Fundamental frequency of phonation and perceived emotional stress

Athanassios Protopapas[a] and Philip Lieberman
Department of Cognitive and Linguistic Sciences, Brown University, Box 1978, Providence, Rhode Island 02912

(Received 27 December 1995; revised 18 November 1996; accepted 22 November 1996)

Nonlinguistic information about the speaker's emotional state is conveyed in spoken utterances by means of several acoustic characteristics and listeners can often reliably identify such information. In this study we investigated the effect of short- and long-term F_0 measures on perceived emotional stress using stimuli synthesized with the LPC coefficients of a steady vowel and varying F_0 tracks. The original F_0 tracks were taken from naturally occurring speech in highly stressful (contingent on terror) and nonstressful conditions. Stimuli with more jitter were rated as sounding more hoarse but not more stressed, i.e., a demonstrably perceptible amount of jitter did not seem to play a role in perceived emotional stress. Reversing the temporal pattern of F_0 did not affect the stress ratings, suggesting that the directionality of variations in F_0 does not convey emotional stress information. Mean and maximum F_0 within an utterance correlated highly with stress ratings, but the range of F_0 did not correlate significantly with the stress ratings, especially after the effect of maximum F_0 was removed in stepwise regression. It is concluded that the range of F_0 *per se* does not contribute to the perception of emotional stress, whereas maximum F_0 constitutes the primary indicator. The observed effects held across several voices that were found to sound natural (three male voices and one of two female ones). An effect of the formant frequencies was also observed in the stimuli with the lowest F_0; it is hypothesized that formant frequency structure dominated the F_0 effect in the one voice that gave discrepant results. *© 1997 Acoustical Society of America.* [S0001-4966(97)02504-6]

PACS numbers: 43.71.Bp, 43.70.Gr, 43.66.Lj, 43.72.Ja [WS]

INTRODUCTION

It is well established that the speech signal carries, in addition to linguistic content, information about the speaker's intentions and emotional state, and that listeners are capable of perceiving this information. The nature of speech production and the human vocal apparatus allow for the encoding of emotional and other nonlinguistic information in several ways. The fundamental frequency of phonation (henceforth F_0) and its prosodic patterns, glottal source characteristics, as well as articulatory details may all be involved in conveying information about the emotional state of the speaker. In fact, previous studies have found correlations with speaker mood or style in all of these (see reviews in Murray and Arnott, 1993; Scherer, 1986). Scherer (1986), reviewing acoustic–phonetic findings on vocal affect, proposed a ''sequence theory of emotional differentiation,'' rooted in the physiology of speech production and taking into account the physiological effects of emotional status. According to Scherer's theory, stimuli are evaluated according to functionally defined criteria, such as ''novelty,'' ''need,'' ''coping potential,'' etc. The net result of the outcomes of all evaluation checks affects the nervous system and, in turn, the physiological consequences of the nervous system's response define the changes in voice characteristics that carry the emotional information. For example, unpleasant stimuli cause ''faucal and pharyngeal constriction and tensing as well as shortening of the vocal tract,'' leading to stronger high-frequency resonances, a rise in the first formant, a fall in the second formant, narrow formant bandwidths, etc. (Scherer, 1986, p. 152).

Beginning with the comprehensive study by Darwin (1872) that outlined the principles of emotional expression independently of will, several different speaker moods or emotions and their vocal consequences have been investigated, including workload (or task-induced) stress (Ruiz *et al.*, 1990; Hecker *et al.*, 1968; Streeter *et al.*, 1983), anxiety (Fuller *et al.*, 1992; Smith, 1977), and simulated emotions such as anger, fear, sorrow, happiness, etc. (Lieberman and Michaels, 1962; Williams and Stevens, 1972; Cummings and Clements, 1995). Extreme levels of stress, in particular those of pilots during (often fatal) inflight emergencies have also been examined, and several F_0-related parameters have been identified as good correlates of stress level (Williams and Stevens, 1969; Kuroda *et al.*, 1976). F_0-related parameters, including short-term perturbations, long-term variability, and mean value, are among the measures often reported to correlate with elevated levels of speaker emotional stress, either task-induced or in real-life emergencies. However, in all of the aforementioned studies it was evident that the acoustic correlates of emotions in the human voice are subject to large individual differences (i.e., among speakers). Streeter *et al.* (1983) concluded that there are no ''reliable and valid acoustic indicators of psychological stress'' (p. 1359).

[a] Present address: Scientific Learning Corp., 417 Montgomery St., Ste. 500, San Francisco, CA 94104; Electronic mail: protopap@scilearn.com

In contrast, there appears to be some regularity in the *perception* of the speakers' emotions based on acoustic parameters, such as F_0. In particular, studies have been conducted to assess the extent to which F_0 measures carry emotional information independently of the speaker's intentions and of the semantic content of an utterance. Lieberman and Michaels (1962) used a fixed-vowel synthesizer driven by natural and smoothed F_0 tracks to investigate identification of emotional content by pitch and amplitude information alone. The original amplitude and F_0 information were taken from utterances spoken in various simulated emotional modes. They found that intact F_0 information, including gross changes and fine temporal structure, was crucial for the correct identification of the original (simulated) emotion. The speech envelope amplitude was found to contribute less to the differentiation between emotional modes. Scherer (1977) used synthesized tone sequences with varying prosodic characteristics to investigate the predictive strength of single acoustic parameters and their interactions in emotional state attribution. He reported ''strong systematic effects of the manipulation of acoustic parameters'' supporting ''a linear model of the judges' response system'' (p. 341). More recently, Scherer *et al.* (1984) used speech degraded by filtering, splicing, or time-reversal and found that F_0 and voice quality ''can convey affective information independently of the verbal context.'' They recommended distinguishing ''linguistic'' and ''paralinguistic'' F_0 features by manipulating acoustic stimuli in a systematic way.

In the present study, we investigated the effects of F_0 measures on perceived emotional stress *in the absence* of verbal content. We employed a method similar to that of Lieberman and Michaels (1962) in that we synthesized fixed-vowel utterances with variations of the F_0 track, and asked listeners to rate their perceived level of stress. In contrast to the study of Lieberman and Michaels (1962), we used original speech taken from a real-life highly stressful situation (i.e., no simulated emotions), we used a single ''stress'' gradient as opposed to several ''emotional modes,'' and we employed more advanced methods for manipulating the F_0-related parameters and for resynthesizing the experimental stimuli, which allowed for better control over the acoustics and more natural-sounding speech. The acoustic parameters of interest were short-term F_0 fluctuations and gross F_0 measures, such as peak values, melodic shape, and range. We conducted experiments with speech synthesized using a constant set of LPC coefficients and varying F_0 tracks. The source and articulatory characteristics were thus kept constant and any perceptual effects could be attributed to the F_0 manipulations.

For our measurements and experiments we used natural speech from a male helicopter pilot. Some utterances were recorded during routine communication with a control tower (unstressful condition) and some were recorded shortly thereafter, when the pilot had lost control of the helicopter and was about to crash (highly stressful condition). The utterances were sampled at 20 kHz using 12-bit linear quantization and the waveform peaks that marked pitch periods were located via a semi-automatic procedure. Temporal resolution in the position of the peaks was increased by quadratic interpolation (as recommended by Titze *et al.*, 1987).

I. JITTER

Period-to-period fluctuations in F_0, known as jitter, are always found in natural speech (Lieberman, 1961), and are known to be more pronounced in cases of pathological conditions such as functional voice disorders (Klingholz and Martin, 1985) and growths on or inflammations of the vocal folds (Lieberman, 1963). The F_0 perturbations have been found to differ among ''emotional modes,'' such as anxiety, fear, anger, etc. (Lieberman and Michaels, 1962; Smith, 1977; Williams and Stevens, 1972), and were predicted to increase in such emotional conditions by Scherer's (1986) model of vocal affect. The empirical status of the reliability of jitter as an emotional indicator remains, however, unresolved. Fuller *et al.* (1992) found *increased* jitter to be an ''indicator of stressor-provoked anxiety [of] excellent validity and reliability'' that is not dependent on individual subjects' ''coping styles.'' They concluded that jitter may be a ''more clinically useful indicator of anxieties'' than other acoustic parameters that may vary with people's coping strategies. In stark contrast, Coster (1986) and Kagan *et al.* (1988) reported that vocal perturbations in children's speech *decreased* with increased stress, and that ''inhibited, compared to the uninhibited, children showed a significantly greater decrease.'' In all, the issues of interpersonal variability and emotional distinctions need to be addressed in more detail before the role of vocal jitter as an affective index can be conclusively established.

A. Jitter analysis

We analyzed unstressed and highly stressed segments of speech (as defined above) using the Average Perturbation Contour (APC) index (proposed by J. Mertus of Brown University) which, for a speech segment containing N pitch periods, is given by the formula

$$\mathrm{APC}_{\alpha}=\frac{1}{N}\sum_{i=1}^{N}\frac{1}{1+\dfrac{\alpha}{(p_i-m_i)^2}},$$

where p_i is the length of the ith pitch period, m_i is the corresponding ''mean'' period that is obtained by smoothing the pitch contour, and α is a weighting constant. This formula is an extension of the Pitch Perturbation Quotient of Davis (1976); the APC gives more weight to larger departures from the smooth contour, but is not thrown off by isolated extreme deviations, because the weighting curve gradually levels off, depending on α. For our measurements, we used α values of 0.10, 0.25, and 0.50. Smoothing was done first with a five-point median, and then using low-pass filtering with a triangular, Hamming, or Savitsky–Golay filter (known to preserve higher momentum; Press *et al.*, 1992, pp. 650ff). Analysis of unstressed and highly stressed speech segments in the same recordings (about 13 s of each) showed that their jitter ranges overlapped completely, the APC ranging between 0.00032 and 0.0057 for the unstressed segments and between 0.00042 and 0.0050 for the stressed segments (depending mostly on the shape and length of the smoothing

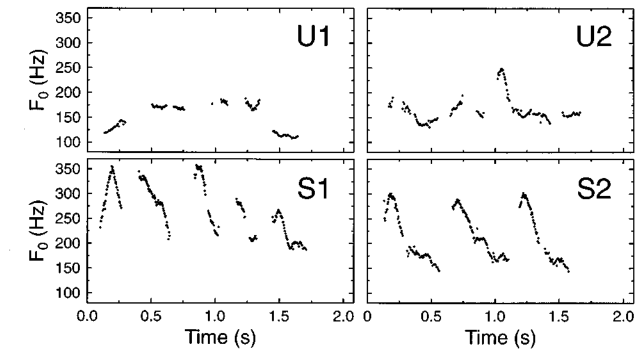

FIG. 1. The F_0 tracks of the four speech segments that were used to synthesize constant-vowel stimuli. U1, U2: unstressed; S1, S2: stressed.

window and on the value of α). Analyses of variance showed that the APC did not differ significantly between unstressed and highly stressed speech [$F(1,76)<1$] for any weighting parameter value of α and for any of the above contour-smoothing windows with lengths between 3 and 15. Identical results were obtained when the instantaneous frequency ($1/p_i$) was used instead of the period and when each period value (p_i) was normalized by the corresponding moving average value (m_i).

From the analysis it appears that, for this speaker, jitter was not an indicator of extreme stress (or terror). Still, it may be that jitter is an indicator of stress in most cases (or other speakers). If this is true, listeners may generally expect jitter to change between various states of stress and, consequently, interpret such changes in their evaluation of the speaker's emotional state. Because of the large individual differences found in vocal indicators of emotion, and because such indicators may result from common underlying sources, it is also possible that jitter may have a perceptual effect only in the context of other acoustic indicators of stress. To test these hypotheses, we presented subjects with F_0 tracks originating from speech produced under the two distinct emotional conditions, in which the jitter was systematically varied but everything else was kept constant.

B. Experiment 1: Perceived stress

The F_0 tracks of two unstressed (U1 and U2) and two stressed (S1 and S2) segments (ranging in length from 1.6 to 2.0 s) were used to synthesize stimuli with varying degrees of jitter. Figure 1 plots the four F_0 tracks that were used. Ten listeners were then asked to rate the stimuli according to the ''emotional stress of the speaker.'' We expected that, if speech with more jitter sounds more ''stressed,'' stimuli with higher degrees of jitter would get higher ratings. If jitter has an effect only in the context of additional acoustic indicators of stress, we should observe a perceptual effect of jitter in the ratings of S1 and S2 variants but not in those of U1 and U2.

1. Method

The four F_0 tracks were smoothed, first with five-point median smoothing and then linearly with a five-point triangular window. The differences, for each pitch period, between each smoothed track and the corresponding original track were then multiplied by 0.0, 0.5, 1.0, 1.5, and 2.0, to create five ''jitter-tracks,'' which were separately added to the smoothed track to create five new F_0 tracks. Thus the smoothed F_0 track plus the 0.0 jitter-track was identical to the smoothed F_0 track, the smoothed plus 1.0 was identical to the original F_0 track of the utterance, and the remaining combinations corresponded to lower (0.5) or higher (1.5, 2.0) degrees of jitter than in the original. Variation of jitter in this manner has the advantage that the spectral distribution of the F_0 perturbations remains constant (and therefore natural) across all degrees of jitter. Using the particular perturbation pattern of each utterance for synthesis also means that, if different ''kinds'' of jitter are somehow present under different emotional conditions, these distinctions are preserved in their acoustic context and will facilitate the desired perception (if jitter has the expected perceptual effect).

A 20-ms segment corresponding to the middle portion of the vowel [a] was excised from the word ''top'' spoken by a male native speaker of American English. The digitized waveform was upsampled to 200 kHz for increased temporal resolution (in particular, for precise control of jitter by means of fine resolution placement of the impulses prior to resynthesis) and analyzed using 200-pole LPC analysis. The analysis program used the autocorrelation method with Durbin's recursive algorithm for solving the LPC equations (Rabiner and Schafer, 1978, pp. 411–413). The resulting coefficients were combined with the jittered F_0 tracks using LPC synthesis to create five (constant–vowel) synthetic stimuli from each of the four original utterances. Synthesis was done by direct implementation of the recursive LPC filter, driven by constant-amplitude impulses. Finally, the stimuli were low-pass filtered with a 1001-tap FIR filter at 9.5 kHz and downsampled to 20 kHz. Calculation of the APC index of the synthesized stimuli indicated that jitter was indeed varying as intended.

In this and in all following experiments, subjects were recruited from the Brown University community (ten for each experiment, ranging in age between 18 and 40 years, mostly undergraduate students) through announcements at local bulletin boards and were paid for their participation. In this experiment, subjects were asked to listen to the synthetic stimuli and were told that ''an 'ah' sound had replaced all the words so [they] could concentrate on the voice and would not be influenced by what had been said.'' Their task was to rate each utterance according to the ''emotional stress'' of the speaker, from 1 (calm) to 7 (very stressed) by pressing the appropriate button on a seven-button response box. The direction of the rating scale, indicated by labels on the response box, was counterbalanced between subjects, and the order of the trials was randomized separately for each participant. Each subject rated each stimulus twice.

2. Results

Listeners did not find it difficult to imagine that real utterances, spoken in different situations of stress, had been ''masked'' with [a] for the purpose of the experiment. Figure 2 (top) shows the ratings for each utterance as a function of relative jitter. Each utterance received different ratings, in accord with its recording situation, but jitter differences did not seem to affect the stress judgments.

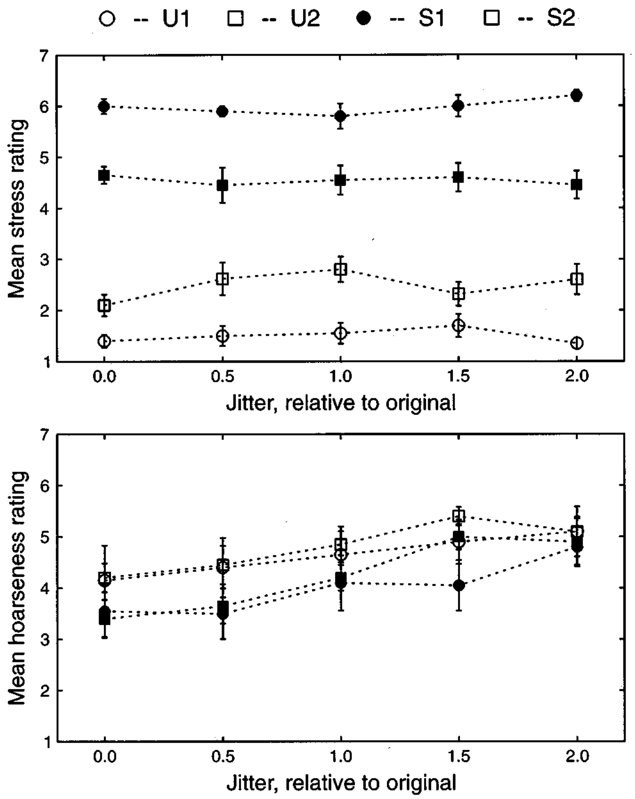

FIG. 2. Mean ratings of speaker's stress (top) and speaker's voice hoarseness (bottom), averaged across subjects, for the four utterances, as a function of relative amount of jitter. The rating scale was 1 to 7; error bars show standard error.

In a 4×5 two-way ANOVA (4 utterances × 5 jitter levels) there was a significant main effect of utterance [$F(3,27)=275.53$, $p<0.00005$], but neither a main effect of jitter [$F(4,36)<1$] nor an interaction between the two [$F(12,108)=1.15$, $p>0.25$]. Thus the four original F_0 tracks indeed reflected very different levels of speaker emotional stress, but the amount of jitter had no effect on the perceived stress level. The average ratings by utterance were (on a scale of 1 to 7) 1.5, 2.5, 6.0, and 4.5 for U1, U2, S1, and S2, respectively.

C. Experiment 2: Perceived hoarseness

In order to rule out the possibility that the null result of experiment 1 was due to a failure of the synthesis method or to other methodological reasons, it was necessary to verify that the jitter differences in the stimuli were perceptible as intended. Since voice hoarseness is known to be a perceptual correlate of jitter (Lieberman, 1963; Muta *et al.*, 1988); particularly in synthesized voices (Hillenbrand, 1988), we conducted an experiment identical to Experiment 1, in which the only difference was in the instructions to the participants: instead of the ''emotional stress of the speaker,'' listeners were now asked to judge the ''hoarseness of the speaker's voice.''

1. Method

The stimuli and procedure for this experiment were identical to those of experiment 1, with the exception of instructions, as described above. Ten subjects from the same population who had not participated in experiment 1 rated the synthesized stimuli for perceived voice hoarseness.

2. Results

Figure 2 (bottom) shows the mean hoarseness ratings, averaged across subjects, for the five levels of jitter. This time the ratings for the four utterances overlapped completely, indicating that the jitter levels were comparable among utterances, as intended, in that the hoarseness ratings were mainly affected by jitter level, equally so for all utterances. However, there is now a strong linear effect of jitter on hoarseness ratings, as expected, that is nearly identical in the four utterances.

Note that the range of hoarseness ratings is relatively small, most of the ratings being around the midpoint of the available scale. Presumably, it would take much more extreme levels of jitter to obtain a mean hoarseness rating closer to 6 or 7.[1] The ''zero jitter'' condition did not give rise, on average, to very low hoarseness ratings (1 or 2) because the ''smooth'' contour is a smoothed version of the original F_0 track and not a perfectly smooth artificial contour. In other words, there is no ''zero jitter'' condition, but only a ''minimal jitter'' condition, relative to the other conditions.

In a 4×5 ANOVA (4 utterances×5 jitter levels) there was no main effect of utterance [$F(3,27)<1$] but there was a significant main effect of jitter [$F(4,36)=11.88$, $p<0.00005$] which did not interact with utterance [$F(12,108)<1$]. Trend analysis of the data indicated that there was a significant linear trend [$F(1,9)=41.85$, $p=0.0001$] that did not interact with utterance ($F<1$), and that there was no quadratic trend ($F<1$). Therefore the jitter differences between the stimuli were perceptible, equally so in all four utterances. In particular, the synthesis method was appropriate in that increasing amounts of jitter led to monotonically increasing hoarseness ratings.

3. Discussion

Our findings indicate that jitter does not affect perceived emotional stress. Experiment 2 clearly showed that the intended jitter gradation was indeed present in our stimuli, so the interpretation of the results of experiment 1 is rather straightforward. However, it must be noted that the type of stress we examined and individual differences in the acoustic correlates of emotional stress may have played an important role. In particular, since jitter was not a factor, in this speaker's voice, that conveyed the emotional distinction under investigation, it is possible that the distribution or some other characteristic of the natural perturbations of his voice was not of the kind that can lead to perception of an utterance as stressed. Alternatively, jitter may be an indicator of other emotional distinctions, as previous studies have suggested, but perhaps not a consistent correlate of extreme stress or terror and thus our subjects ignored it in their interpretation of the stimuli. In particular, jitter may serve to distinguish between states of low level anxiety or task-induced stress, as previous findings have indicated (cf. Fuller *et al.*, 1992;

Coster, 1986). Both of these explanations are compatible with Scherer's (1986) model of vocal affect, since perturbation variations are optional for this emotional condition (''fear/terror,'' pp. 158, 161). On the other hand, our subjects were not instructed as to the kind of stress to pay attention to and had no reason to consider only terror as a stress condition. Nonetheless, it may still be the case that jitter conveys subtle distinctions that were washed out in the context of the extreme F_0 excursions that were present in the recordings from the highly stressed condition. Furthermore, due to the individual differences often found in vocal affect (Hecker *et al.*, 1968) and in vocal jitter measurements in particular (Coster, 1986; Nittrouer *et al.*, 1990), jitter may be too unreliable an indicator to be used by listeners in emotional assessments when the ''normal'' voice of a particular speaker is not known.

II. MELODIC CHARACTERISTICS

In addition to short-term F_0 variability, long-term F_0 measures have also been found to correlate with emotional stress. Scherer (1986) reported in his review that the mean F_0 and the variability of F_0 have been found to increase in situations of fear/terror; his model of vocal affect predicted such changes through the stimulus evaluation checks and their physiological consequences. However, the F_0-related parameters that have been investigated are highly interrelated in natural utterances, and it is not clear whether some of them convey the actual emotional information or whether the whole acoustic constellation is necessary for correct perceptual interpretation. For example, utterances with higher mean F_0 also have higher F_0 range. Does either the high F_0 or the wide range of its values signify a high degree of emotional stress, or is the whole F_0 pattern perceived as a holistic stress indicator?

A. Experiment 3: Perceived stress

In order to investigate the individual contributions of F_0 measures to perceived emotional stress, we examined several parameters. From the F_0 tracks of the original utterances we calculated the mean and maximum F_0, as well as the F_0 range, Max(F_0)$-$Min(F_0), and what we call the ''geometric range,'' Max(F_0)/Min(F_0). S1 and S2 gave higher values in all these measures than U1 and U2, as expected, but the small sample and the relations between these measures precludes conclusions about their relative importance in general. The perceptual effects of each of the F_0-related measures that were found to differ between stressed and unstressed utterances were examined in an experiment using stimuli, synthesized as before, whose F_0 tracks were manipulated to contrast mean, maximum, range, and geometric range of F_0.

1. Method

For each of the four utterances, four F_0 tracks were used: (a) the *original* F_0 track, as measured from the natural speech; (b) the *time-inverted* track, in which the order of pitch periods was the inverse of that in the original, but their length was unchanged; (c) the *scaled* track, in which each pitch period was multiplied by a constant; and (d) the *shifted* one, in which a constant was added to the inverse of each pitch period. Figure 3 illustrates the four manipulation conditions using the U1 utterance. In order to preserve the melodic shape and the duration of the utterance in the scaled and shifted versions, the actual pitch periods that were used in LPC synthesis were calculated by interpolation from the scaled or shifted values, respectively.

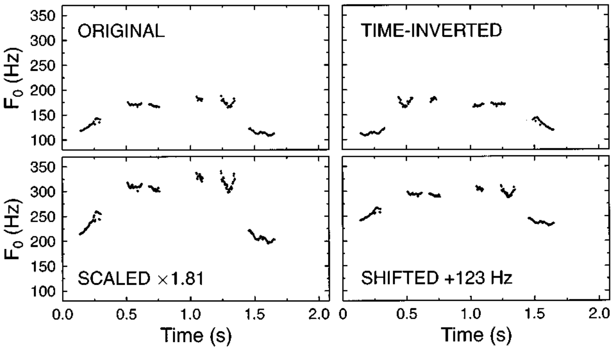

FIG. 3. The F_0 tracks of the four stimuli from experiment 3 that were based on utterance U1.

Each unstressed utterance was paired with a stressed one (U1 with S1 and U2 with S2). The shift and scale constants for each utterance were chosen so that the altered F_0 tracks of one member of each pair resulted in a mean F_0 approximately equal to that of the original F_0 track of the other member of the pair. For example, the pitch periods of U1 were scaled by 1.81 in the scaled condition and shifted by 123 Hz in the shifted condition, the resulting F_0 tracks having a mean F_0 approximately equal to that of S1 (277 Hz). Conversely, the pitch periods of S1 were scaled by 0.55 and shifted by -123 Hz, the resulting mean F_0 being approximately equal to that of U1. Table I shows the F_0 mean, maximum, range, and geometric range of each stimulus. The same LPC coefficients for a male [a] were used as in the previous experiments, and all stimuli were synthesized with jitter equal to that of the corresponding original utterances.

Because it is impossible to completely separate the parameters under investigation, multiple comparisons between the stimuli are necessary. For example, increasing the mean F_0 value to a given value by multiplication and by addition leads to stimuli with matched mean F_0 and different ranges and geometric ranges, respectively. The original and the scaled stimuli are matched in geometric range but differ in F_0 mean and range, whereas the original and the shifted ones are matched in range but differ in F_0 mean and geometric range. Examination of the pattern of results should thus indicate which parameters are most closely related to differences in perceptual judgements of stress.

Ten subjects from the same population who had not participated in the previous experiments rated each stimulus five times, in a procedure identical to that of Experiment 1 (including instructions).

2. Results

Figure 4 shows the ratings of the original and time-inverted stimuli for each utterance. The stress ratings of utterances with time-inverted F_0 tracks were not significantly different from the ratings of the original utterances [$F(1,9)$

TABLE I. The F_0 measurements of the stimuli used in experiment 3. Data for time-inverted stimuli are not shown, as they are identical to those of the original ones. Mean, maximum, and range of F_0 are in Hz, geometric range is a ratio (no units).

	F_0 measurements			
F_0 track	Mean	Maximum	Range	Geometric range
U1				
Original	151.2	188.4	80.0	1.739
Scaled	275.2	340.9	145.4	1.744
Shifted	273.1	311.4	80.4	1.348
U2:				
Original	169.4	248.5	117.6	1.899
Scaled	225.4	331.3	156.8	1.898
Shifted	224.2	303.8	117.4	1.630
S1:				
Original	276.8	355.1	167.6	1.894
Scaled	151.7	198.0	94.3	1.910
Shifted	160.6	232.7	166.8	3.530
S2:				
Original	222.6	302.0	156.7	2.078
Scaled	166.6	226.5	117.8	2.084
Shifted	170.8	247.0	156.9	2.742

<1], and there was no interaction between track-direction and utterance [$F(3,27)<1$]. Therefore, for the stimuli we used, the direction of the melodic patterns (rising versus falling, breath-group slope, etc.) did not affect the perception of stress. In the following analyses the ratings of the time-inverted stimuli were not used, because they were identical to those of the original F_0 tracks (as were also their F_0 measures) and, if used, they would effectively duplicate the corresponding points, thus artificially inflating correlation coefficients.

Figure 5 shows the mean ratings of the stimuli (excluding time-inverted stimuli) plotted against their (a) maximum F_0, (b) mean F_0, (c) range of F_0, and (d) geometric range of F_0. Mean and maximum F_0 correlated well with stress ratings (mean F_0: $r=0.82$, $p=0.001$; maximum F_0: $r=0.89$, $p=.0001$), but range and geometric range of F_0 did not (range: $r=0.51$, $p=0.09$; geom. range: $r=-0.29$, $p=0.37$). In stepwise regression analysis, F_0 range did not correlate significantly with stress ratings after the linear effect of maximum F_0 had been removed (partial $r=0.22$, $p>0.5$) but approached significance after the linear effect of mean F_0 had been removed (partial $r=0.56$, $p=0.056$). The multiple-r correlation coefficient using both mean and range was 0.89, equal to the correlation between ratings and maximum F_0 alone.

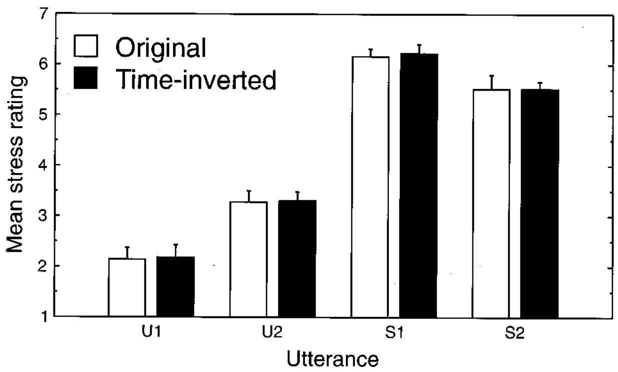

FIG. 4. Stress ratings of the original and time-inverted stimuli for each utterance. Error bars show standard error.

Variants of the stressed utterances received higher ratings than the corresponding (matched) variants of unstressed utterances. Although such differences were generally not quite significant (using Tukey's procedure for *post hoc* pairwise comparisons, as described in Maxwell and Delaney, 1990, pp. 181–184), some aspect of the melodic patterns of stress utterances seemed to have perceptual effects beyond gross statistical measures. For example, the original F_0 track from S2 was rated more stressed than the ''matched'' F_0 track of scaled U2 although the latter had the same mean and range of F_0, higher maximum F_0, and lower geometric range.

3. Discussion

The strong correlation between maximum (and mean) F_0 and the stress ratings comes as no surprise, given previous reports on speech production under various emotional conditions. The lack of a perceptual effect of range and directionality, however, stands in contrast to popular belief that increased F_0 range also conveys such information. Melodic directionality, as defined for our purposes by such parameters as rising versus falling melody and breath-group slope,[2] did not affect perceived emotional stress for any of the four utterances. However, other aspects of the melodic pattern seem to have some influence as mentioned above. Since the exact nature of the salient patterns is not known it is not possible at this stage to systematically vary them in order to investigate them in more detail.

Close inspection of Fig. 5, in conjunction with the results of the regression analysis, leads to the conclusion that mean and maximum F_0 are *the* salient F_0 measures that convey emotional information, at least for the extreme kind of emotional stress that was investigated in this study. Note that, in Fig. 5(a) and (b), ratings of variants of each utterance

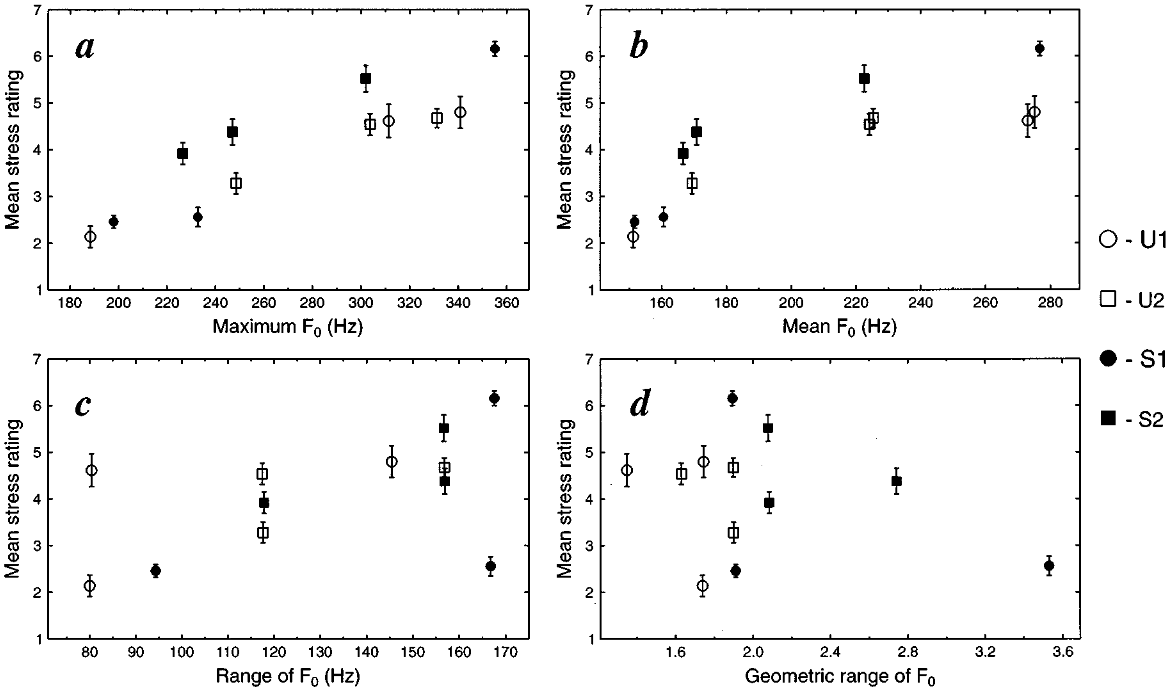

FIG. 5. Stress ratings to the resynthesized stimuli in experiment 3 as a function of (a) maximum F_0, (b) mean F_0, (c) range of F_0, and (d) geometric range of F_0 (excluding ratings to time-inverted stimuli). Refer to Table I for identification of individual stimuli on the basis of their F_0 measures. Error bars show standard error.

(represented by identical markers) lie approximately on straight lines parallel to each other, indicating the gradual, almost linear, effect of maximum and mean F_0 on perceived emotional stress. In contrast, in Fig. 5(c) and (d), ratings of variants of each utterance form right angles with one vertical and one horizontal side, one stimulus pair having almost identical range (or geometric range) but very different ratings, and the other pair having very different range (or geometric range) and almost identical ratings. The apparent weak correlation between F_0 range and the stress ratings is entirely due to the correlation between F_0 range and F_0 maximum (and mean). After removing the linear effect of maximum F_0, there is no other significant correlation. After removing the linear effect of mean F_0, the apparent correlation of the (normalized) stress ratings with F_0 range is an artifact that results from the unequal range of shift and scale of the two utterance pairs. As shown in Fig. 6, the points of each utterance still lie approximately on right angles with a vertical and a horizontal side (except S2), but the higher minimum range of S2 and U2 (compared to that of S1 and U1) combines with their higher ratings to produce a spurious correlation that approaches statistical significance when all points are considered together.

A consideration for the manipulation of F_0 range has been to implement both the arithmetic range, which is calculated by subtraction of the lowest from the highest value, and the geometric range, which is the result of the division of the highest by the lowest value. Although the arithmetic range is the parameter usually examined, the logarithmic nature of human frequency representation might lead one to expect that the geometric mean would correlate better with perceptual effects. Our use of both parameters effectively counters this possibility as it allows for control of each of the two using the other one. It should be clear that no range parameter affects perceived emotional stress, and this finding could be of use to speech synthesis systems, when a high level of stress needs to be conveyed. Apparently, the perceptual system evaluates the *effort* of the speaker, which is higher in order for higher F_0 to be produced (deriving from higher subglottal pressure and laryngeal muscle tension), to assess the degree of emotional stress the speaker is under. The relative importance of maximum F_0 is also evidenced by the fact that high-F_0 variants of unstressed utterances received higher ratings than low-F_0 variants of stressed utterances, i.e., F_0 information was enough to override all other prosodic cues that might have been present in the highly stressful recordings.

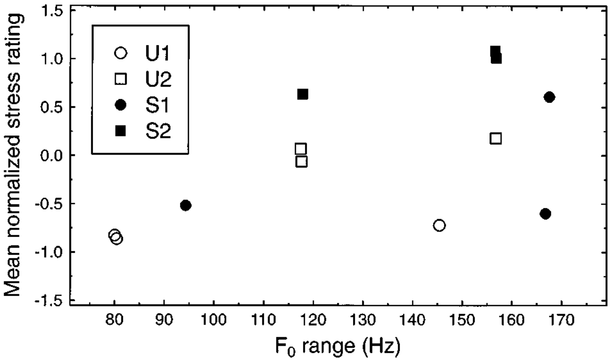

FIG. 6. Normalized stress ratings in experiment 3 as a function of F_0 range, after subtracting the linear effect of mean F_0. (Ratings to time-inverted stimuli are not included.)

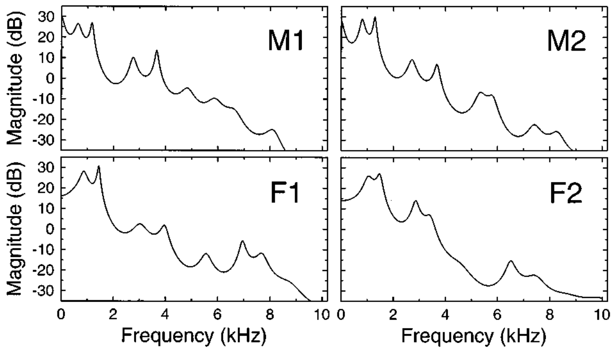

FIG. 7. All-pole power spectra calculated from the LPC coefficients of the four vowels that were used to synthesize the stimuli in experiment 4. M1, M2: male voices; F1, F2: female voices.

B. Experiment 4: Different voices

Given the findings of experiment 3, it was of interest to investigate whether the information conveyed by the F_0 measures we examined varies with voice quality or is speaker independent. Previous studies have identified ''voice quality'' (or timbre) as a primary acoustic carrier of emotional information (Scherer *et al.*, 1984; Scherer, 1986). In the context of the present study, the question is not so much that of distinguishing between different emotions as it is of assessing the degree of a particular emotional state given particular F_0 information. Therefore, it is not of primary importance to systematically examine the effects of acoustic energy distribution but, rather, to establish the F_0 effects in a wide range of voice qualities. To this end, we repeated experiment 3 using four different voices by recording the [a] vowel from four new speakers.

1. Method

Four speakers were recruited from the same population as the listeners, including one relatively large and one relatively small person of each sex (to cover a larger range of formant frequencies). Each was asked to say the word ''top'' and 20 ms of the vowel [a] were excised from its center portion (after digitizing at 20 kHz and upsampling to 200 kHz, as for experiment 3). The four vowels were subjected to 200-pole LPC analysis and each set of parameters that was generated was used in conjunction with the 12 F_0 tracks to synthesize a set of stimuli as in experiment 3 (excluding the time-inverted tracks, which showed no effect). Figure 7 shows the LPC spectra of the four vowels (for the frequencies 0–10 kHz only, since all stimuli were downsampled to 20 kHz after synthesis). For each of the 12 F_0 tracks there were now four versions, labeled M1, M2, F1, and F2, corresponding to the four speakers, bringing the total number of stimuli for this experiment to 48.

The testing procedure and instructions were identical to those for experiment 3. Ten new subjects were recruited from the same population and each one rated each stimulus three times (as opposed to five times in experiment 3, which had quite fewer stimuli) in different random orders. The mean of the three ratings was used for the analysis.

2. Results

With one exception, participants said that they did not find the voices particularly unnatural and that they could imagine utterances spoken with these intonations and voices in various stressful conditions. One listener reported that she found some of the stimuli very unnatural, sounding like a musical instrument. Most listeners correctly identified four distinct voices in the experiment, but three of them thought there were maybe ten or fifteen different voices. These concerns regarding stimulus naturalness will be further addressed below, in experiment 5.

Although the same F_0 tracks were used with all voices, there was a significant effect of voice [$F(3,27)=20.09$, $p<0.0005$], with M1 receiving the lowest ratings (mean 3.78), F2 the highest (mean 4.24), and M2 and F1 intermediate ratings (means 3.88 and 3.89, respectively). Note that this ordering pattern parallels that of the voices' first two formant frequencies (see Fig. 7), which are lowest for M1, highest for F2, and intermediate for M2 and F1. The mean rating for each voice was subtracted from the ratings to all utterances of the same voice, in order to make the correlations meaningful, independently of the voice effect.

Multiple regression analysis of the ratings onto the four predictor F_0 measures gave results similar to those of experiment 3: after subtracting the voice mean from each stimulus, the ratings correlated best with maximum F_0 (partial $r=0.69$, $p<0.00005$) and mean F_0 (partial $r=0.65$, $p<0.00005$), weakly with F_0 range (partial $r=0.37$, $p=0.01$), and not at all with geometric range (partial $r=-0.26$, $p=0.08$). Again, the correlation of the ratings with F_0 range was owed to the interrelation between maximum F_0 and range of F_0 and was not significant after the linear effect of maximum F_0 had been removed (partial $r=0.06$, $p=0.7$). It was, however, weak but still significant after the removal of the linear effect of mean F_0 only (partial $r=0.29$, $p=0.049$), as in experiment 3. The multiple-r correlation coefficient after inclusion of mean F_0 and F_0 range was 0.69, equal to the partial correlation of the normalized stress ratings with maximum F_0 alone. In all, the pattern of results is identical to that of experiment 3 and the same considerations lead us to conclude that F_0 range did not contribute to the perception of emotional stress whereas maximum F_0 is once again the critical parameter.

Additional correlational analyses were performed using the stimuli generated from each voice separately. Table II shows the partial correlation coefficients between the stress ratings of each utterance and the F_0 measures separately for each voice. Note that maximum F_0 correlated most strongly with stress ratings for M1 and M2, followed by mean F_0, whereas maximum F_0 correlated only slightly less strongly than mean F_0 with the stress ratings for F1. The correlation of the stress ratings with range of F_0 approached significance only for M1, and from the pattern of results from M1 we may safely attribute this to the correlation between F_0 maximum and F_0 range, as before. The geometric range of F_0 failed to correlate significantly with the stress ratings of any voice. Surprisingly, none of the F_0 measures correlated significantly with the stress ratings of the stimuli with the F2 voice (which received the highest overall ratings).

TABLE II. Correlation coefficients between the F_0 measures and the stress ratings of the utterances listed separately for each voice. F_0 maximum, mean, and range were measured in Hz, F_0 geometric range is a ratio (no units).

	F_0 measurements			
Voice	Maximum	Mean	Range	Geometric range
M1	0.845[a]	0.782[a]	0.561	0.177
M2	0.763[a]	0.741[b]	0.469	0.159
F1	0.632[c]	0.659[c]	0.109	0.449
F2	0.454	0.353	0.224	0.339

[a] $p<0.005$.
[b] $p<0.01$.
[c] $p<0.05$.

3. Discussion

The stress ratings of the utterances that were synthesized with LPC parameters derived from male voices corroborate the findings of experiment 3 (whose stimuli were also based on a male voice). In addition, there seems to be a correlation between formant frequencies and perceived stress, because stimuli with identical F_0 tracks but higher formants were judged to sound more stressed. The ratings of the female voice stimuli, however, correlate less strongly (F1) or not at all (F2) with F_0 mean and maximum. Possible explanations, other than women's vocal affect being unrelated to F_0, include precedence of voice-specific characteristics and stimulus quality considerations. In particular, it is possible that somehow the voice quality of F2 (and perhaps, to some extent, F1) is such that any utterance sounds equally stressed. If voice quality is a more salient cue for vocal affect, then it may override F_0 measures under certain unknown circumstances that were present in the case of F2. Alternatively, the speech synthesis method may have been inconsistent in that some LPC parameter sets may have led to higher quality (more natural sounding) speech stimuli than others, and this difference may have affected the stress ratings. Since the latter option is much easier to investigate than the former, we examined it in a subsequent experiment. Further discussion of the stress ratings across voices is deferred until the results of the study on the quality of the stimuli for each voice are presented.

C. Experiment 5: Naturalness of stimuli

One issue that needed to be investigated before firm conclusions could be drawn from the results of experiment 4 was whether the stimuli that were given different stress ratings sounded equally natural. It may be the case that more natural stimuli sounded more (or less) stressed than more synthetic-sounding ones, or that the gradual F_0 effects were an artifact of the synthetic character of the stimuli. In particular, we also needed to examine whether the lack of an F_0 effect for the F2 stimuli was a result of that group of stimuli sounding less natural than those of the other voices. The one subject's difficulty imagining real voices with some stimuli suggested that naturalness varied among the stimuli. We therefore conducted an experiment in which subjects were asked to rate the naturalness of the stimuli, and we looked for differences between naturalness ratings that would correlate with differences in F_0 effects on the stress ratings.

1. Method

All the stimuli from experiment 4 were used, along with an equal number of lower quality stimuli that were added in order to create a wider range of naturalness. The new stimuli were created in the exact same way as those for experiment 4, but using 50-pole LPC analysis and synthesis (as opposed to 200-pole LPC for the original ones). This had the effect of maintaining the intonation, intensity, and some of the vowel quality, but giving a clearly synthetic quality to the sound. Thus, subjects could get a better idea of what ''natural'' and ''synthetic'' meant for the purposes of this experiment. Given that even real voices would not be judged to be ''perfectly natural'' if they only said [a] with some intonation, we considered it necessary to make the distinction more salient. It should be noted that what was of interest is not whether our stimuli sounded perfectly natural (which they certainly did not, mainly because people don't generally say ''ah'' with sentential intonation) but, rather, whether there were any correlations between the degree of naturalness and the observed F_0 effects that might render the interpretation of the findings of experiment 4 less meaningful.

Ten new subjects were recruited (from the same population) for this experiment. They were seated in front of a seven-button response box, as before, with the endpoints labeled ''natural'' and ''synthetic.'' Half the participants had ''natural'' at the rightmost end and the other half at the leftmost end. The participants were instructed to rate each stimulus for naturalness on a scale from 1 (natural) to 7 (synthetic), based on whether ''a real person would sound like that if (s)he were to say 'ah' with the same intonation and intensity.'' The order of the stimuli was randomized separately for each participant. Each stimulus was rated twice by each participant in a single session that lasted about 15 min; the mean of the two ratings was used for the analysis.

2. Results

Stimuli synthesized using 50-pole LPC received a mean naturalness rating of 5.7, which was significantly different from the 3.2 mean rating of the stimuli that were synthesized using 200-pole LPC and were used in experiment 4 $[F(1,9)=142.65,\ p<0.0005]$, as expected. The ratings to low-quality stimuli were not considered further in the analysis, as the sole purpose of those stimuli was to expand the naturalness range.

Mean naturalness by voice was 2.88, 2.80, 3.34, and 3.93 for M1, M2, F1, and F2, respectively. Mean naturalness by utterance ranged between 2.91 and 3.78. In a two-way analysis of variance (4 voices×12 utterances) there was a significant effect of voice on the naturalness ratings $[F(3,27)=14.21,\ p<0.0005]$ but no significant effect of utterance $[F(11,99)=1.15,\ p=0.33]$. There was an interaction between voice and utterance $[F(33,297)=1.79,\ p=0.007]$, indicating that the voice effect was different across utterances.

In particular, there was a significant effect of voice (after Bonferonni adjustment, described in Maxwell and Delaney, 1990, pp. 177–180) for the shifted and scaled S1 and S2. Voice pairwise comparisons (with Bonferonni adjustment) using the naturalness ratings of these four utterances only indicated that F2 was rated significantly less natural than either of the other three voices, F1 was judged significantly less natural than either male voice (and more natural than F2), and the two male voices did not differ significantly from each other in naturalness.

3. Discussion

The pattern of the naturalness ratings of the stimuli parallels the strength of the correlation between maximum (and mean) F_0 and the stress ratings: the two male voices sounded more natural and showed a strong linear relationship between maximum F_0 and perceived emotional stress, the F1 stimuli sounded somewhat less natural and their maximum F_0 correlated less well with their stress ratings, and the F2 stimuli sounded the most synthetic and their maximum and mean F_0 did not predict their stress ratings at all.

The synthetic quality of the F2 stimuli makes the evaluation of the results of experiment 4 regarding F2 more difficult. It is not possible to conclude that the observed correlation between maximum F_0 and perceived emotional stress does *not* hold for all voices, because it may well hold for all *natural* voices. The conclusion that this correlation holds for *any* voice is also unwarranted, because we cannot prove that it is the synthetic quality of the F2 stimuli that was responsible for the lack of correlation. However, since the strength of the correlation follows the same pattern as the degree of naturalness of the stimuli, we suspect that stimulus quality is probably the reason that differences in the correlations were found between voices. Because the F2 stimuli that were rated as sounding most unnatural were those with the lowest *minimum* F_0, we suggest that the low-F_0 stimuli were too low in F_0 for women's voices and were thus not perceived as intended. In particular, the low F_0 may have led subjects to interpret the problematic stimuli as male, but the high formants then imposed an interpretation of an abnormally small male or, most likely, a male with a vocal tract shortened by an expression of terror (tightened larynx, mouth wide open, and retracted lips). This may have served as an overriding cue to perceived emotional stress that countered the F_0 effect so that low-F_0 F2 stimuli were perceived as highly stressed, i.e., in the opposite direction from the expected correlation.

III. CONCLUSION

In agreement with previous findings by Scherer *et al.* (1984), we conclude that vocal F_0 carries emotional information independently of the verbal content of an utterance, in fact, even in the absence of verbal content. Lieberman and Michaels (1962) showed in a similar manner that amplitude and F_0 information alone can be utilized by listeners to distinguish between different emotional modes of the speaker, although they found that F_0 perturbations were important for the emotional distinctions and we found no evidence for such a role of jitter. Again, it should be emphasized that we were not concerned with the distinctions between various emotions but with the gradual perception of an undifferentiated ''emotional stress'' which, given the source of our F_0 tracks, was closer to terror than to task-induced anxiety. It is possible that vocal perturbations are a cue to low levels of stress. We believe that the great individual differences found between speakers in jitter studies make jitter an unlikely indicator of emotional state (or stress level), except in cases where a particular speaker's ''normal'' voice is well known, so that departures from it can be reliably evaluated. It appears more promising to concentrate on the diagnostic potential of perturbation measurements, given recent advances in our understanding of voice disorders and in automated voice-analysis systems (Laver *et al.*, 1992).

For all three male voices (one in experiment 3 and two in experiment 4) we found maximum F_0 to be the single best predictor of emotional stress ratings, independently of voice, melodic shape, F_0 range, and jitter. The F_0 range failed to correlate with emotional stress ratings, and it was shown that its often-reported correlation with speaker emotional stress owes to its correlation with maximum F_0. In order to be able to compare the effects of F_0 in different voices we used identical F_0 tracks with male and female voices, resulting, in some cases, in female-voice stimuli that sounded unnatural, possibly due to their very low minimum F_0. Because of the strong and robust correlations found with all male voices, we conclude that maximum F_0 is the most important F_0-related parameter in vocal affect for all voices, and we would expect to find the same pattern with female voices if more appropriate F_0 ranges were used. Although our stimuli were not designed to assess the effects of formant frequencies on perceived emotional stress, our findings indicate that voices with higher formants sound more stressed. The contribution of higher formants in the stress ratings may in fact be quite significant, if the alternative interpretation of the findings with the female voices, particularly F2, is correct. That is, if the formant structure dominated the F_0 effect and caused it to all but disappear.

The fact that high correlations are obtained between some acoustic parameters (here, maximum F_0) and stress ratings is in agreement with the claim of Streeter *et al.* (1983) that ''listeners view certain vocal behaviors as indicative of particular emotional states'' (p. 1359). Consequently, findings on the perceptual role of acoustic parameters in emotional vocalizations can be of practical use in speech synthesis programs, to increase the perceived naturalness, or to convey additional nonlinguistic (emotional) information. The present study clearly shows that the range of F_0, contrary to what is often taken for granted, does not contribute to perceived stress when decorrelated from mean and maximum F_0. This finding is less surprising when one considers the multitude of attention-driving uses of vocalizations with great F_0 excursions, notably including infant-directed speech.

Perhaps more surprising than the lack of an F_0-range effect, reversing the temporal structure of the entire F_0 track resulted in virtually identical ratings of perceived stress. This does not mean that the temporal structure of F_0 variations within an utterance plays no role in conveying emotional information; such an extreme conclusion is unwarranted and probably wrong. It is likely that the speed of change, or other gross temporal characteristics of the F_0 track, affect the perceived emotional state of a speaker. After all, variants of stressed utterances were always rated more stressed (if only slightly) than matched variants of unstressed utterances, therefore something in the overall shape of the F_0 track carries perceptible emotional information. The null result of the time-reversal manipulation suggests that the *direction* of F_0 variations is insignificant with respect to the emotional information of the utterance, and this is a novel and important finding from the point of view of vocal affect research as well as for practical systems of recognition and synthesis of emotional speech.

Since F_0 is also used to convey linguistic (verbal) information, such as lexical stress and some syntactic properties, it is necessary to examine the role of F_0 in normal utterances. Although the present method cannot be dismissed for that purpose, it should be noted that LPC is an inadequate model of most speech sounds, and that it is unlikely to produce natural-sounding stimuli that contain phonemes other than vowels and glides. Consonants, with sound sources not at the glottis, and nasals, with the coupled resonators, are certain to cause problems for the researchers. However, in order to further our understanding of vocal affect, natural utterances with precisely controlled acoustics are necessary. Advanced speech synthesis technology must be used in order to investigate in detail the emotional information conveyed by speech.

ACKNOWLEDGMENTS

This work was supported in part by consortium agreement NCC2-5041 from NASA Ames Research Center to Brown University; portions of it were first reported at the ESCA/NATO Tutorial and Research Workshop on Speech Under Stress, Lisbon, Portugal, September 14–15, 1995. We thank the participants of the workshop, especially R. K. Moore, for their constructive criticism and comments. We also thank David Ascher, Peter Eimas, Laurie Heller, Emily Pickett, Jennifer Utman, and two anonymous reviewers for valuable comments on the manuscript.

[1] Such extreme levels of jitter were judged by the experimenters as sounding overly unnatural in the context of constant–vowel stimuli and were thus not employed in the present experiments.

[2] We did not specifically measure any parameters of directionality. Presumably, inverting the entire F_0 track effectively reversed all directional characteristics.

Coster, W. J. (**1986**). ‘‘Aspects of voice and conversation in behaviorally inhibited and uninhibited children,’’ unpublished Ph.D. dissertation, Harvard University.

Cummings, K. E., and Clements, M. A. (**1995**). ‘‘Analysis of the glottal excitation of emotionally styled and stressed speech,’’ J. Acoust. Soc. Am. **98,** 88–98.

Darwin, C. (**1872**). *The Expression of the Emotions in Man and Animals* (Appleton, London).

Davis, S. B. (**1976**). ‘‘Computer evaluation of laryngeal pathology based on inverse filtering of speech,’’ SCRL Monograph 13, Speech Communications Research Lab, Santa Barbara, CA.

Fuller, B. F., Horii, Y., and Conner, D. A. (**1992**). ‘‘Validity and reliability of nonverbal voice measures as indicators of stressor-provoked anxiety,’’ Res. Nurs. Health **15**, 379–389.

Hecker, M. H. L., Stevens, K. N., von Bismark, G., and Williams, C. E. (**1968**). ‘‘Manifestations of task-induced stress in the acoustic speech signal,’’ J. Acoust. Soc. Am. **44,** 993–1001.

Hillenbrand, J. (**1988**). ‘‘Perception of aperiodicities in synthetically generated voices,’’ J. Acoust. Soc. Am. **83,** 2361–2371.

Kagan, J., Reznick, J. S., and Snidman, N. (**1988**). ‘‘Biological bases of childhood shyness,’’ Science **240,** 167–171.

Klingholz, F., and Martin, F. (**1985**). ‘‘Quantitative spectral evaluation of shimmer and jitter,’’ J. Speech Hear. Res. **28**, 169–174.

Kuroda, I., Fujiwara, O., Okamura, N., and Utsuki, N. (**1976**). ‘‘Method for determining pilot stress through analysis of voice communication,’’ Aviat. Space Environ. Med. **47**, 528–533.

Laver, J., Hiller, S., and Mackenzie Beck, J. (**1992**). ‘‘Acoustic waveform perturbation and voice disorders,’’ J. Voice **6**, 115–125.

Lieberman, P. (**1961**). ‘‘Perturbations in vocal pitch,’’ J. Acoust. Soc. Am. **33,** 597–603.

Lieberman, P. (**1963**). ‘‘Some acoustic measures of the fundamental periodicity of normal and pathologic larynges,’’ J. Acoust. Soc. Am. **35,** 344–353.

Lieberman, P., and Michaels, S. B. (**1962**). ‘‘Some aspects of fundamental frequency and envelope amplitude as related to the emotional content of speech,’’ J. Acoust. Soc. Am. **32,** 922–927.

Maxwell, S. E., and Delaney, H. D. (**1990**). *Designing Experiments and Analyzing Data: A Model Comparison Perspective* (Wadsworth, Belmont, CA).

Murray, I. R., and Arnott, J. L. (**1993**). ‘‘Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion,’’ J. Acoust. Soc. Am. **93,** 1097–1108.

Muta, H., Baer, T., Wagatsuma, K., Muraoka, T., and Fukuda, H. (**1988**). ‘‘A pitch-synchronous analysis of hoarseness in running speech,’’ J. Acoust. Soc. Am. **84,** 1292–1301.

Nittrouer, S., McGowan, R. S., Milenkovic, P. H., and Beehler, D. (**1990**). ‘‘Acoustic measurements of men’s and women’s voices: A study of context effects and covariation,’’ J. Speech Hear. Res. **33**, 761–775.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (**1992**). *Numerical Recipes in C, The Art of Scientific Computing* (Cambridge U. P., Cambridge), 2nd ed., pp. 650–654.

Rabiner, L. R., and Schafer, R. W. (**1978**). *Digital Processing of Speech Signals*, Prentice–Hall signal processing series (Prentice–Hall, Englewood Cliffs, NJ).

Ruiz, R., Legros, C., and Guell, A. (**1990**). ‘‘Voice analysis to predict the psychological or physical state of a speaker,’’ Aviat. Space Environ. Med. **61**, 266–271.

Scherer, K. R. (**1977**). ‘‘Cue utilization in emotional attribution from auditory stimuli,’’ Motivat. Emot. **1**, 331–346.

Scherer, K. R. (**1986**). ‘‘Vocal affect expression: a review and a model for future research,’’ Psychol. Bull. **99,** 143–165.

Scherer, K. R., Ladd, D. R., and Silverman, K. A. (**1984**). ‘‘Vocal cues to speaker affect: testing two models,’’ J. Acoust. Soc. Am. **76,** 1346–1356.

Smith, G. A. (**1977**). ‘‘Voice analysis for the measurement of anxiety,’’ Br. J. Med. Psychol. **50**, 367–373.

Streeter, L. A., Macdonald, N. H., Apple, W., Krauss, R. M., and Galotti, K. M. (**1983**). ‘‘Acoustic and perceptual indicators of emotional stress,’’ J. Acoust. Soc. Am. **73,** 1354–1360.

Titze, I. R., Horii, Y., and Scherer, R. C. (**1987**). ‘‘Some technical considerations in voice perturbation measurements,’’ J. Speech Hear. Res. **30**, 252–260.

Williams, C. E., and Stevens, K. N. (**1969**). ‘‘On determining the emotional state of pilots during flight: an exploratory study,’’ Aerospace Med. **40**, 1369–1372.

Williams, C. E., and Stevens, K. N. (**1972**). ‘‘Emotions and speech: some acoustical correlates,’’ J. Acoust. Soc. Am. **52,** 1238–1250.