In Proceedings of the ESCA/NATO Tutorial and Research Workshop on Speech Under "Stress," (pp. 1–4) Lisbon, Portugal, 14–15 September 1995

Effects of vocal F_0 manipulations on perceived emotional stress

Athanassios Protopapas and Philip Lieberman
protopap@cog.brown.edu, lieberman@cogvax.cog.brown.edu

Department of Cognitive & Linguistic Sciences
Brown University, Providence, RI 02912, U.S.A.

ABSTRACT

In this study we investigated the effect of jitter 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 stressful and not stressful conditions. Stimuli with more jitter were rated as sounding more hoarse but not more stressed. Mean and maximum F_0 within an utterance correlated highly (r>0.8) with stress ratings but range of F_0 did not correlate significantly with the stress ratings, especially after the effect of maximum F_0 was removed by stepwise regression.

1. INTRODUCTION

Vocal indicators of emotional stress have long been the subject of study. Domains of interest range from basic research in the perception of mood and style to speech synthesis and robust recognition by machine. It is well known that the speech signal carries, besides linguistic content, several other kinds of information about the speaker's condition and intentions, and that listeners are naturally capable of perceiving such 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, the glottal source characteristics, and the 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 [5] and [6] for reviews).

In this study we were interested in information about the speaker's emotional state conveyed by short-term F_0 fluctuations and by gross F_0 measures, such as extreme values, melodic shape, and range. We conducted experiments using speech synthesized using a constant set of LPC coefficients and various F_0 tracks. The source and articulatory characteristics were thus kept constant and any perceptual effects could be safely attributed to the F_0 manipulations.

For our measurements and experiments we used natural speech from a helicopter pilot. Some utterances were recorded during routine communication with a control tower (unstressful condition) and some were recorded shortly afterwards, when he 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 resulution in the position of the peaks was increased by quadratic interpolation [8].

2. JITTER

Period-to-period fluctuations in F_0, known as jitter, are always found in natural speech [2], and are known to be more pronounced in some pathological conditions such as growths on or inflammations of the vocal folds [3]. F_0 perturbations have been found to differ between "emotional modes," such as anxiety, fear, anger, etc. [4], [7], [9] and were predicted to increase in such emotional conditions by a model of vocal affect [6].

We analyzed unstressed and highly stressed segments of speech (as defined above) using the Average Perturbation Contour (APC) index, which 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. Analysis of the unstressed and the highly stressed speech segments showed that their jitter ranges overlapped completely. Analyses of variance showed that the APC did not differ significantly between unstressed and highly stressed speech (F(1,76)<1).

2.1 Experiment 1

In the first experiment we investigated the effect of jitter on perceived emotional 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. 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.

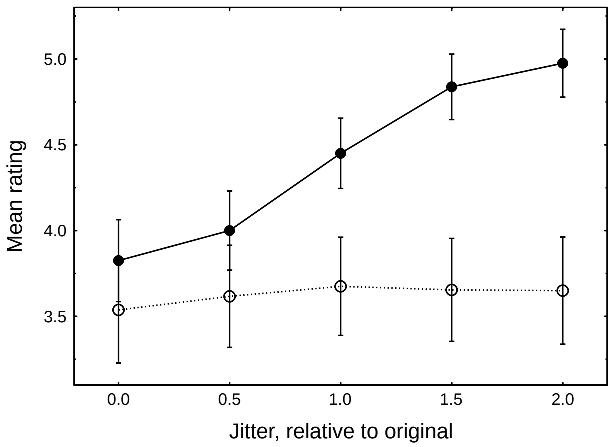

Fig. 1. Mean ratings of speaker's stress (○) and voice hoarseness (●), averaged across subjects and utterances, as a function of relative amount of jitter. The rating scale was 1 to 7; error bars show standard error.

2.1.1 Method

The four F_0 tracks were smoothed, first with 5-point median smoothing and then linearly with a 5-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.

A 20 ms segment corresponding to 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 and analyzed using 200-pole LPC analysis. The resulting coefficients were combined with the jittered F_0 tracks using LPC synthesis to create five synthetic stimuli from each of the four original utterances. Finally, the stimuli were low-pass filtered at 9.5 kHz and downsampled to 20 kHz.

Ten students were asked to listen to some speech where "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. The direction of the rating scale, indicated by labels on the response box, was counterbalanced, and the order of the trials was randomized separately for each participant. Each subject rated each stimulus twice.

2.1.2 Results

Listeners did not find it difficult to imagine that real utterances, spoken in different situtations of stress, had been "masked" with [a] for the purpose of the experiment. 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 from 1 to 7) 1.5, 2.5, 6.0, and 4.5 for U1, U2, S1, and S2, respectively.

2.2 Experiment 2

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, we had 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 [1], 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," ten students (who had not participated in Experiment 1) were now asked to judge the "hoarseness of the speaker's voice."

2.2.1 Results

Figure 1 shows the mean hoarseness ratings, averaged across subjects and F_0-tracks, for the five levels of jitter. The stress ratings from Experiment 1 are also included for comparison. 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

TABLE I

F_0 measurements of the stimuli 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	Geom. 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

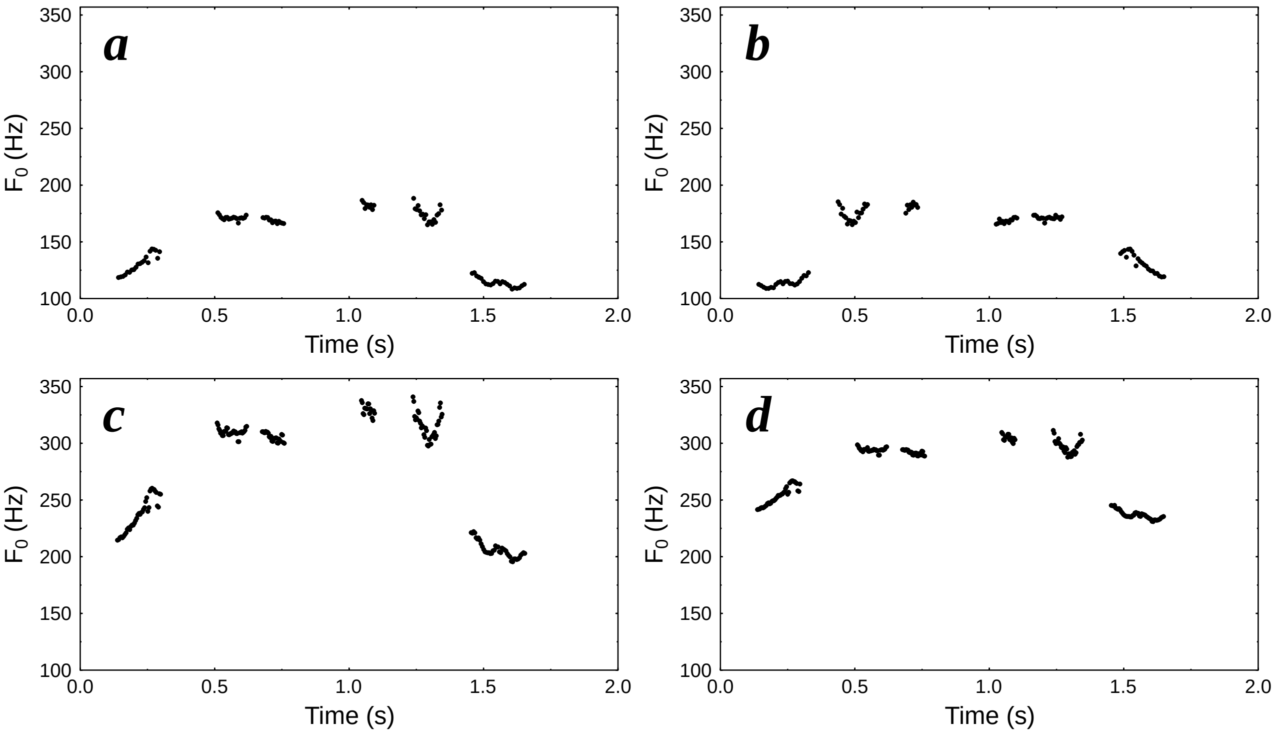

Fig. 2. The F_0 tracks of the four stimuli from Experiment 3 that were based on utterance U1: (a) original, (b) time-inverted, (c) scaled by 1.81 (period scaling by 0.55), and (d) shifted up by 123 Hz.

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 there was no quadratic trend ($F<1$). Therefore the jitter differences between the stimuli were clearly perceptible, equally in all four utterances. In particular, the synthesis method was appropriate in that increasing amounts of jitter led to monotonically increasing hoarseness ratings. We concluded that jitter does not affect perceived stress.

3. MELODIC CHARACTERISTICS

From the F_0 tracks of the original utterances we calculated the mean and maximum F_0, as well as the F_0 range, $\mathrm{Max}(F_0)-\mathrm{Min}(F_0)$, and geometric range, $\mathrm{Max}(F_0)/\mathrm{Min}(F_0)$. 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 safe conclusions.

3.1 Experiment 3

In order to investigate the perceptual effects of each of the F_0-related measures that were found to differ between stressed and unstressed utterances, an experiment was conducted with stimuli, synthesized as before, whose F_0-tracks were manipulated to contrast mean, maximum, and range of F_0.

3.1.1 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 2 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 were calculated by interpolation from the scaled or shifted values, respectively.

Each unstressed utterance was paired with a stressed one. The shift and scale constants were chosen so that the altered F_0-tracks of one member of each pair would result in a mean F_0 approximately equal to that of the original F_0-track of the other member of the pair. 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.

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

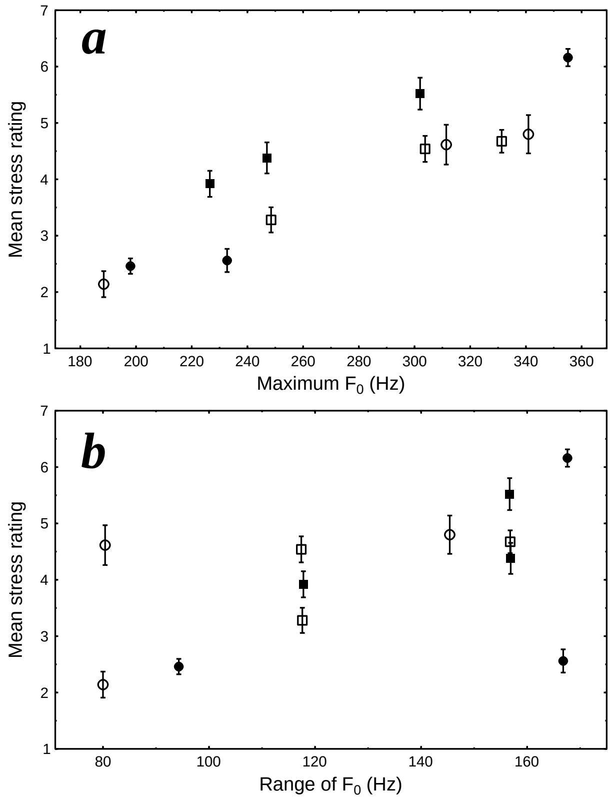

Fig. 3. Stress ratings to variants of U1(○), U2(□), S1(●), and S2(■) in Experiment 3 as a function of (a) maximum F_0 and (b) range of F_0 (excluding ratings to time-inverted stimuli). Refer to Table I for identification of individual stimuli. Error bars show standard error.

3.1.2 Results

The stress ratings of utterances with time-inverted F_0-tracks were not significantly different from the ratings of the original utterances ($F(1,9)<1$), 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 vs. falling, breath-group slope, etc.) did not affect the perception of stress.

Figure 3 shows the mean ratings of the stimuli (excluding time-inverted stimuli) plotted against their (a) maximum value and (b) 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=0.0001), but range and geometric range of F_0 did not (range: r=0.51, p=0.09; geom. range: r=−0.29, p=0.367). 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 (r=0.22, p>0.5).

Variants of the stressed utterances received higher ratings than the corresponding (matched) variants of unstressed utterances. Although such differences were generally not quite significant, some aspect of the melodic patterns of stress utterances seem to have perceptual significance beyond gross statistical measures. For example, the original F_0-track from S2 was rated significantly more stressed than the "matched" F_0-track of scaled U2 ($F(1,9)$=6.48, p=0.031), although the latter had the same mean and range of F_0, higher maximum F_0, and lower geometric range.

4. CONCLUSION

We have examined the statistical measures and perturbations of F_0 and their effects on the perception of the speaker's emotional stress. Jitter and F_0 range, i.e., the variability measures, had no such effect. Mean and maximum F_0 within an utterance correlated significantly with higher stress ratings. Previous findings of higher F_0 range in stressed utterances were presumably counfounded by the correlation of F_0 mean and maximum with F_0 range.

A few important points need to be clarified: First, we do not claim that maximum F_0 alone signals emotional stress in speech. Source spectrum and articulatory characteristics probably carry information at least as important [6]. Second, although we have not asked our subjects about a particular kind of stress, it is obvious that our S1 and S2 utterances were spoken in a situation very different from those often used to induce stress in a laboratory. Thus our findings may not be generalizable to emotions other than extreme stress and terror. Third, since women have generally higher F_0 than men, our results might be taken to imply that women sound more stressed than men, and we would certainly not endorse this implication based on results obtained using a single voice.

Acknowledgements

This work was supported in part by consortium agreement NCC2-5041 from NASA Ames Research Center to Brown University. We thank David Ascher for comments on the manuscript.

5. REFERENCES

[1] J. Hillenbrand. Perception of aperiodicities in synthetically generated voices. *J Acoust Soc Am*, 83(6):2361–2371, 1988.

[2] P. Lieberman. Perturbations in vocal pitch. *J Acoust Soc Am*, 33(5):597–603, 1961.

[3] P. Lieberman. Some acoustic measures of the fundamental periodicity of normal and pathologic larynges. *J Acoust Soc Am*, 35(3):344–353, 1963.

[4] P. Lieberman and S. B. Michaels. Some aspects of fundamental frequency and envelope amplitude as related to the emotional content of speech. *J Acoust Soc Am*, 32(7):922–927, 1962.

[5] I. R. Murray and J. L. Arnott. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *J Acoust Soc Am*, 93(2):1097–1108, 1993.

[6] K. R. Scherer. Vocal affect expression: a review and a model for future research. *Psych Bulletin*, 99(2):143–165, 1986.

[7] G. A. Smith. Voice analysis for the measurement of anxiety. *Br J Med Psychol*, 50:367–373, 1977.

[8] Ingo R. Titze, Yoshiyuki Horii, and Ronald C. Scherer. Some technical considerations in voice perturbation measurements. *J Speech Hear Res*, 30:252–260, 1987.

[9] C. E. Williams and K. N. Stevens. Emotions and speech: some acoustical correlates. *J Acoust Soc Am*, 52(4):1238–1250, 1972.