ANALYSIS AND PREDICTION OF HEART RATE USING SPEECH FEATURES FROM NATURAL SPEECH

Jennifer Smith, Andreas Tsiartas, Elizabeth Shriberg,
Andreas Kathol, Adrian Willoughby, Massimiliano de Zambotti

SRI International, Menlo Park, CA, USA

ABSTRACT

Interactive voice technologies can leverage biosignals, such as heart rate (HR), to infer the psychophysiological state of the user. Voice-based detection of HR is attractive because it does not require additional sensors. We predict HR from speech using the SRI BioFrustration Corpus. In contrast to previous studies we use continuous spontaneous speech as input. Results using random forests show modest but significant effects on HR prediction. We further explore the effects on HR of speaking itself, and contrast the effects when interactions induce neutral versus frustrated responses from users. Results reveal that regardless of the user's emotional state, HR tends to increase while the user is engaged in speaking to a dialog system relative to a silent region right before speech, and that this effect is greater when the subject is expressing frustration. We also find that the user's HR does not recover to pre-speaking levels as quickly after frustrated speech as it does after neutral speech. Implications and future directions are discussed.

Index Terms— autonomic nervous system, heart rate, speech features, frustration, dialog system

1. INTRODUCTION

Speech to a personal assistant or dialog system carries more information than just the words that were spoken. The same signal may offer new opportunities for sensing changes in a speaker's psychophysiological state. Such information could be used for health-based applications when additional sensors (such as fitness trackers [5]) are either not available, or add too much complexity to process, align, or store. To this end, we investigate the feasibility of predicting heart rate (HR) from only the speech signal. In contrast to prior work (for example, work using sustained vowels [21, 15, 16, 13, 17]), we focus on natural continuous speech, since the end goal is not to change what speakers are saying but rather to use their naturally produced signal. In particular, we use continuous speech data from the SRI Biofrustration Corpus [9], which includes samples collected while subjects were interacting with a misbehaving automated customer service system. We investigate the feasibility of HR detection during frustrating interactions with an automatic dialog system.

Relatively little research has focused on the relationship between speaking and HR in the context of continuous spontaneous speech. A large number of studies have described effects of emotion or stress on HR and other physiological measures; for a summary see for example [11]. A smaller number of studies examined HR (or blood pressure or skin conductance) with respect to other types of speech or voice input, including sustained vowels [21, 15, 16, 13, 17], read speech [10], acted speech [14], an arithmetic task [19, 22, 20, 8, 24], exercise [18] and breathing [1].

In a previous study [25] using the SRI BioFrustration corpus [9], we conducted a first analysis of HR change prediction from speech, using a crude definition of HR change (positive or negative) and temporal regions that comprised multiple utterances to a dialog system. Results showed that HR was higher during interactions with the dialog system that were designed to frustrate the user as compared with interactions that were designed to be neutral. We also found that the direction of change in HR could be classified using speech features to train random forest classifiers with accuracy greater than chance.

In the present work, we extend the investigation in several ways. First, we examine the time course of HR changes before, during and after the speaker engages in speaking to the dialog system. This is important for the interpretation of HR changes in real time, over regions that comprise both speech and nonspeech intervals. Specifically, we compare pre-speech to speech regions and pre-speech to post-speech regions for both neutral and frustrated contexts. We also predict normalized HR, rather than predicting only the direction of change, and we consider the effect of user frustration on HR prediction.

2. DATA

2.1. Corpus & speech data

This study uses the SRI BioFrustration Corpus, which contains data from 53 native English speakers (20 males, 33 females) aged 18 to 75, with an average age of 36.4 years. Participants were SRI employees, their relatives, and Stanford University undergraduates. The corpus combines state-of-the-art audio, video, and physiological signals recorded while subjects interacted with a misbehaving computer dialog system. In this data set, HR was derived from electrocardiogram (ECG) signals.

The *Returns* dialog system used to collect this data was built with SRI's Virtual Personal Assistant (VPA) technology and designed as an automated customer service system for returning purchases. Subjects were told that the *Returns* system was capable of understanding their speech and responding to their emotional tone. However, the system was designed to mostly follow a pre-defined work-flow that ensured the system would reject their attempts to return items. The set-up was designed to frustrate subjects by offering them a $100 reward at the beginning of the session and telling them that failure to satisfactorily complete a minimum number of tasks (i.e., returns) would lead to confiscation of the reward.

The tasks and the predefined system response were designed to elicit frustration responses from participants at particular points in the exchange. For instance, while the first return task presented the participants with a successful problem-free return experience, in "Task 2", the system rejected the refund request ("Let me check... Unfortunately, according to company policy these are insufficient

978-1-5090-4117-6/17/$31.00 ©2017 IEEE

reasons to reconsider our decision."), even after participants had been prompted three times to provide reasons why they should be eligible to return a defective watch. Thus, participants were not only unsuccessful in their refund request, but the system behaved in a particularly uncooperative way in the rejection.

Furthermore, to remind subjects to convey their emotional state to the system, they were given unsolicited "feedback" on their "inadequate" performance at the end of "Task 2". For example:

- System: For your information, compared to other people in the study you are currently performing below average. To keep your reward, you can only afford one more unsuccessful return in the remaining six tasks.

Participants were again faced with an uncooperative system in "Task 3", which (at one point) blatantly disregarded their input:

- System: Did you use the item in any way that would void the warranty?
- User: No
- System: Could you elaborate on what you did that voids the warranty?

Additionally, it claimed to not understand the user:

- System: I'm sorry, <PAUSE> I did not fully understand what you just said. Could you repeat please?

Finally, it again rejected the appeal for a refund, after having repeatedly asked the the participant for justification.

The system behavior was designed to be more cooperative in subsequent tasks. Participants were given feedback that suggested that this increased cooperation was in part due to their improved ability to express their emotional state. At the same time, since there was a certain amount of repetitiveness to the individual return tasks, it became apparent that user engagement started to decrease in subsequent tasks, which in turn affected the expression of emotional state. For that reason, we concentrate our analysis in the present study on user behavior in "Task 2" and "Task 3," rather than considering all 7 tasks.

2.2. Annotation

In contrast to previous modeling of this corpus which averaged values for utterances across pre-trigger (cooperative system behavior) and post-trigger (uncooperative system behavior) regions for each of 7 tasks [25], this study considers each utterance separately across only "Task 2" and "Task 3". Each utterance was hand coded by an experienced human annotator. The utterances were coded as either "neutral," "slightly frustrated," or "frustrated." From Table 1, we see that there is considerable variation in the number of utterances recorded per subject and also in the number of utterances for which the subject displayed each perceived affect.

Table 1: *Distribution of Annotations by Subject*

	No. of Utterances	Neut.	Slightly Frust.	Frust.
Mean	37.2	20.9	9.7	6.7
Std. Dev.	9.5	5.0	4.1	5.4

3. SPEECH FEATURES

We analyzed a variety of speech features that captured the spectral, temporal, and prosodic differences over background environment. The features are measured at the 20-40ms windows with a frame shift of 10ms. Frame level features include Mel frequency cepstral coefficients (MFCC) [4] which capture the energy of cepstral domain frequencies. In addition, we use Mel Frequency Bands (MFB) [3] which capture the spectral energy over longer windows of 40ms. Our set also includes features that measure the prosodic content of the speech signal. In particular, we use the pitch and intensity information for this purpose. To measure the differential energy of voiced and unvoiced regions, we use the harmonic to noise energy ratio (HNR)[26] feature. Finally, we include in our analysis the diverse set of spectral and prosodic features included in OpenSMILE [6] feature set extracted at the frame level. Based on the segmentation output, we post-process the features for each utterance by computing statistics at the utterance level. In particular, for each above-mentioned frame level feature we computed the mean, median, variance, median, interquartile range, kurtosis, and skewness; and the 5th, 25th, 75th, and 95th percentiles.

4. EXPERIMENTAL SETUP

4.1. Analysis of the effects of speech on heart rate

To build a system that can interpret HR changes in real time, we need to understand how HR changes over regions that comprise both speech and nonspeech intervals. To this end, we examine the time course of each subject's ECG-derived-HR changes, before, during and after the speaker engages in speaking to the dialog system. We also examine the ways that these changes differ when the user is frustrated.

First, we compared HR measurments taken during each utterance to that of the 5-second nonspeech region preceding the utterance. If speaking to the dialog system had no effect on HR, we would expect HR to increase for about half of the subjects and decrease for about half. We used a proportion test to see if the proportion of subjects for whom HR increased was significantly greater than half. The test compares the following z-score test statistic to a normal distribution:

$$z = \left(\frac{p - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \right) \tag{1}$$

The observed value, p, is calculated by first averaging the change in HR across all utterances for each subject and then calculating the proportion of subjects for whom the average change is positive. If the change in HR is random, then p_0 should be 0.5.

Using the same method, we examined the effect of frustration on the change in HR from the nonspeech region before engaging in speaking to a dialog system to the speech region. In particular, we tested the proportion of subjects for whom the average change in HR while speaking, relative to before, was greater for utterances coded as *frustrated* than for those coded as *neutral*. For this test, the observed value, p, is calculated by first calculating the average change in HR by subject on two subsets of the data: those coded *neutral* and those coded *frustrated*. Then, we found the proportion of subjects for whom the average change in HR across *frustrated* utterances was higher than the average change in HR across *neutral* utterances.

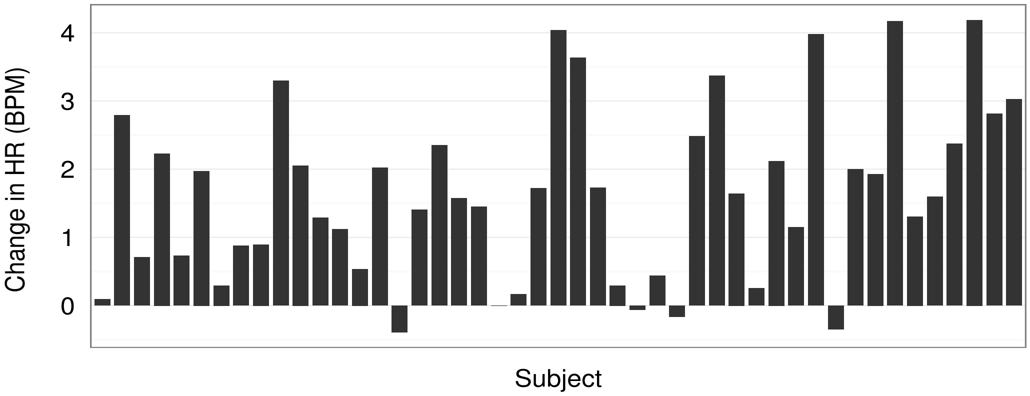

Fig. 1: Absolute change in HR (BPM) from before to during speech, averaged over all utterances, regardless of perceived affect

Lastly, we investigated the effect of frustration on the return of HR to pre-speaking levels after speaking to a dialog system. For this test, the observed value, p, is calculated by first subtracting HR in the 5 seconds before speech from HR in the 5 seconds after speech, and then finding the average difference by subject across all *neutral* utterances and all *frustrated* utterances. We then calculated the proportion of subjects for whom the average difference between HR after and before speech was greater for *frustrated* utterances than for *neutral* utterances.

4.2. Regression Analysis

For prediction, HR and speech features were normalized by subject. For each variable, we normalized by subtracting the subject's mean and dividing by the standard deviation for that variable. Normalized HR values were predicted from normalized speech features using 2 regression methods: random forest modeling [2] and LASSO regression [23].

Random forest models (RF) draw multiple bootstrap samples from the data and, for each bootstrap sample, grow an un-pruned regression tree. At each node of each tree, the algorithm chooses the best split from a random sample of the variables.

LASSO regression creates regression models using a sequence of 100 lambda values over 10 random partitions of the training data and returns the average error for each value of the penalty term, lambda. We compared values of lambda using 10-fold cross-validation and selected the optimal lambda as the largest value of lambda such that model error was within 1 standard error of the minimum model error across folds.

For prediction, the data was partitioned into 5 sets with no speaker overlap. RF and LASSO models were created for each partition and tested on the held-out speakers. For both models, the root mean squared error (RMSE) was computed for each partition, and then the average of the 5 RMSE's was computed as an overall error estimate:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2}$$

5. RESULTS & DISCUSSION

5.1. Effects of speech on heart rate

Using the previously described proportion tests, we observe that regardless of the user's emotional state, average HR is higher while the user is engaged in speaking to a dialog system, relative to a silent region right before speech ($z = 5.40$, $p < 0.001$). From Figure 1, we can see that for the majority of subjects, average HR is higher during engagement in speaking to a dialog system.

To test the difference between the effect of speaking on HR when *frustrated* versus when *neutral*, we considered only the subset of subjects who produced at least 5 utterances that were perceived as *frustrated* by the human annotator so that we could have a more robust average measurement per subject. Using these 27 subjects, we conclude that when a subject is expressing frustration while speaking to a dialog system, their change in HR is, on average, greater than during neutral speech ($z = 3.66$, $p < 0.001$). This result is visualized in Figure 2.

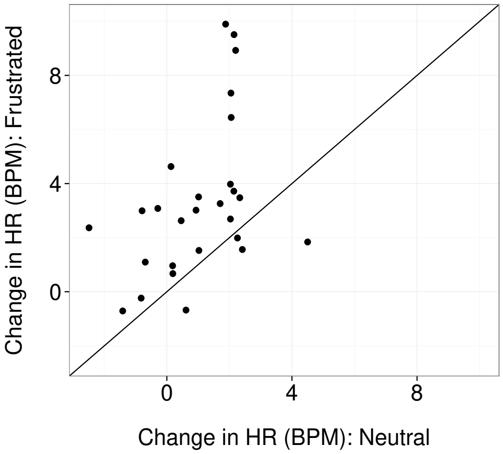

Fig. 2: Comparison of the absolute change in HR (BPM) from before to during speech for utterances coded *frustrated* versus *neutral* (each data point represents one subject).

Each point in Figure 2 represents one subject and any point that falls on the line through the origin is a subject for whom the average change in heart rate from before to during speech is the same for *frustrated* utterances and *neutral* utterances. As seen, most of the subjects fall above the line, supporting our conclusion that the average change in heart rate from before to during speech is higher for *frustrated* utterances than for *neutral* utterances.

We also find that the user's HR does not recover to pre-speaking levels as quickly after *frustrated* speech as it does after *neutral* speech ($z = 2.12$, $p = 0.017$). This result is visualized in Figure 3.

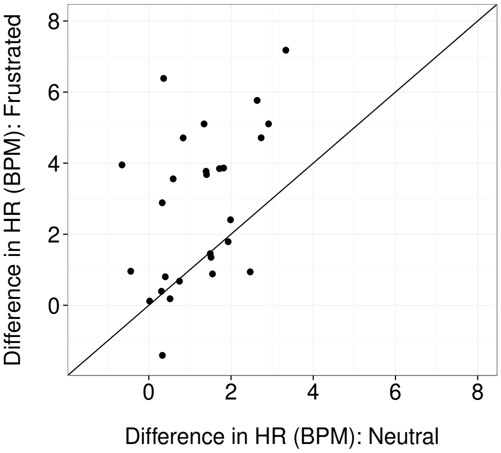

Fig. 3: Comparison of absolute difference in HR (BPM) before and after speech for utterances labeled *frustrated* versus *neutral* (each data point represents one subject).

Each point in the plot represents one subject and any point that falls on the line through the origin is a subject for whom the difference in heart rate before and after speech is equivalent for *frustrated* utterances and for *neutral* utterances. As seen, most of the subjects fall above the line, supporting our conclusion that the difference in heart rate before and after speech is higher for *frustrated* utterances than for *neutral* utterances.

5.2. Prediction results

Figure 4 shows the percentage change in prediction error (RMSE) between the models and baseline. In this case, we set baseline error as the standard deviation of speaker-normalized HR. This is equivalent to knowing the mean HR of each speaker and assigning the mean as the prediction value. When no speech information is available, the mean is the predicted value that minimizes RMSE.

The results in Figure 4 show the results using speech information to predict HR from Lasso and RF models as improvement over baseline. Models were created using the same set of speech features on three different data sets: (1) "All Data" contains all the utterances for 47 subjects (a total of 1746 data points); (2) "Neutral" contains only the utterances perceived as *neutral* in labeling (987 data points); and (3) "Frustrated" contains only the utterances perceived as *frustrated* in labeling (306 data points).

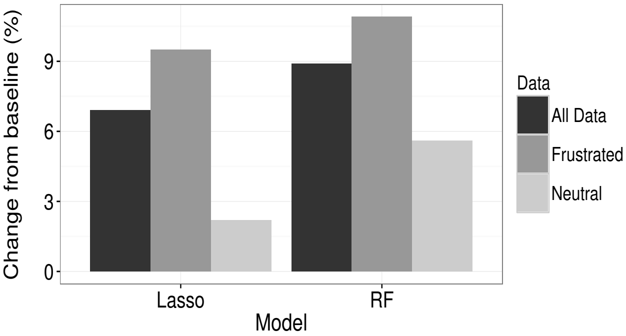

Fig. 4: HR prediction results as improvement over baseline

Although the subset of data perceived as *frustrated* is about 18% the size of the total data set, both RF and LASSO models perform better on the *frustrated* subset. Also, the models trained and tested using the subset of utterances perceived as *neutral* perform much worse. This may be explained, in part, by the fact that some subjects produced more *frustrated* utterances than others, so selecting *frustrated* utterances gave more weight to subjects who were more responsive to the experiment.

6. CONCLUSIONS

Our findings demonstrate the feasibility of predicting changes in user's physiological state when the user is engaged in speaking to a dialog system designed to induce different levels of frustration. We were able to observe statistically significant effects on both HR and speech, despite the fact that the induced frustration was intentionally mild. These results suggest that effects in real-world contexts for more intense levels of frustration are likely to be even more pronounced.

We also found that the act of speaking itself affects physiology and that these effects are greater when a user is frustrated. Using mainly frame-based speech features, we obtained modest but statistically significant patterns in HR prediction both within and across speakers.

These results provide important guidance for future work in the area of detecting user state from natural and spontaneous speech. Future work should analyze the user's comprehensive physiological pattern by including several peripheral indices in addition to HR. For example, information should include skin conductance, beat-to-beat blood pressure and breathing, as well as derived indices of heart rate variability that relate to cardiac autonomic nervous system control. The additional measures will allow us to map a more refined and specific physiological state of activation in response to frustration or other states, with the ultimate goal of predicting a pattern of physiological activation starting from the user's speech features. If successful, such technology could augment or even substitute for specialized wearable sensors in situations in which the sensors are impractical, inconvenient, or not available.

7. REFERENCES

[1] L. Bernardi, J. Wdowczyk-Szulc, C. Valenti, S. Castoldi, C. Passino, G. Spadacini, and P. Sleight, “Effects of controlled breathing, mental activity and mental stress with or without verbalization on heart rate variability,” *Journal of the American College of Cardiology*, vol. 35, no. 6, pp. 1462–1469, 2000.

[2] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 5-32, 2001.

[3] C. Busso, S. Lee, and S. S. Narayanan, “Using neutral speech models for emotional speech analysis.” in *Interspeech*, 2007, pp. 2225–2228.

[4] S. B. Davis and P. Mermelstein, “Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[5] K. R. Evenson, M. M. Goto, and R. D. Furberg, “Systematic review of the validity and reliability of consumer-wearable activity trackers,” *International Journal of Behavioral Nutrition and Physical Activity*, vol. 12, no. 1, pp. 1–22, 2015.

[6] F. Eyben, M. Wöllmer, and B. Schuller, “Opensmile: the munich versatile and fast open-source audio feature extractor,” in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[7] J. Friedman, T. Hastie, and R. Tibshirani, “Regularization paths for generalized linear models via coordinate descent,” *Journal of statistical software*, vol. 33, no. 1, p. 1, 2010.

[8] H. J. Johnson and J. J. Campos, “The effect of cognitive tasks and verbalization instructions on heart rate and skin conductance,” *Psychophysiology*, vol. 4, no. 2, pp. 143–150, 1967.

[9] A. Kathol and E. Shriberg, “The SRI BioFrustration Corpus: Audio, video, and physiological signals for continuous user modeling,” *AAAI Spring Symposium*, 2015.

[10] J. Kaur and R. Kaur, “Extraction of heart rate parameters using speech analysis,” *International Journal of Science and Research*, vol. 3, no. 10, p. 13741376, 2014.

[11] S. D. Kreibig, “Autonomic nervous system activity in emotion: A review,” *Biological psychology*, vol. 84, no. 3, pp. 394–421, 2010.

[12] A. Liaw and M. Wiener, “Classification and regression by randomforest,” *R News*, vol. 2, no. 3, pp. 18–22, 2002. [Online]. Available: http://CRAN.R-project.org/doc/Rnews/

[13] A. Mesleh, D. Skopin, S. Baglikov, and A. Quteishat, “Heart rate extraction from vowel speech signals,” *Journal of computer science and technology*, vol. 27, no. 6, pp. 1243–1251, 2012.

[14] A. J. Pappachen, “Heart rate monitoring using human speech spectral features,” *Human-centric Computing and Information Sciences*, vol. 5, no. 1, p. 1, 2015.

[15] M. Sakai, “Modeling the relationship between heart rate and features of vocal frequency,” *International Journal of Computer Applications*, vol. 120, no. 6, 2015.

[16] ——, “Estimation of heart rate from vocal frequency based on support vector machine,” *International Journal of Advances in Scientific Research*, vol. 2, no. 1, pp. 16–22, 2016.

[17] B. Schuller, F. Friedmann, and F. Eyben, “Automatic recognition of physiological parameters in the human voice: Heart rate and skin conductance,” in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7219–7223.

[18] ——, “The Munich Biovoice Corpus: Effects of physical exercising, heart rate, and skin conductance on human speech production.” in *LREC*, 2014, pp. 1506–1510.

[19] U. Schuri and D. Cramon, “Heart rate and blink rate responses during mental arithmetic with and without continuous verbalization of results,” *Psychophysiology*, vol. 18, no. 6, pp. 650–653, 1981.

[20] P. Seraganian, A. Szabo, and T. G. Brown, “The effect of vocalization on the heart rate response to mental arithmetic,” *Physiology & behavior*, vol. 62, no. 2, pp. 221–224, 1997.

[21] D. Skopin and S. Baglikov, “Heartbeat feature extraction from vowel speech signal using 2D spectrum representation,” in *Proc. the 4th Int. Conf. Information Technology*, 2009.

[22] R. Sloan, J. Korten, and M. M. Myers, “Components of heart rate reactivity during mental arithmetic with and without speaking,” *Physiology & Behavior*, vol. 50, no. 5, pp. 1039–1045, 1991.

[23] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[24] J. Tomaka, J. Blascovich, and L. Swart, “Effects of vocalization on cardiovascular and electrodermal responses during mental arithmetic,” *International Journal of Psychophysiology*, vol. 18, no. 1, pp. 23–33, 1994.

[25] A. Tsiartas, A. Kathol, E. Shriberg, M. de Zambotti, and A. Willoughby, “Prediction of heart rate changes from speech features during interaction with a misbehaving dialog system,” *INTERSPEECH*, 2015.

[26] E. Yumoto, W. J. Gould, and T. Baer, “Harmonics-to-noise ratio as an index of the degree of hoarseness,” *The journal of the Acoustical Society of America*, vol. 71, no. 6, pp. 1544–1550, 1982.