Prediction of Heart Rate Changes from Speech Features During Interaction with a Misbehaving Dialog System

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

SRI International, Menlo Park, CA, USA

{andreas.tsiartas,andreas.kathol,elizabeth.shriberg, massimiliano.dezambotti,adrian.willoughby}@sri.com

Abstract

Most research on detecting a speaker's cognitive state when interacting with a dialog system has been based on self-reports, or on hand-coded subjective judgments based on audio or audio-visual observations. This study examines two questions: (1) how do undesirable system responses affect people physiologically, and (2) to what extent can we predict physiological changes from the speech signal alone? To address these questions, we use a new corpus of simultaneous speech and high-quality physiological recordings in the product returns domain (the SRI BioFrustration Corpus). "Triggers" were used to frustrate users at specific times during the interaction to produce emotional responses at similar times during the experiment across participants. For each of eight return tasks per participant, we compared speaker-normalized pre-trigger (cooperative system behavior) regions to post-trigger (uncooperative system behavior) regions. Results using random forest classifiers show that changes in spectral and temporal features of speech can predict heart rate changes with an accuracy of ~70%. Implications for future research and applications are discussed.

Index Terms: autonomic nervous system, heart rate, speech features, emotion, frustration, dialog systems, speech corpora.

1. Introduction

Most research aimed at detecting the state of a user has been based either on self-reporting or on post-hoc assessments of user state using audio/audio-visual recordings. In the latter case, ground truth is typically based on subjective judgments of the spoken utterances to the system [1, 2]. In this study, we are interested in understanding (1) how undesirable system responses affect people physiologically, and (2) the extent to which we can predict physiological changes based on only the speech signal. These questions are important for attempts to mitigate the negative emotional impact of the technological shortcomings of such systems—for example, offering explicit sympathy, offering rewards (e.g., a discount), preventing inappropriate responses such as upselling during a negative user experience, or passing the user to a human agent when feasible [13, 14].

Little is known about how well we can predict physiological state from speech alone in this context. To investigate this, we analyze a recently collected corpus of simultaneous speech and physiological sensor data during the interaction with a dialog system in the product returns domain. The system is to "misbehave" at specific points in the workflow, thus allowing controlled comparisons across users between speaker-normalized speech features and speaker-normalized physiological signals. In this paper, we focus specifically on the prediction of a speaker's change in heart rate from changes in their speech.

2. Method

2.1. Speech and sensor data

We used the SRI **BioFrustration Corpus** [1] which contains data from 53 native English speakers (20 males, 33 females) aged 18 to 75, with an average age of 36.4. Participants were SRI employees, their relatives, and Stanford University undergraduates.

The corpus offers two advantages for studying the relationship between speech and physiological features. First, it combines state-of-the-art audio, video, and physiological signals. We used a BioDerm Skin Conductance Meter (model 2701; UFI, Morro Bay, CA) to measure skin conductance by means of two electrodes placed on the palm of the subject's non-dominant hand. Portapres Model-2 (TNO TPD Biomedical Instrumentation, Amsterdam, NL) was used to measure systolic, diastolic, and mean beat-to-beat blood pressure. The Electrocardiogram (ECG) was recorded using Medi-Trace Ag/AgCl surface spot electrodes placed in a modified Lead II Einthoven configuration. Thoracic Piezo Grael Rip Bands were used to record the breathing respiration.

Increases in heart rate, systolic blood pressure, skin conductance level and respiration rate generally reflect an increase in physiological activation either positive or negative. The collection was designed to exhibit only neutral (as expected in any particular context) or negative behaviors. We thus interpret increases in physiological signal values as increases in negative experience for the participants.

A second advantage of the corpus is experimental control. The corpus design uses a fixed workflow, even though subjects believed they were interacting with a real system. This design permits meaningful comparisons both within and across speakers, as described in the sections to follow.

2.2. User Incentives

Frustration is typically defined as the emotional response to obstacles in the pursuit of needs or desires. This definition raises the question of how to ethically manipulate desires in a laboratory setting in order to generate authentic frustration experiences and behavior. Because tailoring each session to items that generate the greatest emotional response is

impractical, we chose instead to offer a monetary reward ($100). This reward was given to the participants at the beginning of the session, but they were told that failure to satisfactorily complete a minimum number of tasks would lead to confiscation of the reward. By presenting the incentive in this manner (rather than promising to pay them at the end), we appeal to participants' loss aversion. Additional incentives involved (fake) feedback to subjects during collection that they were performing less well than others in the study.

Participants were instructed that the emotional content of their speech could persuade the system to reverse its earlier refusal to grant a full refund, and the sense that this behavior alone made the difference was important. For that reason, any positive outcome (i.e., the system granting a full return after initial refusal) was accompanied by an "empathetic" statement suggesting that the system was able to sense the frustration, for example: ***SYS****: I am sorry, you don't seem to be very satisfied, let me check if we can help you. Good news! Based on the information you provided, you qualify for a full refund.*

2.3. Tasks

Participants were told to assume the identity of customers who had bought some household items and found them to be defective in some way. The details of that identity (name, address, etc.) and of the item to be returned were given to them on cue cards at the beginning of the session. To return items, subjects spoke to a dialog system (*"Returns"*). The task was to persuade the system to reimburse the full price of each item. Each participant had eight items to return; in order to qualify to keep the monetary reward, they had to return at least six of those items "successfully" (i.e., at full price). Otherwise, the participants were told, they would lose the entire amount.

If the *Returns* system did not immediately offer a full refund, the participants were told that they had at least one further chance to convince it to reverse its decision. We call this a "Justification". Additionally, they were told that the system could detect their emotional state, and that it was more likely to reverse its earlier decision if they succeeded in conveying their emotional state in their speech. This latter instruction served two purposes. First, it tries to counteract any inclination to speak to the computer in a "mechanical" voice devoid of emotional content. Second, by making the system behavior responsive to emotional expressiveness, users were encouraged to use their speech in the same way that they would convey emotions such as frustration with actual human interlocutors.

2.4. Dialog System

The *Returns* dialog system used in this experiment was built with SRI's Virtual Personal Assistant (VPA) technology. VPA systems are designed to cover a wide range of possible user intents and dialog states; for this study, we chose a system with rather limited functionality. To make the user believe that the system was capable at some level to understand their speech, the speech recognition output needed to match the expected responses for a number of narrowly defined prompts, such as questions about name, address, etc. In other instances, the system completely ignored the user input and instead proceeded according to a predefined workflow. Additionally, the outcome for each returnable item was entirely deterministic. Thus, contrary to what participants were told, their speech had no bearing on how the system would deal with their return request. In this sense, the dialog system implemented a "Wizard-of-Oz" (WOZ) protocol: the system had the very functionality that we hoped to add by using the data being collected. However, unlike most WOZ studies, no human interaction was required. The various tasks and workflows were sufficiently complex (with actual speech recognition inserted at strategic points) to render plausible the existence of an intelligent system sensitive to emotional tone.

2.5. Frustration Inducers

All but the first of the eight return tasks used involved various "frustration inducers" from these classes:

"**Ugly policy**". A reasonable assumption is that any customer who has bought a defective product is entitled to a full refund. However, this expectation is violated when the system insists on reducing the refund amount by subtracting items such as shipping fee, restocking fees, or price drops.

Throwback. Another reasonable assumption is that the computer would accumulate the information that the participants were asked to provide. However, in a number of tasks, the system flouted this expectation by "forgetting" the immediately preceding utterances and prompting the participant for the information that had already been collected previously.

Lack of understanding. In two tasks, additional loops were inserted in which the system declared its inability to understand the user and then prompted the user for a repetition. Again, we considered this device a frustration trigger because it kept the participant from achieving their goal with no clear indication of the cause for failure in understanding.

3. Method

3.1. Analysis Regions

The carefully controlled workflow allows comparison of thematically linked regions of user behavior across tasks and subjects. Regions that we compare are illustrated in Figure 1.

Figure 1 : Workflow regions used in analysis.

Each task started out with *Initiation* utterance(s) and included as a feedback prompt about the session at the end. For the purposes of the present study, those regions were ignored. After *Initiation*, subjects were given a series of prompts for buyer- and product-related information such as name, address, purchase date, etc. This "*Pre-justification*" region was not designed to induce frustration in subjects; thus, it served as a local baseline for each task. *Pre-justification* was followed by a prompt for an explanation for the return and a sequence of *Justification* prompts, which asked the subjects to state reasons why they thought they deserved a full refund. In two tasks, the flow was interrupted without explanation, taking the subject back to the *Pre-justification* stage. As a result, for those two tasks there were two instances of the *Pre-justification*, *Explanation*, and *Justification* regions. In contrast to the *Pre-*

justification regions, the *Justification* regions (especially those after a system throwback) presented participants with the strongest reason for frustration. In the analysis described in the next section, we concentrate on the differences in speech and physiological features observed in these two regions.

3.2. Speech Features

We examined a range of features that capture spectral, temporal, and prosodic characteristics of speech. This set included **Mel frequency cepstral coefficients (MFCC)** [15] to capture the spectral energy of the speech signal in the cepstral domain; **Mel Frequency Bands (MFB)** to capture the energy information in the frequency domain; **Energy contours (EC)** [17] to capture long-term information in the speech signal, including rhythmicity and speaking rate without use of speech recognition; **Spectral tilt (TILT)** [18] to capture the vocal effort and its independence of the session variability; **DLE** [18] to capture the local vocal effort changes without requiring speaker normalization; **Harmonic-to-noise ratio (HNR)** [16] to capture the harmonic versus non-harmonic energy of the signal; and **Intonation-related features (IR)** (F0, F0 peaks and F0 peaks statistics) to capture longer-term information pertaining to pitch, pitch peaks, and intensity. OpenSMILE features [19] are extracted at the frame level based on the set-up described in [20].

Features were computed for *Pre-justification* and *Justification* regions. For each region and for each task, we computed the mean, variance, median, interquartile range, kurtosis and skew for each of the speech features. Then, for each feature, we computed the percent change from *Pre-justification* to *Justification* regions. This process was repeated by computing the same values from *Justification* to *Pre-justification* which represent the inverse relation, and augment the training data leading to a balanced two class data set. Physiological signals were aligned to the speech signals by time-correlation alignment of two microphone signals (one aligned with heart rate; the other with speech capture. Heart rate was extracted by finding the interval between the signal peaks.

3.3. Physiological Features

Due to space constraints, we focus on heart rate only; compared to the other collected measures, heart rate is less sensitive to artefacts. The average heart rate per task is categorized as increasing or decreasing. All features are represented as percentage change from *Pre-justification* to *Justification* and vice versa. Note that the baseline for the *Pre-justification* to *Justification* direction is the *Pre-justification* features, while the baseline for the opposite direction is the *Justification* features. Hence, the percentage change is different for each direction, even though the same source features are used.

3.4. Classification

Figure 2 shows the approach used to predict heart rate change. Overall statistics were extracted for each feature and the percentage of change from *Pre-justification* to *Justification* (and vice-versa) was computed for each feature, after alignment of the multiple microphone signals.

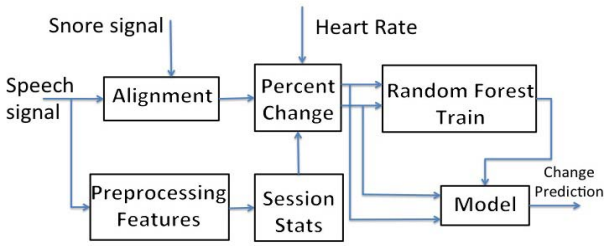

Figure 2 : Flow diagram for prediction of the heart rate direction from change from the speech signal

Speech features were then modeled by a Random Forest classifier (RF) [21] trained to predict when the heart rate was increasing or decreasing. Speakers were partitioned into 10 sets with no speaker overlap, for cross-validation. Of the total number of sessions, the usable sessions were 50. After experimenting with random forest classifiers with 20 to 50 trees and 5 to 20 features per tree, we picked a RF with 50 trees and 5 features per tree with 70% data overlap based on the results on a small held out set of 40 task sessions (Note that each subject has 6-7 task sessions). Finally, we computed accuracy as the number of correctly predicted heart rate direction changes. For fusion, we used a linear voting method in which each system carries equal weight, to avoid further data splitting that would be required for a tuning set.

4. Results

To answer the first question in the study, on *the effect of dialog system misbehavior on heart rate,* an analysis was performed within subject. For each subject and task, the difference in mean heart rate between the justification and pre-justification regions was computed. An initial analysis compared heart-rate change between two neutral pre-justification regions; there was no system misbehavior between these regions and overall heart-rate change was random -- suggesting that lack of system misbehavior is associated with lack of heart-rate change.

This yielded seven values per subject for 40 of the participants (for other participants, some values were missing). Using a strict binary distinction, the distribution of rises versus falls for each subject is plotted in Figure 3. We also tried various methods for thresholding the amount of change; and found that the basic trend in Figure 3 does not change. Also, plotted are the per-subject expected values from the binomial distribution if the occurrence of rise versus fall were random, with a 0.5 prior for each outcome.

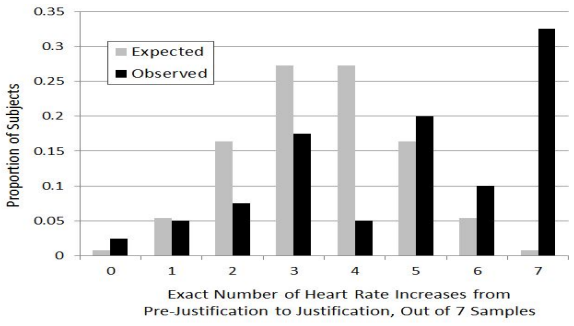

Figure 3: Expected versus observed heart rate increases from pre-justification to justification region, by subject.

Two important results are visible. First, there is a strikingly higher proportion (.325) than expected (1/2^7 or .0078) of subjects who show all seven out of seven possible heart rate increases from pre-justification to justification. Second, the tendency for within-subject consistency in direction is not symmetrical. It is only true for raised heart rate; observed versus expected values for outcomes with consistently lower heart rate, i.e. x=0 or x=1, are more in line with expected values. Overall, the misbehavior of the system had a significant effect in raising subject heart rates from pre-justification to justification within each return task.

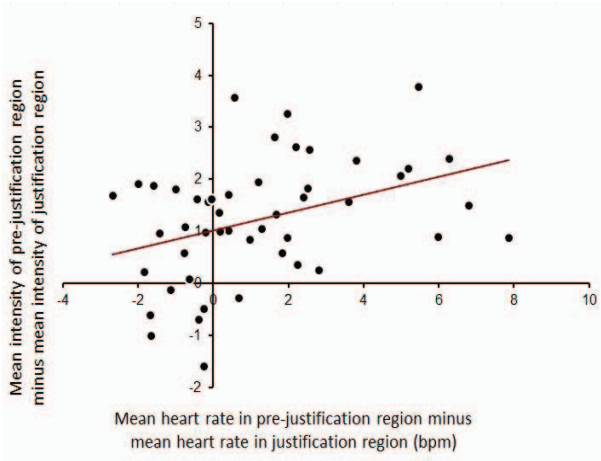

Figure 4. Heart rate change vs. Intensity change by speaker

To explore the second question in the study, on *how well heart rate can be predicted from speech*, two types of results are relevant. One set of results examines the relationship between specific speech features and heart rate, again within subject. An example outcome is shown in Figure 4, for intensity. On the x-axis, we plot the actual difference between pre-justification and justification regions rather than only the direction of difference. The y-axis provides the delta energy between the same two regions for that speaker. Values are averaged over all tasks for each speaker. As can be seen, the majority of data points (one per subject) lie in the upper right quadrant, indicating that both heart rate and intensity are increased in justification regions compared with pre-justification regions. But many data points show negative values for both heart rate and intensity; in these cases there is a weak but non-negligible correlation between the two. That is, sometimes speakers have a lower heart rate after the system misbehaviors, but when they do, they tend to also lower their intensity. The reasons for lowered heart rate need further explanation but we observed "sighing" behaviors (which could result in large temporary heart rate deflections), and effects involving the interaction between respiration, speech, and heart rate.

The other set of results relevant for the second question on predicting heart rate from speech use multiple features and machine learning techniques, as presented in Figure 5. The goal of the results is to provide a first look at how well features perform rather than to optimize specific results. It should be kept in mind that the data were collected in a quiet room with no other acoustic stimuli and while the subject was sitting in a chair with fixed microphone locations. Features such as intensity, MFCCs, and other frame level features affected by the spectrum are likely to be less robust in real-world environments; for this reason additional features such as pitch are also included for comparison.

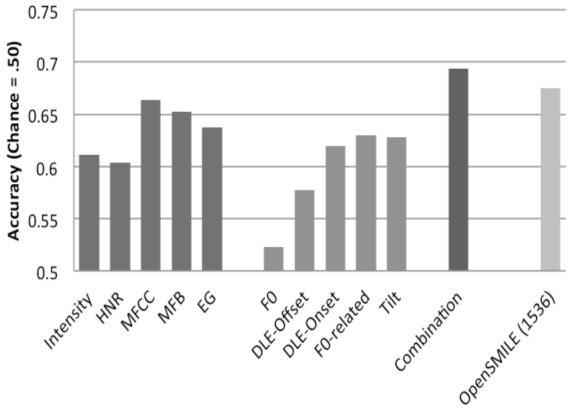

Figure 5. Prediction of change in direction of heart rate from diverse sets of speech features.

Results using random forest classifiers in cross-validation mode as described earlier, are shown in Figure 5. The "Combination" system contains the 10 feature types to the left in the plot. All systems perform better than chance, although only the MFCC, Combined, and OpenSMILE systems exceed chance with the result being significant even at $p<0.01$ using a Wilcoxon Signed-Rank test. The MFCC and OpenSMILE results are similar; the latter set includes 1536 total dimensions and may not be warranted given the small performance difference on this data set. By combining all the systems except OpenSMILE, results are marginally better than MFCC and OpenSMILE. Given that the current analysis uses majority voting to avoid further data partitioning, it is possible that better fusion results can be obtained using a larger data set.

5. Summary and Future Work

The present study found overall increased heart rate effects as a result of system misbehaviors in a highly controlled experiment. The increases in heart rate are subtle (as might be expected or exceeded in real-world contexts) but are statistically significant. The changes are partially predicted by changes in a range of speech features, suggesting that speaker state prediction could be partially achieved using features only from speech. Subjects appear to differ in both heart rate changes and speech features changes, a topic that requires further investigation. The modeling of physiological effects should also be further explored, controlling for whether or not speech is present, breathing effects, body movement, and gender and age differences, among other factors. In future work, we plan to explore the predictability of other physiological measures such as skin conductance and blood pressure, as well as to examine the use of video signals for multimodal prediction of physiological changes. Also, we plan to evaluate the effect of speaker variables, for example, gender and age.

6. References

[1] Kathol, A., E. Shriberg, "The SRI BioFrustration Corpus: Audio, Video, and Physiological Signals for Continuous User Modeling", AAAI Spring Symposium, March 2015.

[2] Cacioppo, J., Tassinary, L., Berntson, G.. "Handbook of Psychophysiology". New York, NY. Cambridge University Press, 2007.

[3] Kahneman, D., Tversky, A. "Choices, values, and frames." *American Psychologist* **39** (4): 341–350. , 1984.

[4] Cellini, N., de Zambotti, M., Covassin, N., Sarlo, M., Stegano, L.. "Working memory impairment and hyperarousal in young primary insomniacs." *Psychophysiology*, 51(2), 206–214, 2014.

[5] Hone, K. "Empathic agents to reduce user frustration: The effects of varying agent characteristics." *Interacting with Computers* 18 (2), 227–245, March 2006.

[6] Klein, J, Moon, Y., Picard, F.W.. "This computer responds to user frustration: Theory, design, and results." *Interacting With Computers 14*, no. 2 (February 2002): 119–140.

[7] Riseberg, J., Klein, J., Fernandez, R., Picard, R.W.. "Frustrating the user on purpose: Using biosignals in a pilot study to detect the user's emotional state.", *CHI 98 Conference Summary on Human Factors in Computing Systems* 227–228.

[8] Scheirer, J. Fernandez, R., Klein, J., Picard, R.W. "Frustrating the user on purpose: A step toward building an affective computer. *Interacting With Computers* 14 (2) (February 2002): 93–118.

[9] Schuller, B., Friedmann, F., Eyben, F. "The Munich Biovoice Corpus: Effects of physical exercising, heart rate, and skin conductance on human speech production", *Proc. LREC*, 2014.

[10] Sethu, V., Epps, J., Ambikairajah, E. "Speech-based emotion recognition." In Ogunfunmi, T., Togneri, R., and Narasimha, M. (eds), *Advances in Speech and Audio Processing for Coding, Enhancement and Recognition*, Springer, 2014.

[11] Ward, N. G., DeVault, D. 2015. "Ten challenges in highly-interactive dialog systems", *AAAI 2015 Spring Symposium*.

[12] https://www.bbvaopenmind.com/en/bbva-and-sri-international-debut-the-first-intelligent-virtual-personal-assistant-vpa-for-banking.

[13] C. Busso, M. Bulut, C.C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, and S.S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database", Journal of Language Resources and Evaluation, vol. 42, no. 4, pp. 335-359, December 2008.

[14] F. Ringeval, A. Sonderegger, J. Sauer, D. Lalanne, "Introducing the RECOLA Multimodal Corpus of Remote Collaborative and Affective Interactions", 2nd International Workshop on Emotion Representation, Analysis and Synthesis in Continuous Time and Space (EmoSPACE), In Proc. of IEEE Face & Gestures 2013, Shanghai (China), April 22-26, 2013.

[15] S.B. Davis, and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", in *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4), pp. 357–366, 1982.

[16] Yumoto, E., Gould, W.J., Baer, T., "Harmonics-to-noise ratio as an index of the degree of Hoarseness", Journal of the Acoustical Society of America **71**: 1544-1550, 1982.

[17] E. Shriberg, A. Stolcke, D. Hakkani-Tur, L. Heck, "Learning when to listen: Detecting system-addressed speech in human-human-computer dialog," in *Inter-speech*, 2012.

[18] E. Shriberg, A. Stolcke, S. Ravuri, "Addressee Detection for Dialog Systems Using Temporal and Spectral Dimensions of Speaking Style," Proc. Interspeech, 2013.

[19] Florian Eyben, Martin Wöllmer, Björn Schuller: "openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor", *In Proc. ACM Multimedia (MM), ACM, Florence, Italy*, ACM, ISBN 978-1-60558-933-6, pp. 1459-1462, October 2010.

[20] Schuller, B., Steidl, S., Batliner, A., Nöth, E., Vinciarelli, A., Burkhardt, F., Weiss, B. "The INTERSPEECH 2012 Speaker Trait Challenge". In *INTERSPEECH,* Sep, 2012.

[21] Liaw, A., & Wiener, M. (2002). "Classification and regression by random Forest". *R news*, *2*(3), 18-22.