Detection of time-pressure induced stress in speech via acoustic indicators *

Matthew Frampton, Sandeep Sripada, Ricardo Augusto Hoffmann Bion and Stanley Peters
Center for the Study of Language and Information
Stanford University, Stanford, CA, 94305 USA
{frampton@,ssandeep@,ricardoh@,peters@csli.}stanford.edu

Abstract

We use automatically extracted acoustic features to detect speech which is generated under stress, achieving 76.24% accuracy with a binary logistic regression. Our data are task-oriented human-human dialogues in which a time-limit is unexpectedly introduced partway through. Analysis suggests that we can detect approximately when this event occurs. We also consider the importance of normalizing the acoustic features by speaker, and detecting stress in new speakers.

1 Introduction

The term *stressed speech* can refer to speech generated under psychological stress (Sigmund et al., 2007). Stress alters an individual's mental and physiological state, which then affects their speech. The ability to identify stressed speech would be very valuable to Spoken Dialogue Systems (SDSs), especially in "stressful" applications such as search-and-rescue robots. Speech recognizers are usually trained on normal speech, and so can struggle badly on other speech. Techniques exist for making ASR robust to noise/stress (Hansen and Patil, 2007), but knowing when to apply them will in general require the ability to detect stressed speech. This ability is clearly also needed when the user's stress level should affect how the SDS responds. An SDS should sometimes generate stressed speech itself—for example, to impart a sense of urgency on the user.

This paper investigates spectral-based acoustic indicators of stress in human-human, task-oriented dialogues in which stress is induced in the latter stages by the unexpected introduction of time-pressure. Unlike previous studies, we detect stress in whole utterances in the raw audio, which is more realistic for applications. We also consider the importance of normalizing the features, and detection of both the introduction of the stressor, and stress in new speakers.

2 Related work

Stressors and clip sizes: The stressors in previous studies include logical problems, images of human bodies with skin diseases/severe accident injuries (Tolkmitt and Scherer, 1986), loss of control of a helicopter (Protopapas and Liberman, 2001), university examinations (Sigmund et al., 2007), and an increasingly difficult air controller simulation and verbal quiz (Scherer et al., 2008). Sigmund et al. (2007) detect stress in approximately 2000 voiced segments of 5 vowels. Tolkmitt and Scherer (1986), Protopapas and Liberman (2001) and Scherer et al. (2008) detect stress in whole utterances, but these are respectively, read from a card, quiz answers, and with verbal content removed. Studies on the Speech under Simulated and Actual Stress (SUSAS) corpus (Hansen and Bou-Ghazale, 1997) detect stress in words. These include (Hansen, 1996; Zhou, 1999; Hansen and Womack, 1996; Zhou, 2001; Casale et al., 2007). The SUSAS corpus contains aircraft communication words from a common highly confusable vocabulary set of 35, and they are divided into different speaking styles.

Acoustic cues: The most widely investigated acoustic cues relate to *fundamental frequency* (F0, also called pitch), formant frequencies and spectral composition e.g. (Tolkmitt and Scherer, 1986; Hansen, 1996; Zhou, 1999; Protopapas and Liberman, 2001; Sigmund et al., 2007; Scherer et al., 2008). Mel-Frequency Cepstral Coefficients

* The research reported in this paper was sponsored by the Department of the Navy, Office of Naval Research, under grant number N00014-017-1-1049. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Office of Naval Research.

Proceedings of SIGDIAL 2010: the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 253–256,
The University of Tokyo, September 24-25, 2010. ©2010 Association for Computational Linguistics

Category	Examples
F0-related	Median, mean, minimum, time of minimum as % thr' clip, max, time of max as % thr' clip, range (max-min), standard deviation, mean absolute slope, mean slope without octave jumps, number of voiced frames.
Intensity-related	Median, mean, minimum, time of minimum as % thr' clip, max, time of max as % thr' clip, range (max-min), standard deviation.
Formant-related (for F1-F3)	Mean, minimum, time of minimum as % through clip, max, time of max as % through clip, range (max-min).
Spectral tilt-related	Mean, minimum, maximum, range (max-min).

Table 1: The acoustic features which are extracted from the audio clips using Praat (Boersma and Weenink, 2010).

(MFCCs)[1] and Teager Energy Operator (TEO)[2] (Kaiser, 1990) based features have also been considered e.g. (Hansen and Womack, 1996; Zhou, 2001; Casale et al., 2007).

Features of all these types have proved useful in detecting stressed speech. The classification methods employed are various, including a traditional binary hypothesis detection-theory method (Zhou, 1999) and neural networks (Hansen and Womack, 1996; Scherer et al., 2008), while Casale et al. (2007) used genetic algorithms for feature selection. Of the two more recent studies which detected stress in whole utterances, Protopapas and Lieberman found that mean and maximum F0 within an utterance correlate highly with subject stress ratings, and Scherer et al.'s neural network outperformed a human baseline. Note that findings/results in these and other previous studies are not directly comparable with our own, because we detect stress in whole utterances in raw audio.

3 Data

The original data (Eberhard et al., 2010) are 4 task-oriented dialogues between 2 native English-speaking participants. Hence there are 8 speakers in total (7 male, 1 female), and the dialogues contain 263, 172, 228 and 210 utterances respectively.

During a dialogue, the participants (the *director* and *member*) are on a floor with corridors and rooms that contain various colored boxes. The director stays in one room, and gives task instructions via walkie-talkie to the member, providing directions with a map which is partially complete and accurate for box locations. The tasks are locating boxes which are unmarked on the map, and transferring blocks between and retrieving specified boxes. Initial instructions do not mention a time limit, but at the end of the 7^{th} minute, the director is given a timer and told there are 3 minutes to complete the current tasks, plus one new task.

We use the Nuance speech recognizer (V. 9.0) to end-point each dialogue's audio signal, and the resulting clips are mostly 1 to 3 seconds. In preliminary experiments (not reported), denoising seemed to remove acoustic information which is indicative of stress. Hence we use raw audio.

Stressed speech: For present purposes, we assume that all speech after the introduction of the time limit is stressed. Hence 448 of the 663 audio clips in our experimental data are unstressed, and 215 are stressed. In future we plan to use the *Amazon Mechanical Turk* to obtain perceived stress ratings on a scale with more gradations.

4 Experiments

Acoustic features: We use Praat (Boersma and Weenink, 2010) to compute *F0*, *intensity*, *formant* and *spectral tilt-related* features for each clip (Table 1). F0 (pitch) corresponds to the rate of vocal cord vibration in Hertz (*Hz*), and Intensity, to the sound's loudness in decibels (*dB*), (derived from the amplitude or increase in air pressure). A formant is a concentration of acoustic energy around a particular frequency in the speech wave. There are several, each corresponding to a resonance in the vocal tract, and we consider the lowest three (*F1-F3*). Spectral tilt measures the difference in energy between the 1^{st} and 2^{nd} formants, and so estimates the degree to which energy at the fundamental dominates in the glottal source waveform.

Comparing different normalization methods: We evaluate binary logistic regression models with 10-fold cross-validation, and try the following 4 methods for normalizing each clip's acoustic features according to its speaker.

- *Maximum normalization*: Due to the possibility of outliers, we divide each feature value

[1] MFCCs model the human auditory system's nonlinear filtering in measuring spectral band energies.

[2] The TEO is a nonlinear operator which uses mechanical and physical considerations to extract the signal energy.

Normalization	% Accuracy		US %correct		S %correct		MCB	
Maximum normalization	74.4	(74.25)	86.67	(85.05)	48.5	(54.3)	67.8	(67.1)
Z-score	73.5	(73.78)	84.89	(84.12)	49.53	(52.7)	67.8	(67.1)
US Average	75.61	(76.24)	86.63	(86.2)	53.5	(55.9)	67.8	(67.1)
S Average	75.31	(75)	84.67	(84.375)	55.6	(55.9)	67.8	(67.1)
No normalization	68.52	(70.45)	84.34	(82.8)	37.4	(45.2)	67.8	(67.1)

Table 2: Binary logistic regression 10-fold cross validation with different feature normalization approaches: Scores within brackets are when the female speaker data is removed; S = Stressed, US = Unstressed, MCB = Majority Class Baseline.

by the 95^{th} percentile value for that feature, rather than the maximum.

- *Z-score*: Using the mean and standard deviation for each feature, the feature vector is converted to Z-scores[3].
- *Unstressed (US) average*: Each feature is normalized by its mean value in the unstressed region.
- *Stressed (S) average*: Each feature is normalized by its mean value in the stressed region.

Table 2 shows the results. All those generated with feature normalization are significantly better ($p < 0.005$) than the majority class baseline (MCB), (i.e. classifying all utterances as unstressed). Without normalization, the overall accuracy drops about 5—6%, and the stressed speech class about 11—18%. Different normalization methods do not produce very different results, but US average gives the best overall accuracy (75.61%). When we remove the female speaker, this increases to 76.24%, and feature normalization remains important.

We also tested our assumption that the speech before and after the introduction of time-pressure is unstressed and stressed respectively, by checking that they really are different. As before, we considered 7 minutes unstressed, and 3 stressed, and used US average normalization. However we now assigned different minutes to the unstressed and stressed categories: first we swapped the 6^{th} and 8^{th}, then also the 5^{th} and 9^{th}, and then also the 7^{th} and 10^{th}. As a result, classification accuracy dropped, (to 75%, then 68.71%, then 67.66%), which supports our assumption.

Feature contribution analysis: Table 3 shows the US average normalized features with *information gain* greater than zero. Intensity and pitch features are ranked most predictive (i.e. maximum and mean intensity, and mean and median pitch), but *Spectral tilt mean* and a couple of formant features are also predictive. In general, higher values for the most predictive pitch and intensity features (e.g. *Intensity max* and *Pitch mean*) seem to indicate stress. An interaction term for *Intensity max* and *Pitch mean* caused a significant improvement in the fit of the model—the χ^2 value (or change in the -2 Log likelihood) was 4.952 ($p < 0.05$).

Feature	Info. Gain
Intensity max	.101
Pitch mean	.099
Intensity mean	.099
Pitch median	.088
Pitch max	.059
Intensity min	.046
Spectral tilt mean	.042
Pitch min	.041
F1 min	.038
Intensity range	.034
Intensity std. dev.	.033
F3 range	.033
Intensity median	.031

Table 3: Unstressed average normalized features ranked by information gain.

Detecting the introduction of the stressor: Figure 1 shows the percentage of audio clips in each minute that were classified as stressed. As we would hope, there is a dramatic increase from the 7^{th} to the 8^{th} minute (around 20% to over 50%). Such an increase could be used to detect the introduction of the stressor, time-pressure.

Detecting stress in new speakers: To detect stressed speech in new speakers, we evaluate the logistic regression with an 8-fold cross-validation, in each fold training on 7 speakers, and testing on the other. We apply US average normalization, initially with the average values for the new speaker's unstressed speech, and then with the average values in unstressed speech across all "seen" speakers (speakers in the training set). Evaluation scores (Table 4) are now lower, especially for the latter approach, but the former remains significantly

[3] A Z-score indicates the number of standard deviations between an observation and the mean.

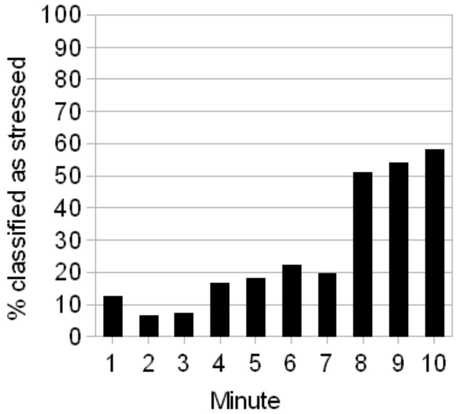

Figure 1: The percentage of clips in each minute of the dialogues which our classifier marks as stressed, (note that time-pressure is introduced at the end of minute 7).

better than the MCB. Since the female speaker's stress class F-score is 0, we tried normalizing the 7 male speakers based on only seen male data, and then average accuracy for a male rose from 67.09% to 68.02% (not statistically significant).

Spkr	% Accuracy		F-unstress		F-stress	
1	62.5	(62.2)	.77	(.74)	.07	(0.34)
2	75	(75)	.84	(.84)	.09	(0.44)
3	57.6	(72.9)	.62	(.81)	.49	(0.5)
4	71.73	(74)	.84	(.83)	0	(0.43)
5	71.62	(73.0)	.77	(.77)	.62	(0.68)
6	77.6	(80.4)	.86	(.88)	.51	(0.52)
7	64.36	(65.6)	.74	(.77)	.4	(0.35)
8	60.97	(71.7)	.67	(.8)	.49	(0.46)
Av.	67.67	(71.9)	.76	(.80)	.33	(0.46)

Table 4: Predicting stress in new speakers: New speaker features are normalized based on unstressed speech for all speakers in training set (unbracketed) and on their own unstressed speech (bracketed). Speaker 4 is the female.

5 Conclusion

For detecting stressed speech, we demonstrated the importance of normalizing acoustic features by speaker, and achieved 76.24% classification accuracy with a binary logistic regression model. The most indicative features were maximum and mean intensity within an utterance, and mean and median pitch. After the introduction of time-pressure, the percentage of clips classified as stressed increased dramatically, showing that it is possible to detect approximately when this event occurs. We also attempted to detect stressed speech in new speakers, and as expected, results were poorer.

In future work we plan to expand our data-set with more dialogues, and test accuracy for detecting the introduction of the stressor. We want to use MFCCs and TEO features, and also non-acoustic features such as disfluency features. As mentioned previously, we also hope to move beyond binary classification, by acquiring perceived stress ratings on a scale with more gradations.

References

P. Boersma and D. Weenink. 2010. Praat: doing phonetics by computer (version 5.1.29). Available from http://www.praat.org/. [Computer program].

S. Casale, A. Russo, and S. Serrano. 2007. Multistyle classification of speech under stress using feature subset selection based on genetic algorithms. *Speech Communication*, 49:801–810.

K. Eberhard, H. Nicholson, S. Kubler, S. Gundersen, and M. Scheutz. 2010. The Indiana "Cooperative Remote Search Task" (CReST) Corpus. In *Proc. of LREC*.

J.H.L. Hansen and S. Bou-Ghazale. 1997. Getting started with SUSAS: a Speech Under Simulated and Actual Stress database. In *Eurospeech-97: International Conference on Speech Communication and Technology*.

J. Hansen and S. Patil, 2007. *Speaker Classification I: Fundamentals, Features, and Methods*, chapter Speech Under Stress: Analysis, Modeling and Recognition, pages 108–137. Springer-Verlag, Berlin, Heidelberg.

J. Hansen and B. Womack. 1996. Feature analysis and neural network based classification of speech under stress. *IEEE Transactions on Speech & Audio Processing*, 4(4):307–313.

J. Hansen. 1996. Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition. *Speech Communications, Special Issue on Speech Under Stress*, 20(2):151–170.

J.F. Kaiser. 1990. On a simple algorithm to calculate the energy of a signal. In *Proc. of ICASSP*.

A. Protopapas and P. Liberman. 2001. Fundamental frequency of phonation and perceived emotional stress. *Journal of the Acoustical Society of America*, 101(4):2267–2277.

S. Scherer, H. Hofmann, M. Lampmann, M. Pfeil, S. Rhinow, F. Schwenker, and G. Palm. 2008. Emotion recognition from speech: Stress experiment. In *Proc. of LREC*.

M. Sigmund, A. Prokes, and Z. Brabec. 2007. Statistical analysis of glottal pulses in speech under psychological stress. In *Proc. of the 16th European Signal Processing Conference*.

F. J. Tolkmitt and K. R. Scherer. 1986. Effect of experimentally induced stress on vocal parameters. *Journal of Experimental Psychology*, 12(3):302–313.

G. Zhou. 1999. *Nonlinear speech analysis and acoustic model adaptation with applications to stress classification and speech recognition*. Ph.D. thesis, Duke University.

G. Zhou. 2001. Nonlinear feature based classification of speech under stress. *IEEE Transactions on Speech & Audio Processing*, 9:201–216.