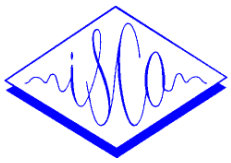

ESCA-NATO Workshop on
Speech under Stress
Lisbon, Portugal
September 14-15, 1995

Analysis of the Acoustic Correlates of Stress from an Operational Aviation Emergency

Peter Benson
ITT Aerospace/Communications Division
10060 Carroll Canyon Road
San Diego, CA 92131
(619) 578-3080
benson@acdca.itt.com

ABSTRACT

The acoustic correlates of psychological stress have largely been examined through artificial means. Task overloads and loud noise have been used to generate changes in a person's voice which, it has been hoped, are similar to actual changes found in an operational environment. Purely artificial approaches such as pretending to be angry have also been used. These approaches offer no guarantee that the measured changes are indeed the changes found in the real situation.

An audio tape of pilot's speech during a serious aircraft malfunction, engine failure of the single-engine F-16 was obtained and analyzed. The tape records speech both before and after the incident. Analyses were made of the linguistic structure of the speech and acoustics. Acoustic measures included pitch, spectral slope and formant frequencies. Additionally, time-warped versions of pre- and post-incident examples of the same words were compared using a cepstral-distance measure, in an effort to determine how stress might effect recognition performance.

[This work was supported by the US Air Force through Rome Laboratory on Contract F30602-88-C-0007].

1. INTRODUCTION

The information about the nature of speech produced under stressful conditions which was acquired has been largely from speech recorded under laboratory conditions. A large portion of this speech demonstrated not real stress but rather simulated stress. It is not clear how much simulated stress speech is similar to real stress speech. The generality then of the results found in the acoustic phonetic characterization of stressed speech is not established. To this end, another source of data was sought.

An audio tape of a serious aircraft malfunction was acquired from Maj. B. J. Stanton of the US Air Force Academy. The tape contains approximately 10 minutes of interchanges among an F-16 pilot, his wingman, and the tower during an in-flight emergency. During the course of normal flight, the pilot noticed his oil pressure dropping. These conditions were life-threatening and might engender the kind of changes in a person's speech commonly associated with stress.

2. WORD AND SENTENCE LEVEL ANALYSIS

The pilot's voice during this incident was remarkably calm. Only on two occasions did he depart audibly from maintaining "an even strain." In all, his voice sounds excited but not extraordinarily so. During the pre-flameout portion of the tape, he speaks in normal sentences and follows the check-list mode of speaking, during which a pilot is called upon to read off the values of various indicators in the cockpit, in a fixed order with set descriptions of the values. Thus, there are fairly long sentences or monologues in the pre-emergency portion where the pilot is reporting the status of items on his checklist.

The perceptible differences between the two sections are that the post-emergency speech is less rambling and more concise. The pilot does not have to search for words, as indicated by the fact that there are no pause or hesitation fillers ("ah","uh", or "oh") in the post phase whereas there are 9 in the pre phase. The average sentence length is much shorter in the post-emergency phase versus pre-emergency phase: 5.6 words versus 3.7 words The number of unique words in the pre-emergency phase is 160; in the post phase it is 53. Thus, there are shorter sentences with fewer words after the emergency.

3. ACOUSTIC PHONETICS

The sound of the pilot's speech after the incident is similar in large part to the excitatory speech described in the Acoustic Phonetic Characterization report [1]. To give quantitaive acoustic phonetic evidence for this, the speech was measured for funda-

mental frequency, amplitude, spectral slope and formant location.

4. PROCEDURES

The speech was received on standard audio cassette. It was filtered, digitized and stored in pcm files. The start and stop points of the pilot's speech were marked by hand. The excerpts between these start and stop points were then assembled into a set of 44 files.

The key question is what acoustic phonetic differences are exist between the pre- and post-emergency phases. We approached this question two ways: we looked at identical parts of words before and after flameout and we looked at the acoustic phonetics across the entire corpus of the pilot's speech. The first way has the benefit of comparing like things analytically; the second uses a large amount of data to get stable estimates.

4.1 Acoustic Measures

Pitch estimates were made using Maximum Likelihood Pitch Estimation(ML) [2]. Spectral slope estimates were made by computing a linear regression line of the log of power spectrum, derived by FFT, between 300Hz and 3300 Hz. The 300 Hz is a good estimate of the location of the first formant and therefore a major energy peak in the lower spectrum. Formant values were obtained via a formant tracking algorithm based on peak-picking and line spectral pairs (LSP). Following Hunt[3], we used a dynamic programming approach for formant tracking, augmenting this basic approach by using the LSP approach for formant assignment, close formants, and formant transitions. Amplitude was measured as the root mean square of the pcm

4.2 Single Word Acoustic Phonetics

First, we identified those words that were spoken both before and after the flameout. Then we chose a subset of these for analysis by excluding the function words. Next we isolated a section of the voiced portion in these words corresponding to the /er/ in *emergency* the /in/ in *engine* and the /nuw/ in *newt.* For the word *emergency,* the slope is less during the stressed speech, but not out of the normal range. Pitch is approximately 40 Hz higher. F1 is higher. F2 and F3 are too variable for comparison. For the word *engine*, the stressed speech slope is within the range of normal. F0 is 30 to 40 Hz higher. F1 is approximately 80 Hz higher. F2 is over 100 Hz higher. F3 is again too variable. For the word *newt*, the stressed speech slope is flatter, F0 is approximately 40 Hz higher, F1 is higher 20 to 220 Hz, F3 is lower. In all of these the difference in power is small, much smaller than would be expected in normal speech, even for the normal speech sections. This lead us to suspect that the speech is being amplitude compressed during some portion of the recording process.

These words show a pattern of acoustic phonetic measures similar to the ones found in the earlier study, i.e., excitement yields increases in pitch and flattening of spectral slope. Additionally, there were relatively large changes in F1 and sometimes F2.

4.3 Gross Acoustic Phonetic Analyses

Since the flameout tape consists of a large amount of relatively poor quality data, an alternative approach to estimating phonetic events was undertaken. In this approach, the speech for the entire set of utterances was measured on a frame-wise basis. The same measures as in the previous section were taken, described below in the section on acoustic measures. Averages of these measures were taken selectively and the results of these averages are presented in Table I. Frames were selected on a figure of merit that measured amplitude and strength of voicing.

The simple observations to be taken from the gross acoustic measures are that

1. the pitch is higher - 115 Hz. for normal versus 163 Hz. for stressed.
2. F1 is approximately 25 Hz higher, 510 Hz. for normal versus 537 Hz. for stressed
3. F2, F3, and F4 are not much different, relative to their absolute values
4. slope is flatter under stress by about 2 dB per octave.
5. power does not vary significantly.

The upward movement of F1 is small and may represent some corruption of the measurement of F1 with F0. The lack of effect for power may be due to the limiting circuits placed on the microphone audio.

These results are very similar to the results from the Acoustic Characterization study. This extensive study of three stress databases presents evidence to suggest that it is reasonable to talk of an *excited* speech quality. The second portion of I shows measurements from the simulated stress conditions from the Lincoln Laboratories Stressed Speech Database [4] are presented. Note especially cond70, which is a high work load task, and lombard, a condition with competing noise. Some differences obtain between these simulated stress conditions and the real stress condition: amplitude is not comparable due to differences in recording, and some systematic differences are occasioned by differences in speaking fundamental frequency and vocal tract length. An inspection of the overall patterns however suggests that the pi-

Flameout Tape						
Type f0	*f1*	*f2*	*f3*	*slope*	*power*	
Normal	115	510	1476	2555	-7.377	65.767
Stressed	163	537	1430	2517	-5.248	65.740
Lincoln Labs Database						
Angry	221	536	1539	2483	-3.39	20.78
Clear	118	447	1572	2483	-5.17	20.83
Cond50	116	451	1549	2435	-5.30	21.39
Cond70	118	456	1562	2444	-5.14	21.51
Fast	117	441	1561	2396	-4.50	1564
Lombard	147	482	1556	2461	-4.61	21.00
Loud	188	524	1540	2482	-3.70	23.04
Normal	113	440	1526	2443	-4.95	18.71
Question	151	434	1520	2475	-5.25	19.00
Slow	109	443	1517	2460	-5.84	24.50
Soft	105	392	1536	2442	-5.47	15.52

TABLE I
Acoustic Phonetic Values for the noirmal and stressed utterances from the pilot's speech 03.

lot's stress is similar to the comparable conditions in a simulated stress database.

5. CEPSTRAL DISTANCES

In making acoustic phonetic measurements, we are serving two purposes. The first is general knowledge; we would like to characterize the effects of stress and other forms of speaker variation for purely scientific purposes. The second is application; we would like to characterize the properties of speech spoken under stressful conditions to be able to position our speech recognition algorithms to accommodate for the changes. In this second purpose, we believe that there is some correlation between phonetic effects and recognizer performance.

A widely-used parameterization of speech for recognition purposes is the mel ceptrum. This representation allows some of the aspects of speech most correlated with stress to be partialed out and removed. That is, by discardiung the first two cepstral parameters, the major efects of amplitude and spectral slope are diminished.

To examine the effects of the stress in the flameout tape through the 'lens' of the cepstrum, a set of distances were computed among the various examples of the words that appear before and after the flameout. The words were the ones noted in the previous section: *emergency, engine and newt.* These words were parameterized into the mel cepstral domain, (i.e., a cepstral representation wherein the spacing of the cepstral estimates is based on the mel frequency scale.) The mel frequency scale is a psychophysical estimate of equal frequency spacing. It is roughly linear up to 1KHz and is logarithmic thereafter. Mel cepstral values have been found to produce the best speech recognition results in normal operating circumstances.

The distances were computed using a dynamic programming algorithm which forced a match between the words so that the hand-marked endpoints were aligned. That is, the frames of the two words which were being compared are matched up so that the minimal distance is found subject to the constraints that there can be no more than a 2:1 change in the time bases of the comparison and that the marked beginnings and endings are fixed. It is of note that the distances are not symmetric: the distance between a and b is not the same as the distance between b and a. This is due to the use of the asymmetric dynamic time warping algorithm, again a choice forced by virtue of its superior performance in automatic speech recognition. The matrices were averaged so that four classes would be apparent: normal compared with normal, normal compared with stress and vice versa, and stressed compared with stressed. These compressed matrices are presented in Table II Note that the 0.0 in the S-S cell for emergency is due to the fact that only one example of the word

Comparing Cepstral Distances without c0 and c1

Word	*N-N*	*N-S*	*S-N*	*S-S*
Emergency	25.98	28.33	29.61	**
Engine	30.83	35.48	36.76	34.65
Newt	37.65	36.92	38.94	37.74

Comparing Cepstral Distances without c0

Word	*N-N*	*N-S*	*S-N*	*S-S*
Emergency	30.68	34.33	32.04	**
Engine	34.72	39.54	39.88	41.85
Newt	51.34	43.48	40.57	40.69

TABLE II

Comparing cepstral distances within and between stressed and normal utterances. The Normal values are from Utterances 01-28; the stressed speech values are from 29-44.

was found in the post-incident phase. The distance between a word and itself is 0.0. This value should therefore be disregarded.

There are three major observations to be drawn from these compressed matrices. First, there is only a small difference between the normal-normal and the mixed stressed-normal and normal-stressed distances. This suggests that normal speech is a good model for this example of operational stressed speech and could make good templates for this speech.

Second, following the adage that there are many ways to be wrong and few ways to be right, one might expect that the stressed speech would differ from one another in many different ways and have larger distances than the normal. Given that the ordering of the magnitude of differences seems to be

$$normal - normal < (normal - stressed, stressed - normal) < stressed - stressed$$

stressed would appear to have greater variability than normal.

Third, there are no large differences in distance scores within the three matrices. The distances between the stressed-stressed speech is only slightly higher than the distances between normal-normal. Here either the stressed speech does not differ much or the ways in which it differs are not apparent through the cepstral transformation. We re-did these computations but this time the cepstral component associated with spectral slope was kept in. The result was to increase distances. in an almost linear fashion in most cases: there is a 4 point increase in relevant cells for emergency and engine, a 3-6 point increase for 3 cells of the compressed scores for *newt*. The increase for the Normal-Normal cell seems due to the large difference in slope for the two words in this cell (Cf. Table II). The slope measures in this table is only for one vowel in the word, but could serve as an indicator of the overall slope changes. Thus, there are stable increases when c_1 is re-added to the parameterization which seem only to increase all the distances linearly.

6. GENERAL CONCLUSIONS ABOUT THE FLAMEOUT TAPE

The flameout tape is a good example of speech spoken under stressful conditions. The characteristics of that speech are principally that the speech becomes simpler with shorter sentences and fewer hesitation phenomena. Acoustically, the major difference is that pitch and slope change, in much the same way as they changed in the simulated stress conditions in the LLDB and the ITT Vocal Effort databases. Comparisons based on recognizer distance show that stressed speech is further from normal speech than other normal speech, but also that there can be some large variation even among normal speech.

7. REFERENCES

[1] Peter Benson, *Acoustic Phonetic Characterization of Three Stressed Speech Databases*, RADC, Contract F30602-88-C-0007, June 1989.

[2] W. J. Hess, *Pitch Determination of Speech Signals-Algorithms and Devices*, Springer-Verlag, Berlin, 1983.

[3] M. J. Hunt, *A Robust Formant-Based Speech Spectrum Comparison Measure*, Proc. ICASSP'85, pp. 1117-1120.

[4] E. A. Martin, R. P. Lippmann and D. B. Paul, *Multi-style Training for Robust Isolated-Word Speech Recognition*, Proc. ICASSP'87, Vol. 2, pp. 705-708.