M.I.T Media Laboratory Perceptual Computing Section Technical Report No. 513

Analysis and Classification of Stress Categories from Drivers' Speech

Raul Fernandez and Rosalind W. Picard

MIT Media Lab. Room E15-391
20 Ames St., Cambridge, MA 02139
{galt,picard}media.mit.edu

Abstract

In this paper we explore the use of features derived from multiresolution analysis of speech and the Teager Energy Operator for classification of drivers' speech under stressed conditions. The potential stress categories are determined by driving speed and the frequency with which the driver has to solve a mental task while driving. We first use an unsupervised approach to gain some understanding as to whether the discrete stress categories form meaningful clusters in feature space, and use the clustering results to build a user-dependent recognition system which combines local discriminants of 4 discreet stress categories. Recognition results are reported for 4 subjects.

1 Introduction

Much of the current effort on studying speech under stress has been aimed at detecting several stress conditions for improving the robustness of speech recognizers; typical categories of speech under stress have targeted perceptual (e.g. Lombard effect), psychological (e.g. timed tasks), as well as physical stressors (e.g. roller-coaster rides, high G forces) [1]. In this work we are interested in modeling speech in the specific context of driving where the speech has been produced under varying conditions of cognitive load which are hypothesized to induce a level of stress on the driver. The results of this research may be not only relevant to building recognition systems that are more robust in the context described, but also applicable to and inspired by applications that may infer the underlying affective state accompanying an utterance. We have chosen to simulate the scenario of driving while solving a stressful task on the phone as an application in which knowledge of the driver's state may prove relevant to the dynamics of driving and may provide benefits ranging from a more fluid interaction with a speech interface to more serious safety concerns.

The recent literature discussing the effects of stress on speech applies the label of *stress* to different phenomena surrounding the production of the acoustic signal. Following the taxonomy proposed by Murray et al. [2], we are investigating the effect on speech of what the authors call "third-order stressors," that is, the effect of external stimuli as well as underlying affective conditions.

2 Speech Database

The speech data was collected in an experiment in a driving simulator at the Nissan's Cambridge Research Lab. Subjects were asked to complete a series of rounds while engaged on a simulated phone task: while the subject drove, a speech synthesizer prompted the driver with a math question consisting of adding up two numbers whose sum was less than 100. We controlled for the number of additions with and without carry-ons in order to maintain an approximately constant level of difficulty across trials. The two independent variables in this experiment were the driving speed and the frequency at which the driver had to solve the math questions. Subjects drove at 60 m.p.h. in the low speed condition and at 120 m.p.h. in the high speed condition (the perceptual speed in the simulator is approximately half). When a subject complained of motion sickness in the high speed condition, the speed was reduced to 100 m.p.h. The frequency at which the driver was prompted for an answer was once every 9 seconds in the slow condition, and once every 4 seconds in the fast condition. The driver's answers were captured by a head-mounted microphone and recorded in VHS format.

3 Feature Set

Nonlinear features of the speech waveform have received much attention in the context of studying speech under stress; in particular, the Teager Energy Operator

This work was supported by the MIT Media Lab Digital Life Consortium.

(TEO) has been the subject of several studies which have proposed its robustness to noisy environments and usefulness in stress classification [3],[4], [5]. Another useful approach for analysis of speech and stress has been subband decomposition or multi-resolution analysis via wavelet transforms [6],[7]. Multi-resolution analysis and TEO-based features have also been combined in the context of recognizing speech in the presence of car noise and shown to yield superior rates [5]. In this work we investigate a feature set consisting of variants of features proposed in [5] and [7] based on the TEO and multi-resolution analysis and apply it to the task of modeling different categories of drivers' stress.

3.1 Subband Based Feature Extraction

After the speech signal has been sampled at 8kHz, a wavelet packet decomposition is applied in this approach to the discrete signal $x[n]$ in order to obtain a multiresolution analysis into $M = 21$ bands corresponding to the frequency division shown in Figure 1. This process can be viewed as filtering the speech through the branches of a tree-structured filter bank, or as performing a wavelet packet decomposition and then reconstructing the subband signals from the wavelet coefficients obtained at particular scales. The wavelet packet decomposition in this implementation is based on repeated iterations of the minimum-phase 8-tap low and high pass filters associated with the orthogonal *Daubechies-4* [8].

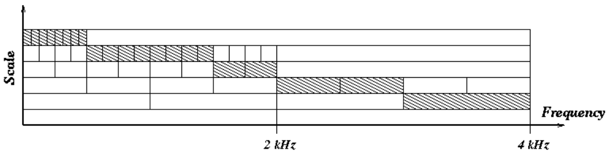

Figure 1: Subband Decomposition

Following the decomposition, the average Teager energy is found for every subband signal according to

$$e_m = \frac{1}{N_m} \sum_{n=1}^{N_m} \left| \Psi(x[n]) \right| \quad m = 1, \cdots, M \tag{1}$$

where N_m is the number of time samples in the m^{th} band and $\Psi(\cdot)$ is the discrete Teager energy operator:

$$\Psi(x[n]) = x^2[n] - x[n-1]x[n+1] \tag{2}$$

An inverse DCT transform is then applied to the log of the energy coefficients to obtain the TEO-based "cepstrum coefficients" E_l [5]:

$$E_l = \sum_{m=1}^{M} \log(e_m) \cos\left[\frac{l(m-0.5)\pi}{M}\right] \quad l = 1, \cdots, L \tag{3}$$

The extraction of the cepstral coefficients defined in (3) is applied to the speech waveform at every frame. Define then $\mathbf{E}^{[r]}$ as the $L \times 1$ vector containing the cepstral coefficients from the r^{th} frame: $\mathbf{E}^{[r]} = \left[E_1^{[r]}, \cdots, E_L^{[r]}\right]^T$. In order to reflect frame-to-frame correlations within an energy subband, the following autocorrelation measure has been proposed [7]:

$$ACE^{[r]}_{l,\tau} = \frac{\sum_{n=r}^{r+T} E_l^{[n]} E_l^{[n+\tau]}}{\mathrm{argmax}_j\left(ACE^{[j]}_{l,\tau}\right)} \quad l = 1, \cdots, L \tag{4}$$

where τ is the lag between frames, T is the number of frames included in the autocorrelation window, and j is an index which spans all correlation coefficients within the same scale along all frames to normalize the autocorrelation. Define the vector containing the logarithm of the L autocorrelation coefficients as $\mathbf{ACE\_L}^{[r]}_{\tau} = \left[\log ACE^{[r]}_{1,\tau}, \cdots, \log ACE^{[r]}_{L,\tau}\right]^T$ We define the frame-based feature vector as the set of L cepstral coefficients and the log of the L autocorrelation coefficients:

$$\mathbf{FS}^{[r]} = \begin{bmatrix} \mathbf{E}^{[r]} \\ \mathbf{ACE\_L}^{[r]}_{\tau} \end{bmatrix} \tag{5}$$

Taking the log of (4) is done to avoid modeling a finite support density distribution (which results from the normalization of (4)) with a single or a small number of Gaussians in the learning stage where the log transformation might provide for a better fit to the distribution. The values of the constants for this implementation are $M = 21$, $\tau = 1$, $T = 2$, and $L = 10$ (resulting in a feature vector of dimensionality 20). The frame features are derived from 24 msecs. of speech and are computed every 10 msecs.

4 Analysis

The speech data was collected under four distinct categories resulting from combining the slow and fast speech conditions with the slow and fast frequency of solving the math tasks. One of the hypotheses of this data collection scheme is that the combined effect of driving while engaged on the solution of a cognitive problem might produce a level of stress which might be reflected in the driver's speech. It is not clear, however, to what degree, if any, the four different conditions yield four distinct vocal states, and whether such states are reflected in the speech features under consideration. For this reason, the first section of this data analysis takes an unsupervised approach to investigate whether we can find clusters of time series which show some homogeneity with respect to a given class to gain insight as to whether the stress categories defined *a priori* are also delineated in feature space. The ultimate goal of this research, however, is to be able to discriminate between these stress conditions if they indeed are relevant to the vocal data collected. In the second part of this section, we use the results of the unsupervised clustering to train classifiers to learn and predict the categorical data.

Since it is not known how the categorical patterns may vary across different subjects and whether the experimental paradigm was equally successful in inducing the desired level of stress in the response, the analysis that follows is speaker dependent. In Section 5, we present the results for four subjects.

4.1 Clustering

In this section we investigate whether homogeneous clusters emerge when we apply unsupervised clustering techniques to the data. To handle the temporal nature of the data, we model each cluster with a Hidden Markov model (HMM). HMM parameters and cluster memberships are iteratively estimated by embedding the HMM training algorithm (which learns the parameters of a cluster given its data assignment) within a K-means algorithm (which assigns time series to clusters according to the probability of membership to each cluster). The algorithm is outlined below:
Given K clusters and a data set consisting of N time series $\{X\} = \{\mathbf{x}_t^1, \cdots, \mathbf{x}_t^N\}$, let $\lambda_k^{(l)}$ $(k = 1, \cdots, K)$ be the parameters of the k^{th} HMM at the l^{th} iteration and let $\hat{k}_n^{(l)} = \text{argmax}_k P\big(\mathbf{x}_t^n | \lambda_k^{(l)}\big)$ be the cluster that maximizes the probability of the n^{th} time series at the l^{th} iteration and $\lambda_{\hat{k}_n}^{(l)}$ its parameters.

1. Initialize cluster memberships. Randomly assign time series to clusters to obtain data sets for each cluster $\{X\}_k^{(0)}$. Set $l = 0$.

2. Find initial total log likelihood of the assignment: $P^{(0)} = \sum_n \log P\big(\mathbf{x}_t^n | \lambda_{\hat{k}_n}^{(0)}\big)$.

3. For $k = 1, \cdots, K$, apply the Baum-Welch algorithm [9] to $\{X\}_k^{(l)}$ to obtain the estimates $\lambda_k^{(l+1)}$.

4. For $n = 1, \cdots, N$ find $\hat{k}_n^{(l+1)} = \text{argmax}_k P\big(\mathbf{x}_t^n | \lambda_k^{(l+1)}\big)$ (via the forward-backward or Viterbi algorithms) [9].

5. For $k = 1, \cdots, K$, let $\{X\}_k^{(l+1)} = \{\mathbf{x}_t^n\}$ for all $\mathbf{x}_t^n$ whose $\hat{k}_n^{(l+1)} = k$.

6. Find $P^{(l+1)} = \sum_n \log P\big(\mathbf{x}_t^n | \lambda_{\hat{k}_n}^{(l+1)}\big)$.

7. If $d\big(P^{(l+1)}, P^{(l)}\big) > \epsilon$ (where $d(\cdot, \cdot)$ and ϵ define some convergence criterion), let $l = l + 1$ and go to 3; otherwise, stop.

This algorithm was implemented using similar HMM structures for all clusters; the model consisted of a fully connected 5-state structure with single Gaussian full covariance output densities. After convergence of the algorithm, we need to establish whether there exists a dependence between the data labels and the cluster identities. Our approach to evaluate this is to consider the outcome of the clustering algorithm in terms of two multinomial variables: the class of the data sequence $\mathbf{x}_t^n$ (ω_n) and the cluster to which it is assigned (c_n). The clustering algorithm may then be viewed as yielding a data set $\{\omega_n, c_n\}_{n=1}^N$ to which we want to apply a hypothesis test to determine whether the set of labels and the set of clusters were generated by different multinomial distributions or by the same multinomial. More formally, we would like to know the probability that the sets $\Omega = \{\omega\}_n$ and $C = \{c\}_n$ were generated by the same distribution:

$$\begin{aligned} p(s|\Omega, C) &= \frac{p(\Omega, C|s)p(s)}{p(\Omega, C|s)p(s) + p(\Omega, C|d)p(d)} \\ &= \frac{1}{1 + \frac{p(\Omega, C|d)}{p(\Omega, C|s)} \frac{p(d)}{p(s)}} \qquad (6) \end{aligned}$$

where the labels s and d indicate same or different distributions. The main quantity involved in computing (6) is the ratio of evidence of the data sets under different and same distributions $\frac{p(\Omega, C|d)}{p(\Omega, C|s)}$, a quantity which may be written in terms of factorized and joint evidence $\frac{p(\Omega)p(C)}{p(\Omega, C)}$. Let the class ω take on one of J outcomes, and let the cluster c take on one of K outcomes. Define the following counts $N_{j,k} = \sum_{n=1}^N \delta_{w_n, j} \delta_{c_n, k}$, $N_j = \sum_{k=1}^K N_{j,k}$ and $N_k = \sum_{j=1}^J N_{j,k}$ for $j = 1, \cdots, J$, $k = 1, \cdots, K$. It can be shown [10] that, under the assumption of multinomial distributions, the evidence ratio in (6) is given by

$$\frac{p(\Omega)p(C)}{p(\Omega, C)} = \frac{\Gamma(JK)}{\Gamma(N+JK)} \prod_j \frac{\Gamma(N_j+K)}{\Gamma(K)} \prod_k \frac{\Gamma(N_k+J)}{\Gamma(J)} \prod_{j,k} \frac{\Gamma(1)}{\Gamma(1+N_{j,k})} \qquad (7)$$

The quantity in (7) also has an interpretation as the mutual information between the variables ω and c [10]. Equation (6) may be used to determine whether the clustering procedure has introduced some dependencies between labels and clusters. Furthermore, it may be used together with the clustering algorithm above to select the number of clusters which establishes the largest dependency between variables.

4.2 Classification

The unsupervised learning procedure described above may be used to identify time series which form clusters in feature space and may be used as a preamble for building cluster dependent supervised learners which exploit the "locality" of data sets in regions of the space. The learners at this classification stage are therefore trained with only a portion of the categorical data which corresponds to those time series assigned to a common cluster. HMMs have been used to implement the cluster-dependent class-conditional models at this stage. We applied the same HMM structure and output distribution forms from the unsupervised learning stage. Using the clustering with the number of clusters K which maximized the dependency between classes and clusters, the Baum-Welch algorithm [9] is applied to learn the HMM parameters. The posterior probability of the class given an observation is given by

$$p(\omega|\mathbf{x}_t) = \sum_c p(\omega, c|\mathbf{x}_t) = \sum_c p(\mathbf{x}_t|\omega, c)p(c|\omega)p(\omega) \quad (8)$$

where the quantity $p(c|\omega)$ can be estimated from the output of the clustering. Assuming equal priors on all classes, and a maximum a posteriori classification scheme, the following decision rule is then obtained:

$$\hat{\omega} = \text{argmax}_l\, p(\omega_l|\mathbf{x}_t) = \text{argmax}_l \sum_c p(\mathbf{x}_t|\omega_l, c)p(c|\omega_l) \quad (9)$$

The quantity $p(\mathbf{x}_t|\omega, c)$ can be efficiently evaluated using the Viterbi algorithm [9].

5 Results

The speech data of 4 subjects was first divided into a training and testing set comprising approximately 80% and 20% of the data set respectively. Unsupervised clustering was first applied to each subject's training data with the number of clusters ranging from 2 to 6. We shall use the following labels to denote the four categories of data: FF, SF, FS, SS. the first letter denotes whether the data came from a fast (F) or slow (S) speed condition; the second indicates the frequency of questions, every 4 seconds (fast) (F) or every 9 (slow) (S). Table 1 shows the results of the unsupervised clustering with the cluster that maximizes (6).

Subject	1	2	3	4
K	4	2	4	6
$p(s\|\Omega, C)$	0.436	0.014	0.971	0.001

Table 1: Results of Unsupervised Clustering

Table 1 shows that the results are greatly dependent on the subject. Subject 3 shows a very high correlation between the clusters and the categories; there is very little correlation, however, in the case of subject 4. (One should note that the results in Table 1 are not biased by the case $K = J$ since (7) models the interaction of these two variables.) The dependency between the four discrete categories of stress we have established and the output of the unsupervised clustering –as evidenced by the results of subjects 1 and 3, for instance– suggests that we may want to retain this distinction between labels to build a supervised system that can discriminate between them.

The results of such classifications are summarized in tables 2 and 3 for the training and testing phases for each subject.

Subject	Training Rec. Rates (%)				
	FF	SF	FS	SS	All
1	100	97.87	100	100	99.21
2	100	100	100	100	100
3	97.14	94.29	70.0	85.71	89.19
4	100	100	94.44	88.89	97.39

Table 2: Classification Results (Training Set)

Subject	Testing Rec. Rates (%)				
	FF	SF	FS	SS	All
1	83.33	50.00	25.00	0	42.86
2	100	69.23	0	40.00	69.70
3	100	100	83.33	100	96.15
4	16.67	91.67	0	0	36.11

Table 3: Classification Results (Testing Set)

Once again the dependency on the subject is evident from these results. Subject 4 shows poor generalization whereas subject 3 generalizes well. In all cases, the overall recognition rates for these labels exceed random classification (25%).

The approach of building a classifier that combines these local models has the added cost of an increase in parameter estimation and computational power. It is therefore desirable to compare its performance with that of a single classifier. Using the same HMM structure, we obtained a simple subject dependent classification scheme without pre-clustering the data. The overall results for this classifier are shown in Table 4 for the training and testing set.

Subject	1	2	3	4
Train. Rec. Rate (%)	98.63	94.31	86.49	99.13
Test. Rec. Rate (%)	50.00	51.52	42.31	55.56

Table 4: Recognition Rates (No Pre-clustering)

Although higher individual recognition rates were obtained in some cases with this classification scheme, it is interesting to note that the generalization results shown in Table 4 are more uniform than those summarized in Table 3. This illustrates how using local models can boost the performance considerably for some subjects.

6 Conclusions

In this paper we have investigated the use of features based on subband decompositions and the TEO for classification of stress categories in speech produced in the context of driving at variable speeds while engaged on mental tasks of variable cognitive load for a set of 4 subjects. To establish whether the resulting discrete categories constitute meaningful labels in feature space, we have first used an unsupervised approach to uncover underlying clusters and then correlated the cluster membership to the class labels. We have used these results to build cluster-dependent discriminants to exploit local subsets of the data and then combined the results to yield a decision rule. We report recognition rates that are greater than random for all subjects.

ACKNOWLEDGMENTS

The authors would like to thank Nissan's CBR Lab and Elias Vizas for their help with data collection, and Thomas Minka for valuable technical discussions and for suggesting the significance test.

References

[1] Herman J.M. Steeneken and John H.L. Hansen. Speech under stress conditions: Overview of the effect on speech production and of system performance. In *Proceedings ICASSP '99*, volume 4, pages 2079–2082, 1999.

[2] I.R. Murray, C. Baber, and A.J. South. Towards a definition and working model of stress and its effects on speech. *Speech Communication*, 20:1–12, November 1996.

[3] Guojun Zhou, John H.L. Hansen, and James Kaiser. Classification of speech under stress based on features derived from the nonlinear Teager energy operator. In *Proceedings ICASSP '98*, volume 1, pages 549–552, 1998.

[4] Guojun Zhou, John H.L. Hansen, and James F. Kaiser. Methods for stress classification: Nonlinear TEO and linear speech based features. In *Proceedings ICASSP '99*, volume 4, pages 2087–2090, 1999.

[5] Firas Jabloun and A. Enis Çetin. The Teager energy based feature parameters for robust speech recognition incar noise. In *Proceedings ICASSP '99*, volume 1, pages 273–276, 1999.

[6] Ruhi Sarikaya and John N. Gowdy. Wavelet based analysis of speech under stress. In *Southeastcon '97. Engineering new New Century. Proceedings IEEE*, pages 92–96, 1997.

[7] Ruhi Sarikaya and John N. Gowdy. Subband based classification of speech under stress. In *Proceedings ICASSP '98*, volume 1, 1998.

[8] Ingrid Daubechies. *Ten Lectures on Wavelets*. SIAM, 1992.

[9] L. Rabiner and B-H. Juang. *Fundamentals of Speech Recognition*, chapter Theory and Implementation of Hidden Markov Models. Prentice Hall, 1993.

[10] Thomas P. Minka. Bayesian inference of a multinomial distribution. http://www.media.mit.edu/~tpminka/papers/tutorial.html.