Accession Number : ADA142996

Title : The Discrimination of Pitch in Pulse Trains and Speech

Descriptive Note : Technical rept.

Defense Technical Information Center

Corporate Author : MASSACHUSETTS INST OF TECH LEXINGTON LINCOLN LAB

Personal Author(s) : Mack, M. A. ; Gold, B.

Handle / proxy Url : http://handle.dtic.mil/100.2/ADA142996 Check NTIS Availability...

Report Date : 12 APR 1984

Pagination or Media Count : 33

Abstract : Much research has been conducted on the discrimination of pure and complex tones. Yet relatively little work has been carried out on the discrimination of pitch in speech. Thus, the present experiment was designed to explore listener's ability to discriminate the pitch to three types of acoustically complex stimuli - pulse trains with monotone pitch, vocoded speech with monotone pitch, and vocoded speech with natural pitch. Results revealed that the discrimination of naturally intoned sentences was worse than that of the pulse trains or monotone sentences. Implications for speech synthesis and processing are discussed.

Descriptors : *SPEECH, *DISCRIMINATION, *VOCODERS, *AUDIO TONES, STIMULI, PULSE TRAINS

Subject Categories : NON-RADIO COMMUNICATIONS

Distribution Statement : APPROVED FOR PUBLIC RELEASE

Technical Report
680

The Discrimination of Pitch in Pulse Trains and Speech

M.A. Mack
B. Gold

12 April 1984

Lincoln Laboratory
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
LEXINGTON, MASSACHUSETTS

Prepared for the Department of the Air Force
under Electronic Systems Division Contract F19628-80-C-0002.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
LINCOLN LABORATORY

THE DISCRIMINATION OF PITCH IN PULSE TRAINS AND SPEECH

M.A. MACK
B. GOLD
Group 24

TECHNICAL REPORT 680

12 APRIL 1984

LEXINGTON MASSACHUSETTS

ABSTRACT

Much research has been conducted on the discrimination of pure and complex tones. Yet relatively little work has been carried out on the discrimination of pitch in speech. Thus, the present experiment was designed to explore listeners' ability to discriminate the pitch of three types of acoustically complex stimuli--pulse trains with monotone pitch, vocoded speech with monotone pitch, and vocoded speech with natural pitch. Results revealed that the discrimination of pulse trains was superior to that of monotone sentences, while the discrimination of naturally intoned sentences was worse than that of the pulse trains or monotone sentences. Implications for speech synthesis and processing are discussed.

CONTENTS

I. INTRODUCTION

Human pitch perception, as a scientific discipline, has a rich history, dating back at least to the pure-tone experiments of Seebeck (1841) and Helmholtz (1863). Consistently, experiments in the perception of the pitch of pure tones have revealed that under certain conditions listeners can discriminate differences as small as about .3 Hz in the region of 100-250 Hz (e.g., Boring, 1940; Rosenblith and Stevens, 1953; Nordmark, 1968; Moore, 1974). However, most researchers have not attempted to examine pitch discrimination in speech or speech-like stimuli. Yet as Watson (1976) indicates,

> a reasonable hypothesis is that as sounds are made temporally and spectrally more complex, they place proportionally greater demands on . . . central, or psychological mechanisms. This may be one of the reasons why it is impossible to predict accurately people's ability to extract information from the complex sounds of speech or music from their corresponding abilities with the simpler sounds which could serve as the components of the more complex ones (p. 175).

Two important studies which have dealt with the issue of the discrimination of speech-like sounds are Flanagan and Saslow (1957) and Klatt (1973).

In their study of the discrimination of pitch in synthetic steady-state vowels, Flanagan and Saslow derived difference limens (DLs) that were actually slightly smaller than those previously established for pure tones. They suggest that "discriminations of vowel pitch are not being made solely on the basis of changes in frequency of the fundamental component. [Rather] listeners make some use of frequency changes in the . . . harmonics (p. 440). This hypothesis seems valid in light of work

done by Ritsma (1967) who found that the lower harmonics do convey critical pitch information.

Klatt (1973) likewise found relatively small DLs for the discrimination of pitch in synthetic speech. For the steady-state vowel /ɛ/ and for the glide-vowel syllable /ya/, his subjects' average DLs ranged from .18 to .5 Hz. Only when the pitch of the stimuli changed fairly rapidly from beginning to end, as in the steeply ramped fundamental frequency of a vowel, were the DLs as large as 4 Hz.

These experiments suggest that--at least if stimuli are steady-state vowels--listeners can perceive very slight differences in their pitch. It appears, though, that if the fundamental frequencies of the stimuli change over time, discrimination worsens. Klatt's experiment also suggests that pitch discrimination is probably worse for speech than for pure tones or steady-state vowels, since the pitch of speech is normally time varying. Yet his experiment does not address the following questions: Is it simply the presence of time-varying frequency that induces a decrement in discrimination? Or might other factors contribute? For example, does discrimination worsen if additional demands are placed upon speech processing, as they are in fully elaborated linguistic stimuli consisting of phonetic, syntactic, and semantic information?

We have addressed these questions by devising an experiment using speech and non-speech stimuli. Data derived from this experiment are of considerable relevance for work with vocoded speech communication. For, as an initial step in the determination of how precisely the fundamental frequency must be specified in algorithms used for speech synthesis, it is necessary to establish the degree to which pitch can be discriminated.

That is, if listeners can only discriminate relatively large pitch differences in vocoded speech, it may be assumed that their tolerance for pitch errors (in vocoded speech) might be fairly high. On the other hand, if they prove capable of discriminating differences on the order of those found for pure tones and steady-state vowels, it may be reasonable to assume that a high degree of accuracy in the synthesis of pitch is necessary. The work reported here was undertaken as part of a broader program to investigate the problem of vocoder robustness. In noisy environments, vocoded speech is unacceptable to many users. Degradations in both pitch and spectral fidelity contribute to this unfortunate result. Some recent experiments (Gold and Tierney, 1983) have established quantitative correlations between acoustic noise environments and vocoder performance, mainly with respect to loss of spectral fidelity. Moreover, there is general agreement that incorrect generation of the excitation signal in an LPC or channel vocoder is unacceptable to many users.

If pitch errors do not in a practical sense affect intelligibility as measured, e.g., by diagnostic rhyme test (DRT) scores (Voiers, 1983), one might be tempted to vocode speech using either a monotone pitch pulse train or pure noise as the excitation source. However, pitch is an important carrier of prosodic information. In fact, we conducted an informal experiment to determine whether sentences could be identified as declarative or interrogative when they were syntactically acceptable as either (e.g., "The news was very bad") and when they were vocoded with a monotone pitch track. Sentences were recorded with both the declarative and interrogative readings. Although listeners had no difficulty labelling

the stimuli as statements or questions when they heard them with intonation, they found it impossible to label them correctly when the pitch was monotone. (They heard all sentences as declaratives.) For noise excitation, the results were not as straightforward, but listeners still encountered great difficulty. Also, recent DRT results (Gold and Tierney, 1983) show substantial loss in acoustic noise when vocoder noise excitation is used.

If a monotone pitch and pure noise excitation are both unacceptable, what can be done to alleviate pitch errors? If the gross pitch contour can somehow be maintained by synthetic means at the expense of less overall accuracy, would this be of practical use? This report is a first step towards answering these questions.

II. EXPERIMENT

A. Subjects

Subjects were 6 employees at the M.I.T. Lincoln Laboratory. Two were female; 4 were male. Subjects ranged in age from 26 to 49 and all reported normal hearing.

Previous researchers have found that, on tests of tone or melody-sequence perception, musically trained subjects perform better than individuals who have not had musical training (e.g. Stücker, 1980; Raz and Brandt, 1977; Zatorre, 1979). For this reason, subjects were asked to describe the nature and extent of their musical training. This information was not used for screening the subjects. Rather, it was intended to be viewed in conjunction with their discrimination performance so that, if clear differences in their ability to discriminate pitch did emerge, it

could be determined whether such differences correlated with the subjects' musical background. Of the 6 subjects, only 1 (NB) had had essentially no musical training or experience; 4 (JF, JT, RM, and SC) had had several months to several years of musical training; and 1 (MA) had had extensive musical training.

Subjects were screened for the experiment based upon their performance on one of the tests used in the experiment. That is, only those who could discriminate at a specified criterion level on the monotone-sentence test were included. (It was believed that, since the monotone sentences were at least theoretically intermediate in complexity to the pulse trains and the naturally intoned sentences, they would serve as an appropriate diagnostic for determining subjects' inclusion.) Listeners had to be able to discriminate (perceive as different) an average of 75% or more of the stimuli consisting of the 2 largest pitch increments (5.5 and 6.0 Hz). On the basis of this criterion, 3 listeners were excluded from the experiment and the above-mentioned 6 were retained.

B. Stimuli

Three types of stimuli were used in our experiment. These consisted of (1) those which were acoustically complex and which had monotone pitch but no linguistic information (pulse trains); (2) those which had monotone pitch and linguistic information (monotone sentences); and (3) those which had time-varying frequency and linguistic information (naturally intoned sentences). Stimuli (2) and (3) were generated using the real-time channel vocoder programs described by Gold and Tierney (1983). Associated with each of the stimulus types were 2 separate discrimination tests--one

involving discrimination when the reference fundamental frequency (F0) was low (125 Hz) and one when the reference F0 was high (200 Hz). In the case of the naturally intoned sentences, the F0 was that of the speaker, and the values of 125 and 200 Hz were approximate. Thus, there were 6 different conditions in all. ΔF ranged from 0 to + 6.0 Hz in .5-Hz increments, yielding a total of 13 randomized pitch-pair combinations. One item from each pair always had the reference F0.

The stimuli used in our experiment were generated with the Lincoln Digital Signal Processors (LDSPs) (Blankenship and Sferrino, 1977). These processors are simple programmable computers of a von Neumann architecture; they use very high-speed digital circuits to perform computationally intensive real-time algorithms.

The pulse trains (with pulses of width 100 μsec) were generated entirely automatically by computer. The ordering of the stimuli was obtained from a random number table off-line and entered into computer memory along with the program. Then the computer was programmed to sequence through this table while the computer-generated pulses were being recorded on reel-to-reel audio tape.

The generation of the vocoded sentences was somewhat more involved. In order to produce tapes containing pairs of sentences with identical members, we used a "mimic" program which operated in real time. A more detailed description is given below. A set of 130 sentences (consisting of 10 different sentences read 13 times each in random order) was read by a female (MM) and by a male (DM) in separate recording sessions. Each of the 130 sentences was of approximately equal length and all had identical

syllable structure. Each sentence was spoken on cue to a small light which flashed on every 8 seconds. (Test sentences appear in the appendix.) MM and DM did not vary their intonation much from one sentence to the next, yet they did not attempt to speak in a "flat" voice. The consistency with which they spoke may be observed in the pitch plots presented in Figure 1. This figure displays 6 tokens of the sentence, "These rings are made of gold"--3 produced by DM and 3 by MM. It is important to emphasize that, although different tokens of a given sentence were used throughout the test, the 2 members of all pairs were identical, with the exception of changes made in the fundamental frequency as described below.

Each utterance was filtered, sampled at 20 kHz, analog-to-digital converted, and then stored in the LDSP's peripheral memory. Simultaneously, it was directed to an output port through a digital-to-analog converter followed by a low-pass post-sampling filter, after which it was recorded. Then, 3.5 seconds after the onset of the original utterance, its stored copy was also directed to the tape recorder in the same manner. Once 2 tapes of speech--each containing 130 pairs of sentences--had been recorded, it was still necessary to vary the F0 systematically and to vocode and re-record the sentences. This was done on-line, with the F0 being set manually for each sentence just before the utterance was directed through the vocoder.[1] The output was then recorded on reel-to-reel audio tape. On the final test tapes, the F0 of one member

[1]Due to real-time constraints in the synthesis of the sentences, the sampling rate of the system was limited to 10 kHz, or a sampling interval of 100 μsec. This did not allow for sufficiently small frequency increments and made it necessary to use a method based upon one described by Klatt (1973). This approach involved computing the vocoder algorithm at the 10-kHz rate, but generating the excitation signal with reference to a 100-kHz rate. Thus, pitch pulse intervals were obtained with 10-μsec precision. The resultant pulse train was then filtered and downsampled to 10 kHz. Because of the special nature of the input, the filtering operation could be performed without compromising real time.

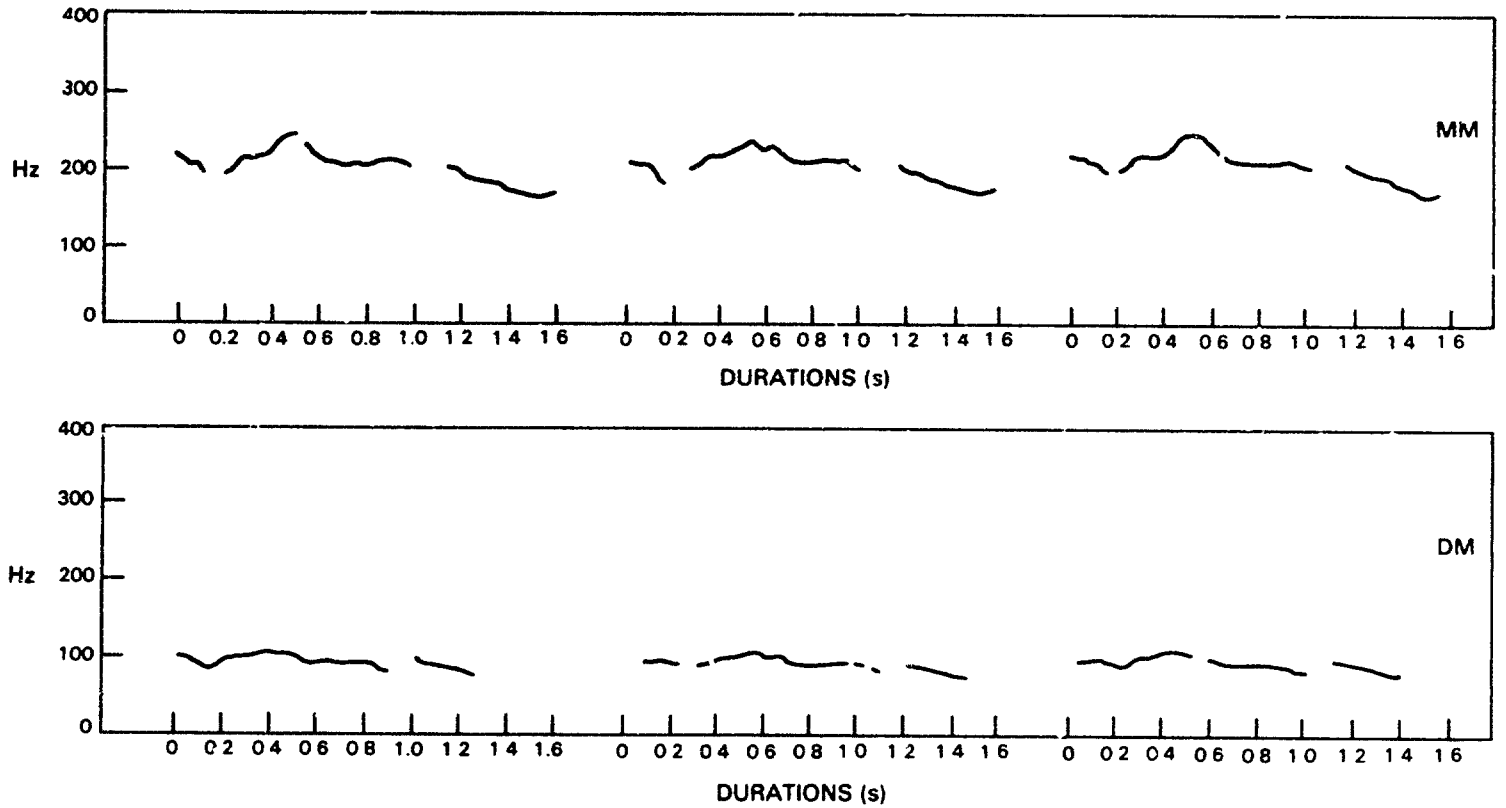


Fig. 1. Fundamental frequency plots of 6 tokens of "These rings are made of gold"--3 produced by MM and 3 by DM--used as stimuli in the experiment. This figure illustrates the degree of similarity maintained by each reader with respect to pitch and duration.

of each pair was the reference F0; the F0 of the other was F0 + ΔF. In the pulse-train conditions, the F0 was set to 125 Hz and 250 Hz. In the monotone-sentence conditions, MM's F0 was set to 200 Hz, and DM's to 125 Hz. In the intoned sentence conditions, the baseline F0 was not set by computer. Instead, the speaker's F0 was retained within a pitch-period accuracy of 100 μsec. In the comparison stimulus, the F0 was raised over the entire stimulus by 0 to 6.0 Hz. Each of the 13 pitch-pair combinations occurred 10 times. One of each of these 13 pairs was associated with each of the 10 different sentences. Half of the pairs were presented in the order AX, the other half in the order XA.

In the pulse train test, each stimulus was 2 seconds long. There was an inter-stimulus interval of 1 second, an inter-trial interval of 3 seconds, and an inter-block interval of 11 seconds. In the tests consisting of sentence stimuli, these intervals were slightly longer, for MM's and DM's utterances were generally somewhat shorter than 2 sec. Prior to the beginning of each test, 3 stimulus pairs were presented to acquaint subjects with the task. Each test lasted approximately 20 minutes. For all subjects, the first test consisted of monotone sentences, the second pulse trains, and the third naturally intoned sentences. The presentation of the high and low F0 tests was counterbalanced. For each subject, tests were separated by at least 1 day.

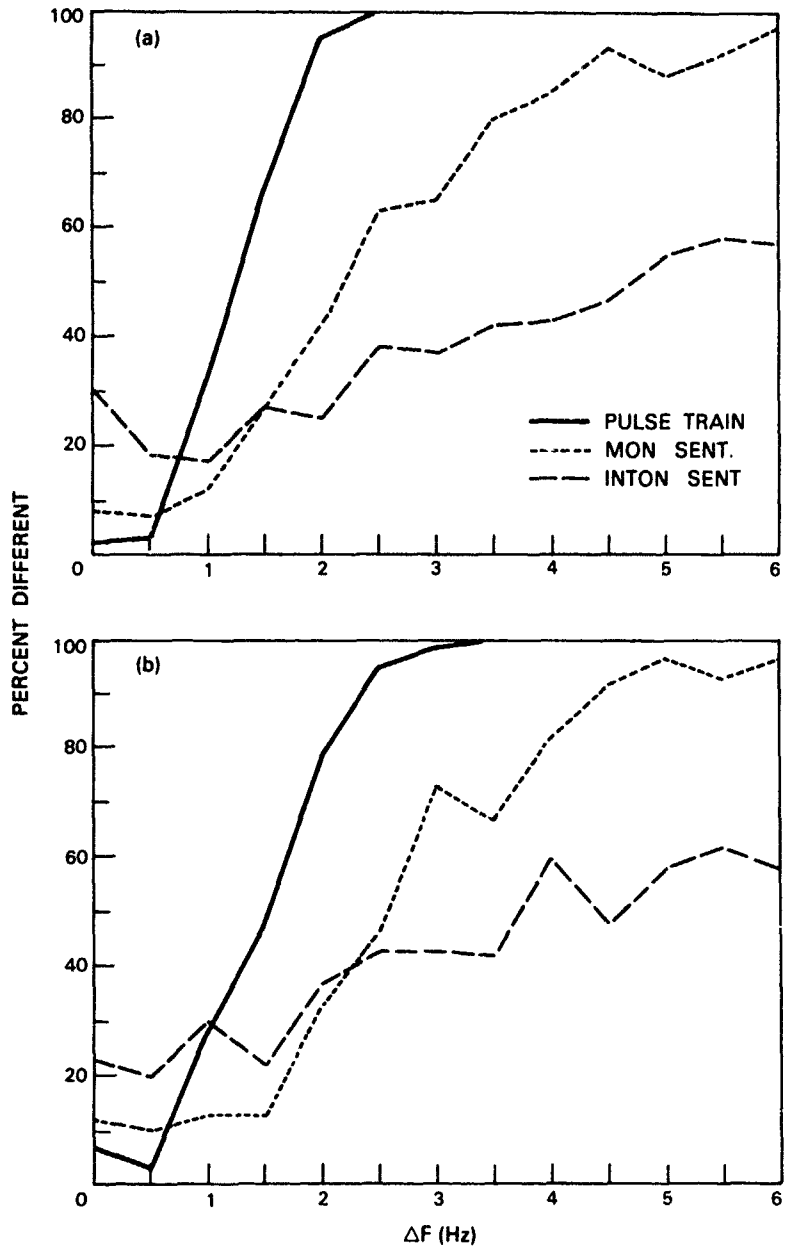
C. Procedure

Subjects were tested individually or in small groups in a sound-attenuated listening room at Lincoln Laboratory. Prior to each test, subjects were told that they would hear pairs of sounds (or sentences) and that they were to circle "S" on their answer sheet if they did not perceive

a difference in the pitch of the sounds in each pair and "D" if they did perceive a difference--however slight. They were told to guess if they were uncertain. Subjects were not told how many of the stimuli would be the same or different. Test tapes were played on a TEAC A2340SX tape recorder with a Hewlett Packard 467a amplifier. Subjects used AKG headphones. Amplitude was set at a comfortable listening level.

D. Results

Results of our experiment indicated that pitch discrimination was strongly related to the type of stimuli presented. Figures 2a and b present discrimination scores averaged across all subjects for each stimulus pair in all conditions. Figures 3a, b, and c show the responses of each subject to each stimulus pair in all conditions. As is apparent from Figures 2a and b, differences in the pitch of pulse train pairs were most discriminable and differences in the pitch of naturally intoned sentence pairs were least discriminable. In order to calculate average DLs and to carry out a statistical analysis of the data, we converted to z-scores the percent of "different" responses given by each subject to each stimulus pair. A linear regression analysis with a least-mean squares solution was carried out on the z-scores (Woodworth and Scholsberg, 1965). The DL was associated with that stimulus having a z-value of .67 (and hence a performance score of 75%). Thus calculated, the average DL for the pulse trains was 1.5 Hz, while for the monotone sentences it was 3 Hz. Because most subjects had considerable difficulty discriminating the pitch of naturally intoned sentences, a linear regression analysis could not be carried out for all subjects. Yet, in order to obtain a general impression of subjects' average performance in this condition, we conducted



135113-N

Fig. 2. Average percent different for each condition with both low (a) and high (b) reference F0. Discrimination of pulse trains is markedly better than that of sentences. Discrimination of monotone sentences is approximately intermediate to that of the other stimulus types. There is little difference in the discrimination of stimuli with a low and high reference F0.

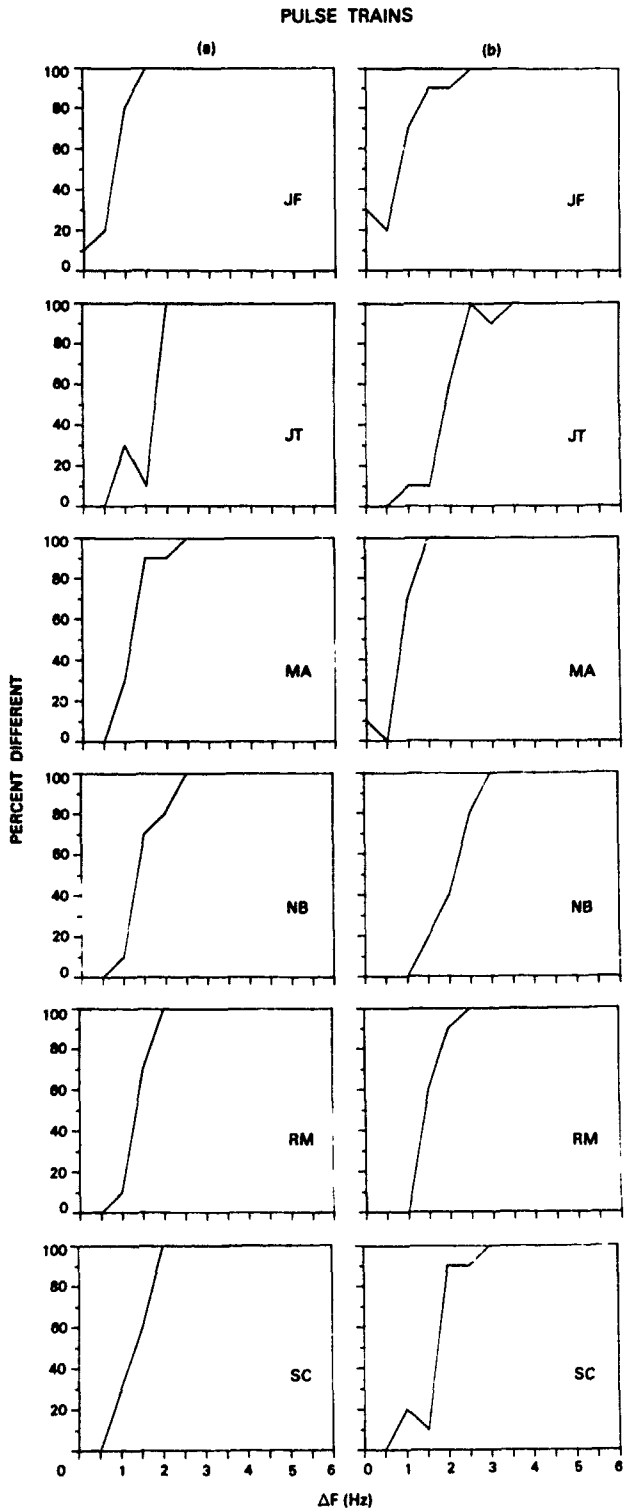
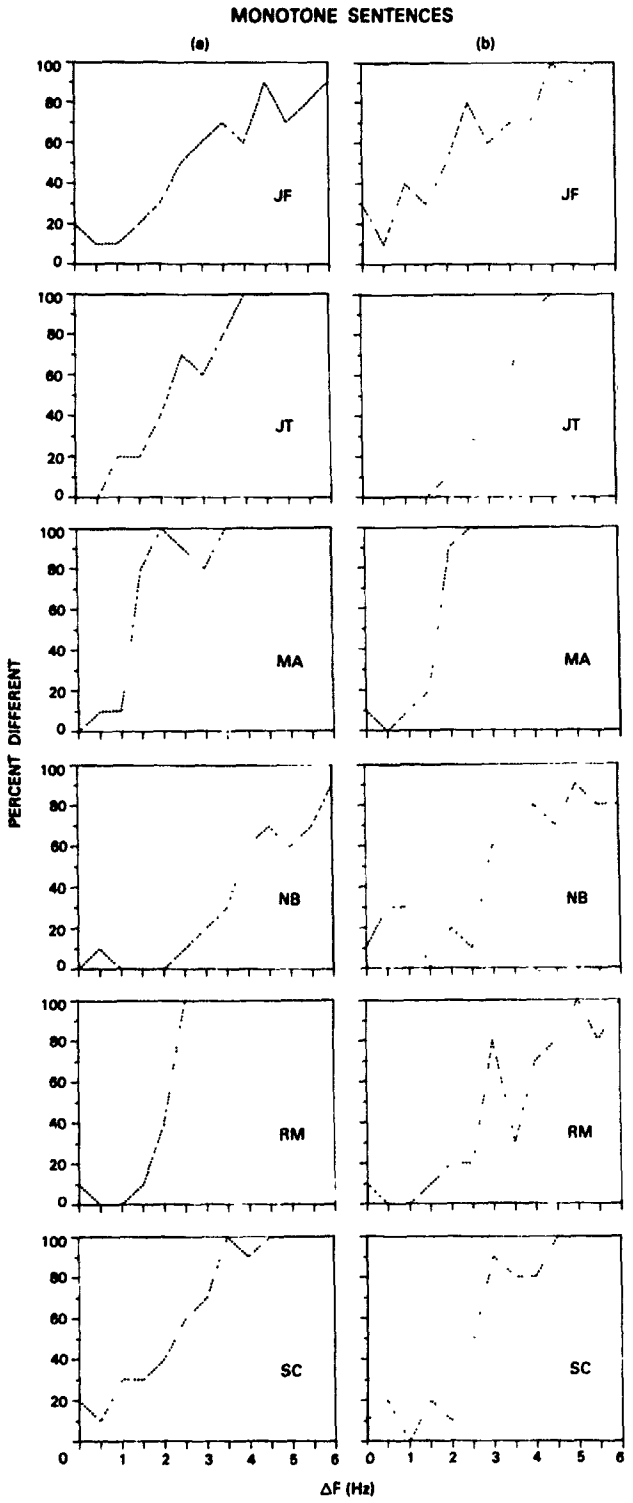
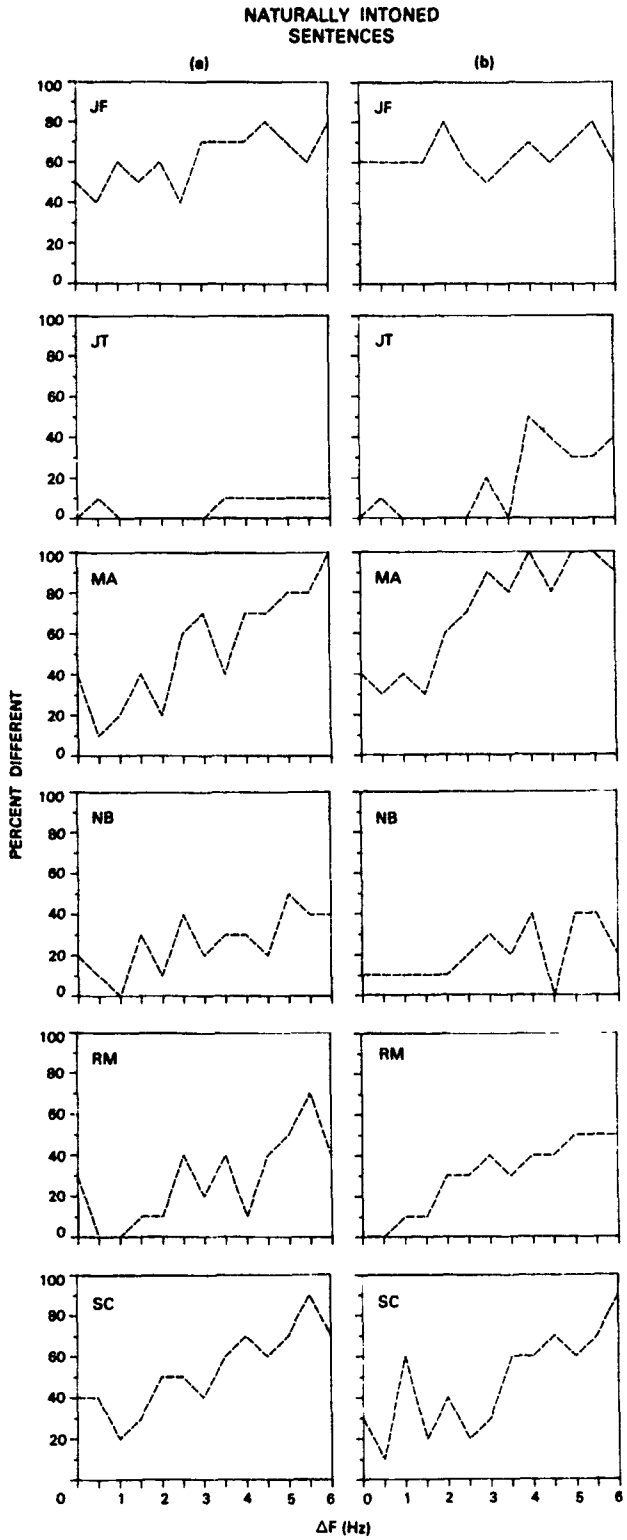


Fig. 3a. Subjects' percent different for pulse trains with low (a) and high (b) reference F0.



135115-N

Fig. 3b. Subjects' percent different for monotone sentences.



135116-N

Fig. 3c. Subjects' percent different for naturally intoned sentences.

a linear regression analysis on the average z-values associated with each stimulus increment. This revealed that the mean DL for the naturally intoned sentences was about 6 Hz. A two-way repeated-measures analysis of variance (ANOVA) (Hays, 1980) was conducted on the pulse train and monotone sentence data. Since, for most of the subjects, a straight line could not be fit to the data in the intoned-sentence condition, data from this condition were not used in the ANOVA. The ANOVA revealed a highly significant main effect due to stimulus type (pitch train vs. monotone sentences) ($F\ [1,5] = 18.26$, $p < .008$) and a non-significant main effect for reference frequency (low vs. high) ($F\ [1,5] = .63$, $p < .464$). There was no significant stimulus type x reference frequency interaction ($F\ [1,5] = .020$, $p < .892$). Although statistical analysis of the responses to the naturally intoned sentences could not be undertaken, it was obvious that performance on these was considerably worse than on the other stimulus types.

We also calculated the percent of "different" responses given to AX stimuli (i.e., to those in which the first stimulus was the reference [lower] frequency). If subjects were responding at chance, 50% of their "different" responses should have been given to AX stimuli (and 50% to XA). Indeed, for the pulse trains and monotone sentences, the average percent of "different" responses given to AX pairs was 51.90 and 54.97, respectively. However, for the naturally intoned sentences, the average percent was 71.97.

Another finding to emerge was that a considerable amount of inter-subject variability existed in all test conditions, as Figures 3a, b, and c reveal. For example, while JF's DL for pulse trains with a 200-Hz

reference pitch was about .5 Hz, SC's was about 2 Hz; and while MA's DL for monotone sentences with a 125-Hz reference pitch was about 1.5 Hz, NB's was about 5.5 Hz.

III. DISCUSSION

It is quite clear that subjects' ability to discriminate the pitch of pulse trains, monotone sentences, and naturally intoned sentences differed. Moreover, there was considerable inter-subject variability in response accuracy, and there was a systematic difference in the number of "different" responses provided for AX and XA stimuli. Finally, the derived DLs were larger than those previously cited in the literature. It is important to consider why these results emerged.

The most salient difference between the pulse trains and the monotone sentences involved the presence of linguistic information in the latter. Unlike the pulse trains, the monotone sentences conveyed phonetic, syntactic, and semantic information. We conjecture that subjects could not process pitch without also attending to the linguistic properties of the signal--and this "division of labor" resulted in a decrement in discrimination. It is of interest that a study on pitch transcription (Lieberman, 1965) derived related results. In this experiment, stimuli consisted of sentences with natural amplitude and F0 contours and non-sentences (a fixed vowel /a/) with natural (sentential) amplitude and F0 contours. Two linguists transcribed the pitch of these stimuli. Lieberman states, "The linguists' ears were remarkably good so long as they did not hear words of the message" (pp. 49-50). That is, the linguists' transcriptions more nearly approximated the actual pitch track of the

utterances when they heard vowel stimuli than when they heard complete sentences. It seems that listeners find it extremely difficult to attend solely to prosodic information when it is associated with a linguistically meaningful segment, even when the demands of the task require that they do so. An additional example of this is the Stroop color-naming task (e.g. Stroop, 1935; Preston and Lambert, 1969). In such a task, subjects are required to label the color of linguistically congruent and incongruent stimuli. It has been found that subjects' reaction times are significantly slower when they are required to name the color of a color word--e.g., when they must say "red" in response to the word "blue" printed in red ink--than when they are required to name the color of an asterisk or of a neutral word (e.g., "dog"). Similar results have also been obtained in an auditory version of this task in which subjects are required to label the pitch of pitch words--e.g., when they must say "high" in response to the word "low" presented at a high pitch (Vaid and Lambert, 1979; McClain, 1983).

Another possible reason for the large differences in the discrimination of the 3 stimulus types may involve the notion of stimulus uncertainty. In a test of frequency resolution, Watson (1976) demonstrated that, when stimulus uncertainty was minimized, response accuracy increased significantly for complex word-length tonal patterns. He concludes that while "extremely accurate selective attention can be achieved when listeners know what frequencies may bear information and when in time to look for them," it is apparently much more difficult for listeners to discriminate when frequency and temporal patterns cannot be predicted with certainty (p. 185). We may consider the detection of differences in the

frequency of pairs of naturally intoned sentences a relatively extreme example of stimulus uncertainty, for subjects were presented with a range of sentences with frequency changes both within and between sentences. A related explanation for the fact that pitch discrimination was worse for the monotone sentences than for the pulse trains could be the following: Because the sentences were fully elaborated linguistically--and because the F0 of normal sentences varies throughout--subjects may have found it difficult to treat the monotone sentences as if they exhibited no fluctuations in frequency. This may be due to the fact that F0 changes often co-occur with variations in amplitude and/or duration, as in the production of stressed syllables or words (Fry, 1955; Lehiste and Peterson, 1960; Lieberman, 1967; Adams and Munro, 1978). Because both amplitude and duration were intact in the monotone sentences, subjects may have responded as if changes in the fundamental frequency existed, since other acoustic cues commonly associated with pitch were present. To the extent that they expected or believed they experienced pitch variations, they may have been unable to treat these stimuli as sentences which were physically monotonic.

It must also be recognized that, while the pulse trains were uninterrupted for the 2-second duration of the stimuli, the voicing of the sentences was not. For although we attempted to use as many voiced sounds as possible in the sentence stimuli, voiceless segments did occur. This was due in part to the nature of consonantal voicing (i.e., even in consonants with short-lag voice-onset times there may be a brief period of aspiration). If it is the case that a continuous periodic signal enhances pitch detection, it is possible that the presence of non-continuous voicing

in the sentences made discrimination at least slightly more difficult in these stimuli than in the pulse trains.

Our finding that the DL's for naturally intoned sentences were largest is in accord with Klatt's data (1973) on the discrimination of pitch in vowel stimuli with a rapidly falling F0. Our data indicate that the human auditory system has considerable difficulty comparing the pitch of two sentences when that pitch varies throughout and when the overall contours of the sentences are identical. It seems likely that, in the perception of the pitch of normal sentences, the overall pitch contour is a more critical cue than is absolute pitch. And, in his study on the transcription of pitch levels, Lieberman (1965) found that the transcribers "apparently responded to the general form of the [utterance] contour rather than to any pitch levels" (p. 53).

It may be too that the pitch of sentences was not discriminated with great acuity because the sentences were synthetic. As Luce, Feustel, and Pisoni (1983) found, free recall of synthetic test words was significantly worse than that of natural test words. They conclude that "at least some of the observed difficulties in the perception and retention of synthetic speech are clearly due to increased processing demand for these items in short-term memory" (p. 29). How increased processing demands might be involved in the discrimination of pitch in synthetic stimuli is not clear; this is a topic worth further exploration.

Another finding of interest was that in AX pairs (those in which the F0 of the first member was lower than that of the second) stimuli were heard as different much more often than they were in XA pairs--but only in

the intoned sentences. A reasonable explanation is this: In normally intoned declarative sentences, the F0 falls at the end. This means that when the F0 is raised on the second member of a stimulus pair (as it was in our AX presentation), there is a greater difference between the F0 of the last segment of the first member and the first segment of the second member than there is in XA pairs. This is illustrated in Figure 4. It seems that subjects compared the last segment of the first stimulus in each pair to the first segment of the second. This points to the effect of short-term memory constraints for, if subjects had been able to retain (or compute the average of) the fundamental frequency of the entire first sentence, they would not have exhibited such a preponderance of "different" responses to AX stimuli. This finding is also in accord with Brady, House, and Stevens (1961) who found that pitch tended to be identified with the terminal frequency in tones with a ramped F0.

Another notable finding of this experiment was the large degree of inter-subject variability. One explanation for this difference lies in the amount of musical training subjects had had. In the present study, the subject who exhibited the best discrimination overall (MA) was the subject who had also received the most musical training. (She had been, in fact, a singer, violinist, and pianist since childhood.) She was also the only subject whose "different" responses to AX and XA pairs in the intoned sentences were nearly equal in number, suggesting that she had an excellent memory for pitch. And the subject who exhibited the worst overall discrimination (NB) had received virtually no musical training. He also displayed very strong context-dependent effects in his "different"

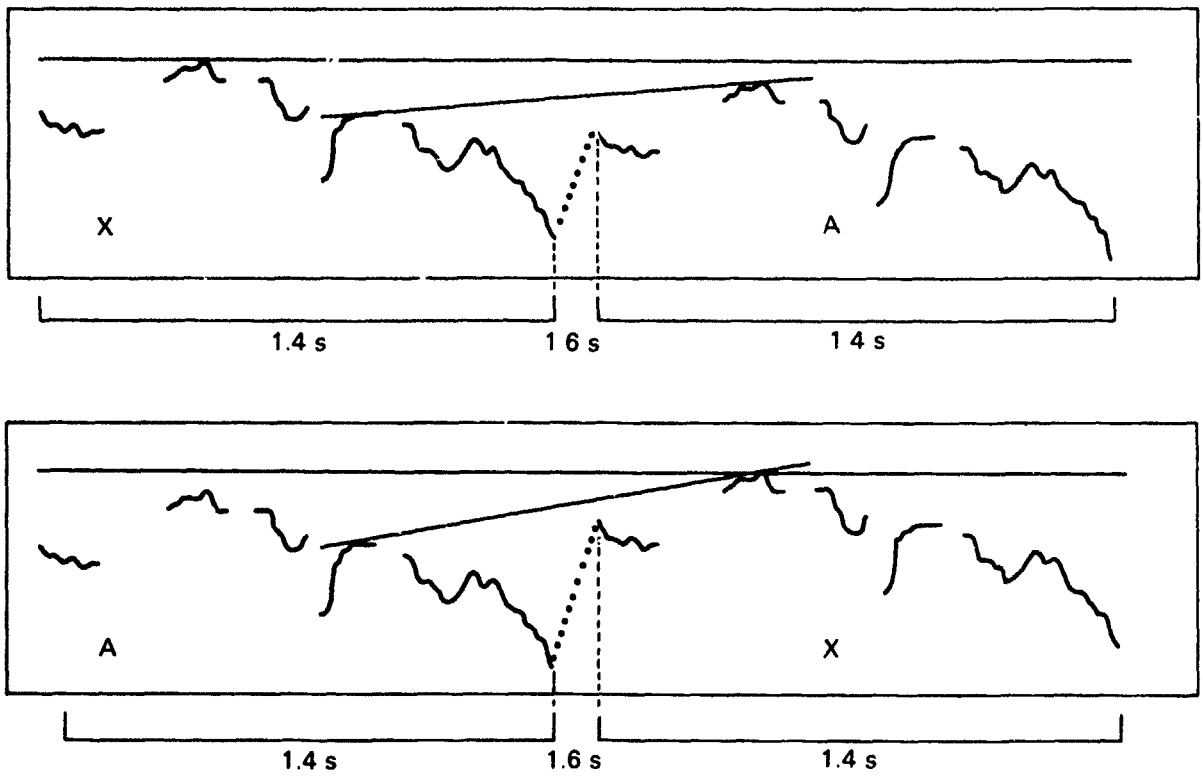


Fig. 4. Fundamental frequency plots of two stimulus pairs whose overall absolute F0s differed by the same amount (6 Hz). One is an XA pair (the first member of the pair is higher in overall F0 than the second) and the other an AX pair (the first member of the pair is lower in overall F0 than the second). Note that, when the final portion of the first member is compared with the initial portion of the second member, the frequency difference is greater in the AX pair, as illustrated by the slope of the diagonal lines and the length of the dotted lines. (This is a plot of an utterance produced by DM. The scale of this display was expanded so that variations in the F0 would be more apparent.)

responses to intoned sentences, for in this condition his average percentage of "different" responses to AX stimuli was 93.55 (and to XA, 6.45).

What is implied here is that there exists a causal relationship between musical training and pitch discrimination. Yet it is possible that those individuals who are inherently predisposed to discriminate pitch accurately are also inclined to pursue musical studies. It must also be acknowledged that other undetermined factors influence pitch discrimination. As the individual response functions clearly reveal, even among those who had received approximately the same amount of musical training there were large differences in discrimination. And it will be recalled that there were 3 individuals who, because they were unable to discriminate to criterion, were excluded from the experiment. It is important to recognize that large individual differences in pitch discrimination do exist; it is also important to attempt to determine the range of such differences.

A final notable result of this experiment involves the size of the DLs which emerged. For even in the pulse train condition, the average DL was much larger than that previously estimated by others for pulse trains (Nordmark, 1968) or steady-state vowels (Flanagan and Saslow, 1957; Klatt, 1973). This may have been due, at least in part, to the method of stimulus presentation and to the properties of the stimuli. For example, in Nordmark's study, subjects adjusted the frequency of an oscillator until its pitch matched the frequency of another oscillator. Moreover, subjects received simultaneous visual input, for they could view the adjustment

knob. (And all of these subjects had received musical training.) In Flanagan and Saslow's study, subjects were tested for over 50 hours, and they received training for approximately 20 hours before final data were collected. Klatt's subjects were likewise tested for many hours (each experiment was repeated 5 times) and they received feedback after each 20-trial block (i.e., they were informed of the number of correct responses they had made). Klatt's stimuli were also only 250 msec long, and they were separated by intervals of 250 msec--durations far shorter than ours. In addition, it is possible that, in the synthetic vowel stimuli, slight but perceptible changes in timbre occurred with changes in the fundamental frequency.

To investigate one possible reason for differences between our results and Klatt's we re-recorded our pulse train stimuli using stimulus durations and intervals identical to his. The responses of two subjects (BG and JT) were tabulated. Their DLs still proved to be considerably larger than those presented by Klatt. (BG's was about 2.5 Hz and JT's was about 1.5 Hz). So conclusive reasons for differences between the DLs in our study and those of others have yet to be established.

Since little work has been conducted on pitch discrimination for sentences, we may surmise that the DLs in the present study were not especially large. Note that the average for Klatt's subjects was 4.0 Hz for a vowel with a steep rate of F0 change. It is also wise to bear in mind that "there is no such thing as a simple, single, differential frequency threshold" (Kling and Riggs, 1971, p. 249). In fact, it could be argued that what is of most interest in the examination of pitch perception

are relative differences in discrimination in various types of speech and non-speech stimuli. And our data did reveal that the subjects' average DL for monotone sentences was twice that of pulse trains, while their average DL for naturally intoned sentences was about twice that of monotone sentences. It should also be stressed that our main purpose in this experiment was to gain insight into pitch deviations in a vocoder communication environment. It is reasonable to assume that the DLs obtained could be interpreted as the lower bounds on acceptable pitch deviations in such environments. In fact, given that our subjects were diligently attempting to discriminate the pitch of adjacent stimuli, they might prove even more tolerant of deviations in a non-test situation not requiring a conscious effort to discriminate. This is not to say that pitch differences in normally intoned speech would go unnoticed if a specific smaller segment (syllable, word, etc.) within the sentence were changed in one of the stimulus pairs, or if the overall contour were somehow anomalous.

Nonetheless, the above results suggest that accuracy of pitch detection in a vocoder is not critical. In fact, if some accuracy is sacrificed in the interests of avoiding gross pitch and voicing errors, the vocoder may gain in user acceptability. Assuming that a relatively high tolerance for pitch changes becomes even higher in an acoustic noise environment, we can go one step further. We can try to design the vocoder pitch detector to minimize gross pitch and voicing errors even if this entails the loss of overall accuracy.

In our future work, we intend to (a) test the hypothesis that listeners have greater tolerance for pitch deviations in an acoustic noise

environment than in a quiet environment, and (b) experiment with methods of partially synthesizing the pitch track in a vocoder when the pitch detector falters. In work such as this, we hope to gain further insight into the nature of human pitch perception given a variety of stimulus types and acoustic environments.

REFERENCES

C. Adams and R.R. Munro, "In Search of the Acoustic Correlates of Stress: Fundamental Frequency, Amplitude, and Duration in the Connected Utterance of Some Native and Non-native Speakers of English," Phonetica 35, 125 (1978).

P.E. Blankenship and V.J. Sferrino, Personal Communication.

E.G. Boring, "The Size of the Differential Limen for Pitch," Am. J. Psychol. 53, 450 (1940).

P.T. Brady, A.S. House, and K.N. Stevens, "Perception of Sounds Characterized by a Rapidly Changing Resonant Frequency," J. Acoust. Soc. Am. 33, 1357 (1961).

J.L. Flanagan and M.G. Saslow, "Pitch Discrimination for Synthetic Vowels," J. Acoust. Soc. Am. 30, 435 (1958).

D.B. Fry, "Duration and Intensity as Physical Correlates of Linguistic Stress," J. Acoust. Soc. Am. 35, 765 (1955).

B. Gold and J. Tierney, "Voccder Analysis Based on Properties of the Human Auditory System," Technical Report 670, Lincoln Laboratory, M.I.T. (22 December 1983),DTIC AD-A138660.

W.L. Hays, Statistics (Holt, Rinehart and Winston, New York, 1981).

H.L.F. von Helmholtz, Die Lehre von den Tonempfindungen als Physiologische Grundlage für die Theorie der Musik (Friedrick Vieweg und Sohn, Braunschweig, 1863).

D.H. Klatt, "Discrimination of Fundamental Frequency Contours in Synthetic Speech: Implications for Models of Pitch Perception," J. Acoust. Soc. Am. 53, 8 (1973).

J.W. Kling and L. Riggs, Wordworth and Schlosberg's Experimental Psychology (Holt, Rinehart and Winston, New York, 1971).

I. Lehiste and G.E. Peterson, "Vowel Amplitude and Phonemic Stress in American English," J. Acoust. Soc. Am. 31, 428 (1959).

P. Lieberman, "On the Acoustic Basis of the Perception of Intonation by Linguists," Word 21, 40 (1965).

P. Lieberman, Intonation, Perception, and Language (The M.I.T. Press, Cambridge, Mass., 1967).

P.A. Luce, T.C. Feustel, and D.B. Pisoni, "Capacity Demands in Short-Term Memory for Synthetic and Natural Speech," Human Factors 25, 17 (1983).

L. McClain, "Stimulus-Response Compatibility Affects Auditory Stroop Interference," Percept. and Psychophys. 33, 266 (1983).

B.C.J. Moore, "Relation Between the Critical Bandwidth and the Frequency-Difference Limen," J. Acoust. Soc. Am. 55, 359 (1974).

J.O. Nordmark, "Mechanisms of Frequency Discrimination," J. Acoust. Soc. Am. 44, 1533 (1968).

M.S. Preston and W.E. Lambert, "Interlingual Interference in a Bilingual Version of the Stroop Color-Word Task," J. Verbal Learning and Verbal Behavior 8, 295 (1969).

I. Raz and J.F. Brandt, "Categorical Perception of Nonspeech Stimuli by Musicians and Nonmusicians," J. Acoust. Soc. Am. 62, S60 (abstract (1977)).

R.J. Ritsma, "Frequencies Dominant in the Perception of the Pitch of Complex Sounds," J. Acoust. Soc, Am. 42, 191 (1967).

W.A. Rosenblith and K.N. Stevens, "On the DL for Frequency," J. Acoust. Soc. Am. 25, 980 (1953).

A. Seebeck, "Beobachtungen uber einige Bedingungen der Entstehen von Tonen," Ann. Phys. Chem. 53, 417 (1841).

N. Stücker, "Über die Unterschiedsempfindlichkeit für Tonhöhen in verschiedenen Tonregionen," Zsch. f. Sinnesphysiol. 42, 392 (1908).

J.R. Stroop, "Studies of Interference in Serial Verbal Reactions," J. Exp. Psych 18, 643 (1935).

J. Vaid and W.E. Lambert, "Differential Cerebral Involvement in the Cognitive Functioning of Bilinguals," Brain and Language 8, 92 (1979).

W.D. Voiers, "Evaluating Processed Speech Using the Diagnostic Rhyme Test," Speech Technology 1, 30 (1983).

C.S. Watson, "Factors in the Discrimination of Word-Length Auditory Patterns," in S.K. Bush, D.H. Eldredge, I.J. Hirsch, and S.R. Silverman (Eds.), Hearing and Davis: Essays Honoring Hallowell Davis (Washington University Press, New York, 1976), pp. 175-189.

R.S. Woodworth and H. Scholsberg, Experimental Psychology (Holt, Rinehart and Winston, New York, 1965).

J. Zatorre, "Recognition of Dichotic Melodies by Musicians and Nonmusicians," Neuropsychologia 17, 607 (1979).

APPENDIX

Test Sentences

A bird lay on the ground.
The moon is far away.
These rings are made of gold.
I loved the big brown dog.
The news was very bad.
They gave the girl a doll.
Weeds grow wild in my yard.
Those boys would never yell.
Their baby will be born.
The man was all alone.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER ESD-TR-84-01	2. GOVT ACCESSION NO. AD-A142996	3. RECIPIENT'S CATALOG NUMBER
4. TITLE *(and Subtitle)* The Discrimination of Pitch in Pulse Trains and Speech		5. TYPE OF REPORT & PERIOD COVERED Technical Report 6. PERFORMING ORG. REPORT NUMBER Technical Report 680
7. AUTHOR*(s)* Molly A. Mack and Bernard Gold		8. CONTRACT OR GRANT NUMBER*(s)* F19628-80-C-0002
9. PERFORMING ORGANIZATION NAME AND ADDRESS Lincoln Laboratory, M.I.T. P.O. Box 73 Lexington, MA 02173-0073		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS Program Element Nos. 33401F and 63735F Project No. 7820
11. CONTROLLING OFFICE NAME AND ADDRESS Air Force Systems Command, USAF Andrews AFB Washington, DC 20331		12. REPORT DATE 12 April 1984 13. NUMBER OF PAGES 36
14. MONITORING AGENCY NAME & ADDRESS *(if different from Controlling Office)* Electronic Systems Division Hanscom AFB, MA 01731		15. SECURITY CLASS. *(of this report)* Unclassified 15a. DECLASSIFICATION DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

None

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

pitch discrimination
vocoders
perception
monotone speech

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

Much research has been conducted on the discrimination of pure and complex tones. Yet relatively little work has been carried out on the discrimination of pitch in speech. Thus, the present experiment was designed to explore listeners' ability to discriminate the pitch of three types of acoustically complex stimuli — pulse trains with monotone pitch, vocoded speech with monotone pitch, and vocoded speech with natural pitch. Results revealed that the discrimination of pulse trains was superior to that of monotone sentences, while the discrimination of naturally intoned sentences was worse than that of the pulse trains or monotone sentences. Implications for speech synthesis and processing are discussed.

DD FORM 1 Jan 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*