The SRI BioFrustration Corpus: Audio, Video, and Physiological Signals for Continuous User Modeling

Andreas Kathol and Elizabeth Shriberg
Speech Technology and Research Laboratory, SRI International
{kathol,ees}@speech.sri.com

Abstract

We describe the **SRI BioFrustration Corpus**, an in-progress corpus of time-aligned audio, video, and autonomic nervous system signals recorded while users interact with a dialog system to make returns of faulty consumer items. The corpus offers two important advantages for the study of turn-taking under emotion. First, it contains state-of-the-art ECG, skin conductance, blood pressure, and respiration signals, along with multiple audio channels and video channels. Second, the collection paradigm is carefully controlled. Though the users believe they are interacting with an empathetic system, in reality the system afflicts each subject with an identical history of "frustration inducers." This approach enables detailed within- and across-speaker comparisons of the effect of physiological state on user behavior. Continuous signal recording enables studying the effect of frustration inducers with respect to speech-based system-directed turns, inter-turn regions, and system text-to-speech responses.

Introduction

Successful coordination of human-computer interactions is greatly aided if the computer can model user states, including emotion (Sethu et al., 2014; Schuller et al., 2014; Ward and DeVault, 2015). Since studies have found (Klein et al., 2002; Hone, 2006) that user frustration can be mitigated, detecting user frustration can facilitate more successful human-computer interactions.

We present an initial description of a new, ongoing data collection, the **SRI BioFrustration Corpus**, which we believe offers two advantages for studying turn-taking phenomena for user-state-aware interaction with a dialog system. First, the corpus offers rich, state-of-the-art audio, video, and physiological signals. Multiple audio and video recordings are time-aligned with autonomic nervous system (ANS) signals, including ECG, skin conductance, blood pressure, and respiration. Increases in each of the four ANS sensor signals are generally correlated with increases in physiological activation (e.g., Cacioppo et al., 2007; Cellini et al., 2014). The measures are designed to capture user responses to points in the dialog that are specifically constructed to elicit frustration. The present study sets itself apart from the growing body of efforts in this area in the range of measurements taken and the quality of the recordings (see below for more details).

A second advantage of the SRI BioFrustration Corpus is experimental control. Although having a flexible system that can respond to many different user utterances is generally desirable, such flexibility also means that users may traverse sessions in different ways. Thus, comparison across subjects is problematic because a different conversational history may precede a particular dialog state for each speaker. To overcome this issue, our system is instead designed to allow minimal variation in system responses across speakers. This approach permits both meaningful inter-speaker comparison and examination of within-speaker variation due to exposure to different emotional triggers during a session.

The collection paradigm is carefully designed to motivate users to convince a seemingly empathetic system to provide the desired outcome. However, in reality the system afflicts each subject with the same history of "frustration inducers"—thereby enabling controlled study of the relationship between physiological correlates of frustration in both speech and video.

Continuous signal recording enables studying the location and timing of frustration-predictive features with respect to speech-based system-directed turns, inter-turn regions, and system text-to-speech responses, as illustrated in **Figure 1** below, including asking the following questions:

The authors would like to thank Edgar Kalns, Massimiliano de Zambotti and Ajay Divakaran for numerous valuable suggestions.

- What is the effect of physiological state on turn length, turn latencies, and barge-in rates?
- Is speech and/or gesture-based behavior between system-directed turns (for example, off-turn talk, grunts, or facial or head movements) useful for prediction of current or future physiological state?
- What effects are seen in the speech signal under frustration?
- What effects are seen in the video signals under frustration?
- How are effects in speech and video related to the user's specific history of and magnitudes of changes in physiological signals?
- How do individual speakers vary in the relationship between physiological and A/V signal changes?

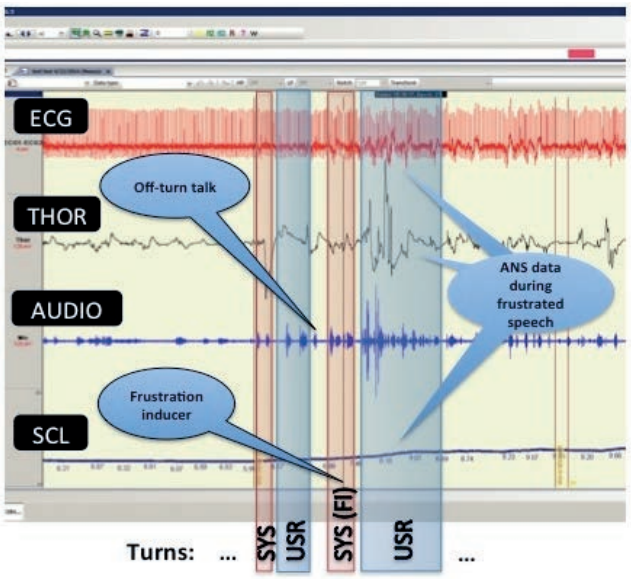

Figure 1: Illustration of turns overlaid onto (subset of) ANS tracks (ECG, breathing (THOR), skin conductance level (SCL)); examples of off-turn talk, frustration inducers.

In the following section, we describe the experimental setup (including the dialog system used), triggers of frustration, and methods for recording ANS and other data.

Corpus Design and Collection

If frustration is the emotional response to obstacles in the pursuit of needs or desires, then this definition raises the question what kinds of needs can be manipulated ethically in a laboratory setting to generate authentic frustration experiences and behavior. Because customizing each session to the items that generate the greatest emotional response in each subject is impractical, we chose instead to use a substantial monetary reward ($100). This reward is given to the subjects at the beginning of the session, but they are told that failing to satisfactorily complete a minimum number of tasks will lead to having the entire reward revoked. By presenting the incentive in this manner (rather than promising them pay at the end), we appeal to subjects' loss aversion (Kahneman & Tversky, 1984). Additionally, the subjects are told that other subject have had no trouble completing the tasks, which sets the expectation that the tasks will be easy to accomplish.

Tasks

The subjects are told to assume the identity of customers who have bought some household items and have found them to be defective in some way. The details of that identity (name, address, etc.) and of the item to be returned are given to them on cue cards at the beginning of the session. To return the items, the subjects need to talk to a dialog system (*"Returns"*). The task is getting the system to take back each item and reimburse the customer at full price. Each subject has eight items to return; in order to qualify for keeping the monetary reward, they must return at least six of those items "successfully" (i.e., at full price). Otherwise, the subjects are told, they will lose the entire amount.

If the *Returns* system does not immediately offer a full refund, the subjects are told that they have at least one chance to convince it to reverse its decision. Additionally, they are told that the system can detect their emotional state and is more likely to reverse its earlier decision if they succeed in conveying their emotional state in their speech. This latter instruction serves 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 seemingly responsive to emotional expressiveness, users are encouraged to use their speech in the same way that they would convey emotions such as frustration with actual human interlocutors.[1]

Dialog system

The *Returns* dialog system used in this collection was built with SRI's internal VPA technology.[2] VPA systems are designed to cover a wide range of possible user intents and dialog states; but for this work, we opted for a system with rather limited functionality. To make the user believe that the system is capable of understanding at some level, the speech recognition results must match the expected responses for a number of narrowly defined prompts, such as questions about name, address, etc. In other instances, the system completely ignores the user input and instead

[1] While this may be taken by subjects as an invitation to overact, we have not actually found such behavior in our data so far.
[2] See e.g., https://www.bbvaopenmind.com/en/bbva-and-sri-international-debut-the-first-intelligent-virtual-personal-assistant-vpa-for-banking/

proceeds according to a predefined workflow. Additionally, the outcome for each returnable item is entirely deterministic. Thus, contrary to what subjects are told, their speech has no bearing on how the system will deal with their return request. In this sense, the dialog system implements a "Wizard-of-Oz" (WOZ) protocol: the system has the very functionality that we hope to add by using the data being collected. However, unlike most WOZ studies, no human interaction is required. The various tasks and workflows are sufficiently complex (with actual speech recognition inserted at strategic points) to render plausible the existence of an emotionally intelligent system.

Frustration inducers

All but one of the eight tasks presents subjects with a variety of what we call "frustration inducers." These elements have been deliberately inserted in the workflow to hinder the subjects in their goal of a full refund for the item to be returned. Three frustration inducers are discussed in the following.

"Ugly policy"

A reasonable assumption is that any customer who has acquired a defective product is entitled to a full refund. However, this expectation is thwarted when the system insists on reducing the refund amount by subtracting certain expenses such as shipping fees or restocking fees. Additionally, a sudden price drop is presented as a reason why only a partial refund can be awarded.

Throwback

Another reasonable assumption is that the computer will accumulate the information that the subjects are asked to provide. However, in a number of tasks, the system flouts this expectation by simply "forgetting" the immediately preceding *n* utterances and prompting the subject for the information that was already collected a few turns earlier.

Lack of understanding

In a number of tasks, additional loops have been inserted in which the system declares that it is unable to understand the user and prompts the user for a repetition. Again, we consider this setup a frustration trigger because it keeps the subject from achieving his/her goal with no clear indication of what caused the failure to communicate.

A schematic example of a dialog follows. Speaker turns that are ignored by the system are listed as "<anything>". System prompts designed to frustrate the user are annotated as "FI" (frustration inducer):

SYS: Welcome. Please tell me what you'd like to do.
USER: I need to return hiking boots.
SYS: Sure, I can help you with returning your hiking boots. First, tell me your name please.
USER: <name from card>
SYS: … (more prompts)
SYS: Tell me in your own words what the problem is.
USER: <anything>
SYS: Did you use the item in any way that would void the warranty?
USER: No
***FI** (user input ignored):*
SYS: Could you elaborate on what has voided the warranty?
USER: <anything>
***FI** (user input ignored, "ugly" policy):*
SYS: Due to the voided warranty, you have to pre-pay the restocking fee of $35 to return the hiking boots. Hence, you only qualify for a partial refund of $63. ... If you are not satisfied, you may qualify for a higher refund if you can state your case convincingly in a few sentences.
USER: <anything>
***FI** (initial explanation rejected):*
SYS: Okay. Please elaborate in greater detail why you think you are eligible for a full refund.
USER: <anything>
***FI** (understanding failure):*
SYS: I'm sorry, I did not fully understand what you just said. Could you repeat please?
USER: <anything>
***FI** (understanding failure, again):*
SYS: My apologies, but I still have trouble understanding you. Could you repeat that please?
USER: <anything>
***FI** (final rejection):*
SYS: Let me check ... Unfortunately, according to company policy, these are insufficient reasons to reconsider our decision. You still only qualify for a partial refund of $63.

The workflow for a given task ignores user input at the same places for each subject. This ensures that the dialog history is fully comparable across subjects at any given point, including frustration inducers.

Nonfrustration regions

Because the subjects are instructed that the emotional content in their speech is what will get the system to reverse its earlier refusal to grant a full refund, providing a sense that precisely that behavior has made the difference is important. For that reason, any positive outcome (i.e., the system granting a full return after initial refusal) is accompanied by an "empathetic" statement suggesting that the system was able to sense the frustration:

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.

Tasks with a positive final resolution also serve to establish a nonfrustrated baseline that frustrated speech can be compared against. Additionally, we add half a minute of relaxation between tasks to reduce any spillover effect from an emotionally charged task to the next task.

Recording setup

The *Returns* application runs on a dedicated MacBook Pro laptop. The subjects interact with the application by using a button that activates speech recognition in a push-to-talk mode. The audio files for individual utterances are written out to disk. Simultaneously, a continuous video and close-talking audio recording of the subject is being made by using QuickTime in "movie recording" mode, as well as by a separate Kinect device for distant audio and high-quality video. Although the video track is not being processed at in this collection, the audio track constitutes a full record that can capture the speech that occurs between turns. That is, subjects have been found to express frustration outside of the explicit input turns to the dialog system, including spontaneous vocalizations such as "oh, come on!" As such, we hope to eventually test the hypothesis that some subjects may be so used to modulating their emotional response when talking to a computer that off-turn vocalizations may be far better (possibly the only) indicators of emotional state.

ANS measurements

The idea of measuring autonomic nervous system data in computers interactions that invoke frustration goes back at least to Riseberg et al. (1998). That study proposed using ANS data as a way to gauge the ground truth with respect to the presence and degree of frustration, as a potentially superior method over self-reporting or annotation by observers. Scheirer et al. (2001) showed how data can be used to predict the frustration/non-frustration status for given sections of a game from ANS data with above-chance accuracy. We follow a similar path in the present study in that we aim to correlate speech behavior with some non-subjective indication of the presence and degree of frustration. To this end, the subjects are instrumented with a number of ANS sensors during the sessions. In addition to skin conductance and blood pressure as in Riseberg et al. (1998), we use two more measurements.

Skin conductance

A BioDerm Skin Conductance Meter is used to measure skin conductance level by means of two electrodes placed on the palm of the subject's non-dominant hand.

Blood pressure

A Portapres Model-2 cuff is placed on the intermediate phalanx of the middle finger of the non-dominant hand to measure systolic, diastolic, and mean blood pressure.

Electrocardiography

ECG recordings are performed by using Medi-Trace Ag/AgCl surface spot electrodes placed in a modified Lead II Einthoven configuration through a ProFusion nexus platform using Grael amplifiers.

Respiration rate

Thoracic Piezo Grael Rip Bands are used to record the breathing signal through the ProFusion 3 software platform using Grael amplifiers with a sample rate of 64 Hz.

The analog outputs from the Portapres Model-2 and the BioDerm Skin Conductance Meter are connected via optically isolated ExLink DC inputs to Compumedics Grael amplifiers and sampled at 64 Hz. All signals are recorded using ProFusion 3, and all indices are analyzed with a beat-to-beat time resolution.

Conclusion

Although the effort described here is still in its early stages, we hope that this corpus, with a target of 50 subjects, will prove valuable to the study of turn-taking, in particular with respect to the correlation between speaker (frustration) state and dialog behavior such as barge-in or response latency.

References

Cacioppo, J., Tassinary, L. and Berntson, G. 2007. *Handbook of Psychophysiology*. New York, NY. Cambridge University Press.

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

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

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

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

Riseberg, J., Klein, J., Fernandez, R., Picard, R.W. 1998. "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.

Scheirer, J. Fernandez, R., Klein, J. and Picard. R.W. 2001. "Frustrating the user on purpose: A step toward building an affective computer. *Interacting With Computers* 14 (2) (February 2002): 93–118.

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

Sethu, V., Epps, J., and Ambikairajah, E., 2014. "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.

Ward, N. G. and DeVault, D. 2015. "Ten challenges in highly-interactive dialog systems." to appear in *AAAI 2015 Spring Symposium*.