AFRL-HE-BR-TR-2003-0159

United States Air Force Research Laboratory

FATIGUE ASSESSMENT THROUGH VOICE ANALYSIS

Joel MacAuslan

Speech Technology and Applied Research Corp.
4 Militia Drive
Lexington, MA 02421

HUMAN EFFECTIVENESS DIRECTORATE
BIOSCIENCES AND PROTECTION DIVISION
FATIGUE COUNTERMEASURES BRANCH
2504 GILLINGHAM DRIVE
BROOKS CITY-BASE TX 78235-5100

February 2004

Distribution authorized to U. S. Government agencies only (Proprietary information, Dec. 03). Other requests for this document shall be referred to AFRL/HEDM, 8315 Hawks Road. Brooks City-Base, TX 78235.

WARNING – This document contains technical data whose export is restricted by the Arms Export Control Act (22 U.S.C. 2751, et seq.) or the Export Administration Act of 1979, as amended (Title 50 U.S.C. 2401, et seq.) Violating these export-control laws is subject you to severe criminal penalties. Dissemination of this document is controlled under DoD Directive 5230.25.

NOTICES

This report is published in the interest of scientific and technical information exchange and does not constitute approval or disapproval of its ideas or findings.

This report is published as received and has not been edited by the publication staff of the Air Force Research Laboratory.

Using Government drawings, specifications, or other data included in this document for any purpose other than Government-related procurement does not in any way obligate the US Government. The fact that the Government formulated or supplied the drawings, specifications, or other data, does not license the holder or any other person or corporation, or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report has been reviewed and is approved for publication.

//SIGNED//

JEFFERY N. WHITMORE
Contract Monitor

//SIGNED//

THOMAS L. CROPPER, Col, USAF, BSC
Chief, Biosciences and Protection Division

DESTRUCTION NOTICE: Destroy by any method that will prevent disclosure of contents or reconstruction of the document.

WARNING-This document contains technical data whose export is restricted by the Arms Export Control Act (22 U.S.C. 2751, et seq.) or the Export Administration Act of 1979, as amended (Title 50 U.S.C. 2401, et seq.) Violating these export-control laws is subject to severe criminal penalties. Dissemination of this document is controlled under DoD Directive 5230.25.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*
February 2004

2. REPORT TYPE
Final

3. DATES COVERED *(From - To)*
01 June 2002 – 28 February 2003

4. TITLE AND SUBTITLE
Fatigue Assessment Through Voice Analysis

5a. CONTRACT NUMBER
F33615-02-M-6057

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
65502F

6. AUTHOR(S)
Joel MacAuslan

5d. PROJECT NUMBER
3005

5e. TASK NUMBER
HP

5f. WORK UNIT NUMBER
07

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Speech Technology and Applied Research Corp
4 Militia Drive
Lexington MA 02421

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Human Effectiveness Directorate
Biosciences and Protection Division
Fatigue Countermeasures Branch
2504 Gillingham Drive
Brooks City-Base, TX 78235-5100

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/HE

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-HE-BR-TR-2003-0159

12. DISTRIBUTION / AVAILABILITY STATEMENT
Distribution authorized to US Government agencies only, (proprietary information, Dec 03). Other requests for this document shall be referred to AFRL/HEDM, 8315 Hawks Road Brooks City-Base Texas 78235-5147.

13. SUPPLEMENTARY NOTES

14. ABSTRACT
The primary objective of this research project is to design a system to accurately assess the human operator's fitness for duty via voice analysis. Ideally, such a system will be non-obtrusive, real-time, and viable across a variety of platforms, thus facilitating automated assessment in field locations ranging from the cockpit to the automobile. Such a system will permit maximally convient monitoring and maximally broad markets, and the opportunity to assess an opponent's fatigue state. It is unlikely that a single combination of speech-related measures will be appropriate for all tasks. The Phase I effort evaluated a wide range of voice characteristics to distinguish between rested and fatigued states. Repeated landmark-based speech measures appear useful for subject-dependent fatigue assessment from 5 to 10 second samples. Speech non-fluencies in 3 minute samples appear to identify fatigue independent of the speaker. Automating the analysis and developing a suitable user interface for a deployable system are Phase II goals.

15. SUBJECT TERMS
SBIR, fatigue, alertness, voice, speech

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Maj Brandon Doan
Unclass	Unclass	Unclass	Unclass	45	19b. TELEPHONE NUMBER *(include area code)* (210) 536-8129

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

1 Executive Summary

The Phase I project on Fatigue Assessment through Voice Analysis is now complete. This Final Report details the project itself and its results.

We begin the report with a summary of the problem and its significance and an explanation of the analysis techniques used. We then review the Work Plan and the changes necessitated by the complications with the fatigue data. Finally we detail the results of the project and recommend avenues for continuation and further research and development. Analyses based on landmarks and on non-fluencies appear especially promising. Some of the other analyses that were expected to show fatigue effects, such as roughness of voice, did not. We provide some speculations about these absences, the implications for analysis in realistic environments, and the conditions under which these analyses might be useful.

Due to the time-constraints imposed by late receipt of the fatigue data, and the difference between the actual data and the fatigue measures we proposed for Phase I, we were unable to produce a Phase II proposal. However, we completed Phase I using the available data and suggest directions in which this work might be usefully continued.

2 Significance of the Problem

2.1 Problem

Modern military operations require individuals to handle a multitude of dynamic challenges in high-pressure situations. Pilots, for instance, must operate on the edge of human performance in intense, sustained mission environments. Likewise, command and control operators must make rapid situation assessments, fuse data from multiple sources, and provide dynamic battle management during time-critical situations. Indeed, in such environments there is little room for loss of alertness (due to fatigue or overload) that might lead to performance lapses, errors of judgment, system breakdowns, or mission failures.

Loss of alertness can have equally devastating consequences in areas such as surface transportation, commercial flight, and medicine. Each year in the United States there are approximately 56,000 motor vehicle crashes related to drowsiness or fatigue, as cited in police reports. These crashes result in an estimated 40,000 nonfatal injuries and 1500 fatal injuries each year; in fact, these values are likely to underestimate the actual magnitude of the problem [National Highway Traffic Safety Administration, 1999]. Similarly, the issues of extended workdays and around-the-clock operations in a variety of settings have led researchers to focus on human performance in light of human circadian rhythms, time-of-day effects, and work schedules [Tepas, Paley, & Popkin, 1997]. As a result, there are established links between these variables and changes in human fatigue and alertness, and reviews of the literature suggest that safety decrements are associated with these changes (e.g., Duchon & Smith [1993]).

Given the critical importance of fatigue-induced changes in human performance, there exists a significant need to develop a system to accurately assess the human operator's fitness for duty. Ideally, such a system should be non-obtrusive and viable across a variety of platforms, thus facilitating automated assessment in field locations ranging from the cockpit to the automobile. Accordingly, the primary objective of this research project is to design such an assessment system, and in particular, to explore the viability of voice and speech analysis as a method of fatigue assessment.[1]

2.2 Opportunity

There has been scant research on the topic of fatigue (or alertness) assessment from analysis of speech, but what there is suggests that there may be a solution. Like ethanol and hypoxia, fatigue causes changes ranging from the cellular to the cognitive and behavioral. Some of these are evident in (acoustic) speech signals. This is fortunate because collecting speech signals, by reciting words in a set protocol, for instance, is generally non-intrusive. In some situations, it can even be completely passive, if it relies on speech that occurs for unrelated reasons; *e.g.*, status reporting. Such cases permit not only maximally convenient monitoring — and maximally broad markets — but also the opportunity to assess an opponent's fatigue state.

[1] Throughout this document, *voice* refers to sound generated in the larynx (both normal voicing and whispering or the phoneme /h/). *Speech* refers to the articulatory patterns, both temporal and spectral, that are imposed on this, primarily by the mouth.

For speech signals, fatigue-related changes can be manifest:

- acoustically, e.g. as for example, roughness of voice;
- phonetically, e.g. as imprecise articulation (*ataxia*, slowness and inaccuracy in the force, range, timing, and direction of speech movements); and
- stylistically, e.g. as inappropriate vocabulary, phrasing (prosody), and even reasoning.

This project focused on the first two levels, acoustic and phonetic. This avoided many of the complications of speech recognition and deep inferencing, including the confounding influences of emotional state and communication goals, which may be expressed through such subtleties as the connotation of words.

In the acoustic domain, several acoustic measures have been studied. Most may be characterized as measures of "voice", especially those based on the fundamental frequency (F0) of the laryngeal vibrations during speech [Whitmore & Fisher, 1996] — the objective correlate of pitch. Several phonetic measures of speech have also been studied for effects of fatigue, intoxication, or ataxia, such as syllable or word duration [Fell *et al.*, 2001; Whitmore & Fisher, 1996].

2.3 General measurement and assessment issues:

In keeping with the ultimate goal of an assessment tool, this project will focus on automated, objective measures. A useful assessment system must address several ancillary issues: confusion and data-collection protocol, signal quality, confounding influences, and the nature of the task, among others. First, most measures are sensitive to "confusion". The presence of other sounds in the recording, especially multiple speakers, is one such confusion. Assessment reliability is generally higher for speech recorded with a fixed, well-structured protocol, particularly one that minimizes extraneous sounds and excludes all other speakers. It is important to understand the trade-offs of intrusiveness versus reliability, because many desirable applications do not permit highly constrained protocols. Determining the nature and size of this effect, however, has been largely deferred to Phase II.

Each of the other issues has its own impact on the project or on the resulting product. For example, some of the proposed speech-related measures, such as articulatory precision (see below), can be more sensitive to bandwidth and other issues of signal quality than others, but may yield better correspondence with fatigue. Phase I addressed this straightforward but widespread issue. Confounding influences are those that can produce cognitive and/or vocal changes unrelated to fatigue; these include ethanol intoxication and hypoxia (cognitive) and even antihistamine use (sometimes cognitive [Eddy *et al.*, 1992] and often laryngeal, due to drying of tissues). The problems of confounding influences have been deferred to Phase II. Finally, the nature of the task determines the nature of the fatigue measure(s) that the system must predict: reaction time versus reasoning skills, for instance. It is unlikely that a single combination of the speech-related measures will be equally appropriate for all tasks; however, given the nature of the expected data, we did not study task-based prediction in Phase I.

3 Overview of Phase I

Primary deliverables for Phase I are speech and fatigue measures that show reliable associations, with an assessment of which measures of each are appropriate for future use in a system to gauge fatigue. A further deliverable is a description of the architecture and requirements of such a system, with due regard for the operator tasks, intrusiveness of signal collection, etc., appropriate to its use.

3.1 Phase I Scientific Approach

Several speech-analysis techniques are appropriate to the cognitive level chosen here; i.e., acoustics, vocal control, and speech articulation (rather than the more subtle factors such as semantics). We focus here on the techniques most investigated during Phase I.

3.1.1 Fundamental Frequency

Fundamental frequency (F0) represents a most basic measure of vocal behavior. It reflects the lowest frequency component in the acoustic spectrum of speech and derives from the rate of vibration of the vocal folds ("vocal cords") during speech. It can be measured as the reciprocal of the period of the acoustic wave.

Changes in fundamental frequency during speech are perceived as changes in the pitch of the voice, and are mainly effected by three variables: the mass of the vocal folds, the velocity of air-flow through the vocal folds, and the tension of the laryngeal muscles. While mass is largely a function of body size and gender, airflow velocity and laryngeal muscle tension are minutely controlled and coordinated millisecond by millisecond. Typically, raising F0 requires more energy than lowering F0. Thus, measures such as a decrease in long-term pitch average (LTAP), and/or overall pitch range, have typically been used to chart overall level of tiredness [Whitmore & Fisher, 1996] and lack of newness [Cooper and Sorenson, 1978], as well as boredom, depression, etc. [Bachorowski, 1999]. An increase in long-term F0 average and/or pitch range has been shown to accompany heightened emotional states such as anger [Bachorowski, 1999], as well as to mark the introduction of new information. In particular, Bachorowski [1999] suggests that F0 may act as a marker of cognitive arousal level. In the typical speech context, when most utterances are intended to provide information or to comment, variation in LTAP is derived from changes in pitch range. Thus, pitch range and LTAP are related measures, though pitch range is considered to be more sensitive and LTAP more robust.

LTAP is typically calculated over several sentences. Pitch range, on the other hand, is valid as a measure when calculated over any sentence or phrase of speech that reflects a single utterance. We proposed to calculate both LTAP and pitch range for each speaker over all utterances per fatigue condition.

Hollien *et al.* [2001] found LTAP to be a reliable correlate of ethanol intoxication. These researchers recorded a standard paragraph used in speech research and therapy, the Rainbow Passage, plus 30 sec of conversational speech, for measurements of F0. Subjects repeated these speech tasks at various (precisely calibrated) intoxication levels. For each testing session, F0

was measured and averaged over the length of the speech samples. The mean and standard deviation were computed for each subject at each testing session.

3.1.2 Landmarks

In order to detect where syllable onsets occur and where speech sounds begin and end, the speech signal must be scanned for acoustic *landmarks*. They mark perceptual foci and articulatory targets. Landmarks are points in an utterance around which one may extract information about the underlying distinctive phonetic features [Stevens *et al.*, 1992]. One type of landmark is linked to glottal activity and can be used to identify vocalic segments of the speech signal. Other landmarks identify intervals of sonorancy; i.e., intervals when the oral cavity is relatively unconstricted. The most common landmarks are acoustically abrupt and are associated with consonantal segments; e.g., stop closure and release.

3.1.2.1 Modified Landmarks (MLM)

MLM is syllable-detection software based on work by Liu [1995] to detect landmarks in adult speech. It was developed as part of a speech-feature recognition system. Liu's work is founded on Stevens' acoustic model of speech production [Stevens *et al.*, 1992].

MLM first sends speech input through a general processing stage in which a spectrogram is computed and divided into several frequency bands. An energy waveform is constructed in each of the bands, the time derivative of the energy computed, and peaks in this derivative are located. These peaks represent times of abrupt spectral change in the six bands. In type-specific processing, the localized peaks direct processing to find three types of landmarks:

1. **g**(lottis). Marks the time when the vocal folds transition from freely vibrating (+g) to not vibrating (-g) or vice-versa.
2. **s**(onorant). Marks sonorant consonantal releases (+s) and closures (-s).
3. **b**(urst). Designates stop or affricate bursts (+b) and points where aspiration or frication ends (-b) due to a stop closure.

Because MLM detects when the glottis (laryngeal opening) begins to vibrate and when it stops, it is a good syllable detector. To do this, a syllable is defined as a (+g/-g) interval ignoring voicing breaks shorter than 20 ms (about four glottal pulses). The length of the interval is the syllable length. The mean of all of the syllable lengths in the utterance gives a measure of average syllable length, and the standard deviation gives us a measure of variability. Finally, counting the number of syllables in the utterance (the interval of time between the first (+g) and the last (-g)) yields a measure of rate in syllables per second. Figure 1, below, illustrates the output of syllables detection on normal adult speech.

3.1.2.2 Landmark-Based Classification (LMC)

LMC is phonetic-feature-classification software, also built on Liu's work [Liu, 1995] to detect landmarks in adult speech and is founded on Stevens' acoustic model of speech production [Stevens et al., 1992]. It was not used in Phase I, but we anticipate that it would be highly useful in further work. In fact, ongoing research is designed to extend its capability to analyze dysfluent speech, which our results (4.7) show is most promising for fatigue assessment.

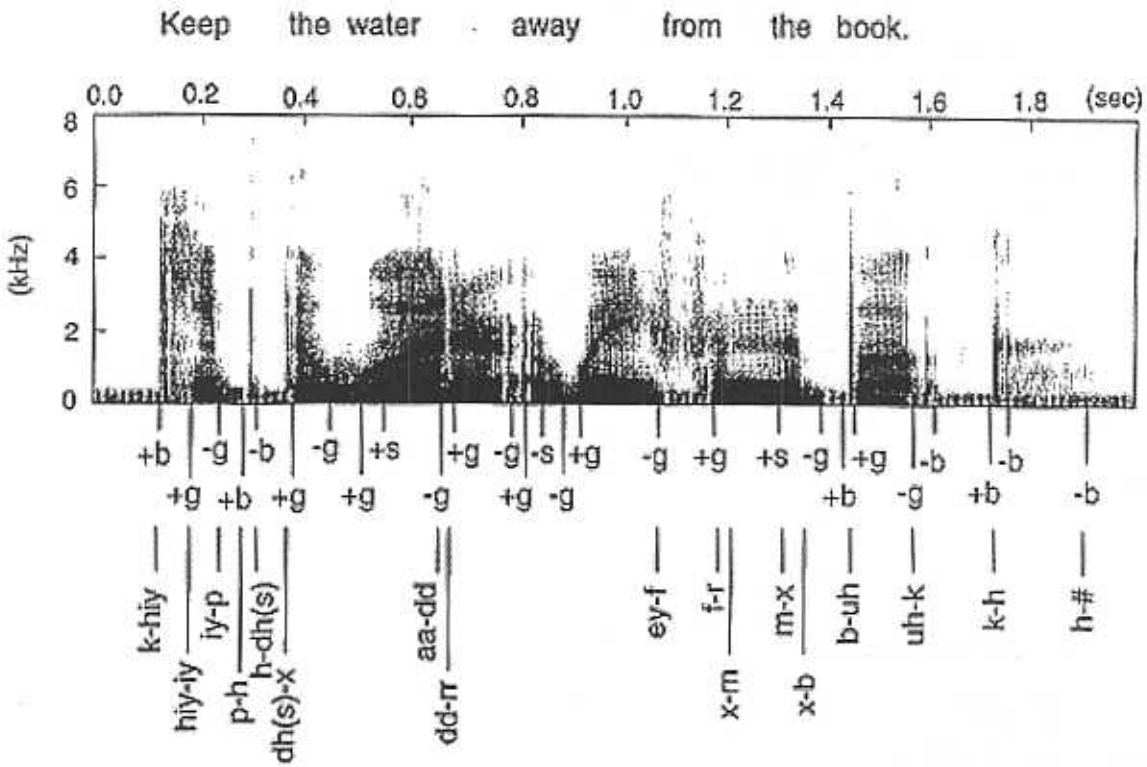


Figure 1. Spectrogram with landmark detection in adult speech (Liu-Stevens)

Bitar and Espy-Wilson [1996] applied landmark theory, described briefly in the previous section, and developed a knowledge-based signal representation based on phonetic features, which allows segmentation of speech into 10 speech classes, or "classifiers", built on top of the landmarks. Bitar and Espy-Wilson's signal representation consists of acoustic parameters that are measures of the speech signal or its time-frequency representation; these parameters provide evidence for the acoustic correlates of phonetic features. An example of a simple acoustic parameter is the measure of energy in the 100-400 Hz band which can be used to capture the signature of the feature `sonorant`.

To date, acoustic parameters used in the LMC have been developed for the manner-of-articulation phonetic features `sonorant`, `syllabic`, `fricative`, and `consonantal`; the place-of-articulation features `labial`, `alveolar`, and `velar` for stops; and `palatal` and `alveolar` for strident fricatives. Some of the acoustic events, such as the ones associated with the phonetic feature `sonorant`, segment the speech signal into intervals. Others, such as those associated with `nonsyllabic`, mark particular instants in time[2]. From these acoustic parameters, Bitar and Espy-Wilson detect 10 speech classes from the physical signal: `syllabic` (vowels, syllabic nasals ["hid-d'n"], and syllabic liquids ["lit-tle", "her"]), `sonorant consonant` (semivowels and nasals), `palatal fricative`, `alveolar fricative`, `affricate`, `labial stop`, `alveolar stop`, `velar stop`, `weak fricative`, and `silence`.

The robustness of the acoustic events based on acoustic parameters has been illustrated in a series of recognition experiments [Bitar & Espy-Wilson, 1996]. Results show that, compared to

[2] Instead of *intervals*.

traditional speech-recognition processing (cepstral, with a modern Hidden Markov Model), the acoustic parameters (a) better target the linguistic information in the speech signal, and (b) reduce interspeaker variability.

These attributes need not be relevant for assessment of fatigue, which affects non-linguistic components of speech. However, based on the results found in Phase I (below), reducing interspeaker variability may be particularly valuable for a fatigue assessment system that relies on little or no prior speech data from the speaker being assessed.

3.1.3 Other Basic Signal Processing

Multipurpose signal analysis packages can detect the acoustic features underlying perceptual features like low pitch. For example, the program "get_f0" [Entropic, 1993] looks for periodicity to calculate fundamental frequency in voiced regions. It returns not only a probability of voicing but also a pitch track for these regions, which can be analyzed further.

3.1.4 Perturbation Measures: Jitter and Shimmer

Jitter is short-term (cycle-to-cycle) variability in fundamental frequency F0 (or in its reciprocal, the period). [Titze, 1995] A considerable literature exists to support the usefulness of jitter measurements in evaluating laryngeal pathology, which suggests its usefulness in detecting fatigue as well. Jitter may be measured from running speech or a sustained vowel, though the latter is preferable. If a sustained vowel is not available, then systematic changes in F0 (for stress or intonation, for example) must be eliminated from the jitter measurements. In either case it can easily be measured using commercially available equipment such as the Computerized Speech Laboratory (CSL™) and Multi-Dimensional Voice Profile (MDVP™) from Kay Elemetrics. Because smaller absolute amounts of cycle-to-cycle variation are expected from voices of lower F0, jitter is generally reported as a *jitter ratio* or *jitter factor*, where the amount of variation is relative to the F0. If jitter is useful as a measure of fatigue, we would expect it to increase with increasing fatigue.

Shimmer, on the other hand, is short-term amplitude perturbation. It is based on the peak amplitude of each vibratory cycle and is at least as important as or more important than jitter is to the perception of the vocal quality of hoarseness. Shimmer is also easily measured with CSL. However, as with any intensity-based measure, mouth-to-microphone distance must be very carefully controlled: this may make the measurement of shimmer impractical for pre-existing data. A calibrating signal of known amplitude is also useful to provide a reference to standard or absolute intensity values.

Each of these has variations and extensions. For example, tremor is a periodic variation in pitch. When present, it produces a high value of jitter. However, it also produces other perturbation features with parameters of their own, such as the period (in sec) and "amplitude" (in Hz) of the variations. The Kay Elemetrics systems routinely measure all of these.

Although these perturbation measures are popularly associated with fatigue, it has been unclear whether the association is justified. Caffeine consumption can cause these measures to rise, as can fatigue of the laryngeal muscles (*vocal* fatigue). Neither of these need be associated with

cognitive fatigue. In fact, the analyses reported later (4.6) indicate that the perturbation measures are uncorrelated with cognitive fatigue.

3.1.5 Harmonics-to-Noise Ratio

Another part of the perception of vocal hoarseness or roughness is the replacement of harmonic energy by broadband energy, or "noise". Harmonics-to-noise ratio (HNR) evaluates the energy in the periodic part of the speech signal, i.e., the harmonics of F0, compared to the remainder, i.e., between the harmonics. There are many algorithms for computing the HNR of a voice signal; they vary according to the frequency bands evaluated and the methods for computing the harmonics. This makes HNR a flexible measure, one which could be used for band-limited speech such as telephone or radio speech. The CSL speech-analysis environment provides a measure of HNR. The same reservations apply to HNR as to the perturbation measures (above): It may measure vocal fatigue instead of cognitive fatigue.

3.1.6 Nonlinear dynamics ("chaos")

The techniques of nonlinear dynamics or "chaos theory" have recently been shown useful [Titze *et al.*, 1993; MacAuslan *et al.*, 1997] in analyzing voice signals, the signals generated by laryngeal dynamics. Fundamental to such analysis are the notions [Peitgen et al., 1992] of:

1. *state space*, a collection of interacting variables, such as components' locations and momenta, that provide a description of the system and of its evolution. The evolution depends *only* on the instantaneous values of these variables; i.e., position in this state space and on external "environmental" constants like muscular tension and lung pressure;
2. *orbits*, or (smooth) paths in state space along which a system evolves, perhaps repetitively;[3] and
3. *attractors*, regions of state space which a given orbit fills, after any initial (transient) behavior of the system has died away.

A fundamental feature of nonlinear dynamical systems is the following remarkable fact: if a system depends on N interacting variables and virtually *any one* of them ("X") is sampled in time at regular intervals to provide a sequence of data $<X_1, X_2, \ldots>$, then a "copy" of the state space may be reconstructed by forming "delay vectors" $\mathbf{V}_k$ having (at most) 2N+1 elements:

$$\mathbf{V}_k = <X_k, X_{k+1}, X_{k+2}, \ldots, X_{k+2N}>.$$

A sequence of these vectors forms an orbit. We could even plot a picture of such an orbit in two or three dimensions, if N were sufficiently small. The orbit fills out an attractor which is, in a sense, *dynamically equivalent* to the original attractor: generally warped but having the same dimensionality, the same topology, the same structure of the orbits, and even similar measures of distance between corresponding neighboring points.

This last item is particularly useful. It means that if we measure the larynx by some scalar function of time, such as instantaneous pressure (after filtering to remove formants, the resonances of the mouth), then we may construct a simple picture of the structure of *repetitions of the larynx's*

[3] Indeed, an orbit must therefore never intersect itself, unless the system is precisely periodic. This holds because the path leading from the intersection -- the subsequent evolution of the system -- depends *only* on the state-space position of the intersection point itself, not on any other aspect of the system's history.

states [MacAuslan *et al.*, 1997]. These repetitions need not be perfect; thus, the laryngeal attractors need not be precisely periodic, although this is approximately the usual case.[4]

In these pictures, "good" periodicity, whether perfect or not, appears as a nearly horizontal stripe: at every measurement time T (x-axis), the state of the larynx approximately repeats after some fixed interval or period P (y-axis). The brightness of a point reflects the closeness of the repetition. If environmental constants, such as airflow and muscular tension, remain fixed, then the state will repeat again at 2P, at 3P, etc., producing more bright stripes, vertically stacked at equal intervals P.

Such a picture can show immediately the instant at which the larynx changes period or even becomes non-periodic. The horizontal stripes rise or fall across the picture as the repetition interval rises or falls (as the pitch *falls* or *rises*, respectively), or even end altogether as periodicity stops. Thus, stacked "wavy" horizontal stripes are characteristic of vocal tremor; and discontinuous changes in the stripes' height reflect pitch breaks. (The parameters of these features, e.g., the instant and magnitude of a pitch break, are readily computed and do not necessitate displaying the picture itself.)

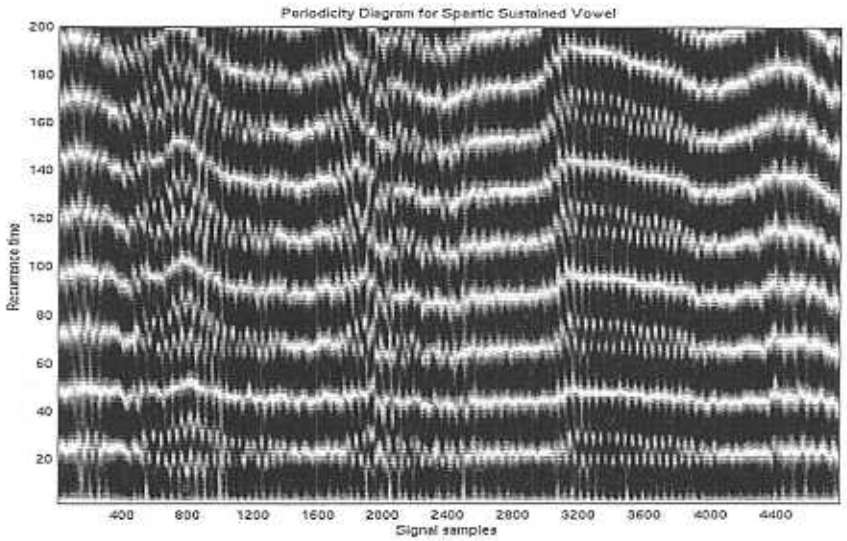


Figure 2: Periodicity Diagram for a Disordered (Spastic) Sustained Vowel. In this picture, periodicity appears as a nearly horizontal stripe. At every measurement time T (horizontal axis), the state of the larynx approximately repeats after some fixed interval or period P (vertical axis): one sample on each axis represents ½ msec. The brightness of a point in one of these pictures reflects the closeness of the repetition. The patterns near 800 and 3200 samples show that the true recurrence time (height of the first bright stripe from the bottom of the picture) has suddenly doubled. Such a period-doubling (octave jump) indicates a qualitative change in the system's dynamics.

[4] Intriguingly, mathematically *chaotic* laryngeal motion, in which states *never* repeat, also shows "periodicity" structure. The reason is that attractors in such cases are filled with periodic orbits. Although the actual orbit of the system cannot in practice be periodic, it always finds itself near some periodic orbit, and thus nearly repeats for some number of times. Gradually, the two diverge and the system finds itself near some other periodic orbit having a generally different period.

If the larynx passes through one cycle of values and then, in the same period, through another that is similar but *not* identical, before returning to a repetition of the first cycle, then the picture will show this as vertically alternating bright and less-bright horizontal lines. Such "period-doubling" or "octave-jump" behavior by the larynx is dynamically significant. Depending on the attractor, the laryngeal system may pass through several near-repetitions before a true repetition, producing "period-three", "period-four", etc., orbits. In these cases the perceptual pitch, which often sounds constant, differs from the dynamically significant fundamental frequency.

Dynamical changes that alter the true period by large factors (as measured by the system state) often correspond to imperceptible changes in the voice. Thus, time-series analysis such as this can detect changes in laryngeal function that listeners cannot. Furthermore, the dynamics are controlled by the muscles of the larynx; this underlies the hypothesis that fatigue will affect the dynamics, producing pictures (and corresponding numerical measures) with qualitative differences from non-fatigued voices.

3.1.7 Speech Rate

People speak more slowly when tired. Slower speech rate consists not only in an increase in pause duration between utterances, but also in longer word durations and in different ratios of duration between vowel-like sounds and consonants [Whitmore & Fisher, 1996; Wayland *et al.*, 1994; Bachorowski, 1999].

Hollien *et al.* [2001] used duration of a recited passage as one of their indices of ethanol intoxication. They calculated the time it took each subject to read the Rainbow Passage and to complete 15 repetitions each of three *diadochokinetic* tasks, i.e., ones characterized by alternating vowels and single consonants. The diadochokinetic tasks included repeating the nonsense words "pataka", "buttercup", and "shapupie" as many times and as accurately and quickly as possible on one breath. Because these words involve sequential changes of place and manner of articulation, they are useful tasks from which to assess changes in speech over time. At each testing session, Hollien's subjects performed each diadochokinetic task with each word (separately) fifteen times. Therefore, it was possible to determine the average duration of each nonsense word for each subject.

Syllable duration is another measure of speaking rate and is generally measured in terms of the length of a vowel in a stressed or isolated syllable. For example, Fell *et al.* [2001] constructed a landmark-based software system that detects the onset and offset of laryngeal vibration, signal amplitude, and other features to measure syllable length automatically. (This system was used in the results presented below.) Change in speaking rate, as reflected in syllable duration, is also a common clinical measure of speech change, whether improvement due to therapy or degradation due to disease.

This system measures duration of voiced regions in running speech and is fairly robust in the face of noise. Sudden percussive noise will generally not affect its ability to detect voicing, nor will moderate levels of white noise. This makes it a good candidate for use in adverse recording conditions.

3.1.8 Articulatory Precision

By articulatory precision, we mean the accuracy of vocal tract geometry that a speaker achieves during speech. Many measures are available. One, which is linked in the literature to boredom and lack of effort, is vowel centralization. When vowels are centralized, the acoustic distinctions between vowels are lessened as a consequence of a smaller range of movement by muscles of the tongue. This variable is typically tracked by measuring the first three resonant frequencies (or *formants*) of the vocal tract for the vowels "iy" (as in "beat"), "aa" (as in "bot"), "uw" (as in "boot"), and "ae" (as in "bat) in stressed monosyllabic words, and plotting formant 1 (F1) vs. formant 2 (F2), F2 vs. F3, and so on. Other vowels may be added as well. Stressed vowels are used because they are easiest to measure and reflect the speaker's best attempt to be intelligible. However, because stressed vowels are most necessary to intelligibility, they may show effects of fatigue more slowly than unstressed syllables. We therefore measured the above vowels, plus additional English vowels and diphthongs, in unstressed as well as stressed contexts.

3.1.9 Non-fluencies

Non-fluencies were assessed in Hollien *et al.* [2001] by two trained phoneticians on the Rainbow Passage, on sentences from a standardized test of articulation, and on conversational speech. Nine types of non-fluencies were identified:

1. Phoneme or word additions
2. Hesitations or pauses
3. Repetitions
4. Phoneme or word substitutions
5. Voicing substitutions
6. Devoicing substitutions
7. Phoneme or word omissions
8. Distortions
9. Lengthening.

In Phase I, the most reliably noted qualitative features in this list were evaluated: #1, 3, 4, and 7.[5] In part, this reflects the inherently complex, "fuzzy" decisions needed to decide whether, e.g., the degree of distortion (#8) or lengthening (#9) of a phoneme is large enough to be notable. As for #2, the DCIEM subjects, unlike Hollien's, were not reciting text, they were explaining a map; as a result, it was difficult (and perhaps even inappropriate) to decide whether a pause reflected a non-fluency or normal "think time". Such decisions require substantial data-driven analyses of their own, beyond the scope of Phase I.

3.2 Phase I Technical Objectives

To validate speech/voice analysis as a fatigue assessment tool, predictions generated from measures of speech and voice must be correlated to standard measures of fatigue and alertness. <u>This is the primary technical objective.</u>

[5] The contractor has direct experience using the landmark-based classifier LMC with *automatic* detection of certain non-fluencies, both qualitative and quantitative (specifically, those for stuttering). This could be important for an automated Voice-based Fatigue Assessment System, though it was not used here.

In a significant change from the proposed plan, the recorded speech of the DCIEM Map Task corpus was substituted for a protocol containing direct measures of fatigue. The corpus represents the careful control of many variables that confound fatigue,[6] but it provides only three dialogues from each subject, each a few minutes long, over a 64-hour sleepless interval. See Pigeau *et al.* (1995). Cognitive performance data for the subjects are available for the full interval but were not included with the speech corpus. Consequently, they have not been correlated with the speech measures evaluated in Phase I. Instead, the speech measures were correlated with the simple proxy, hours of continuous activity without sleep. (The corpus also includes matched subjects using stimulants; their dialogues were not analyzed in Phase I.)

A secondary aim is to learn which of the many available speech-based measures predict fatigue. The goal is to avoid using speech-based measures that are poor or redundant predictors, that limit future system options (such as intrusiveness) without offsetting advantages, or that are unlikely to be useful in the expanded range of speakers and environments of Phase II and beyond. For example, Phase I was restricted to English-speaking subjects, an undesirable limitation. Phase I also focused on relatively high-bandwidth data (20 kHz), another such limitation. Furthermore, the DCIEM data come from a relatively stable, low-noise environment, an unrealistic assumption for many of the intended applications or markets.

From these results, Phase I also provided a third objective: a description of the architecture and requirements of an assessment system. Several technical issues will arise that are important for the system, although minimally dependent on the processing. These concern the target environment(s), the degree of intrusiveness of the signal collection, the training and speed required of a human operator for the system, and the need for quantitative vs. qualitative predictions. Discussions with customer suggest that an initial version of the system may have more sophisticated controls and displays, for the benefit of fatigue specialists determining its capabilities and limits, than later versions that could be deployed for wider use.

3.3 Phase I Work Plan

Establishing feasibility consisted of three stages. First, we extracted several speech measures (discussed below) that were expected to show fatigue effects. Second, we determined the available fatigue data in consultation with the customer. Third, we performed a comparative study between the retained speech measures and the available fatigue data. Fourth, we produced a document (Section 5) describing the procedure and remaining steps to design the proposed system architecture in consultation with the customer.

The Phase I Work Plan originally called for investigation of linear (quantitative) correlations between speech measures and fatigue measures. However, we only received data consisting of dialogues recorded at three occasions in a 64-hour sleepless period, with no associated fatigue measures. Using time as a direct proxy for these fatigue measures is not fully appropriate, because circadian-rhythm influences cause time to be nonlinearly related to performance measures of fatigue. See Figure 3.

[6] Perhaps too successfully: As discussed later, some otherwise measurable behavioral correlates of fatigue may have been suppressed by the protocol.

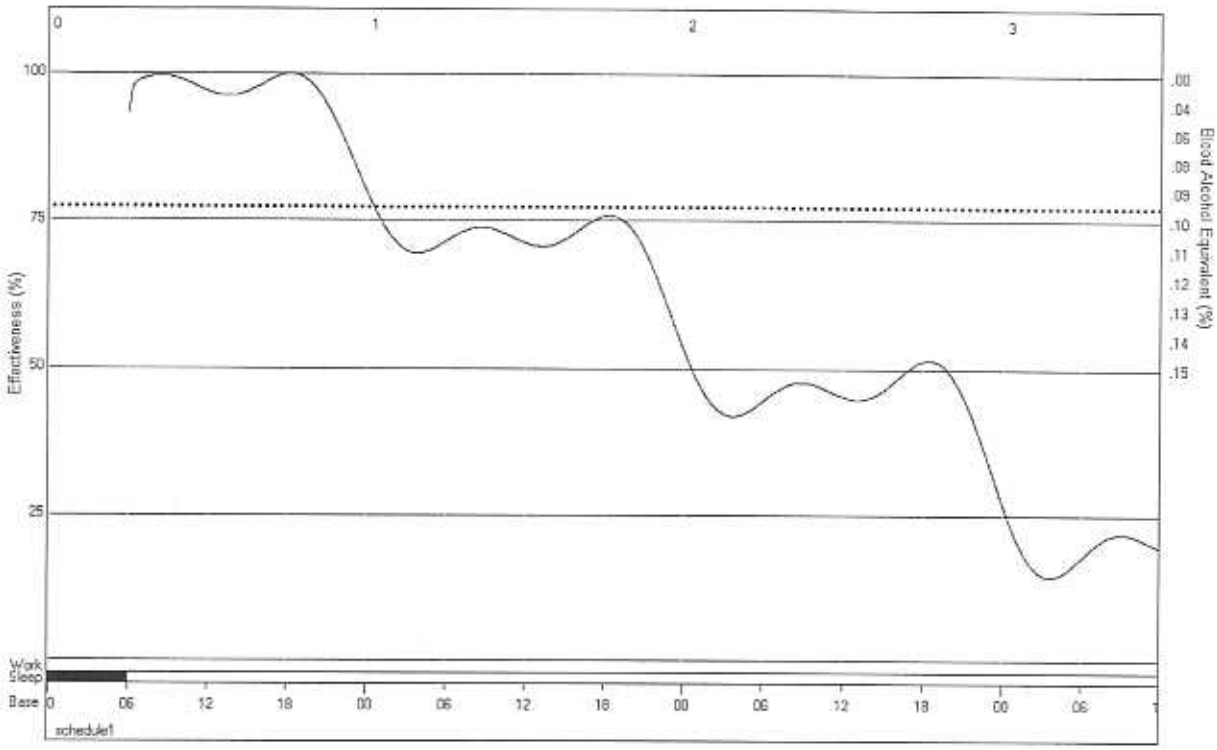


Figure 3. Performance effectiveness vs. time. Note the strong circadian effect, which reduces effectiveness highly nonlinearly with time. Effectiveness is approximately 50% at 64 hours, the duration of the DCIEM study, and throughout the interval 54-68 hours.

We therefore investigated qualitative differences in the speech measures across the dialogues, here labeled *fresh*, *intermediate*, and *fatigued*, particularly between the first and last of these in order to identify a maximal effect. This has the benefit of permitting greater use of distribution-free, or robust, statistics and statistical tests.[7] The Kolmogorov statistic D^* and the Kolmogorov-Smirnov test (Press *et al.*, 1992) that two sets of samples come from the same distribution — *regardless of what that distribution is* — will be particularly useful. Correspondingly, the conclusions will be less precise but more robust.

In each dialogue, one subject explained a map to another subject. Because the instruction-giver therefore provided most of the speech, and some of the instruction-receivers dropped out and were replaced with experimenters during the protocol, only the speech of the "givers" (down-sampled to 16 kHz) was analyzed. The "givers" were all healthy, male, aged 20-45 years.

It should be noted that the two subjects in each dialogue were given close-talking microphones but were in physical (and acoustic) proximity. As a result, the "giver" recordings often contain faint but clear speech from the receivers. However, the second speaker never produced confusions for the processing described below. (This therefore serves as a useful, though weak, test of the sensitivity of the results to such contamination.)

[7] The median is perhaps the most familiar robust statistic. Like the mean, it is a measure of location; but unlike the mean, it exists for *all* univariate distributions, including the important Cauchy distribution ~ $1/(1+x^2)$, which has no mean because of its "fat tails".

4 Phase I Results: Analysis of Speech

The dialogues of the DCIEM Map Task corpus were subjected to several speech and voice analysis techniques. Some of the earlier-explained techniques (vowel formants, "chaos") that were expected to show fatigue effects did not, and we briefly describe the results and the inferences we make about the technique and the corpus. Some (LTAS, F0 distribution) showed highly subject-specific results. And others (landmarks, non-fluencies) provide strong results — in the case of landmarks, perhaps implausibly strong results. Both of these last measures show "universal" or subject-*independent* thresholds that distinguish fresh from fatigued speech; that is, they do not appear to need a direct comparison with the subject's own speech. (In realistic use, such subject-independent thresholds would permit assessing even an unknown person's or opponent's state of fatigue.)

We begin with landmarks and describe several processing variations intended to explore its limits.

4.1 Landmark processing

The results for landmark processing were sufficiently — implausibly — strong that its limits were systematically explored. In this processing, the speech of each dialogue was divided into between four and ten "sections" of slightly less than one minute durations (~45-50 seconds), which were then divided further into segments up to 10 seconds long, about 1-2 sentences. Only the first section was analyzed for the six intermediate dialogues; all sections were analyzed for the fresh and fatigued dialogues, a total of 43 and 34 sections, respectively. Analysis:

- These were analyzed in the standard manner (described above) for the basic landmarks ±**g**, ±**b**, and ±**s**.
- Post-processing then automatically grouped these into "syllables", of which 38 types (specific landmark patterns) are recognized; and "utterances", speech surrounded by ~½ sec of silence.
- Each dialogue, or each section within a dialogue (depending on the test), was summarized by noting the relative frequencies of the types; of the relative frequencies of utterances with 1, 2, 3, or 4+ syllables; etc.[8]
- For each measurement, the mean and irreducible "noise" or fluctuation level σ was estimated, based on the measured value and the nature of the measure (*not* on other data values). For example, for any measure (e.g., number of syllables of a given type) that is based on N event counts, an estimated mean of N+1 counts and σ of $(N+1)^{1/2}$ counts were adopted; these are the Poisson estimates. Notice that this does *not* "use up" any additional degrees of freedom: only the measured value (N) itself is used.
- The measures were transformed to approximately Gaussian-distributed variables of zero mean and unit standard deviation (still representing the irreducible noise level).

Deliberately, the total numbers of syllables and of utterances in each dialogue are *not* included among the measures. The Map Task might be sufficiently structured that these measures could

[8] The words "aah", "bah", "bat", "batch" correspond to a sequence of increasing pattern complexity in a single syllable. Phonemes are grouped very broadly, however, so "oo", "go", "grid", and "that's" would show the same patterns.

show a detectable fatigue effect. However, they would not generalize well to dictations or less structured dialogues and were therefore suppressed in all analyses here. Thus, the results should apply relatively well to many other speech tasks, perhaps even spontaneous conversation.

It is important to recognize that each segment must (and did) represent contiguous speech data. However, the summarization step does *not* require contiguous *segments*. For example, the summary of each section (~50 seconds) could, in principle, have been a summary of arbitrary segments, of the same total duration from the same dialogue.

Finally, all dialogues were analyzed as a batch to determine a least-squares fit to the data (a "calibration" of the coefficients). In this stage, a principal-components (PC) analysis of the measures showed that some of the PCs (of the 12 that are uniquely determined from the 12 dialogues: six subjects when fresh and the same six when fatigued) should be suppressed because their contributions[9] to fitting the data were less than twice the unit "noise" level. That is, these PCs simply described effects that model primarily the noise, so that retaining them would certainly over-fit the data. (However, some subject-specific or other random effects, larger than the irreducible measurement errors, are surely present among the retained PCs.)

In the interest of robustness, some additional PCs were also discarded, those corresponding to effects slightly higher than the noise level, though still accounting for at most ~3% of the total variance of the data. For convenience, the results were then expressed on an arbitrary scale of ~ 10-40, with larger "point scores" (ideally) representing the fatigue condition.

4.1.1 Base case

Retaining 8 PCs from all fresh and fatigued dialogues provides *complete* separation of the fresh scores from the fatigued: the maximum fresh score = 18.4, the minimum fatigued score = 24.1, so a simple threshold, e.g., 20 points, could be used to separate the two distributions. That is, this result provides a subject-*independent* threshold for fatigue, an implausibly strong result for any broader population of subjects. (If such a threshold exists, it would at least need to be based on far more subjects.)

A formal test for the distinctness of the two distributions yields a strong result ($D^*(6,6) = 1.73$, $p < .002$, Kolmogorov-Smirnov [K-S] test).

A separate test, that all subjects' scores merely move in the same direction (up, in this case), is also positive ($p = .03$, sign test). However, this test, though assuredly robust, is quite weak: The strongest possible statistical significance is only $1/2^5 = .03$, the result achieved.

4.1.2 Segregated subsets: halves

Next, each of the 12 fresh or fatigued dialogues was split into two halves, 'a' and 'b', which were treated as 24 data sets. Observe that, although this does not provide any additional speakers, it enables the least-square fits to accommodate some intra-speaker variations. In this case, several additional PCs could be discarded without changing the qualitative conclusions above, because they modeled effects on the order of the noise level. Specifically, retaining 18 PCs (of the 24 available) produced fully separated distributions. More aggressive smoothing,

[9] I.e., the corresponding singular values from a singular-value decomposition.

PROPRIETARY

retaining only 13 PCs, forces the fresh and fatigued distributions to overlap, but it still causes each subject's score to rise when fatigued, by a minimum of 10 points. See Figure 4a.

This splitting of the data (or rather, the less complete summarization of the dialogues) could obviously be increased. However, there is no reason to suppose that further measures of *within*-speaker variation, which this would permit, would provide any better accommodation of *between*-speaker variations. The latter simply requires more subjects than the six available here.

4.1.3 First-minute subset

In order to assess the robustness of this apparently strong result, i.e., the complete separation of the two distributions (and, more realistically, the uniformly upward movement in the scores from the fresh to the fatigued condition), the processing was repeated in several variations. The first variation, using only the first minute of each fresh or fatigued dialogue (about 20% of the total data) also showed the complete separation of the distributions.

Thus, the conclusion does not seem to depend on specific details that require many observations to model adequately.

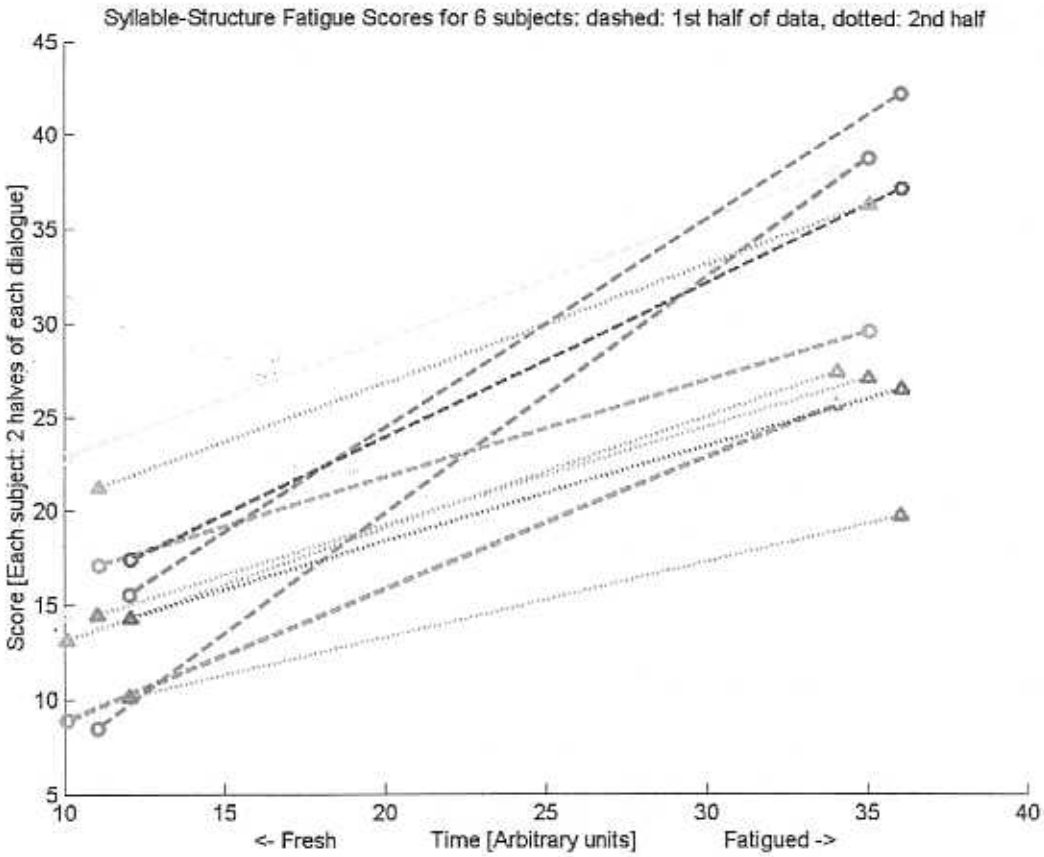


Figure 4a. Landmark-based fatigue scores. Each subject's dialogue was split into two halves, which were analyzed separately. Each subject's score rises by a minimum of 10 points in each half, from the fresh condition to the fatigued. The same color is used for each subject's first and second halves. (Circles and dashed lines: first half; triangles and dotted lines: second half.

4.1.4 "Leave two dialogues out": unbalanced

In an independent check of the results, two of the fresh dialogues or two of the fatigued dialogues were removed, a fit to the remaining 10 dialogues was generated, and the fit was assessed for separation of the distributions. By removing two of the same group (fresh or fatigued), this tests whether the fit is overly sensitive to the numerical equality of the number of sessions.[10] For simplicity (and with no known biases), the two omitted dialogues were the first (D10) of each Run. This protocol was repeated with the second (D11) of each, etc., up to the sixth (D36) of each. Thus, all dialogues were removed once.

Again retaining 8 of the PCs produces the following degree of separation of the two distributions, displayed as [maximum score of fresh, minimum score of fatigued]:

Omitted dialogue	Separation
D10:	[12.6, 30.0]
D11:	[13.8, 21.6]
D12:	[12.5, 23.7]
D34:	[19.2, 28.8]
D35:	[14.8, 36.4]
D36:	[15.5, 34.0]

In all cases, the two distributions do not overlap ($D^*(4,6) = 1.55$ for each case, $p < .005$, K-S test). For example, the value 20 can serve as a threshold for all of them.

From this test, it does not appear that the effect requires a design carefully balanced with fresh and fatigued dialogues, even from the same subjects.

4.1.5 "Leave one subject out"

As a final, most severe, check on the robustness of the landmark results, each of the six subjects was deleted from the set, a fit was derived from the remaining five, and this fit was both assessed for separation of the distributions and tested on the omitted subject.

Retaining 8 PCs once more produces the following degrees of separation:

Omitted subject	Separation
#1:	[15.7, 27.0]
#2:	[16.9, 25.8]
#3:	[9.0, 27.7]
#4:	[13.3, 33.3]
#5:	[11.5, 33.0]
#6:	[14.6, 32.0]

In all cases, the distributions again separate completely ($D^*(5,5) = 1.58$ for each case, $p < .004$, K-S test). Again, in fact, a single threshold such as 20 can serve for all of them (indeed, any value between 16.9 and 25.8 can serve). This result is to be expected: This aspect of the present test is slightly weaker than the above "unbalanced" test, which itself shows the effect.

How did the omitted subject's scores compare to the group's? Specifically, did they at least rise from fresh to fatigued? And did they preserve the non-overlap criterion, i.e., was their fresh

[10] This would obviously be inappropriate if the results were subject-*dependent*, since that would require the same subjects to be represented in both groups (or neither).

score at least as low as their own and everyone else's fatigued score, and was their fatigued score at least as high as their own and everyone else's fresh score?

Here the results are much less favorable. Using the same 8-PC fits as above yields the following:

Omitted subject	Separation
#1:	[42.2, 13.1]
#2:	[11.9, 21.2]
#3:	[37.0, 42.9]
#4:	[23.7, 22.4]
#5:	[22.9, 25.7]
#6:	[22.1, 11.5]

In this case, subjects #1 and 6 produce outright failures of the corresponding fits. Subject #3 has an uncharacteristically high score in the fresh condition, though his score, like the other five used in the fit, rises in fatigue. Subject #4 receives essentially the same score in both conditions. Finally, subjects #2 and 5 agree well with the corresponding fits. We might conclude that subject #2 is very well represented by the other five, as is #5; but that #1 and #6 are quite unlike the other five — in some undetermined way.

In conclusion, the fits should *not* be regarded as robust, subject-*independent* or "universal" fits. Out-of-group subjects, i.e., those not used in the fits, will generally not produce results consistent with in-group subjects — at least for fits defined by just five subjects in the group, each represented by a single, aggregated set of measures for each dialogue.

4.1.6 Segregated subsets: sections

The data were split one step further, into individual sections (or rather, they were less completely summarized). Observe that, although this does not provide any additional speakers, it enables the least-square fits to accommodate within-speaker variations much more fully.

All sections — 43 fresh, 6 intermediate, and 34 fatigued — were analyzed as a batch to calibrate the coefficients. The coefficients were determined by retaining 19 PCs, which accounted for 96% of the initial variance. (See Figure 4b.) This represents suppression of all PCs that describe variations less than 3.0 times the measurement error. Compared to the above analysis, this provides a fuller accounting for within-subject variations. Across-subject variations are still limited by the use of only 6 rather similar subjects in the dialogues.

4.1.6.1 Distribution differences

Interestingly, as the Figure shows, the clear separation already shown between fresh and fatigued conditions is apparent between the fresh and intermediate conditions as well, but *not* between the intermediate and fatigued. Specifically, for fresh vs. intermediate, $D^*(43,6) = 1.92$, $p < .001$; but for intermediate vs. fatigued, $D^*(6,34) = 0.62$, $p > .5$.[11] With only six samples for the intermediate sections, however, the K-S test has low statistical power, i.e., even rather large differences in the underlying distributions might go unnoticed. (As the Figure shows, the same is likely to hold for most other tests as well.)

[11] Such a finding would permit merging the two indistinguishable distributions and re-analyzing the data. The result would likely have slightly greater statistical significance *and* better separation between the fresh and intermediate/fatigued distributions. However, this was not attempted.

A fuller evaluation of the fresh vs. fatigued results produces $D^*(43,34) = 3.31$, $p << .001$. Focusing on the two replications of the experiment (Run 120 and Run 140) yields:

Run	d.f.	D*	p
120	22,16	2.65	«.001
140	21,18	2.00	<.001

Even the single-subject results show the effect, although the quantity of data (number of sections) limits the statistical significance that can be achieved:

Subject	d.f.	D*	p
#1	10,5	1.46	<.02
#2	5,4	1.49	<.01
#3	7,7	1.87	<.001
#4	8,7	1.41	<.02
#5	5,5	1.26	.036
#6	8,6	1.31	.032

The subjects in Run 120 (#1,2, and 3) appear to produce stronger results than those in Run 140 (#4, 5, and 6). It is not clear why this should be so. Testing the similarity of fresh or fatigued distributions between the two Runs suggests (weakly) that they may differ in their fresh-condition distributions; it shows no evidence for such a difference in the fatigued condition:

Condition	d.f.	D*	p
Fresh	22,21	1.19	.089
Fatigued	16,18	0.63	> .5

4.1.6.2 Thresholds

As the Figure makes clear, the median (like the mean) increased for every subject. In fact, this considerably understates the strength of the result. There are several ways of expressing this. First, for half of the subjects, every section ("minute") in the fatigued condition produced higher scores than *any* section in the fresh. The remaining subjects had a single outlier in one or both conditions. (Thus, the second-highest fresh score was lower than the second-lowest fatigued score, in every subject.) One version of a two-way screening test might therefore be stated:

- If two new scores from a subject are *both* lower than his highest recorded fresh score, then the subject is fresh.
- If *both* are higher than the lowest fatigued score, then the subject is fatigued.
- Other cases require more information.

In fact, this case might be strengthened further. For *every* subject, the second-highest fresh score was lower than 22.5, and for the second-highest fatigued score was higher than 22.5: This is, again, a "universal" threshold. The contingency table for this case is simple:

Subject criterion	**# Subjects, Fresh**	**# Subjects, Fatigued**
Score ≤ 22.5	6	0
Score > 22.5	0	6

This yields $p = .001$ (Fisher Exact Test [Sheskin, 1997] for equality of two Binomial distributions).

The gap is, however, as narrow as 0.5 unit for one subject.[12] This suggests that the threshold would not apply with additional subjects, nor perhaps even for additional data from the same subjects.

Another way to express the strength of the increase in scores is the following: Not merely the 50th %ile (median) of each subject's fatigued scores is higher than the 50th %ile of his fresh; even the 25th %ile[13] of the fatigued scores is higher than even the 75th %ile of the fresh. Indeed, the statement appears to be true of the 15th and 85th %iles, although the number of sections for some subjects is too small to be confident of this result.

4.1.6.3 Interpretation of coefficients

Analysis of the coefficients derived for the fit shows that they are of two types. First, 78% of the sum of squared coefficients (a measure of their contribution to the goodness-of-fit) corresponds to the probability distribution of syllable type. Most of this, 60% of the total for the fit, constitutes a contrast among just six specific syllable types, three being more frequent in the fatigued condition and three in the fresh.

More Frequent in Fresh:	+b/-b/+g/+s/-g	+g/+s/-s	+g/+s/-s/-g
More Frequent in Fatigue:	+b/+g/-s/-g	+g/-s/-g/-b	+s/-s/-g

It has not been feasible to determine which, if any, specific words or syllables in the corpus correspond to this contrast of landmark patterns. (It is likely, however, that they are stressed syllables.)

A further 20% — i.e., virtually all of the remainder — corresponds to the probability distribution of syllables per utterance, described simply as the frequency of occurrence of utterances of 1, 2, 3, or 4-or-more syllables. Specifically, the coefficients measure a contrast between the fraction of 1- or 2-syllable utterances and the fraction of 4^+-syllable ones.

Interestingly, a large fraction of short utterances characterize the *fresh* dialogues. It is possible that this reflects the specific notion of "syllables" here, defined merely as certain patterns of landmarks. Fatigued speech might not contain more true (linguistic) syllables but simply more non-fluencies. In fact, *this non-fluency effect was noted directly* (see below), though with only a few examples per subject.

[12] Though as large as 5 units for 3 of the subjects.

[13] I.e., the value separating the bottom 25% of the subject's section scores from the top 75%.

PROPRIETARY

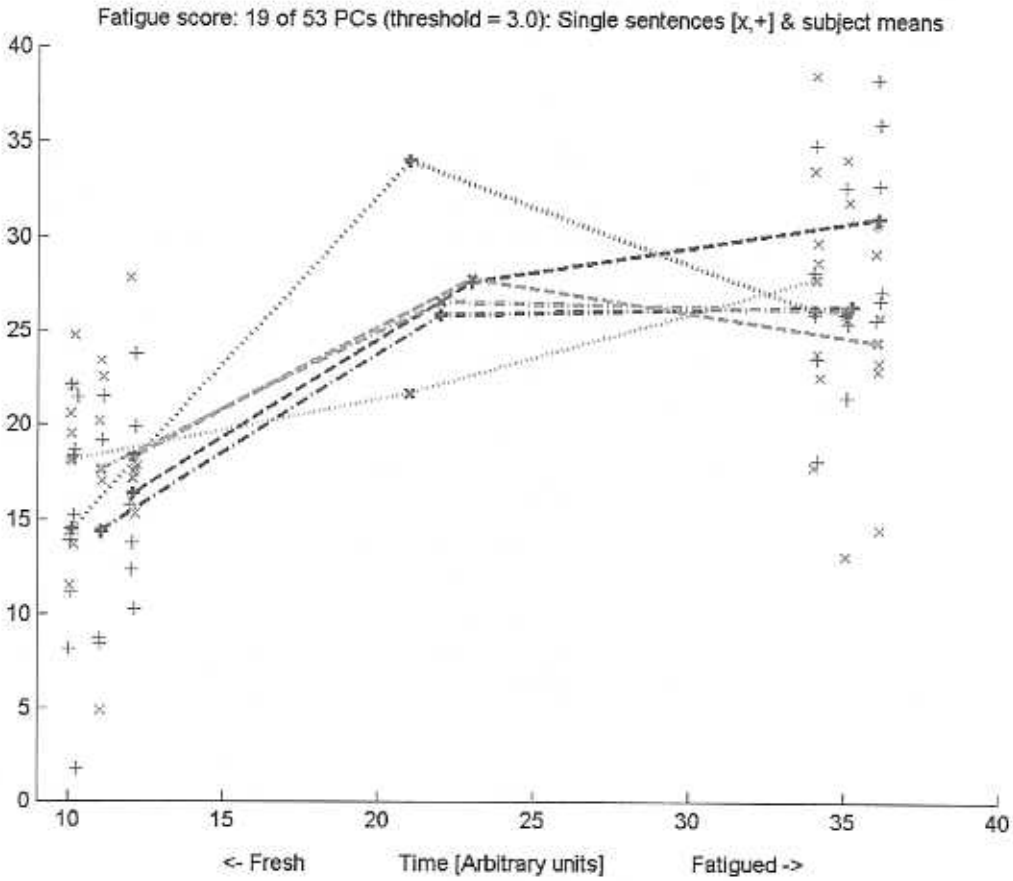


Figure 4b. Landmark-based fatigue scores. Each point represents one section of dialogue, approx. 50 secs: +, subject in Run 120; x, subject in Run 140. Heavy symbols (+,x) with connecting lines represent subject means. For clarity, points for distinct subjects in each Run are shown with slight horizontal offsets. For each subject, only one section of the intermediate dialogue was included (middle points horizontally). The distribution of scores for intermediate dialogues appears to be the same as that of the fatigued ones. In all cases, subjects' mean scores increased from the fresh to the intermediate condition, and from the fresh to the fatigued. There is only slight overlap between the all-subject fresh and fatigued distributions of section scores; and only slight overlap, or none at all, between each subject's individual fresh and fatigued distributions.

4.1.7 Discussion of landmark-based results

The landmark-based measure has been explored in some detail, because of the apparent strength of the initial conclusion. They seemingly provide a subject-*independent* or "universal" threshold to assess fatigue — a very strong result, but one that will be unrealistic in practice. These results provide an important baseline for assessment of more realistic data, that is, data containing more noise, more confounding effects, and perhaps evidence of behavioral correlates (e.g., caffeine consumption) that were suppressed in the DCIEM study.

Certain results, however, appear firm. For example, a few 5-10-second segments of speech with a total duration under one minute are often sufficient to characterize a speaker in either the fresh

or fatigued condition. Two such collections were *always* sufficient — for young, healthy, Canadian-accented male speakers.

4.2 Long-Term Average Spectra

The long-term average spectrum for one subject produced initially intriguing results: Several sentences from the fresh condition produced highly consistent patterns in the contrast between sonorant and nonsonorant spectra, in both shape and magnitude; and several from the fatigued condition produced highly inconsistent ones in both regards.

Specifically, six ~30-second segments were analyzed from each of Run-120's dialogues 11 (rested) and 35 (fatigued), subject "Paul". For each segment:

- The signal was differentiated, to compensate for the generally falling intensity at higher frequencies. (This does not affect the final spectral differences, however.)
- Sonorant intervals were computed by observing voiced spectral power.[14]
- Then adjacent intervals were identified as nonsonorant if they exhibited sufficient broad-band power. (This was generally ~30 dB lower than the sonorant power, however.)
- Short intervals of interruption were ignored, in order to avoid, e.g., splitting the word *maps* into two segments at the brief stop in the /p/.
- The short-time power spectrum (over ~10 msec) was computed about each sonorant or nonsonorant sample, except for those samples within ~5 msec of the edge of a sonorant (resp., nonsonorant) interval.
- The log-power spectra of the sonorant intervals of the segment were averaged, and the log-power spectra of the nonsonorant intervals were averaged. [By using 10 log10, the spectra are expressed in dB.]
- These two averages are subtracted, frequency by frequency.

The result is highly consistent across the segments of the rested speech. This is emphasized by computing the mean difference across the quarters of each spectrum: 0-2, 2-4, 4-6, and 6-8 kHz (nominally, "1, 3, 5, and 7 kHz"). In every case, the mean value rises from 1 kHz to a peak at 3 kHz, falls to a value at 5 kHz that is even lower than that at 1 kHz, and falls slightly further by 7 kHz. With one exception,[15] these mean spectra do not even cross each other; very nearly, they differ only in "height" (i.e., power ratio between sonorant and nonsonorant segments). And at all four nominal frequencies, the spectra span a range of ~15 dB.

In contrast, the results for this subject's fatigued speech are highly variable. First, only two of the six segments show the peak/fall pattern that all six rested-speech spectra showed. Second, four of the mean spectra cross each other (a total of seven times). Finally, at 3, 5, and 7 kHz, the mean spectra span a range of 20-25 dB.

However, evaluation for all subjects failed to show any consistent pattern distinguishing fresh from fatigued speech. It is possible that some subjects might show this pattern routinely, and others not, but that could not be assessed with only six subjects. Obviously, any Fatigue Assessment System using such a measure would need much more research to determine (a)

---

[14] That is, high-power vocal harmonics were noted. This identifies "sonorants": vowels, liquids (/r/, /l/), nasals, semi-vowels (/w/, /y/), and diphthongs/glides (/au/, /oi/, etc.).

[15] The cyan and magenta (second and third highest) spectra in the figure, and only at 3 kHz, i.e., a single crossing: The cyan starts slightly higher at 1 kHz and ends slightly lower at 7 kHz.

whether a particular subject could be evaluated on such a pattern, or at least (b) whether such a pattern, however fortuitously detected in any subject, could be relied upon.

4.3 Pitch Distribution

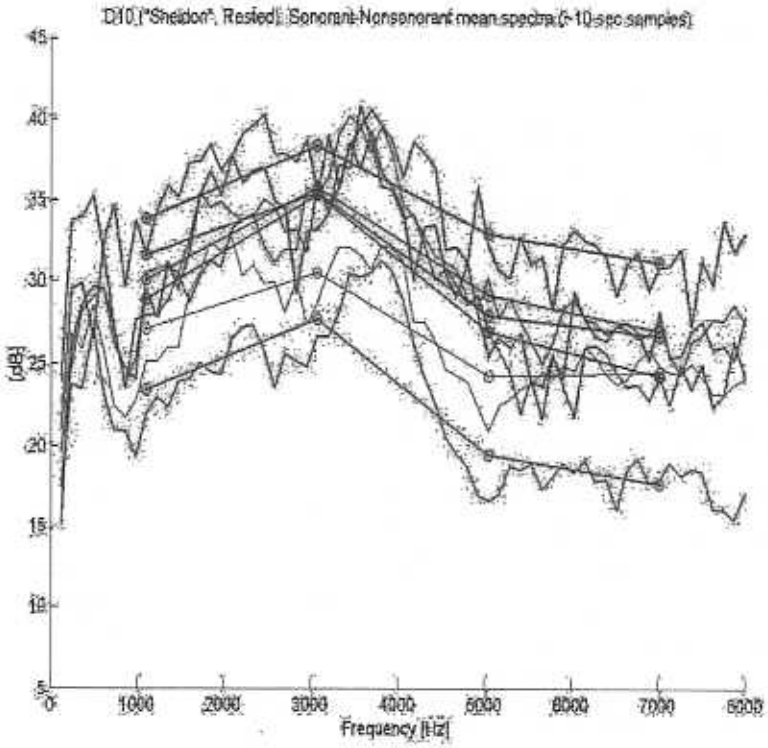
Like LTAS, pitch (F0) assessment for a single subject (Paul) showed an unusually promising pattern that was not replicated for other subjects. Specifically, the mean pitch across long segments of speech with many yes/no questions was unchanged between the fresh and fatigued conditions, but the distribution was markedly narrower in the latter.

This subject was chosen simply because his speech contained many explicit yes/no questions, or invitations to confirm map information (implicit yes/no questions). Segments with such questions were chosen because the rising pitch at the end of a phrase to ask a yes/no question requires extra muscular effort in the lungs and larynx *and* planning for sufficient breath support at the end of the phrase. Both of these effects can be compromised by fatigue, making this rising pitch a potentially valuable discriminator, at least when comparing a known subject across time.

For this subject, the mean pitch was virtually unchanged: 153.6 Hz when rested, 154.4 Hz when fatigued (from ~800 pitch measurements, with $\sigma \sim 25$ Hz). However, the fatigued-speech distribution was noticeably narrower than the rested-speech: $\sigma = 23.8$ Hz vs. 30.1 Hz ($p << .001$, F-test with 752 and 862 d.f.). The figure shows the comparison of distributions. As these make clear, the highest pitch in the fatigued case is indeed typically lower than in the rested case. (From other analysis [not shown], these are known to correspond to the ends of the phrases in the segments.) However, the lowest pitch is *also* closer to the mean for the fatigued case than for the rested one.

This strong, simple pattern was *not* seen for the other subjects, for reasons that are unclear. It is quite possible that Paul uses a yes/no question style in a more consistent manner than the other subjects: It is remarkable that the mean is virtually unchanged between two dialogues ~2 days apart. However, it is equally plausible that differences in the pitch distribution, even if they were highly consistent for any given subject, would vary from subject to subject.

PROPRIETARY



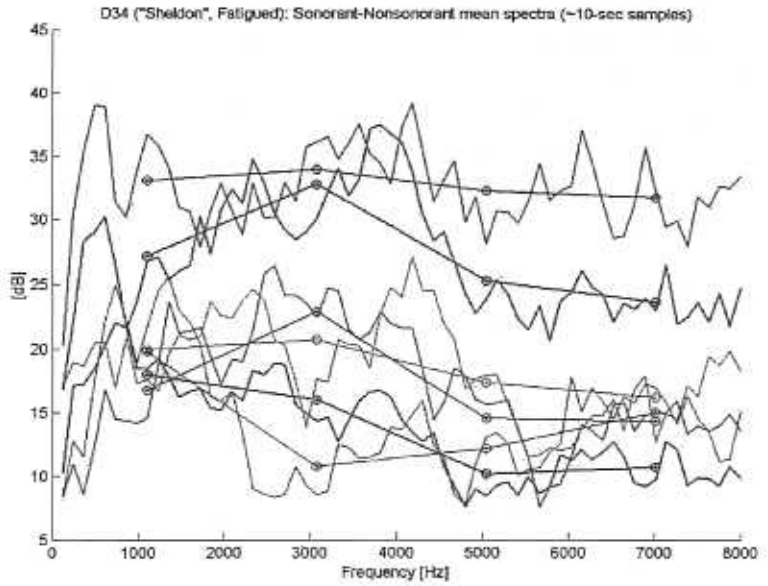


Figure 5. Average difference of log-spectra, sonorant minus nonsonorant, one subject. Straight segments represent averages over 2 kHz (¼ of frequency range). Upper figure shows six segments when fatigued; lower, six when rested. Notice the larger variation of mean values of the six spectra when fatigued, the generally flatter spectra (most visible in the straight-segment versions), and much lower consistency of shape (straight-segment version).

PROPRIETARY

A further investigation, this time of the maximum pitch in each "phrase" (actually a segment of speech interrupted by at least 40 ms, but no more than 100 ms, of silence), likewise showed no consistent variation from fresh to fatigued conditions. This was true whether the measure was the mean, median, or standard deviation of the samples, the entropy of the distribution, or the point-by-point difference of the distributions (normalized histograms).

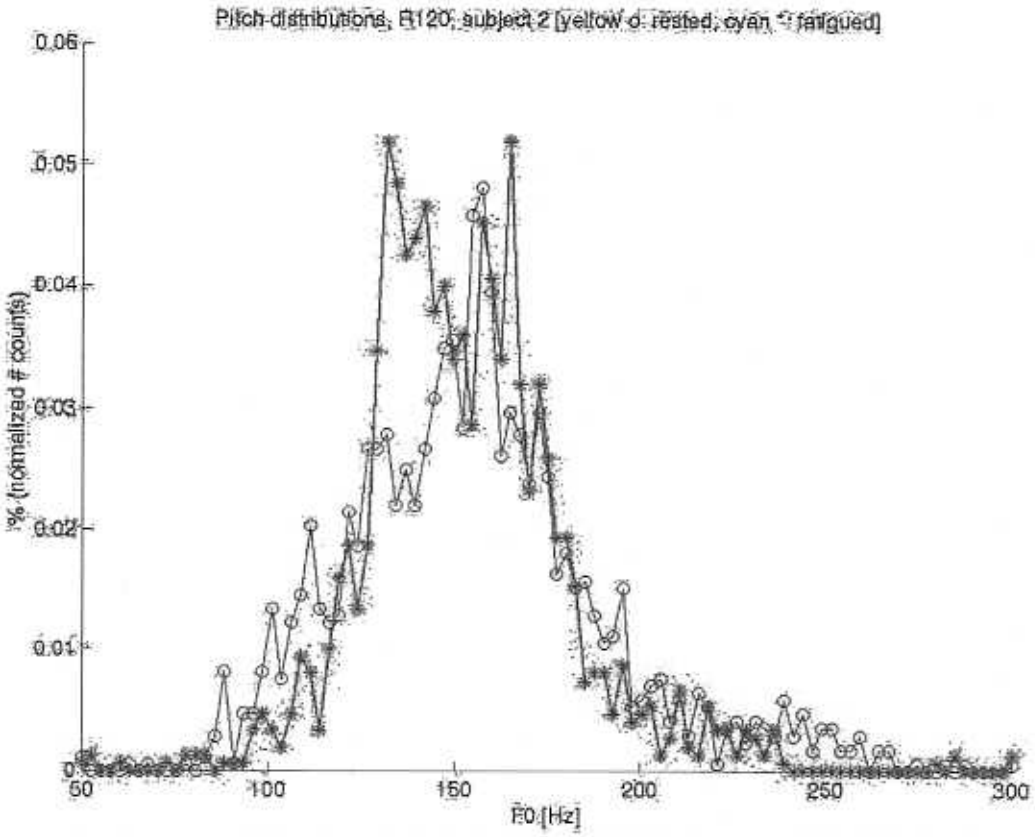


Figure 6. Distribution of pitch across ~one-minute segments of questions, one subject. The light [yellow] lines and open symbols show the distribution of pitch in a segment with many yes/no questions in rested speech (Run 120, dialogue 11, subject "Paul"); the heavy [cyan] line and * symbols show the distribution in fatigue (dialogue 35). Both distributions have a mean pitch of 154 Hz. Notice the narrower distribution for fatigue, especially visible in the absence of samples of high pitch (> 180 Hz) and of low pitch (< 90 Hz).

4.4 Vowel Formant Frequencies

For one subject ("Corey", Run 140), the "vowel space", or range of lowest and second-lowest frequencies of mouth resonances (formants), does not differ substantially between rested and fatigued speech, for either stressed or unstressed vowels. That is, plots of these two frequencies ("F1-vs.-F2") show no evident pattern: see Figure 7.

This is currently a time-consuming, manual operation because of the difficulty — but importance — attached to unstressed vowels. It has been hypothesized that these vowels, being less critical

PROPRIETARY

to communication, might "degrade" more quickly with fatigue than stressed vowels. In this context, "degrade" reflects centralization, i.e., the tendency of the vowel to be pronounced as a *schwa*, the neutral vowel. However, in English (and very few other languages), this is exactly the *correct* articulation of unstressed vowels(!).

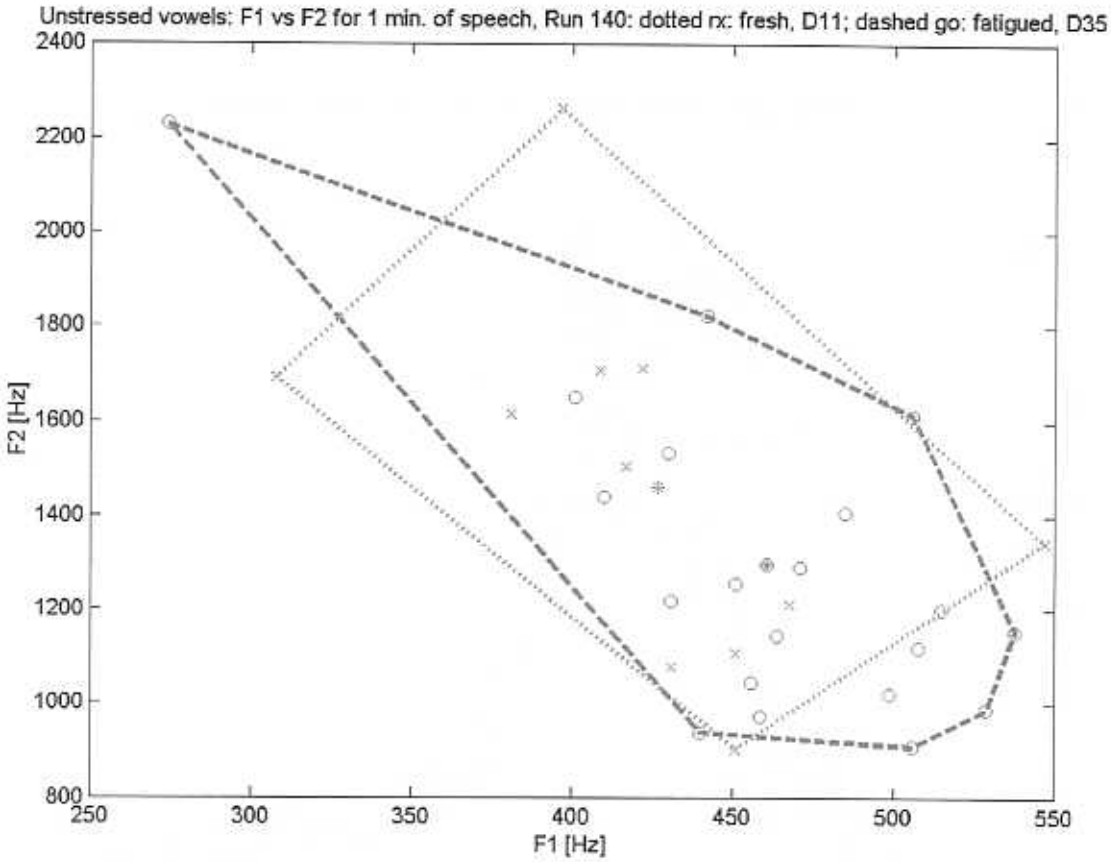


Figure 7. Vowel space for unstressed vowels, one subject. The plot shows the lowest two formant frequencies (mouth resonances) of the unstressed vowels in dialogues from the fresh and fatigued conditions. The centroid of each group is also shown. The lines join the outermost points (convex hull) for each condition. (Fresh: red x's, dotted lines, * at centroid; fatigued: green o's, dashed lines, filled o at centroid.)

Therefore, in the DCIEM corpus, with its all-English speech, it would be problematic to distinguish poorly articulated unstressed vowels from accurately articulated ones. Given the negative finding for the one speaker, and the laborious nature of the task, we elected to investigate no other speakers on this measure. Furthermore, even with high levels of automation, there appears to be no reason to expect meaningful results on this measure for English speech, at least for *native* speakers. (However, English is used by many non-native speakers in safety-critical activities such as commercial aviation. We do not conclude that this measure is necessarily irrelevant for those speakers.)

4.5 "Chaos" Measures

Several short (~30 sec) segments were excised from each subject's speech in each of the 12 dialogues; we particularly included instances of "umm" or "ahh", the simplest vowels that we might find most prolonged. We then down-averaged the data to 2-kHz sampling and analyzed the local periodicity of the waveform for signs of aperiodic laryngeal dynamics. (Such dynamics would be audible as vocal roughness.)

Unexpectedly, we found no evidence that the subjects' voices become rougher, lapsing into episodes of more complex or even chaotic dynamics, during the protocol. Furthermore, prior research has shown that laryngeal abnormalities such as edema or lesions can produce nonstandard vibration modes (period-doubling, 3:2 frequency ratios between the left and right vocal folds, etc.), during running speech; these are not always audible, and they can be found occasionally in normal subjects' speech. Again, however, there was no evidence for these modes in any of the speech samples analyzed.

Anecdotally, a sustained rough mode (known as "pulse register" or "vocal fry") is commonly associated with fatigue. We speculate that the popular association might arise from a behavioral correlate of fatigue, such as lack of hydration or ingestion of the diuretic, caffeine. Both of these are known to produce the drier laryngeal tissues that can cause fry,[16] and, by the construction of the protocol, the subjects have likely avoided both of them.

4.6 Perturbation Measures

Several samples, each ~ 30 sec, from two of the subjects in both conditions (dialogues 210/234 and 212/236) were analyzed with the Kay Elemetrics MDVP™ system. (This required substantial amount of labor, limiting the number of dialogues which could be analyzed.) This analysis provided one interesting negative result.

Several parameters were selected and compared between the conditions: those denoted (in MDVP) as Jitter Jitt, Relative Average (pitch) Perturbation RAP, Smoothed Pitch Perturbation Quotient sPPQ, Fundamental Frequency Variation vFo, Amplitude Perturbation Quotient APQ, Smoothed Amplitude Perturbation Quotient APQ sAPQ, Peak-to-Peak Amplitude Variation vAm, and Voice Turbulence Index VTI. The first four measure pitch perturbations like jitter; the last four, amplitude perturbations like shimmer.

Although some of these showed large values, none was consistently larger (nor smaller) in the fatigued conditions than in the fresh. However, *none* of the pitch-perturbation measures was found to be very large in *any* of the samples. This supports the speculation (above) that every subject (and specifically his larynx) was sufficiently well hydrated at all times that his laryngeal dynamics remained normal throughout each dialogue, producing the highly periodic signal that is characteristic of normal voice.

[16] So can antihistamines.

4.7 Non-fluencies

Because of the established association between non-fluencies and ethanol intoxication, similar non-fluencies were evaluated in Phase I, with the expectation that non-fluencies should increase with fatigue. Specifically, segments totaling ~3 minutes were evaluated for:

- additions of phonemes or words;
- substitution of phonemes;
- repetitions of syllables or words;
- voicing errors;
- omissions of phonemes or words.

In fact, no voicing errors were noted for any subject. The four remaining non-fluency features were totaled, with the following paired results:

Dialogues	Fresh	Fatigued
210/234	1	2
211/235	2	3
212/236	1	4
410/434	0	5
411/435	0	3
412/436	1	2
Total	5	19

The number and consistency of these results is encouraging. Not only did *every* subject's non-fluencies increase from fresh to fatigued ($p = 1/2^6 < .02$, sign test), but the total number of occurrences across subjects is highly significant. (For example, modeling the data as coming from a subject-specific Poisson process is appropriate. The data can be interpreted as Bernoulli trials with an indefinitely large number of "opportunities" to observe non-fluencies, the limit in which the Binomial distribution becomes the Poisson distribution. In this case, $p < .002$: Fisher Exact Test).

One can also construct a subject-*independent* screening test from these scores, i.e., by adopting a threshold such as 1.5 or 2.5: All fresh scores are below 2.5 (i.e., equal to or less than 2); all fatigued scores are above 1.5. (The value 2 is ambiguous.) As a screening test, this has good statistical significance: the data may be recast as a simple 2x2 contingency table, with the following results:

Subject criterion	# Subjects, Fresh	# Subjects, Fatigued
Score ≤ 2	6	2
Score > 2	0	4
	- or -	
Score ≤ 1	5	0
Score > 1	1	6

The first threshold ("2.5") yields $p = .03$ (Fisher Exact Test). The second ("1.5"), $p < .008$.

5 Phase I Results: User Interface Design Issues

This section outlines a proposed process and set of tasks to be used in designing the user interface (UI) for the Voice Analysis-based Fatigue Assessment System. We follow subcontractor Aptima's approach to the design of interfaces for complex "mission critical" systems as customized to meet the requirements of this system.

5.1 Systematic UI Design Approach

The UI Design approach discussed here is a comprehensive and systematic approach whose primary aim is to ensure effective and efficient task performance. See Figure 8 for an overview of this approach. Unlike approaches that focus on enhancing users' experience (perceptual stimulation, the sense of fun, etc.), the design for a mission critical application emphasizes operators' performance and allocates secondary, albeit important, status to operators' experience in the UI design. This approach places greater emphasis on contextualization (i.e., the development of system requirements) and the articulation of a conceptual design than on the presentation aspects of detailed design.

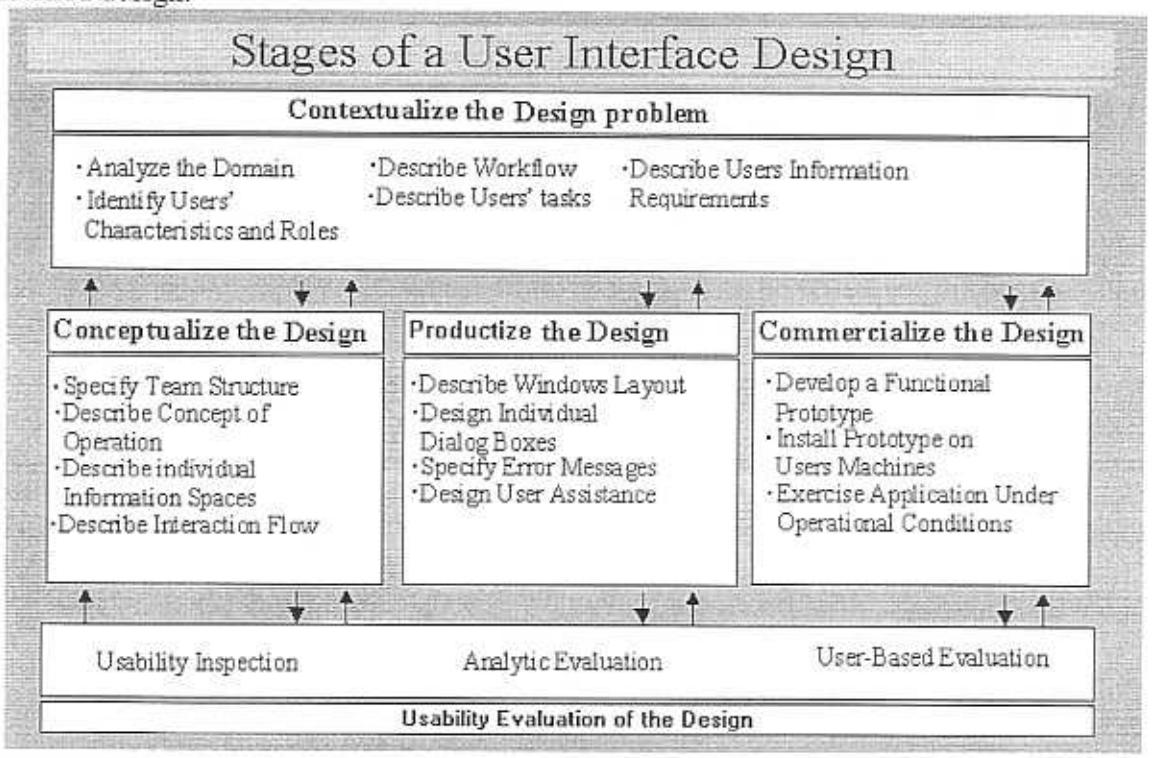


Figure 8. A Systematic UI Design Process.

The stages presented in Figure 8 are described in detail below. It is important to note, however, that the actual method of implementing them will be iterative. A critical outcome of this UI design process is that at the end of the design there is a logical connection between the UI design requirements developed during the contextualization of the design problem and the detailed design.

The UI design (the end product of the UI design process) will be principle-oriented. This means that, to the extent possible, UI design decisions will be driven by a set of empirically validated design principles. Design decisions will be evaluated later during usability testing. The drawback of this approach is that it limits or restricts to some degree the artistic expression of the UI designers and their ability to invent very unique and exciting interfaces. The advantage, however, is that it shortens the design cycle by reducing the number and scope of usability evaluations required to ensure that the proposed design is highly usable and acceptable to end users. It also results in an interface that is more familiar to end-users and, therefore, often easier to learn.

5.1.1 UI Design Tasks for the Voice Analysis-based Fatigue Assessment System

The specific tasks to be performed to *contextualize* the UI design problem for this system, *conceptualize* it in the form of a high level static prototype, and then *productize* it in the form of a dynamic prototype are described below. *Evaluations* of the various representations of the UI will be performed throughout the design process and their results will inform and influence it.

5.1.1.1 Contextualizing the design

The first stage in the UI design effort will be to **contextualize the design problem**; i.e., specify the context within which the UI design decisions should be made. The focus of this stage is on understanding the factors that impact the interface design, developing requirements, and establishing the boundaries within which the UI design must fit.

The design context of the Fatigue Assessment System is comprised of several elements including the "operational" problem the system is addressing, the work its users are expected to accomplish, and the characteristics of the users who would operate it either in the lab under experimental conditions or in the field under operational conditions. Understanding these factors and extracting the appropriate requirements for the interface design are particularly important steps. These steps inform and guide the design to create a system that enables task completion in an effective and efficient way without placing an extra cognitive burden on the end user. Only by performing effective task and environment analyses can we create a system that minimizes the number of steps required for the execution of each task and also minimizes the cognitive load associated with extracting information from the system's display.

5.1.1.1.1 Tasks – Contextualizing the design

The primary tasks associated with specifying requirements to contextualize the design problem are analyzing the domain, identifying user categories, and describing tasks and workflow.

5.1.1.1.1.1 Analyzing the domain and the operational problem

This task has two objectives. The first is to develop an understanding of how a fatigue assessment tool would be employed in research and operational applications. The second is to compile, organize, and explain important domain concepts, terms and abbreviations, as well as common practices, if such exist. This understanding and information is critical to ensure that the interface, which is often perceived by users as the application itself, not only emphasizes the operational objectives of the application, but does so in a manner that is compatible with users' needs and the way they interact with and talk about the domain. An example of an operational

objective could be to alert an operator to the onset of fatigue. Alternatively, it could be to alert the supervisor of an operator to the onset of fatigue of the operator. A research objective could be to use the fatigue assessment tool to measure the effectiveness of a fatigue countermeasure or the effects of medicines on subjective state. To enhance the effectiveness of the application and its value to end users, the operational problems that it attempts to solve need to be clearly articulated and well understood by the UI development team as well as the software development team. This operational problem and the resulting interface design need to be presented using terms and concepts that are common and well understood by the users and the customers of this application.

Developing an understanding of the operational problems and the general domain of fatigue assessment and speech analysis, as well as the concepts used in it will evolve through interviews with potential customers and users of the assessment tool, the available open literature in the area of speech analysis and fatigue assessment, and review of competitive products, if any exist. The output of this task will be a description of a set of potential uses of the fatigue assessment system, and a compendium of domain specific terms, concepts, and abbreviations.

5.1.1.1.1.2 Identifying and describing the users and user categories

This task will explore and describe the high-level roles/categories of end users and the relevant characteristics of users within each user category. This information is critical for determining such issues as how much technical language should and could be included in the interface, how much computer literacy can we assume on the part of the user, and how willing users might be to learn new ways of executing computer tasks.

For example, technical terms such as "F0" and "LTAS" could be quite clear to a speech scientist who is trying to identify which speech parameters will best correlate with fatigue, but confusing to a speech technician who is trying to fine-tune the fatigue assessment tool to fit a specific use environment, and totally meaningless to an end user such as a locomotive engineer who needs to decide when to hand control over to his or her assistant.

In determining the user categories and their characteristics it is important to know not only who the immediate potential users of the Fatigue Assessment System are, but also who the future users of this system will be. Thus it might be reasonable to assume that in the near term, the assessment tool will be fielded within a research setting with speech and fatigue scientists as the primary users able to make fine adjustments to ensure maximum validity of the system based on detailed technical data.

In the long term, as the tool transitions from research to operational use, the ability to understand system operation and adjustment cannot rely simply on technical data. As such, an objective of the UI design is to accommodate both user communities. Identification of near-term and long-term users or user categories will result from reviews of the literature in the area of fatigue assessment methods, market research, and interviews with expected users of the target system.

The output of this task will be several "user personae" which are a structured forms each describing a user category and its main characteristics. These personae serve as the basis for continued UI design.

PROPRIETARY

5.1.1.1.1.3 Describing workflow and tasks associated with using the Fatigue Assessment Tool

Description of the workflow and the tasks associated with analyzing voice samples, determining fatigue levels, and acting on this information are important tasks in developing an understanding and articulating the use-requirements of the Fatigue Assessment System. The workflow description provides a critical input to the UI design in two ways. First, the workflow will specify what tasks need to be supported and therefore the set of functions that the application should include. Second the workflow helps to aggregate the tasks into higher level operator/user functions, which in turn will help structure the UI in a user oriented way.

The expected tasks and their organization will be derived from a set of "To Be" scenarios, often called "use scenarios", that describe in a narrative form some situations in which the Fatigue Assessment System will be used and the ways that developers and users envision it being used. These scenarios will be driven by the operational problems that the Fatigue Assessment Tool is addressing for the different user groups.

The UI storyboards developed in Phase I (see Figure 9) will serve as a starting point for assessing tasks, workflow, and use scenarios. These designs provide a tangible reference point to spawn discussion; they will not be used to restrict functionality or limit design possibilities.

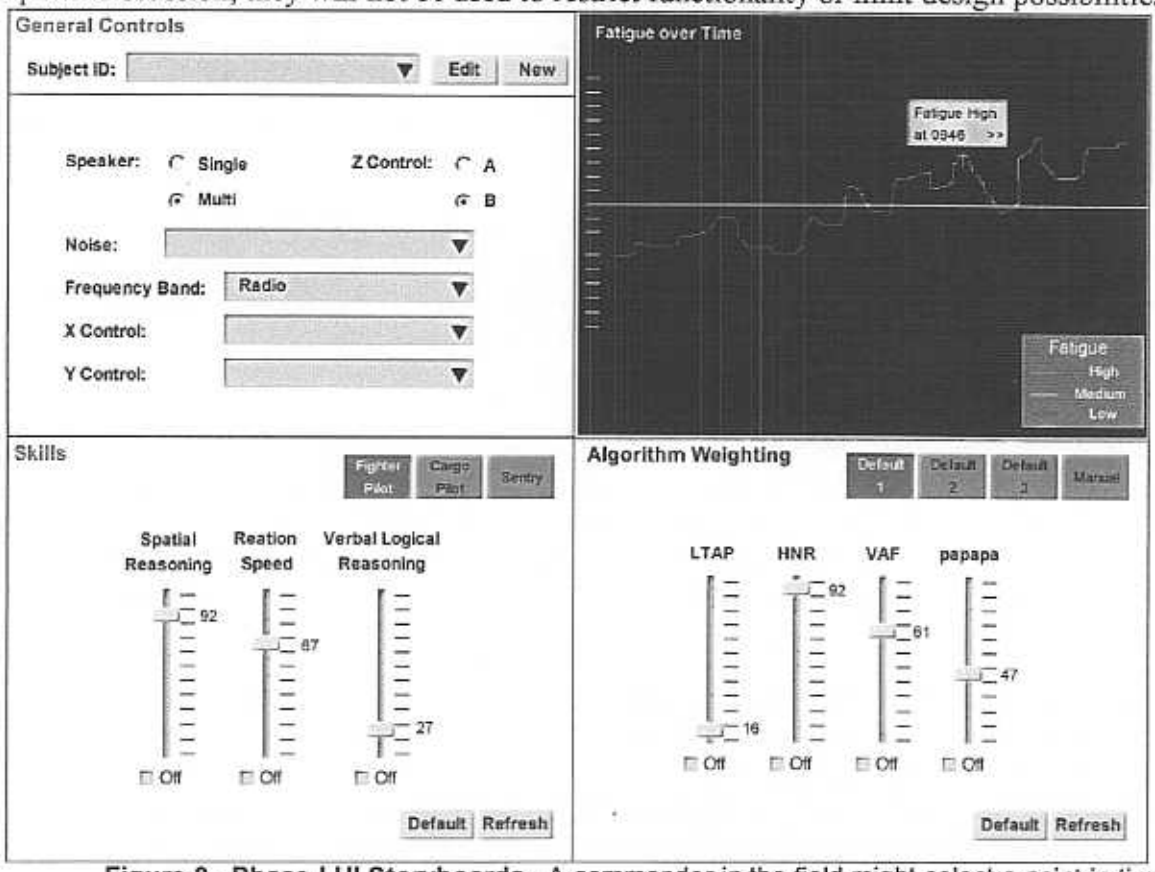


Figure 9. Phase-I UI Storyboards. A commander in the field might select a point in time to obtain detailed information about the measured fatigue level. A researcher in the lab might want to select job categories and have a mechanism to define or adjust the skills associated with the tasks that subjects are performing.

Task description and analysis is an iterative process of generating use scenarios, decomposing them into a set of tasks, and representing the tasks in a hierarchical task flow. After decomposing the scenarios into individual tasks, they are aggregated to define operator functions that will be supported later in the UI by one or more related dialog boxes or windows. For example, in the research mode of using the Fatigue Assessment System, numerous tasks may be related to calibrating different parameters of the tool. These tasks will be aggregated to an operator function of "Application Calibration" and all or most of the controls and displays associated with calibration of the tool will be assigned to a single dialog box or window. The project team will review, evaluate, and refine both the scenario decomposition and task aggregation that will serve as important inputs to the next stage of conceptualization of the interface design.

The output of this task will be a set of "To Be" scenarios and the derived flow diagrams depicting the sequence of task associated with each operator function.

5.1.1.1.1.4 Identify and document information requirements associated with operator tasks

Once the tasks of the different end users have been described and articulated, the information requirements for each task will be identified. The requirements specify the information that the application displays to the user through the UI to support the user in completing tasks. This information will include both the type of data that the user will need for completing a task, the required resolution, and preferred display format; e.g., numerical versus graphical.

The output of this task will be a table of the different tasks and the information requirements for each task.

5.1.1.1.2 Deliverable – Contextualizing the design

The collective output of the contextualization design stage will be a set of formatted user requirements forms, each depicting a single user requirement. The form will include an identifier such as a requirement number; the requirement type such as a functional requirement, an attribute, or a constraint; a description of the requirement; the rationale for the requirement; and the source of the requirement. When aggregated, these requirements will comprise the Requirements Specification document.

5.1.1.2 Conceptualizing the design

The second stage of the UI design will focus on developing the conceptual model of the UI. The conceptual model will be based on and informed by the contextual model developed in Phase I of this project. Thus, the conceptualization efforts will not start from a zero baseline, but rather will leverage the initial design work to refine and expand the concept to fit the additional information and requirements gathered in the contextualization stage.

During the conceptualization stage, the general characteristics of the UI will be articulated, including the high-level organization of the interface and the task flow required to achieve major system functions. The questions addressed during this stage will focus on the nature of the interface in the context of both the research system and the operational or fielded system. Some examples of the questions to be addressed at this stage are:

- Should the interface be process-oriented to lead the user through a set of tasks, or should it be entity-oriented provide a set of tools to enable the operator or user to create the entity?
- What should be the central component of the interface; e.g., the voice analysis results or the fatigue assessment?
- Should the interface capitalize on a common metaphor such as a dashboard?

5.1.1.2.1 Tasks – Conceptualizing the design

The primary tasks associated with developing a viable and effective conceptual model of the user interface for the proposed Fatigue Assessment Tool are to describe: 1) a concept of operation, 2) individual information spaces, and 3) task flow.

5.1.1.2.1.1 Describe a concept of operation

The concept of operation is the high-level notion of how the user will use the application and how tasks will be allocated between the user and the application or service. For example, one concept of operation for the research-oriented application might be that all voice analysis functions will be automated and the operator will monitor the analysis to gain confidence in the results. Another concept might be of a semi-automated system where the operator is primarily monitoring voice analysis and intervening when a drift in system calibration is detected or various filters need to be introduced.

For the operational application, one concept of operation might be to enable the user to calibrate the system based on his or her subjective evaluation of fatigue. Another concept might be to shield the user from the calibration task altogether and only provide him or her with an indication of predicted fatigue.

Based on the selected concept of operation, the overall approaches to the UI design (e.g., organizing displays around a process or making them entity-oriented) will be selected and ideas for metaphors or idiom evaluated.

5.1.1.2.1.2 Describe individual information spaces

The next step in conceptualizing the UI is the organization of tasks, actions, and information into clusters called information spaces. Information spaces will generally correspond to windows, dialog boxes, and menus in the next stage of interface design; i.e., the detailed design or Productization stage of interface evolution.

The organization of actions and information into separate spaces will be determined by the way users think about their tasks rather than, as is often the case, the way developers think about the application. For example, developers might think a logical grouping is to aggregate information about voice parameters into separate dialog boxes based on the nature of the signal; e.g., voice-related versus speech-related. Researchers, on the other hand, might need information aggregated based on whether a parameter is a primary or a secondary predictor of fatigue, or is robust in the presence of transmission noise or background voices. In some cases the developer and the end user perspective will coincide, but when a discrepancy arises the user's perspective will take precedence in the design.

5.1.1.2.1.3 Describe task flow

In parallel to describing the information spaces, we will define the interaction flow. This flow description is a more detailed account than the workflow defined in the previous design stage (contextualizing the problem) and specifies which dialog boxes or tasks can be invoked from a given dialog box as well as the rules governing the flow. For example, the application could "force" the user to select a specific individual or a set of individuals in an operational team before displaying its fatigue parameters. As another example, the application could enable the user to invoke the Save or Print dialog box from the Voice Analysis dialog box. A major objective in this stage is to articulate an interaction flow that is both *efficient* and *effective*. Efficient means the UI requires a minimal number of steps to execute a frequently performed task. Effective means the UI does not overload each dialog box with too many options.

5.1.1.2.2 Deliverable - Conceptualizing the design

The output of this stage will be a conceptual model for both the research-oriented and the operational application of the Fatigue Assessment Tool. These models will be in the form of wire frame representations of the two applications that specify the various menus and dialog boxes, the information content of each information space, and the set of actions that can be invoked from each information space.

5.1.1.3 Productize the design

The third stage of the UI design, sometimes referred to as detailed design, will be to **productize the design**. During this stage the details of the conceptual model of the interface will be developed. Tasks will include selecting appropriate representations for the different UI objects, actions, and information in a given information space, as well as presenting the collection of interface elements in an aesthetic and perceptually efficient manner.

5.1.1.3.1 Tasks – Productize the design

5.1.1.3.1.1 Select ways to represent each UI element

Basic usability principles will drive the selection of appropriate representation methods and labels for each control and display element. For example (see Figure 10), a spin box (right side) rather than a text box is best used when the user's task is to specify a discrete value within a specific range such as when calibrating an instrument; this implementation minimizes users' mental workload and resultant errors.

Figure 10. Alternative representation for frequency setting control; a text box (left side) places the burden on the user to remember the range of acceptable frequencies and increments, and a spin box (right side) places this burden on the user-interface

On the other hand, if the user's task is to name an object (such as a file before saving it) and no restrictions on the selected name are imposed, then a text box will be the more appropriate con-

trol for the action. In naming interface controls such as text box or drop-down list box, we will use labels and/or associated explanatory text that are compatible with the expected knowledge level and experience of the user.

UI presentation principles will guide the detailed design of each dialog box and window. These principles aim at maximizing the ease with which dialog boxes can be scanned for critical information or controls. These principles also aim at maximizing the aesthetic qualities of the dialog box or window to enhance user experience. An example of a detailed design activity and its associated presentation principle is the placement of controls within the dialog box. In order to enhance scanning of a list of items in a dialog box and therefore reduce the perceptual load on the user, the placement of controls should be aligned and related controls should be in close proximity. Repetition of layout formats between and within dialog boxes will enhance control location. For example, repeating the format of the Cancel, OK, and Help button in different dialog boxes and placing this button aggregate in similar locations on different dialog boxes will significantly improve usability of the application and resulting task performance.

5.1.1.3.1.2 Compose error messages

During this stage we will focus on error prevention and also outline the set of error messages and their content to be displayed when the user commits an error. The emphasis in the UI design will be on preventing error by selecting better controls and more appropriate labels, rather than a reactive mode of allowing errors and informing the users after the fact. However, since not all errors can be prevented, when an error occurs we will provide informative error messages to help the user understand what the error was, why it was committed, and how to fix it.

5.1.1.3.2 Deliverable – Productize the design

The primary output of this stage will be a user-interface prototype with limited functionality. The degree of functionality will be determined by usability considerations. For example, a task or a dialog box that supports complex or unfamiliar functionality and might potentially confuse the users, will be made dynamic in the prototype to enable us to test its usability with potential users. Simple dialog boxes or tasks will be represented in the prototype as static elements. A Detailed UI Design Specification will complement the prototype. This specification will include a description of the static and dynamic aspects of each dialog box and an outline of the action flow associated with executing each task.

5.1.1.4 Evaluating the design

This stage of the design process is more functional than structural. Although it appears as the last stage in the UI design process, it is actually invoked throughout the design process. Two of the major forms of evaluation performed during the design stages include *heuristic evaluation*, or walkthrough analysis by a human factors expert, and *usability testing*, or evaluation of a design with actual or representative end users.

5.1.1.4.1 Tasks – Evaluating the design

5.1.1.4.1.1 Heuristic/Expert walkthrough analysis

Heuristic evaluation is an evaluation of an interface that compares a proposed design against a set of usability principles and best practices. These principles and practices are derived from the

user interface design literature and represent the current state of knowledge in effective design. A usability expert who did not participate in the interface design efforts will perform the heuristic evaluation to provide an unbiased, independent perspective. Since the design effort will be driven by these principles, we expect to find few usability issues during the heuristic evaluations. Nevertheless, throughout the design stage we will conduct such evaluations of the different design ideas.

5.1.1.4.1.2 Usability Evaluation

Usability evaluation is an evaluation of the interface that involves actual or representative users performing various user tasks with either a static or a dynamic prototype of the UI. Because usability evaluation focuses on task execution and a complete static or dynamic prototype is required, this evaluation will be performed towards the end of the design phase when actual prototypes are available. Note that heuristic evaluation will be performed on individual dialog boxes as well as constellations of dialog boxes and windows that will support an operator task or function. Usability evaluation in contrast, will be performed on constellations of dialog boxes and windows that support a subject task or function. Participants will be given one or more tasks and will be asked to execute each task using the proposed UI. Participants' comments, actions, and ability to complete each task will be recorded, and results of the usability evaluation will be used in enhancing and improving the UI design.

5.1.1.4.2 Deliverable – Evaluating the design

The output of each evaluation will be a set of ranked and prioritized lists of usability issues and recommended redesigns. These issues and the proposed redesigns will be incorporated into the interface design. The objective of the usability evaluation is to ensure that no "show-stopper" or "critical" usability limitations are present in the final design. Thus the final deliverable of this stage will be a list of pending usability issues, all of which will be "nice to have". These will be non-critical, will not impact users' performance and will only marginally impact learning; however, in combination they could impact the overall impression of the final application.

5.1.2 Summary

The user-interface design process and tasks above represent a highly systematic and thorough approach to user interface design. This approach to user-interface design ensures effective and efficient interaction, a basic requirement for effective application, as well as feature traceability, an important element in a rationally designed application. By tying user-interface design features (or design decisions) to tasks, user characteristics, and operational problems, one can change or modify features in a rational way by visualizing their impact on users' task characteristics and the problem the application is designed to support. Likewise, if any changes occur in the users of the application, such as a decision to employ users with a different skill set; in users' tasks due to new regulations; or in the operational problem, the affected features in the user-interface design can be quickly identified and modified to meet the new requirements. This approach to user-interface design as proposed above is systematic, thorough, flexible, and accountable; therefore, it provides assurances to result in an effective and efficient user interface, one that will support and promote effective and efficient interaction.

6 Recommendations and Conclusions

The experience with the DCIEM Map Task corpus, combined with customer discussions about likely operational use, underlies several recommendations for further development of fatigue assessment through voice and speech analysis.

- The primary recommendation is for analysis of speech data that is paired with standard measures of fatigue, and perhaps of task performance. It would be especially useful to do this with many additional subjects, especially including females. (There were none in the map-"giver" set analyzed for this project.) The subjects of the corpus provide a substantial amount of speech for analysis, several minutes' worth — but there are only six of them.

- Further research of the landmark-based and the non-fluency-based measures appears useful, particularly with many more subjects, though much less speech from each. The former is already highly automated, but the results may be too subject-specific: The several minutes of data from each subject are probably sufficient to characterize his speech well — indeed, far less speech appears sufficient to produce rather strong results —, but variation across subjects is presumably much broader than these six subjects can demonstrate: all were male, healthy, young (20-45 years), etc. Interestingly, landmark-based measures of the subjects' intermediate dialogues were quite similar, overall, to the fatigued dialogues, and very different from the fresh ones. As for the non-fluency measures, these are simple and probably robust, and they apparently corroborate one of the landmark findings. However, they are perceptual. A major issue for processing large numbers of subjects, and perhaps for deployment of a Voice-based Fatigue Assessment System, will be automating the detection of such non-fluencies. A secondary issue with larger numbers of subjects will be a determination of which non-fluency features are predictive of fatigue. The thresholds for detecting fatigue in these subjects are nearly subject-independent, but it would obviously be essential to know if, e.g., a future subject stammered or stuttered when fresh.

- Speech data collected under more realistic circumstances must be analyzed. Although the corpus is very well designed from some points of view — for example, replication, acoustic signal quality, stimulant comparisons (dextroamphetamine, modafinil) — it does not contain data for the most common and casually ingested stimulant, caffeine, a diuretic. Because of this, and perhaps other aspects of the protocol such as constant activity, the subjects' voices may have avoided the roughness that is commonly observed to correlate with fatigue: Neither the nonlinear dynamics techniques nor MDVP's pitch-related perturbation measures show evidence for rough or irregular vibratory modes in the larynx, a finding consistent with good hydration and at least moderate alertness. This should be revisited when more realistic data can be analyzed.

- There appears to be no point to further investigations of formant frequencies in isolation, at least in unstressed vowels — so long as the language is English and the speakers native. It is possible to infer articulation from formant trajectories (some of which are used in LMC), but there are other techniques that can more directly measure articulation. The stress rules for English directly undercut any assessment of fatigue or inattention in unstressed vowels; most

other languages, however, have sufficiently different rules that this conclusion would not apply. Indeed, English from non-native speakers might demonstrate reliable effects (albeit varying with their *native* language), potentially useful for both the pilots and the foreign air-traffic controllers of commercial aviation.

- Measures of pitch (or F0) might be found useful at a later stage. Data collected to study such measures more closely should include specific linguistic structures appropriate to the measures. For example, pitch was expected to show a fatigue effect especially for the ends of yes/no questions, and one subject with many such questions showed an effect. Others, generally with fewer questions, did not. Analysis to investigate this point further should focus on the small number of instances that do occur, or — ideally — should use data from a protocol that causes all subjects to produce many questions. Any pitch-related analyses should obviously include female voices, which were absent here.

- Similar remarks apply to the long-term average spectrum, particularly regarding the production of bursts and fricatives. Passages with many strong fricatives and bursts, such as /sh, ch, t, k/, could be particularly helpful in demonstrating any reliable fatigue effects. Note that LMC already detects these phonemes automatically and reliably, and could be integrated with the algorithms that were developed in Phase I to evaluate LTAS.

- At the user-interface (UI) level, a primary recommendation is that the system be designed with careful attention to the intuitions of the (future) operators. For example, expressing the subject's degree of fatigue on a scale related to alcohol intoxication would provide most operators with immediate intuition about the subject's performance. Such scales must, however, be related to the tasks on which the subject requires evaluation: General speech production *per se* may be expected to degrade with fine motor control, but this might have little correlation with, say, memory tasks or spatial reasoning needed by the subject. Those tasks and capabilities may correlate well with specific speech features (not general production), but this could not be evaluated in Phase I. (Given the somewhat different mean performance by males and females on such tasks, it will again be important to include female subjects.)

7 Literature Cited

Bachorowski, J. (1999). Vocal expression and perception of emotion. *Current Directions in Psychological Science*. Publication of American Psychological Society, 53-57.

Bitar, N. and C. Espy-Wilson. 1996. A knowledge-based signal representation for speech recognition. *Proc. ICASSP '96*, May, Atlanta, GA. pp. 29-32.

Duchon, J. & T. Smith (1993). Extended workdays and safety. *International Journal of Industrial Ergonomics*, **11**, 37-49

Eddy, D., M. Dalrymple, & S. Schiflett (1992). "Comparative Effects of Antihistamines on Aircrew Mission Effectiveness under Sustained Operations". US Army Medical R&D Command and Armstrong Laboratory, Report # AL-TR-1992-0018.

Federal Highway Administration, Office of Motor Carriers (October, 1998). PERCLOS: A valid psychophysiological measure of alertness as assessed by psychomotor vigilance. FHWA publication # MCRT-98-006.

Fell, H., L. Ferrier, C. Espy-Wilson, S. Worst, E. Craft, K. Chenausky, J. MacAuslan, & G. Hennessey (2000). *Automatic Analysis of Infant Babbling in EVA, the Early Vocalization Analyzer,* Pres., Am. Speech-Lang.-Hearing Conf. (ASHA), November, Washington, DC.

Hollien et. al. (2001) Effects of ethanol intoxication on speech suprasegmentals. *J. Acoustical Society of America*, **110** (*6*), p. 3198-3206

MacAuslan, J., K.Chenausky, J.Juda, & I.Manev. (1997). "Identifying Laryngeal Attractor Changes from Periodicity Diagrams", *Proc. Nonlinear Signal and Image Processing Conf.*, Mackinac, MI.

National Highway Traffic Safety Administration. (1999). The NHTSA and NCSDR program to combat drowsy driving: A report to Congress on the Collaboration between the National Highway Traffic Safety Administration and the National Center on Sleep Disorders Research (http://www.nhtsa.dot.gov/people/perform/human/drowsy2/drws-cov.htm).

Peitgen, H-O, H. Jurgens, & D. Saupe. (1992). *Chaos and Fractals*. New York: Springer-Verlag.

Pigeau R., P. Naitoh, A. Buguet, C. McCann, J. Baranski, M. Taylor, M. Thompson, & I. Mack. (1995). "Modafinil, d-amphetamine and placebo during 64 hours of sustained mental work. I. Effects on mood, fatigue, cognitive performance and body temperature", *Journal of Sleep Research*, **4**, p. 212-228

Press, W., S. Teukolsky, W. Vetterling, & B. Flannery. (1992). *Numerical Recipes in C*, 59-70. New York: Cambridge University Press. Also at: http://lib-www.lanl.gov/numerical/bookcpdf.html

Sheskin, D. (1997). *Handbook of Parametric and Nonparametric Statistical Procedures*. Boca Raton, FL: CRC Press.

Tepas, D., M. Paley, & S. Popkin (1997). Work schedules and sustained performance. In Salvendy, G. (Ed.), *Handbook of Human Factors & Ergonomics, Second Edition*, 1021-1058. New York: John Wiley and Sons.

Titze, I. (1995). Workshop on Acoustic Voice Analysis: Summary Statement, National Ctr. For Voice and Speech, Iowa City, IA.

Titze, I., R. Baken, & H. Herzel (1992) Evidence of chaos in vocal fold vibration. In I. Titze (ed.) *Vocal Fold Physiology: Frontiers in Basic Science*. San Diego, Singular Publishing, p. 143-188.

Wayland, S., J. Miller, & L. Volaitis. (1994). The influence of sentential speaking rate on the internal structure of phonetic categories. *Journal of the Acoustical Society of America*, **95**, 5, 2694-2701.

Whitmore, J. & S. Fisher (1996). Speech during sustained operations. *Speech Communication*, **20**, 55-70.