End-To-End Speech Negotiations with Affective Speech Rollout

Victor Ardulov, Arindam Jati, Maury Lander-Portnoy

December 5, 2017

Abstract

Our project aims to explore the application and potential advantages that affective speech processing and speech synthesis will have on machine negotiations. We have constructed an end-to-end speech negotiation pipeline and run preliminary experiments to gauge the functionality of our paradigm. Our approach is to predict emotion class probabilities in a user's speech and then attempt to use it against them in a negotiation using a temporally discounted model to capture a weighted history of emotion-beliefs. We evaluate the performance of the proposed system in interactions with human agents across different experimental conditions with increasing degrees of affective engagement. We also measure engagement and emotional arousal of users by retrospective self-report when negotiating with our system and compare this to users' perceptions of their performance in the negotiations. The pipeline would be a publicly available end-to-end speech based intelligent negotiation agent with embedded affective modeling. We strongly hope the system will serve as a useful tool for building a human-machine negotiation dataset.

1 Introduction

Negotiation is intrinsically an emotionally engaging activity. Simple lexical (word-level) encoding of language fails to capture a wealth of acoustic information that reveals trust, rapport, and agent's emotional state. Unsurprisingly, this acoustic information can have a large influence on an agent's performance in a negotiation. Affect, particularly emotion and mood, can be leveraged to sway a situation. Human negotiation is highly embedded in this domain, and drives a desire to build "affectively-aware" agents that are designed to understand how human emotion contextualizes a negotiation. These agents would provide the ability to take advantage of or, oppositely, protect the emotional vulnerability of human users. In this project, we seek to evaluate the degree to which emotion in speech can be used in a negotiation by creating an artificial negotiation agent that recognizes emotion in the user's voice, and then responds with emotionally charged speech.

The proposed pipeline begins with Automatic Speech Recognition (ASR) and acoustic emotion recognition of the human agent's speech. The ASR's output text is passed to a text-based negotiation model, while emotion class probabilities are inferred by a temporal emotion negotiation model we propose. We utilize the KALDI ASR library which extracts Mel-Frequency Cepstral Coefficients (MFCC) features to perform Speech-to-Text (STT) conversion [1]. Moreover, we use OpenSMILE [2] to extract paralinguistic acoustic features for emotion recognition from speech. With the variety of features extracted, the system is embedded within a rich feature space allowing the use of affective models, and control.

At the center of our system is a reimplementation of the work presented by Lewis *et al.* [3] (FAIR), where a text-based machine negotiation learning model is presented. This system, while designed for dialog roll-out, does not include an explicit model for emotional dynamics. Our contribution lies in extending the FAIR system, first by introducing a speech input-output method and then further introducing an emotional model which colors the spoken response based on the perceived affective states in the user's speech.

To generate and evaluate the role of emotion in negotiation, a variety of affective models are introduced:

- *Neutral* - A speech output system with no affective response

- *Match* - Response speech affect matches emotion detected in the input
- *Custom* (*LAJ*) - A custom affective model, designed by the authors

These models take in the emotional state detected by our acoustic feature extraction, and determine the emotion to be assigned to the outgoing text.

With the emotion generated from the affective model and text provided from the trained FAIR model, Speech Synthesis Mark-up Language (SSML) was used in conjunction with IBM Watson's[4] Text-to-Speech (TTS) to generate and playback emotionally colored text. Affective acoustics were modeled to simulate one of five emotions: Desperate (Previously "Pleading"), Happy, Angry, Sad, and Neutral. Acoustic correlates of affective states were utilized as outlined in §3.4. Previously, affective states have been found to elicit speech production with prosody varying as a function of affective state in observational studies utilizing phonetic analyses. [5][6] These prosodic manipulations to the baseline TTS were performed deterministically for each affective condition produced by the affective model. That is, given identical text from the FAIR model and emotion to be simulated from the affective model, the produced acoustics of the audio file would be identical.

The novelty and the assumed hypotheses of this project are the following. To the best of our knowledge, there is no publicly available system for end-to-end speech based negotiation. Several attempts have been made in the past towards creating such a system. One recent example is a semi-automated (Wizard-of-Oz) system created by DeVault *et. al.* [7]. We believe, the proposed pipeline would be the first open source system for end-to-end speech based negotiation with affect recognition, modeling and synthesis modules embedded into it. Along with creating the full pipeline, we also hypothesize the following as validations of our models:

- Increased emotional activation/intensity and increased user arousal/engagement;
- Equal or improved performance compared to FAIR, and Non-affective voice negotiation.

2 Background

2.1 Negotiation Chat Systems

Chat systems present themselves nicely as a platform for negotiation research since chat systems (particulary text-based) allow for strict discrete interactions that make learning strategies significantly easier for machines, and computational models.

Recently, IAGO by Mell and Gratch [8] has been developed as a platform to help drive the research for affectively intelligent virtual agents. This along with NegoChat by Rosenfold *et al.* [9] have constructed robust platforms. More recently, Lewis *et al.* released *Deal or No Deal* which was an end-to-end negotiation chat bot trained on data collected from deploying tasks similar to IAGO's on Mechanical Turk.

Although designed specifically for text based communications, these systems lay a foundation for the development of our platform.

2.2 Emotions in Negotiation

Several studies in the past have shown the importance of emotions during negotiations. A review of the field is presented in [10]. Happiness and anger have been shown to have effects during negotiation. Positive affect like happiness can be socially induced and reciprocated [11] to achieve favorable outcomes and pave creative ways to solving problems [12, 13]. However, when the opponent identifies strong positive emotion as flattery, the whole strategy can "backfire" [13]. Similarly, negative affect like anger can be both advantageous or not during negotiations. Anger and other negative emotions can be directed at some task done by the opponent (for example at some unfair deal), or at the person's harmful intentions. These can have different effects on the outcome of the negotiation [14, 15]. Sinaceur *et. al.* [16] found that anger can be fruitful for negotiation only when the opponent has relatively poorer alternatives. Deception during negotiation has also been well

studied. Fulmer *et. al.* [17] examined the impact of emotional and informational deception in negotiation. They discovered that the negotiators found emotionally misleading deception strategies as more ethical than informational deception. In a nutshell, past research suggest that understanding the communication of emotions with the other negotiator can help improving the effectiveness of negotiation [18].

3 Affective Speech Pipeline for Negotiation

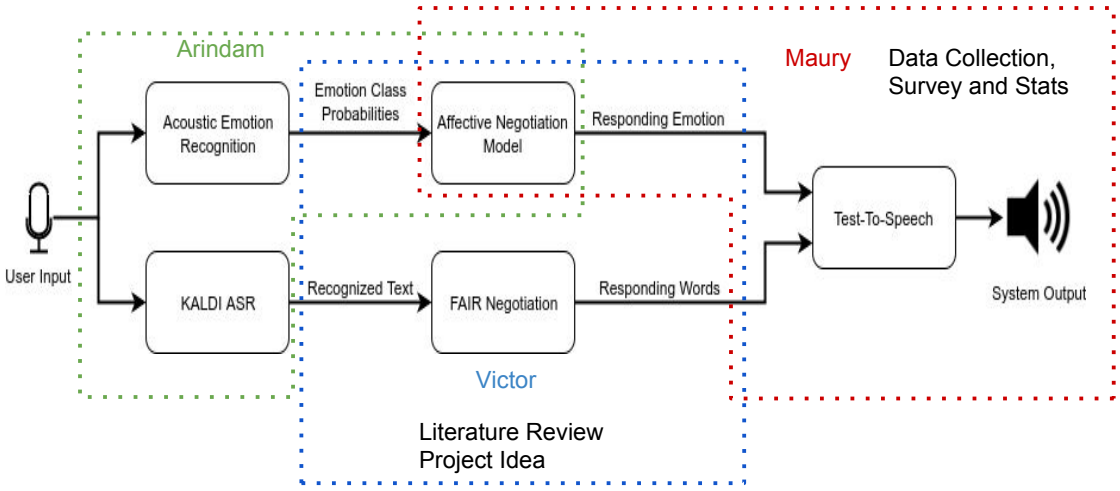

Figure 1: The end-to-end speech based negotiation pipeline along with roles and responsibilities of the authors.

Figure 1 shows the full pipeline along with roles and responsibilities of the authors. The individual modules are described below in detail.

3.1 Automatic Speech Recognition (ASR)

The KALDI ASR [1] has been employed as the speech-to-text module in our pipeline. We have used Deep Neural Network (DNN) based nnet2 models [19]. The acoustic and language models have been trained (pre-trained models available on KALDI website [20]) on Fisher English dataset (English conversational telephone speech (CTS)) [21]. We have utilized i-vector [22] based speaker adaptation for decoding in ASR. An i-vector is a vector of dimension several hundred (one hundred, in this particular context) which represents the speaker properties, and i-vectors were shown to be effective for ASR. For the online decoding of speech into text (to be useful for the pipeline), we have utilized the gstreamer-kaldi system as described in [23], which uses server-client architecture for generating text from speech.

3.2 Emotion Recognition from Speech Acoustics

3.2.1 Acoustic Features

We have employed extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) features [24] for acoustic emotion recognition of the human agent. The eGeMAPS feature set provides a minimalistic set of acoustic features that were shown to have state-of-the-art and/or competitive results on different benchmarks [24]. The minimalism of eGeMAPS features makes it less prone to over-fitting relative to other state-of-the-art feature sets which consist hundreds of features[24]. Furtherstill, the small size of the feature set also allows for faster processing in our online system. Some of the eGeMAPS features are statistical functionals of:

- Frequency related parameters like pitch, jitter, and format frequencies;
- Energy/amplitude related parameters like shimmer, loudness, and harmonics to noise ratio;

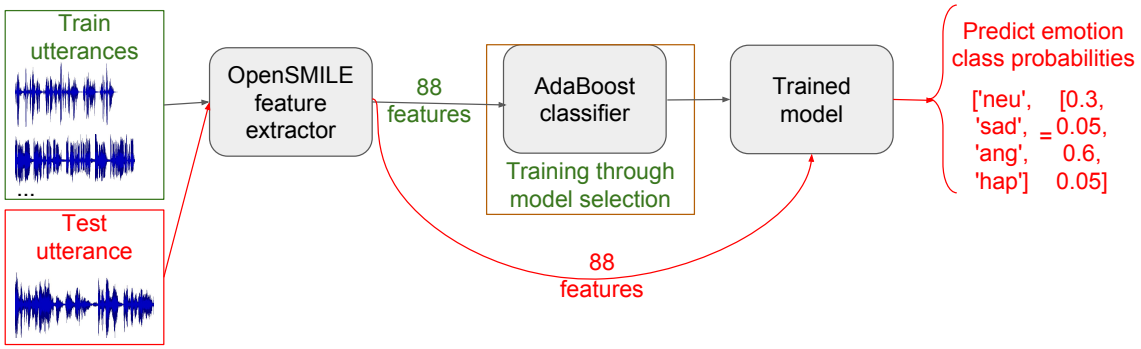

Figure 2: The emotion recognition module.

- Some temporal features like mean length of voiced regions;
- Some cepstral features like MFCC and spectral flux.

For a detailed description of these features please see [24].

3.2.2 The emotion recognition module

Figure 2 shows the emotion recognition module of the pipeline. We use supervised classification for classifying categorical emotion classes. The training utterances have manual annotations of emotion classes. They are passed through the OpenSMILE feature extractor module [25] to extract the eGeMAPS features which are then passed to the classifier for training. The trained model is saved. During test time, the utterance is passed to OpenSMILE, and the extracted features are provided to the trained model for predicting emotion class probabilities.

3.2.3 The Classifier

We have deployed AdaBoost classifier [26, 27] with decision tree classifier as the base classifier in our emotion recognition module. The classification has been done with categorical emotion classes, but class probabilities are used in our affective negotiation model. For model selection, optimal number of base estimators and the learning rate of the classifier have been chosen through leave-one-session-out cross validation on the training set and searching the parameter space by grid search. The leave-one-session-out cross validation ensures no overlap of speakers in training and test sets.

3.3 Affective Negotiation Model

In the pursuit of an end-to-end affectively intelligent negotiation agent, we needed to construct a model that takes in emotion from a user, constructs an output emotion that will be applied to the speech synthesis. This proved to be a non-trivial exercise for a number of reasons.

3.3.1 Obstacles

One of the first issues we encountered was a lack of consensus in literature on the subject of optimal emotional strategies in negotiation. While a number of strategies were outlined on how to conduct one self in a negotiation, these strategies seemed to not directly address the emotional state.

This problem was further backed by a lack of data. While there seemed to be a number of strategies, there was no publicly available data on affect in turn-based negotiation. This also meant that seeding a learned model with human affective data, similar to how FAIR dialog roll-out is learned from pure text, was not an option.

Finally, even if these systems were available the state of the art emotion recognition systems, performed with $< 60\%$ accuracy, meaning that there was a significant amount of noise coupled into the system.

3.3.2 Approach

We begin by using a the soft-max regularized output from our emotion classifier as distributional belief vector. We define a belief vector $\vec{e_t}$ which represent a mean-likelihood that the user is currently experiencing any given emotion. By treating the feature extraction like a noisy sensor signal we can update our belief with the following method.

$$\vec{e_t} = \gamma \vec{e}_{t-1} + (1 - \gamma)\vec{x}_t$$

where,

- γ - is a "discount", such that $0 \leq \gamma < 1$
- $\vec{x}_t$ - is the observed emotion distribution output from the classified at time-step t

Then a hand crafted model W, was designed by weighting how particular emotions from the user should affect whether an emotion is used in the output. These values were generally conform the idea that positive and negative emotional patterns should be matched, with arousal being a small contributing component. Table 3.3.2 describes these relationships in more detail.

We treat the output vector $\vec{o_t} = W\vec{e_t}$ as the output distribution from which we sample the an outgoing emotion. The idea is that emotions are not necessarily deterministically chosen, but clear patterns should emerge. This also means that it is highly unlikely for the agent to repeat the same interactions the same way each time allowing for more strategy exploration in the future.

Detected Emotion	Postively Impacts	Negatively Impacts
Neutral	Happy, Neutral, Anger	
Happy	Happy	Pleading, Sad
Sad	Pleading, Angry, Sad	
Anger	Pleading, Anger, Sad	Happy

Table 1: Relationship between detected emotions and outgoing emotion probabilities

3.4 Affective Speech Synthesis

The affective speech synthesis module takes the target output affect to be simulated (or neutral for control) and combines with the output text from the FB agent. The system then uses custom voice transformation of IBM Watson's TTS [4] using articulatory parameter manipulations to create utterance-long audio files. This is a bit of an exploitation of this feature, as these articulatory parameters are meant to fine-tune an agent's general speaking style (i.e. voice customization for something like a website spoken dialog agent). IBM is beginning to explore the affective TTS space, but currently, it only includes 1 predefined affective parameter set: "good-news". The affective TTS module allowed simulation of the following emotions using the acoustic manipulations described below:

1. **Pleading:** Decreased Glottal Tension, Increased Dynamic Pitch Range, Much Higher F0
2. **Happy:** Increased Breathiness (Decreased Tension), Slightly Fast, Increased Dynamic Pitch Range, Slightly Higher F0
3. **Angry:** Decreased Breathiness, Increased Glottal Tension, Much Faster, Increased Dynamic Pitch Range, Much Lower F0
4. **Sad:** Increased Breathiness, Decreased Glottal Tension, Much Slower, Decreased Dynamic Pitch Range, Much Lower F0
5. **Neutral:** No defining characteristics, slightly faster than default IBM setting

These manipulations exhibit many of the acoustic regularities observed in behavioral observations of affective speech ([5][6]).

4 Experiments

Participants 6 college-age students from the University of Southern California participated in the experiment. Participants were naive to the task prior to beginning the experiment and had no known language, cognitive, or other communicative deficit by self-report.

Procedure Participants were seated at a desk in a quiet room and negotiations took place on a MacBook Pro 13-inch running Mac OSX High Sierra (Version 10.13.1). Negotiations took place with the default microphone level set automatically and dynamically by the operating system and audio output was presented over the built-in laptop speakers at a comfortable volume. Agent output was also printed to screen in case participants missed the response or could not interpret what the TTS had said. Participants were recorded via screen recording using Quicktime 10.4. Verbal consent was obtained for the use of participants' recordings in the class final project.

The experimenter explained the task verbally to participants and then remained in the room in case of participant confusion or malfunction with the negotiation system. No such malfunctions occurred, however, participants did have several questions regarding the task. In the task, participants made offers to and received offers from the negotiation agent to split items in a common pool. The three items were always the same (book, ball, hat), however the quantity of the item in the common ground, and its utility to the agent and participant changed each round. Participants and the agent were unaware of each other's utility functions and participants were instructed to maximize their own utility compared with the agent. While no explicit mention was made to participants to minimize the agents' utility, several participants adopted this strategy and vocalized their thought process of doing so. Upon reaching an agreement with the agent, participants were asked to confirm their apportioning of the common items. If participants agreed with the agent, both scores were revealed. If participants did not agree with the agent, neither received any points.

Negotiations took place in one of four conditions. The full target condition used speech input with emotion recognition, generated output prosodic affect to use in affective TTS using our LAJ affective model, and produced output speech by combining emotion and text using the protocol outlined in §3.4 above. This condition should maximally elicit emotion due to full utilization of the affectively aware pipeline. In the task baseline control, participants interacted with the agent using keyboard and thus no emotion recognition or modeling was performed. Additionally, agent responses were presented only in the neutral affect. This condition provides a baseline for participants' general engagement in and arousal by the negotiation task. As the nature or parameters of the task itself remained constant across conditions, this provides the most basic ability to elicit emotion in users and we thus predict this condition will produce minimal user arousal and engagement. A third condition served as a control for user reaction to responding using speech. In this condition, participants responded to agent propositions using spoken language but were responded to using only neutral affect. This condition simulates no affective state for the agent and allows the measurement of a user's differentially increased arousal and engagement given the opportunity to respond using spoken language. As participants may feed off of their own affectively charged output, we may observe an increase in arousal in this condition, however it is unlikely that participants will experience a large difference in emotional engagement given the emotionless prosody of the agent. The final control condition mirrored users' emotions as classified by the emotion recognition module. In this condition the output emotion used in the TTS module was simply a match of the emotion with the highest likelihood predicted by the emotion recognition system. This allowed us to measure participants' reaction to speech with simulated emotion, even when that emotion simply mirrored their own. As this will allow for emotional feedback mechanisms (such as getting angrier and angrier when being yelled at by the agent), we predict participants will experience a significant increase in arousal on this condition when compared to the previous two control conditions.

After participants completed the negotiation tasks, a quick self report survey was obtained for all participants. In the survey, questions probed the degree of

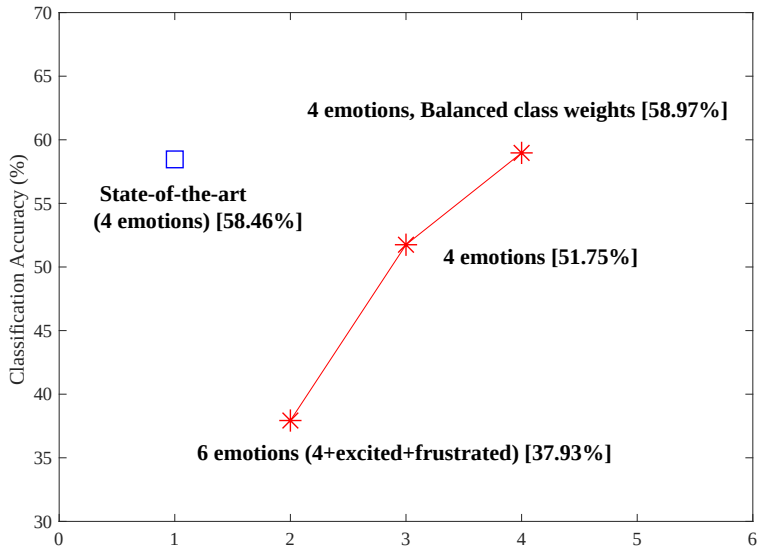

Figure 3: Classification accuracies for emotion recognition in IEMOCAP dataset with different approaches. The state-of-the-art performance is reported from [29].

participant investment in the task, arousal during the task, and overall sentiment about the task. Participants used 7-point-lichert scales to rate engagement and investment, and selected from a multiple choice (with multiple answers allowed) list of which emotions they experienced in which conditions. These questions allowed for user addition of emotions not present on the list. The list was initially populated with several emotions obtained from [28]. These measures were also asked for specific conditions. The results of the negotiation task and post-task survey are described below.

5 Results

5.1 Acoustic Emotion Recognition

The IEMOCAP dataset [30] has been used to train the acoustic emotion classifier model. The dataset has dyadic conversations between 10 actors divided into 5 sessions. The sessions are manually segmented into utterances, and all the utterances have manual annotation of emotion classes.

Figure 3 shows leave-one-session-out emotion classification accuracies in IEMOCAP dataset using different approaches. [29] reported state-of-the-art performance on IEMOCAP dataset for emotion recognition with four classes: Neutral, Happy, Sad, and Angry. We started with six major emotions available [30] in IEMOCAP dataset, which includes 'excited' and 'frustrated' along with the four basic emotions used in [29]. We could achieve 37.93% accuracy. The same architecture with four basic emotions gave 51.75% accuracy. The emotion classes in IEMOCAP dataset have different number of samples or utterances (please see [30] for details). By assigning class weights inversely proportional to number of samples, we could get 58.97% accuracy with four emotions. This configuration has been used in our online system.

5.2 User Engagement Across Input Methods

Supplementally, 5 more negotiations with college-age participants from University of California, Los Angeles were conducted. However, the data associated with the emotional model was corrupted so these participants' data was withheld from further emotional model analysis.

Figure 4 compares average conversation length, and average agreement likelihood for all 10 participants across user input methods (text and voice). The preliminary results seem to indicate that on average there will be longer interactions from a user

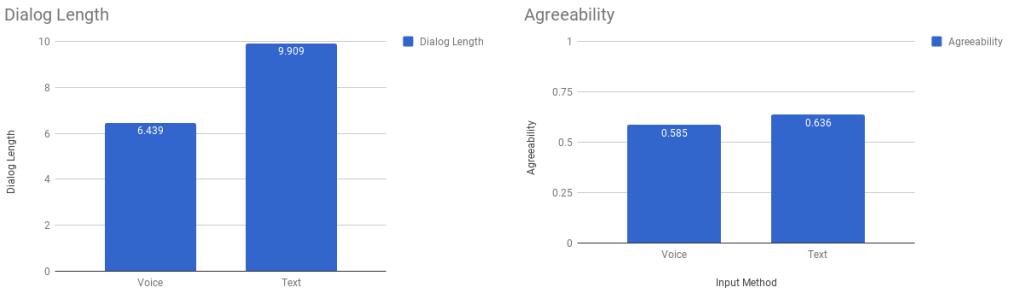

Figure 4: Average Dialog Length (*left*) on average the number of exchanges the user had across different input methods. "Agreeability" representing the likelihood that a user and agent successfully reach a conclusion

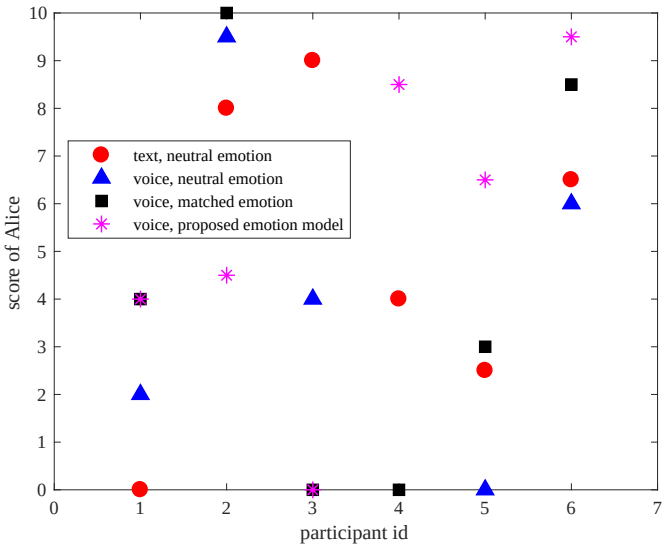

Figure 5: Scores of the computer agent with different participants across different conditions.

during text inputs, meanwhile, the likelihood that an agreement it reached is not largely affected.

This is indicative of speech being a more taxing communication medium within itself (e.g. cognitive effort, time, physical energy). The large difference in number of utterances, with relatively small difference in agreement likelihood, suggests that users are willing to conclude a negotiation earlier, rather than continue bargaining. The actual act of engaging in speech might be strenuous enough for individuals to be more willing to take worse deals in an effort to finish negotiating faster.

This points to values being optimized outside of the context of the negotiation at hand. A deeper investigation into how these costs can be parameterized is warranted.

5.3 Across Conditions

Figure 5 shows the scores obtained by the computer agent with different participants across four different test conditions (as described in Section 4). Although the experiment was done on a few willing participants, we can see that the virtual agent scores the highest most of the time with the proposed naive temporal model of emotion negotiation.

5.4 Survey Results

The findings we obtained from population level statistics on the survey responses were surprising showing that participants report feeling more engaged during typing input than during spoken input and more engaged during spoken input with a neutral response than with simulated affect from the affective model. This is not

only a reversal of our hypothesized engagement, but also seems uncorrelated with subject performance as observed in Figure 5.

We attribute these surprising results to the fact that there is a high amount latency during the conversation due to the increased computational time required to complete the end-to-end processing. This processing time increases proportionally to complexity of the interaction in the increasingly complex conditions. Of note is the robust inversely proportional relationship of latency to users' reported level of engagement. We believe this accounts for the observed survey results and outweigh any increase in engagement participants experience with increasingly complex conditions. Furthermore, it is our belief that the with a large sample-size that the magnitude of this discrepancy will be mitigated.

6 Conclusions and Future Work

Overall, our results raise a plethora of potential research question that can be probed and potentially resolved with the development of an end-to-end speech negotiation platform. Our preliminary experimental results point to advantages in affective models, but a major hurdle exists to making sure that dialog pace is maintained to retain user engagement.

One of the limiting factors thus far, has been the performance of the ASR. Currently, taking the ASR's "best" hypothesis, we search for keywords (the negotiated items, 'ok', 'deal', 'hi' etc.). If no keywords are present in the ASR output, we request the human agent to type the negotiated offer. Since both the ASR and TTS are executed remotely, lag in conversation (1-7s, depending upon utterance duration) seems to limit the emotional engagement users exhibit.

The most promising future directions are as follows:

- **ASR:** Improve ASR performance by evaluating more of the hypotheses and finding the best fit contextually given the dialog; This may also require adapting the language model to be more specific.
- **Emotion recognition:** Creating a richer feature space that accounts for both acoustic and lexical emotional features.
- **Affective model:** Constructing a better, more formal model from already available data. Collecting more data specifically for the purpose of learning emotional dialog models.

Despite its limitations, we believe we have laid a solid foundation for an open source platform and model that supports end-to-end negotiation using speech. It is our desire to present these works to the research community, and hope that the Affective Computing community will find it useful to further their own developments for emotional modeling and virtual agent research.

References

[1] "The kaldi asr." http://kaldi-asr.org/.

[2] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM International Conference on Multimedia*, MM '13, (New York, NY, USA), pp. 835–838, ACM, 2013.

[3] M. Lewis, D. Yarats, Y. N. Dauphin, D. Parikh, and D. Batra, "Deal or No Deal? End-to-End Learning for Negotiation Dialogues," *ArXiv e-prints*, 2017.

[4] "Watson: Text to speech." https://text-to-speech-demo.ng.bluemix.net.

[5] G. A. Adashinskaya and D. N. Chernov, "Acoustic correlates of individual features of functional and emotional states," *Human Physiology*, vol. 37, pp. 790–801, Dec 2011.

[6] K. R. Scherer, "Expression of emotion in voice and music," *Journal of Voice*, vol. 9, no. 3, pp. 235 – 248, 1995.

[7] D. DeVault, J. Mell, and J. Gratch, "Toward natural turn-taking in a virtual human negotiation agent," in *AAAI Spring Symposium on Turn-taking and Coordination in Human-Machine Interaction. AAAI Press, Stanford, CA*, 2015.

[8] J. Mell and J. Gratch, "IAGO: Interactive Arbitration Guide Online," in *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, (Singapore), pp. 1510–1512, International Foundation for Autonomous Agents and Multiagent Systems, May 2016.

[9] A. Rosenfeld, I. Zuckerman, E. Segal-Halevi, O. Drein, and S. Kraus, "Negochat-a: a chat-based negotiation agent with bounded rationality," *Autonomous Agents and Multi-Agent Systems*, vol. 30, pp. 60–81, Jan 2016.

[10] D. Druckman and M. Olekalns, "Emotions in negotiation," *Group Decision and Negotiation*, vol. 17, no. 1, pp. 1–11, 2008.

[11] D. N. McIntosh, "Facial feedback hypotheses: Evidence, implications, and directions," *Motivation and emotion*, vol. 20, no. 2, pp. 121–147, 1996.

[12] S. Kopelman, A. S. Rosette, and L. Thompson, "The three faces of eve: Strategic displays of positive, negative, and neutral emotions in negotiations," *Organizational Behavior and Human Decision Processes*, vol. 99, no. 1, pp. 81–101, 2006.

[13] R. A. Baron, "Environmentally induced positive affect: Its impact on self-efficacy, task performance, negotiation, and conflict," *Journal of Applied Social Psychology*, vol. 20, no. 5, pp. 368–384, 1990.

[14] H. A. Schroth, J. Bain-Chekal, and D. F. Caldwell, "Sticks and stones may break bones and words can hurt me: Words and phrases that trigger emotions in negotiations and their effects," *International Journal of Conflict Management*, vol. 16, no. 2, pp. 102–127, 2005.

[15] J. P. Daly, "The effects of anger on negotiations over mergers and acquisitions," *Negotiation Journal*, vol. 7, no. 1, pp. 31–39, 1991.

[16] M. Sinaceur and L. Z. Tiedens, "Get mad and get more than even: When and why anger expression is effective in negotiations," *Journal of Experimental Social Psychology*, vol. 42, no. 3, pp. 314–322, 2006.

[17] I. S. Fulmer, B. Barry, and D. A. Long, "Lying and smiling: Informational and emotional deception in negotiation," *Journal of Business Ethics*, vol. 88, no. 4, pp. 691–709, 2009.

[18] D. L. Shapiro, "Emotions in negotiation: Peril or promise," *Marq. L. Rev.*, vol. 87, p. 737, 2003.

[19] "Kaldi online decoding." http://kaldi-asr.org/doc/online_decoding.html.

[20] "Kaldi fisher models." http://kaldi-asr.org/downloads/build/2/sandbox/online/egs/fisher_english/s5/.

[21] "Ldc fisher english dataset." https://catalog.ldc.upenn.edu/LDC2004S13.

[22] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[23] "Gstreamer-kaldi." https://github.com/alumae/kaldi-gstreamer-server.

[24] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[25] 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*, pp. 1459–1462, ACM, 2010.

[26] Y. Freund and R. E. Schapire, "A desicion-theoretic generalization of on-line learning and an application to boosting," in *European conference on computational learning theory*, pp. 23–37, Springer, 1995.

[27] T. Hastie, S. Rosset, J. Zhu, and H. Zou, "Multi-class adaboost," *Statistics and its Interface*, vol. 2, no. 3, pp. 349–360, 2009.

[28] R. Plutchik, "The circumplex as a general model of the structure of emotions and personality.," in *Circumplex models of personality and emotions.*, pp. 17–45, Washington, DC, US: American Psychological Association, 1997.

[29] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," *Speech Communication*, vol. 53, no. 9, pp. 1162–1171, 2011.

[30] 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," *Language resources and evaluation*, vol. 42, no. 4, p. 335, 2008.