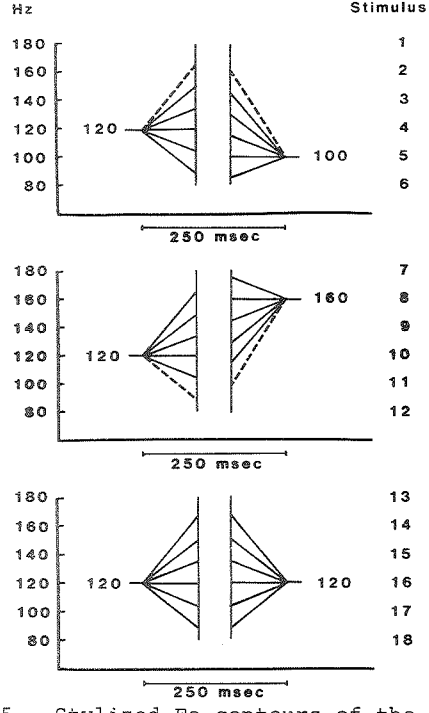

Figure 5. Stylized Fo contours of the fourth version of the test (aba-like stimuli). The dashed lines (stimuli 1 and 12) represent stimuli A and B of the ABX configuration. The gaps represent spectral and intensity manipulations.

C. Procedure

The test tapes were presented in a sound-treated perception lab via a Revox A77 tape recorder and Burwen PMB6 orthodynamic headphones. Each listener was given a printed instruction sheet and an answer sheet. The instructions were also read aloud, the listeners being instructed to listen to the first two sounds of each test item and, upon hearing the third sound, to decide if it was most like the first or the second sound and to circle the corresponding number (1 or 2) on the answer sheet. The tape was stopped after the practice block and the listeners were allowed to ask questions. The listeners were also given a rest pause after half the test.

3. RESULTS

A. Steady-state vowel

The results of the test version consisting of the steady-state vowel were uniform for all listeners. The test was easy to perform, and all listeners categorized the stimuli on the basis of movement pattern recognition, i.e. rise-fall or fall-rise disregarding the differences in end-point frequency.

Figure 6 shows that stimuli 1-4 were categorized as rise-fall (most like stimulus 1) while stimuli 5 and 6 were heard as fall-rise (most like stimulus 12) despite the fact that the end-point frequency was the same in all 6 stimuli. In the group of stimuli where the end-point frequency was 160 Hz, stimuli 7-9 were heard as rise-fall and stimuli 10-12 as fall-rise. The same pattern of results was obtained in stimuli 13-18 with the constant stimulus 16 chosen as most like stimulus 1. No listener chose categories on the basis of end-point frequency, although the difference in responses for stimuli 4 and 10 could be interpreted as being influenced by end-point frequencies.

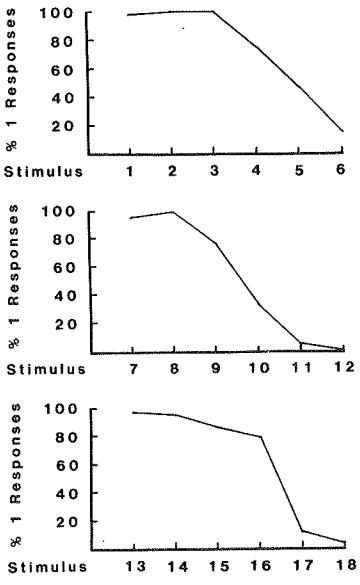

Figure 6. Averaged responses for test 1, steady-state vowel, showing rise-fall and fall-rise categorization.

B. Spectral change early in the vowel.

The averaged results for the test version where spectral changes were placed early in the first part of the vowel show a considerable ambiguity concerning stimulus categories (Fig. 7). In contrast to the results obtained in the steady-state vowel version, stimulus 4 was perceived as most like stimulus 12, stimulus 10 as ambiguous and stimulus 16 as most like stimulus 12. These differences can not be explained in terms of end-point frequencies as the categories seem to be formed contrary to the frequencies of the end-points, i.e. there were

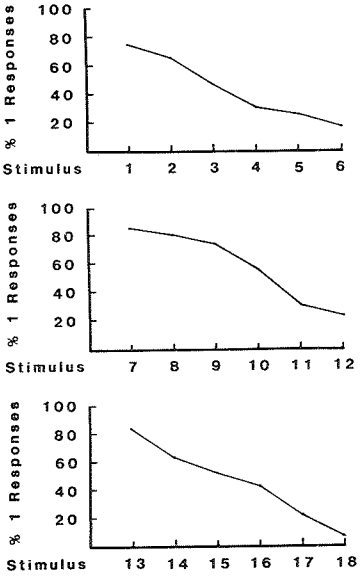

Figure 7. Averaged responses for test 2, /abaa/ stimuli.

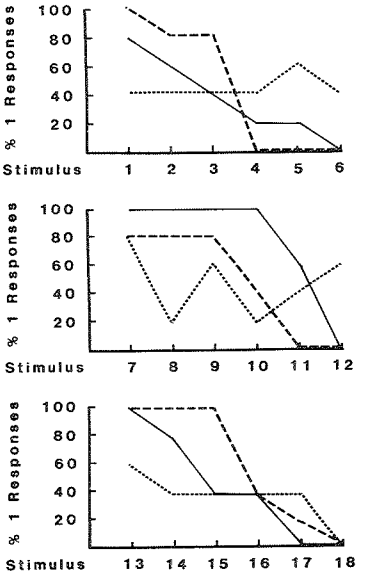

Figure 8. Individual responses (3 listeners) for test 2, /abaa/ stimuli. The solid line represents reverse end-point frequency categorization, the dashed line represents rise-fall, fall-rise categorization, and the dotted line represents no perceived categories.

more stimulus 12 responses to stimuli 1-6 and more stimulus 1 responses to stimuli 7-12. When the end-point frequency was 120 Hz, stimuli 13-18, the averaged responses showed rise-fall, fall-rise categorization, but with much more ambiguity than in the first test.

There was also considerable individual variation among the listeners. Basically, responses fell into three classes of roughly equal size: 1.) those who seemed to perceive categories contrary to end-point frequencies, 2.) those who perceived the same tonal categories as in the first version, and 3.) those who could no longer perceive the categories (Fig. 8).

C. Spectral changes late in the vowel.

In this version of the test, results were similar to those of the second version with the exception that the listeners who perceived categories which were not based on movement pattern recognition tended to choose categories based on the end-point frequency. For example, stimulus 3 which was ambiguous in test 2, was perceived as most like stimulus 1. Stimulus 9, clearly most like stimulus 1 in test 2, was perceived as most like stimulus 12 in this version. Although individual variation was again considerable, the averaged responses show a tendency towards end-point categories (Fig. 9) as there were more stimulus 1 responses to stimuli 1-6 and more stimulus 12 responses to stimuli 7-12. This tendency is more clearly seen in individual responses (Fig. 10).

D. Spectral changes in the middle of the vowel.

In the fourth version of the test, the results were almost identical to the results obtained from the third test where the spectral changes occurred late in the vowel (Fig. 11). There was, however, more of a tendency for listeners to perceive categories based on end-point frequencies in this version than in any other version of the test as exemplified by Fig. 12.

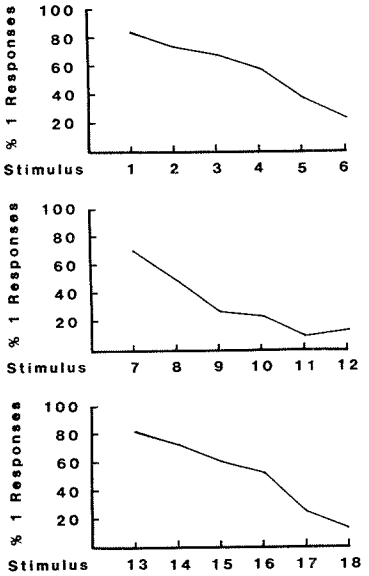

Figure 9. Averaged responses for test 3, /aaba/ stimuli.

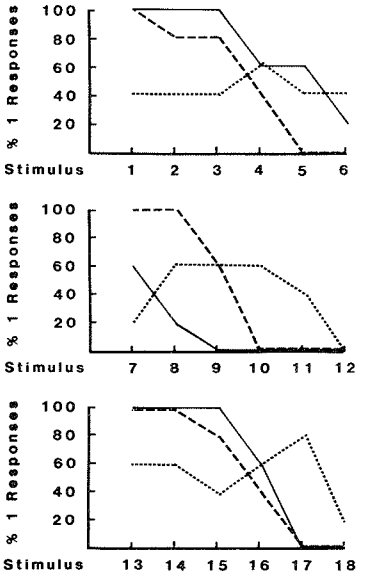

Figure 10. Individual responses (3 listeners) for test 3, /aaba/ stimuli. The solid line represents end-point frequency categorization, the dashed line represents rise-fall, fall-rise categorization, and the dotted line represents no perceived categories.

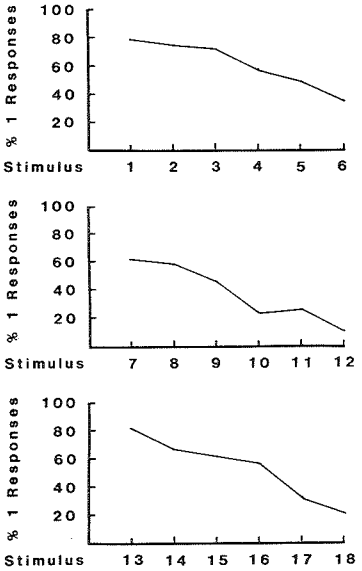

Figure 11. Averaged responses for test 4, /aba/ stimuli.

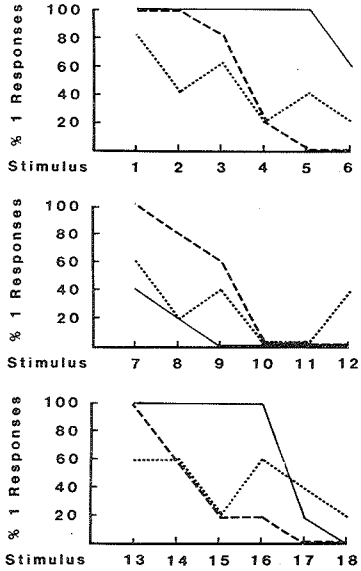

Figure 12. Individual responses (3 listeners) for test 4, /aba/ stimuli. The solid line represents end-point frequency categorization, the dashed line represents rise-fall, fall-rise categorization, and the dotted line represents no perceived categories.

E. Summary of the results

In the first version of the test, comprising steady-state vowels, listeners performed uniformly, categorizing the stimuli in terms of movement direction; rise-fall or fall-rise. In the three other versions, where spectral changes were introduced in the vowel, responses fell into three classes: ambiguous responses, categorization based on movement direction, and categorization based on final pitch frequency. However, in the stimuli where spectral changes occurred early in the vowel, the group of listeners who did not use movement direction but still performed categorizations, seemed to do so based on the reverse of the end-point frequencies.

DISCUSSION

A. Movement pattern as categorical feature.

The first experiment was designed to determine if pitch movement can be used as a categorization feature for tonal patterns in speech perception. The results for stimuli 3,4,9 and 15, all categorized as stimulus 1, and for stimuli 5, 11, and 17, which were categorized as stimulus 12 (Fig. 6), seem to indicate that listeners can use the movement features rise-fall and fall-rise to categorize tonal patterns. The results of the first experiment could also be explained in terms of turning-point configuration where a change in the direction of the pitch movement results in a convex or a concave configuration. It could be that the perceptual mechanism is particularly sensitive to changes in pitch movement direction whereby an increasing number of periods per time unit followed by a decreasing number or vice-versa would be registered by the perceptual mechanism as an event which could then form the basis for categorization. Whatever the mechanism, the results indicate that in the absence of spectral changes, movement pattern rather than end-point frequency level is selected by listeners to form the basis for categorization in this type of forced choice test.

Although it is a large step from synthetic vowel stimuli to actual speech stimuli, these results suggest the possibility that languages could make use of movement direction in the vowel as a distinctive perceptual feature. Furthermore, the ability of about one-third of the listeners (12 listeners) to perform the same movement categories in the stimuli containing rapid spectral changes indicates that some listeners are able to interpolate the movement and arrive at the same kinds of categories as in the continuous pattern categorization. However, the diversity of the responses to the stimuli containing rapid spectral changes indicates that spectral changes do in fact influence pattern categories.

B. Pitch level as categorical feature.

In tests 3 and 4, those listeners who perceived categories which were not based on movement patterns, seemed to perform the categories based on end-point identification. Had it not been for the conflicting results of test 2 where listeners seemed to perform categories in direct opposition to end-point frequencies, the results would seem to point toward some kind of movement masking effected by the rapid spectral changes. This masking would lead listeners to use end-points for forming categories. The results of test 2, however, complicate matters.

In their classic study of pitch discrimination for synthetic vowels, Flanagan and Saslow (1958) found slightly more acute discrimination of changes in fundamental frequency in vowels than in a pure tone. This could mean that listeners use the relatively larger changes in the harmonics present in the vowels as an aid in discriminating Fo changes. Klatt (1973) substantiated these results but found that discrimination performance deteriorated when a linear ramp fundamental frequency contour was used in the place of a monotone. t'Hart, in his study of just noticable differences in pitch movement (1981), found that falls were more difficult to judge than rises and that many subjects compared final pitches instead of sizes of movements. He also reported from Nabelek (reference

omitted) a separate concentration on the initial or the final pitch if there was a pause between the low and high frequency parts of a stimulus rather than a continuous glide.

These findings could be used to help explain the present results. If rapid spectral changes are introduced in the vowel, the load on both the processor and memory is drastically increased. In some cases, an economy measure may be necessary whereby the pitch movement is recoded into pitch levels. In tests 3 and 4, the final pitch levels correspond to the end-point frequencies. If the pitch levels are defined as the average frequency some 30 to 50 msec. after the vowel onset following the stop occlusion, stimulus 1 in test 2 (see Figure 3) would be stored as low-high and stimulus 12 as high-low. Stimuli 3-6 might have therefore been classified as high-low (i.e. most like stimulus 12) while stimuli 7-10 might have been classified as low-high (i.e. most like stimulus 1). This would account for the difference in results between test 2 and tests 3 and 4. Categorization in these tests would therefore not be based on end-point frequency but on averaged frequency level following the vowel onset after the stop occlusion.

C. Interaction with spectral changes.

The results of these experiments indicate that spectral changes occurring during a tonal movement can affect the perception of the tonal contour. During the time period in which rapid spectral changes occur and immediately following this period, the perceptual mechanism may be somewhat less sensitive to pitch movement than during a longer steady-state portion of the vowel. Whether this insensitivity is a result of an increased load on the peripheral mechanism or a result of language acquisition or both remains an open question. However, this insensitivity to movement could require the pitch perception mechanism to increase attention to averaged pitch frequency during this time period. The averaged frequency would then be stored in memory as a pitch level, e.g. High, Low or Mid. These levels could then be used linguistically as categorization

features.

It is possible, then, that rapid spectral changes such as a stop release or any consonant release followed by a vowel can serve as a perceptual boundary marker. This boundary marker would inhibit perception of continuous pitch movement and enhance perception of discrete pitch frequency. The perceptually important feature of the tonal contour immediately following this boundary would be, therefore, a tonal level. Continuous movement as an important feature of the tonal contour would then need to be placed in a longer, steady-state portion of a presumably stressed vowel.

D. Possible linguistic implications

The possible perceptual division of the tonal contour into sections of differential sensitivity for levels and movement might have implications for both production and perception models of tone and intonation. Knowledge of such sections and how they relate to the segmental structure of an utterance could, for example, facilitate the exact placement of Lows and Highs in production models. An aid for perception models would be in determining what part of the tonal contour is relevant for memory storage.

The results of these experiments and an explanation involving differential sensitivity seem to agree with the results obtained in perception tests aimed at studying the perceptual cues necessary to distinguish tone 3 from tone 4 in Modern Standard Chinese, (Gårding, et al. 1985, this issue). It could be that tone languages are prone to use tonal patterns which fit pitch sensitivity differences in the perceptual periphery. These patterns would then be linguistically relevant reinforcing the selective pitch sensitivity.

Finally, the idea of selective pitch sensitivity could be applicable to speech synthesis and speech recognition. In synthesis, the placement of the tonal contour in relation to

the segmental structure could be facilitated, and in recognition, segmental boundaries could be used as markers for extracting critical parts of the tonal contour.

E. Further issues

Some further issues concerning the perception of pitch and spectral change lying beyond the scope of the present study involve perception of pitch immediately prior to spectral change. Another issue could be the use of non-speech noise to see if the perceptual boundary is constrained to spectral changes produced by the articulators. Finally, more work needs to be done using real speech material to more fully understand the processes involved in the perception of the tonal contour in speech.

ACKNOWLEDGEMENTS

I would like to thank Gösta Bruce and Eva Gårding for valuable discussion and comments concerning this study.

REFERENCES

Delgutte, B. 1982. Some Correlates of Phonetic Distinctions at the Level of the Auditory Nerve. In Carlson and Granström (Eds.) The Representation of Speech in the Peripheral Auditory System. Elsevier Biomedical Press, Amsterdam, New York, Oxford, pp. 131-149.

Fant, G. 1973. Speech sounds and features. The MIT Press. Cambridge, Mass.

Flanagan, J.J., and Saslow, M.G. 1958. Pitch discrimination for synthetic vowels, J. Acoust. Soc. Am. 30, 435-442.

Fourcin, A.J. 1980. Speech pattern audiometry. In Beagley, H.A. (Ed.) Auditory investigation: the scientific and technological basis.

Gårding, E., Kratochvil, P., Svantesson, J.O., & Zhang, J. 1985. Tone 4 and Tone 3. Discrimination in Modern Standard Chinese. Working Papers 28 (This issue), Department of Linguistics and Phonetics, Lund University.

Gelfand, S.A. 1981. Hearing, an introduction to psychological and physiological acoustics. Marcel Dekker, Inc. New York and Basel, Butterworths, London.

Goldstein, J.L. 1973. An optimum processor theory for the central formation of the pitch of complex tones. J. Acoust. Soc. Am. 54, 1496-1516.

Green, D.M. 1976. An introduction to hearing. Lawrence Erlbaum Associates, Hillsdale, New Jersey.

Hadding-Koch, K., and Studdert-Kennedy, M. 1964. An Experimental Study of Some Intonation Contours. Phonetica 11, 175-184.

t'Hart, J. 1981. Differential sensitivity to pitch distance, particularly in speech. J. Acoust. Soc. Am. 69, 811-821.

Klatt, D.H. 1973. Discrimination of fundamental frequency contours in synthetic speech: Implications for models of pitch perception. J. Acoust. Soc. Am. 53,8-16.

Klatt, D.H. 1980. Software for a formant synthesizer. J. Acoust. Soc. Am. 67, 971-995.

Plomp, R. 1976. Aspects of Tone Sensation, A Psychophysical Study. Academic Press, London.

Sachs, M.B., Young, E.D., & Miller, M.I. 1982. Encoding of Speech Features in the Auditory Nerve. In Carlson and Granström (Eds.) The Representation of Speech in the Peripheral Auditory System. Elsevier Biomedical Press, Amsterdam, New York, Oxford, pp. 115-130.

Terhardt, E. 1974. Pitch, consonance, and harmony. J. Acoust. Soc. Am. 55, 1061-1069.

Wightman, F.L. 1973. The pattern-transformation model of pitch. J. Acoust. Soc. Am. 54, 407-416.

Young, E.D. & Sachs, M.B. 1979. Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers. J. Acoust. Soc. Am. 66, 1381.