2006 IEEE International Symposium on Signal Processing and Information Technology

Detecting Fatigue From Voice Using Speech Recognition

H.P. Greeley, E. Friets, and J.P. Wilson — Creare, Inc., Hanover, New Hampshire, USA, {hpg, emf, jpw}@creare.com

S. Raghavan and J. Picone — Center of Advanced Vehicular Systems, Mississippi State University, {raghavan, picone}@cavs.msstate.edu

J. Berg — The MathWorks, Inc., Natick, Massachusetts, USA, Joel.Berg@mathworks.com

Abstract *– Military and civilian experience has shown that long-duration assignments present increased risk of performance failures as the mission progresses. This is due to interruption of normal sleep cycles and psychological pressures of the work environment. There continues to be a need for a non-intrusive fatigue assessment system to successfully monitor the level of alertness of personnel during critical missions and activities. Experimental results on human voice show that specific phones have a predictable dependence on fatigue. Hence, precise phonetic identification and alignment are important to voice-based fatigue detection. This paper explores techniques for detecting fatigue from voice using speech recognition to obtain phonetic alignments. A confidence measure was used to filter out less likely word hypotheses from the ASR's output. In this paper we restricted our analysis to dealing with out-of-vocabulary words. The results obtained from voice show strong correlation with other standardized tests such as Sleep Onset Latency and Sleep, Activity, Fatigue, and Task Effectiveness.*

Keywords *– fatigue detection, speech recognition, confidence measures*

1. INTRODUCTION

The unique characteristics of the military and aviation environment make war fighters and civilian pilots particularly susceptible to fatigue. Being able to quickly and non-intrusively monitor an airman's or soldier's level of alertness prior to and during the undertaking of a critical mission activity would provide commanders with critical information regarding personnel assignments and certainly save lives and increase the likelihood of mission success. Unfortunately, there are no cognitive assessment tests that have been proven to be effective in the field under conditions of high stress and limited testing time per subject. This paper describes an approach to the development of a voice-based fatigue prediction system.

Changes in the articulation of voiced sounds due to fatigue could be considered representative of changes in the body's voice production mechanisms. Change in discrete voice parameters (such as fundamental frequency and word duration) has been reported in the literature, however, no single voice characteristic demonstrates a consistent and reliable change as the speakers become fatigued [1][2][3]. Rather than study any one specific voice parameter, our approach is to observe a more holistic representation of the speech signal, the cepstral transformation associated with specific speech phonemes. The coefficients of this transformation, referred to as Mel-frequency cepstral coefficients (MFCCs) [4] are used in association with an automatic speech recognition system (ASR).

Therefore, a straightforward way to automate this process is to use the output of an ASR system to identify the location of key phonemes, as shown in Fig. 1. The time marks produced from the recognition segmentation are used to identify the corresponding MFCC vectors for a given phoneme, and these vectors are in turn used in the fatigue detection system. Since these systems tend to be deployed in extremely noisy environments, the ASR system must be extremely robust, and the fatigue detection system must be tolerant of recognition errors.

Also being a non-intrusive approach, the system must be robust to out of vocabulary words. We used a word posterior-based confidence measure to further improve the overall reliability of the system [5]. A reasonable improvement in fatigue analysis was observed when confidence measures were used for utterances which had out-of-vocabulary words. A threshold to reject incorrect or less probable words was determined by observing the region of convergence for the word posteriors.

2. USING VOICE TO DETECT FATIGUE

Whitmore and Fisher have shown that speech data follow the same trend as the data from cognitive tests and subjective measures of alertness [1]. They also noted a strong circadian trend, as the best voice performances occur during normal waking hours, and the worst performances occur during normal sleeping hours. Satio, *et al*. reported changes in the appearance of sound spectrograms from analysis of specific, repeated utterances as a pilot experienced hypoxia prior to a fatal F-104 accident [3]. These results support the contention that voice characteristics are directly related to the speaker's level of performance which, in turn, is affected by his or her level of fatigue.

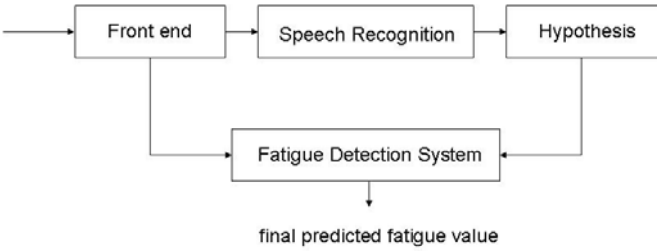

Fig. 1: Integration of the fatigue detection system with an Automatic speech recognition system.

2.1 *Using MFCCs for Fatigue Cues*

An initial data set, referred to as the Phase I data, was developed to further study the effect of fatigue on formant frequencies [6]. Ten volunteers were asked to speak sentences containing words from a set of 37 words. The recordings were made four times a day, before and after a night of sleep depravation. Reaction time was measured just before making the recordings, and sleep latency was measured to determine the general level of fatigue. Approximately 12,000 formant frequencies were analyzed, and 19 of them showed significant correlation with reaction time. Several showed good correlation with the sleep latency tests as shown in Table 1. The results from the table show that the formant frequencies are related to the subject's level of alertness, which is directly related to the fatigue level of the subject.

Initial Phase I analysis confirmed a dependence between formant frequencies and fatigue. Therefore, it became necessary to process the recorded speech signal in a manner that would reveal the required information (*e.g.* vocal tract response which affects the formant frequencies) from the speech data, and this was accomplished by using cepstral analysis techniques.

Cepstral analysis results in the calculation of a discrete number of coefficients called cepstral coefficients. With this analysis, the entire human speech production process can be described by only a few cepstral coefficients. Instead of tracking changes in specific vocal metrics, such as formants, we can use these cepstral coefficients. These cepstral coefficients constitute the Mel Frequency Cepstral Coefficients (MFCC). The MFCC feature vector used for fatigue analysis contains 36 coefficients. The feature vector is comprised of 12 cepstral coefficients, along with their first and second time derivatives. Also of interest is how the MFCC vector changes as a function of the subject's level of fatigue. Fig. 2 shows an example of how the MFCC vector changes over a four-day period of sleep restriction.

From Phase I analysis, it was found that certain formant frequencies are more closely related sleep latency tests than others. Hence, it would be useful to analyze the MFCCs of different phonemes and determine the phonemes that show good variations in the spectral domain due to fatigue. The MFCCs for various sounds were analyzed, and the sounds that were most affected by fatigue were determined.

For example, a fatigue analysis was performed on the sound 't'. A notable change in the MFFC vector was observed as the subject became increasingly fatigued. Fig. 2 gives the correlation of the MFCC vector for different trials with the initial trial, and this is described in the legend of Fig. 2. It can be observed that the correlation between Trial 1 (12 hrs awake) and Trial 10 (39 hrs awake) is higher (0.82) when compared to the correlation between Trial 1 and Trial 21 (78 hrs awake) (0.19). This is an indication that the MFCC components change as the subject gets increasingly fatigued. So, the correlation metric can be used as a prediction metric to determine fatigue.

Table 1: Correlation between formant frequency and performance

Sound	Formants			
	F1	F2	F3	F4
[o] clock	.486	.339	.710	.565
[^] upper	.416	.352	.689	.680
[ay]highly	.356	.359	.3320	.682
[iy] keep	.511	.241	.396	.228
[m] matter	.574	.567	.343	.118
[o] coughing	.367	.071	.487	.310
[n] note	.386	.114	.071	.000
[n] night	.389	.095	.095	.192
[^] fuzzy	.324	.187	.388	.243
[uw] two	.360	.122	.205	.298
[ae] chatter	.359	.152	.316	.351
[ay] time	.326	.045	.326	.045
[ae] cabin	.313	.105	.310	.164
[y] yet	.308	.045	.21	.152
[U] took	.055	.344	.705	.612
[iy] serene	.205	.623	.182	.071
[n] now	.164	.538	.576	.460
[r] rather	.036	.032	.310	.517
[o] not	.045	.265	.164	.109

2.2 *Preliminary Fatigue Experiments*

During a 34-hour period of sleep deprivation, six non-medicated subjects were asked to recite a list of 31 words at six testing times (10:00 AM, 4:00 PM, 10:00 PM, 4:00 AM, 10:00 AM, and 4:00 PM). These testing times were selected to represent circadian high and low points in performance

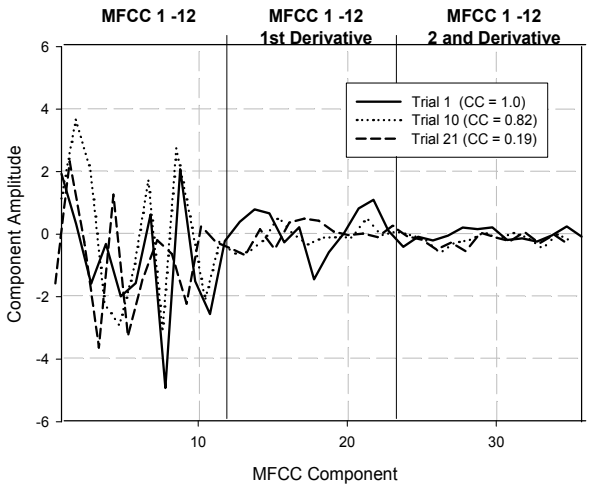

Fig. 2: Changes in the MFCC vector during four days of sleep restriction. The vector during Trials 1 and 10 match more closely than the vector at Trial 21.

[7].

Also measured during these testing times was Sleep Onset Latency (SOL), which is the gold standard for sleepiness testing. This test involves having the test subject lie on a bed in a quiet, darkened room and telling the subjects to go to sleep. The time that it takes them to fall asleep, as measured by an electroencephalogram (EEG), is the sleep onset latency (SOL) [8]. Between tests, subjects were allowed low arousal activities such as reading and watching TV.

Fig. 3 shows the group average change in both SOL and our voice correlation metric for the sounds 'p' and 't' over the 34 hour testing period. The correlation coefficient between SOL and time awake is -0.825, and between voice correlation of sounds 'p' and 't' to time awake is -0.89, and -0.67 respectively. We estimate that time awake accounts for 68%, 79%, and 45% of the variation of SOL, voice correlation of sounds 'p' and voice correlation of 't' respectively.

All three metrics show a circadian peak at 16 hours. However, the SOL peak is significantly larger than the voice metric peaks. Circadian, according to any standard dictionary, means "exhibiting periodicity in a 24-hour period." For example, our sleep cycle can be considered to have a circadian trend (*i.e.*, we sleep better at night than during the day.) Similarly, fatigue levels were observed to be higher during normal sleep hours than at regular working hours, which explained the circadian trend. The circadian pattern has been observed in many alertness versus time experiments [7],[8]. This difference in circadian sensitivity tends to reduce a correlation coefficient-based quantitative comparison.

3. GENERATING PHONE ALIGNMENTS USING ASR

The general architecture of the fatigue detection system is shown in Fig. 1. The ASR system provides time-aligned word (and phone) hypotheses as output. The fatigue prediction software relies on ASR to determine the MFCC vectors for specific phones. It is critical that the correct phones are identified from the input stream of audio data. Eight mixture Gaussian Mixture Models (GMMs) were used to represent the statistics of the training data. Due to the small vocabulary size, a loop grammar was used as a language model. The output of the ASR system was one-best phone alignments with the word likelihood score annotated to each phone hypothesis. The word error rate of the ASR system is dependent on various factors which are beyond the scope of this paper.

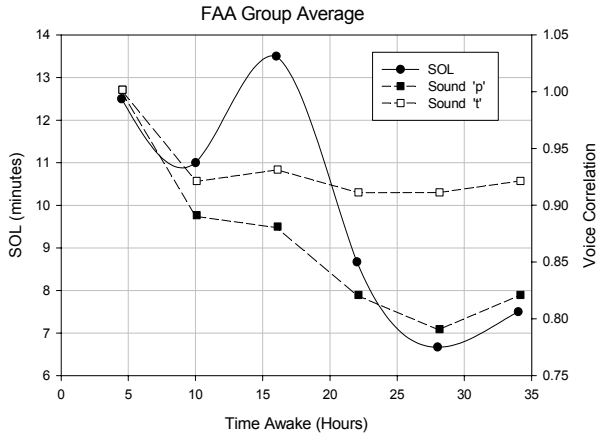

Fig. 3: Change in the voice vector Vs change in Sleepiness, SOL, sleep onset latency

For fatigue detection, our main interest is in finding the presence of certain phones with a high degree of confidence. Word error rate is not a very critical factor. Annotating the one-best hypothesis with confidence measures would help in reducing false hypothesis. A threshold can be used to filter less probable words in the final hypothesis. Word posteriors computed from word graphs were used as a confidence estimate [10]. Experimental results show that the Voice Correlation metric comes closer to the SOL test when word posteriors are used to filter out less likely words.

There is an elegant method to compute posterior probabilities from word-graphs [5]:

$$p(w, t_b, t_e \mid x_1^T) = \sum_{w_b} \sum_{w_e} p(w_b, w, w_e \mid x_1^T) \qquad (1)$$

$w \rightarrow$ Single word

$t_b, t_e \rightarrow$ Start and end time of the word

$x_1^T \rightarrow$ Acoustic vector from time 1 to T

$w_b \rightarrow$ Denotes all word hypothesis sequences preceeding w

$w_e \rightarrow$ Denotes all word hypothesis sequences succeeding w

The probability of passing through the link W is calculated by determining the probability of reaching the start node of the word from the preceding nodes and the probability of leaving the end node to any of the succeeding nodes. The former is called the forward probability and the latter as the backward probability. A forward-backward type algorithm is used to traverse through the lattice and compute the probabilities. The right-hand side term in Eq. 1 cannot be computed directly, but can be decomposed into likelihood and priors using Bayes rule [5] as shown in Eq. 2.

The numerator term of Eq. 2 is calculated by the forward-backward algorithm. The denominator term is the byproduct of the forward-backward computation and is defined as the sum of all paths through the word graph. The purpose of the denominator term is to normalize the posterior values. The posteriors computed in this manner can be used as a confidence measure. The word posteriors are annotated to the one best output generated by the ASR system. The experiments discussed in this paper show the effectiveness of using word posteriors when the input test data contained out of vocabulary words (OOVs).

$$\sum_{w_b}\sum_{w_e} p(w_b, w, w_e \mid x_1^T) = \frac{\sum_{w_a}\sum_{w_e} p(x_1^T \mid w_a, w, w_e).p(w_a, w, w_e)}{p(x_1^T)} \quad (2)$$

$p(x_1^T \mid w_a, w, w_e) \rightarrow$ Acoustic model probability

$p(w_a, w, w_e) \rightarrow$ Languagemodel probability

and

$$p(x_1^T) = \sum_{w}\sum_{w_a}\sum_{w_e} p(x_1^T \mid w_a, w, w_e).p(w_a, w, w_e)$$

4. EXPERIMENTAL RESULTS

Analysis was performed on the voice data from two test subjects who underwent a night of sleep deprivation. At six test epochs, separated by 6 hours, these subjects each recited from two word lists. The ASR system was trained to recognize words from the training list. During fatigue analysis, the speech recognition system was presented words from both the training list and the foreign list which contained words not in the training set. For both subjects, the confidence metric (CM) observed when the speakers recited from the first list had a higher average value and smaller standard deviation than that observed when the speaker recited from the foreign list. Table 2 presents these results.

The data presented in Table 2, represents the degree of overlap between the CM distribution of training words and the CM distribution of foreign words. A receiver operating characteristic (ROC) curve was plotted to predict the true and false "correct word" prediction performance of the CM. A ROC plot for one of the subjects is shown in Fig. 4. Using a CM threshold value of -75 to flag foreign words, the ROC curves had areas of 0.85 and 0.80 for the two subjects. These values are benchmarks indicative of a "good" prediction performance.

The confidence metrics' effect on fatigue prediction performance is illustrated in Fig. 5 andFig. 6. Fig. 5 depicts the subject's normalized sleep onset latency (SOL) and voice-based fatigue prediction for the sound 'p' (Vc) at each of the six trials. As can be seen in the figure, the performance of the voice metric is best when using voice data from only the training set list. Using the voice input of both training set and foreign set words, with no confidence metric, little agreement is seen between voice and SOL and the error rate between them was 0.33. However, by using the CM (with a threshold setting of -75) to filter out foreign words the voice-based prediction is significantly closer to both the SOL and the "training input only curves", and the error rate dropped to 30%, which is a 9% absolute improvement.

Table 2: An analysis of the confidence metric distribution for two speakers on different set of words

	Subject 6	
	Training	Foreign
Average CM	-72.22	-81.51
CM Standard Deviation	3.10	11.34
	Subject 8	
	Training	Foreign
Average CM	-70.24	-82.41
CM Standard Deviation	3.50	15.16

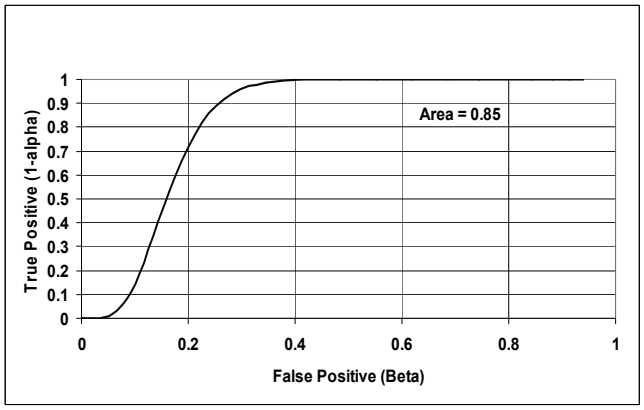

Fig. 4: Receiver operating characteristics plot for subject 6

The cyclic patterns observed in Fig. 5 are due to circadian rhythms. Over the 30 hours between Trial 1 and Trial 6, a full circadian cycle has elapsed. The SOL reflects the circadian influence of an individual's need to sleep. A more direct way to match a speaker's overall performance and circadian influences is to use the speaker's body temperature and his or her time without sleep. This is accomplished using the Sleep, Activity, Fatigue, and Task Effectiveness (SAFTE) model [11].

Fig. 6 shows speaker's SAFTE score and voice-based fatigue prediction for the sound 'p' (Vc) at each of the six trials. As was the case with the SOL, the SAFTE model is closest to the "training list only" words. Using voice input containing mixed words (training set and foreign words), the CM-based word filter provides a significant fatigue prediction improvement over the use of the full mixed word input. The error rate between the curves improved from 0.15 to 0.12, which is a 20% relative improvement. The OOV error rate for this experiment was 61.7%.

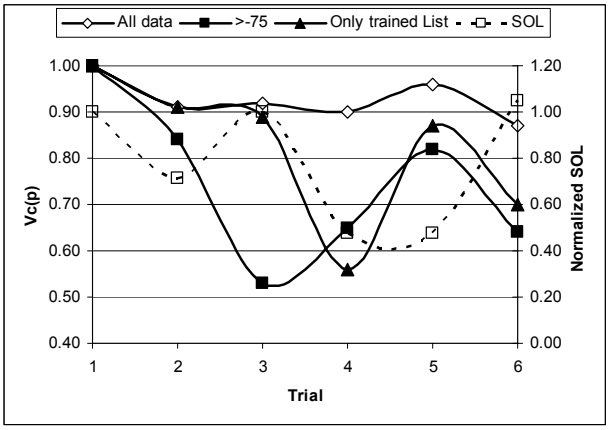

Fig. 5: Comparison of the trend between SOL and voice correlation for sound 'p' with and without a confidence metric

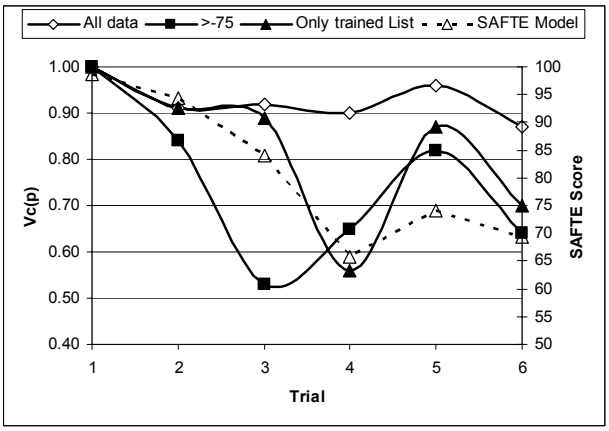

Fig. 6: Comparison of the trend between SAFTE and voice correlation for sound 'p' with and without confidence metric

5. CONCLUSIONS

In this paper we have presented a first attempt to measure fatigue detection using a speech recognition system. The correlation-based voice metric discussed in this paper compares favorably with the gold standard for measuring fatigue, sleep onset latency. The correlation measure also compares favorably to SAFTE measures. Confidence measures played a significant role in fatigue analysis on unseen words. The error between the SAFTE and voice based fatigue metric was decreased by 20% using confidence measures. Future work will be focused on more extensive evaluations on a much larger operational database, and on ways to improve the robustness of the system to recognition errors.

ACKNOWLEDGEMENTS

This study was supported by contract number F33615-03-C-6334 from the Air Force Research Laboratory.

REFERENCES

[1] J. Whitmore and S. Fisher, "Speech During Sustained Operations," *Speech Communication*, Vol. 20, pp. 55–70, 1996.

[2] M. Vollrath, "Automatic Measurement of Aspects of Speech Reflecting Motor Coordination," *Behavior Research Methods, Instruments and Computers*, Vol. 26, No. 1, pp. 35–40, 1989.

[3] I. Saito, O Fujiwara, N Utsuki, C Mizumoto, and T. Arimori, "Hypoxia-Induced Fatal Aircraft Accident Revealed by Voice Analysis," *Avaition, Space, and Environmental Medicine*, Vol. 51, pp. 402–406, 1980.

[4] J. Picone, "Signal Modeling Techniques in Speech Recognition," *IEEE Proceedings*, Vol. 81, No. 9, pp. 1215-1247, 1993.

[5] F. Wessel, R. Schlüter, K. Macherey, and H. Ney, "Confidence Measures for Large Vocabulary Continuous Speech Recognition". *IEEE Transactions on Speech and Audio Processing*, Vol. 9, No. 3, pp. 288-298, 2001.

[6] H. Greeley, J. Berg, E. Friets, J. Wilson, G. Greenough, J. Picone, J. Whitmore and T. Nesthus, "Fatigue Prediction Using Voice Analysis," submitted to the Behavioral Research Methods, February 2006.

[7] T. Roth, T.A. Roehrs, M.A. Carskadon, and W.C. Dement, "Daytime Sleepiness and Alertness in Principles and Practice of Sleep Medicine," Second Edition. In Kryger, M.H., Roth, T., and Dement, W.C. (Eds.), WB Saunders Company, Philadelphia, 1989, pp. 14–23.

[8] T. Roehrs and T. Roth, "Multiple Sleep Latency Test: Technical Aspects and Normal Values," J. Clin. Neurophysiol, Vol. 9, Ch. 1, 1992, pp. 63–67.

[9] J.M. Gregory, X. Xie, and S.A. Megel, "SLEEP (Sleep Loss Effects on Everyday Performance) Model," *Aviation, Space, and Environmental Medicine*, Vol. 75, Sup 1, pp. 125-133, 2004.

[10] L. Mangu, E. Brill and A. Stolcke, "Finding Consensus in Speech Recognition: Word Error Minimization and Other Applications of Confusion Networks," *Computer, Speech and Language* , Vol. 14, No. 4, pp. 373-400, 2000.

[11] S. R. Hursh, D. P.Redmond, M L. Johnson, D. R. Thorne, G. Belenky, T. J. Balkin, W. F. Storm, J. C Miller and D. R. Eddy, "Fatigue models for applied research in Warfigning," *Aviation, Space, and Environmental Medicine*. Vol. 75, No. 3, pp A44 - A53, 2004.