Emotion Elicitation in a Computerized Gambling Game

Vered Aharonson[1], Noam Amir[2]

(1) Department of Electrical Engineering
Tel Aviv Academic College of Engineering, Tel Aviv, Israel
vered@afeka.ac.il
(1) Department of Communication Disorders
Tel Aviv University, Tel Aviv, Israel
noama@post.tau.ac.il

Abstract

We have designed a novel computer controlled environment that elicits emotions in subjects while they are uttering short identical phrases. The paradigm is based on Damasio's experiment for eliciting apprehension and is implemented in a voice activated computer game. For six subjects we have obtained recordings of dozens of identical sentences, which are coupled to events in the game – gain or loss of points. Prosodic features of the recorded utterances were extracted and classified. The resultant classifier gave 78-85% recognition of presence/absence of apprehension.

1. Introduction

Numerous studies have been carried out in recent years, which examine the manifestations of emotion in the speech signal. The most common problem in this type of study is obtaining a corpus of emotional speech. One approach that has been adopted widely is to obtain acted emotions, from professional or non professional actors, as done by Banse and Scherer and others [1,2,3]. Using this type of data always raises the question whether the manner in which emotion is expressed by actors is the same as that in which it is expressed in spontaneous speech or in dialog. This point has been discussed in various papers [4,5,6], though there is no wide agreement between researchers on this subject.

In recent years there appears to be a trend toward studies based on more naturally occurring speech, from various settings: such as television talk shows [4], dialogs recorded during the performance of a given task [6], so called "reality shows", event recollection [5], speech elicited by inducing emotions in non-acting subjects, etc. It is no coincidence that together with the trend toward more natural settings, the definition of the emotions to be studied becomes more troublesome. Whereas actors can be requested to produce speech simulating such basic emotions such as anger, sadness, fear, etc., the emotions found in more naturally occurring speech can be more subtle and also more complex. Due to this very reason these studies propose more flexible means to classify emotions, such as the method proposed by Cowie [7], using dimensional scales rather than the basic emotion labels often found when using acted speech [6].

Several studies attempted to elicit emotional speech in more "naturalistic" environments, i.e. in response to real-life events. Such a real life environment was achieved in studies using driving simulators [8]. The main difficulty in such settings is the lack of control over the textual and emotional content. Therefore, the objective of the current study was to examine a setting in which we could control the range of emotions being elicited, and the text of the subject's responses, without resorting to acted emotions. This was achieved by using a voice-activated computer game, based on Damasio's card game experiment [9].

Damasio performed a card “gambling” experiment during which he measured the skin conductivity of subjects, observing their affective behavior and their physiological responses. His measurements demonstrated strong physiological correlates to the subjects' emotional state (presence/absence of apprehension) in response to the game’s events.

In the present study, we adapted Damasio's card game to a computer based game. Four doors were displayed on the screen, which were opened in response to voice commands from the subject. Sums of money (representing gain or loss) were revealed behind each door, so that the correlation between voice features and the events (gain/loss) in the game could be examined. As opposed to other "natural speech" studies our paradigm ensured that the subjects uttered identical texts repeatedly. Our initial expectation was that the events in the game would elicit a set of emotions such as apprehension, disappointment and contentment. In this initial study we were obliged to limit the classification to presence/absence of emotion, as discussed below.

2. Methods

2.1. Data Acquisition

In this study we developed a computer game based on Damasio’s experiments [Damasio, 1994], that has been

shown to elicit apprehension. The game is voice-activated, consisting of 100 events and yielding 200 short sentences for each subject, labeled according to the events in the game (gain/loss of points).

This paradigm yields a natural speech corpus in which subjects are expected to express a specific set of emotions, that can be determined by the experimenters. Here we expected to find responses that expressed apprehension, disappointment and contentment. The controlled environment enables the experiment to be repeated, if so desired, for verification or in order to obtain additional data. These features make this paradigm a unique and innovative framework for the study of emotion in speech.

Whereas Damasio performed the experiments using cards, in our version of the game, 4 packs of cards were replaced by four doors displayed on the screen. When a door was opened, it displayed a number of points, either positive or negative, thus increasing or decreasing the subject's score.

The amount of points gained or lost by opening a door depended on the doors' position. Two doors (the first and third from the left) were “risky.” Those two doors enabled a larger gain than the other two, but the following loss was much greater, so that opening them could quickly lead to total score loss. The other two doors did not give as large gains, but lost very little. Players needed to learn to choose the latter two doors to be successful in the game

To choose a door, the subjects had to say "open this door" in Hebrew (transcribed "p t ah x d eh l eh t z o") and to indicate the chosen door using a mouse. To close a door, the subjects had to say "close door" (transcribed "s g o r d eh l eh t"). A numeric label and a corresponding vertical bar indicated the subject's accumulated gain. The subjects were told that the game's objective was to accumulate the highest gain and that they could achieve this by figuring out which set of doors to choose and in which order. In this way, subjects were fully concentrated on the game and paid little attention to the lab environment .

For the preliminary study of this paradigm we ran the experiment on six subjects, all males. Each session took approximately 20 minutes to complete.

2.2. Data analysis

The data analysis involved several stages. Initially, the recordings for each speaker were broken into sub-files that contained a single utterance each. All of these were transcribed and labeled according to: a) their text ("open this door" or "close door"); and: b) the event in the game which preceded them – loss or gain of points.

Although the subjects were instructed to utilize one of the two sentences "close door" or "open this door", they sometimes abbreviated these sentences, using forms such as "open door" or "close". Only utterances containing the full forms were retained. This left us with 82, 100, 55, 92, 99, 64 "open this door" utterances for each subject respectively, and 81, 100, 100, 0, 86 and 61 "close door" utterances.

Initial pitch calculation was then carried out automatically over the entire database, at 10 ms. intervals, using Praat software. Pitch detection is a notoriously error-prone procedure, therefore the pitch data for the entire database was inspected visually and corrected where necessary by an experienced research assistant. Typical errors were false detections and octave jumps.

The raw pitch contour, calculated at 10 ms intervals, contains far too many numbers to be used as input to a classification algorithm. On one hand it contains a large degree of redundancy, and on the other hand it contains miniscule variations due to irregularities in the speech production process, which have no perceptual importance. Therefore it is necessary to include an information reduction stage, giving a small number of informative features that contain all the relevant information as concisely as possible. Due to the short duration of the uttered phrases, two approximations to the overall pitch contour were examined: a linear approximation and a second order (parabolic) approximation. An example for a single phrase is shown in Figure 1. An idea of the clustering properties of some of these parameters, Figures 2 and 3 show the linear approximations of the pitch contours for all of the "open this door" and "close door" phrases, for a single speaker (Ami). In Figure 1 ("open this door" phrases) we can see a set of phrases with positive slopes of similar values, and a second set of phrases with negative slopes of similar value. The same behavior is found is Figure 3 ("close door" phrases), though the slopes in the two figures.

From the above approximations a set of normalized parameters was extracted. These included: duration, pitch mean, initial and final pitch, coefficients of linear approximation, and coefficients of the second order approximation.

Since in this pilot study the data size was relatively small, and we extracted 10 features, we could not perform classification to the 3 emotion types we aimed for when designing the experiment, (apprehension, contentment and disappointment). We therefore attempted detection of a single emotion. This choice was also corroborated by the original study of Damassio,

where a single emotion – apprehension – was sought. The essence of a detection process is a hypothesis test – trying to distinguish between "emotion" and "neutral"/ "non-emotion" speech features. This procedure necessitated "tagging" of the events, or stimuli to those that hypothetically would produce the "apprehension" emotion and those that would not. Again, this was not a straight-forward task, since there was no way to determine an a-priori value for the exact values of loss in game where most subjects clearly demonstrated emotion, nor did Damasio's original experiment provide one. We therefore performed a heuristic search: We tried different value of loss in the game which could be considered as the border-line between emotion-eliciting events and non-emotion-eliciting ones. We examined the loss encountered by the subject in each trial (ranging from -2 to -125 points), and the cumulative sum of points (ranging between -1250 and 750). Our first hypothesis, based on Damasio's findings, was that "apprehension" – the emotion Damasio was looking for, would follow each loss and would be manifested in the subject's speech when giving the command to open a door, following the loss. We therefore had two hypothesis, H_0 = Losing less than a threshold (Win) and H_1 = Losing more than a threshold (Loss), and the stimuli were tagged as "0" (non-emotion-) or "1" (emotion eliciting).

Our algorithm checked for each choice of threshold the association rules in the data: Association rules have the form form: *If* A_i *then B,* where A_i are voice features, i=1-10 derived from each voice sample and B is the tagged preceding game event (which equals "0" or "1"). The association is based on calculating how frequently the conjunctions of values $\left(\bigcap_{i=n}^{m} A_i, B\right)$ $n, m = 1,10$) appear in the samples database.

The strength of the rules found was assessed by a) the existence of if-and-only-if rules (as opposed to if-then rules), b) the probability attached to the rules. and c) for if-then rules both the rules probability and the significance level (the probability of an error in the rule). We also checked the improvement of this probability over the prior distribution of the stimuli.

3. Results

Our initial hypothesis was that each loss of points encountered in the game would yield apprehension when next opening a door and would be manifested in the voice parameters of the respective command "open this door". In our first run of the algorithm, therefore, B was the amount of loss. However, checking for all threshold

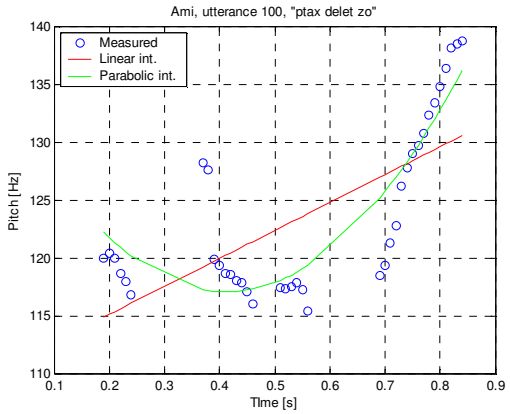

Figure 1: Pitch points and approximations for a single utterance

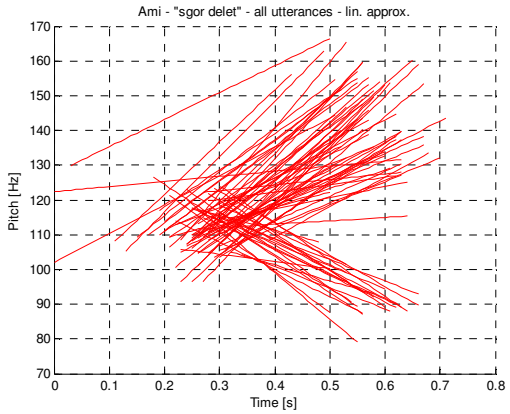

Figure 2: linear approximations for all "sgor" ("close") utterances of subject Ami

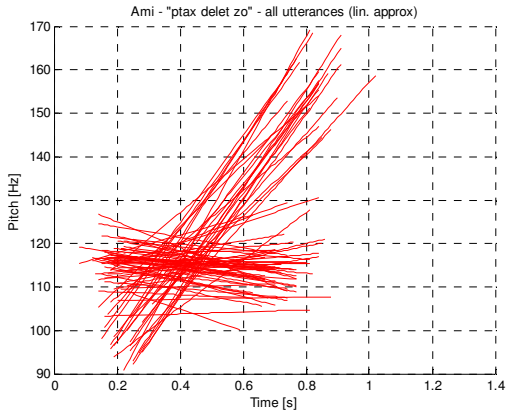

Figure 3: linear approximations for all "ptax" ("open") utterances of subject Ami

levels for tagging "loss", from -2 to -125 points and looking for association rules between B and the voice parameters did not yield strong rules. The rules were of the if-then form and their error probability was 0.1 or more. When looking for rules where B= cumulative sum of points, (tagging the dataset of the 6 subjects with "1" each time the cumulative sum dropped below the threshold), we obtained the strongest rule (if- and – only- if) with probability of **0.847** for "1" when the cumulative sum threshold was less than or equal to -30.0 points. The probability for "0" (a cumulative sum more than -30.000) was **0.787**. The improvement over the prior probability was 2.2 which reflects good detection.

As an example, a typical rule is of the form:

If 1) $-0.10 < f7 < 3.29$
OR 2) $0.46 < f1 < 0.56$ **AND**
$92.90 < f6 < 114.80$
OR 3) $-0.10 < f2 < 0.67$ **AND**
$97.14 < f5 < 106.09$ **AND**
$94.97 < f6 < 121.14$
Then the probability that the Cumulative Loss is less than or equal to -30.00 is 0.847. The *improvement factor* over the prior probability is 2.129 (0.369 / 0.173) where 0.369 is the primary probability that Cumulative Loss is more than -30.00 and (1-0.847=0.173) the complement probability of the rule.

The voice features that were associated with this rule were: utterance duration (f1), linear slope (f2), pitch mean (f5), final pitch (f6), and first parabolic coefficient (f7).

3. Discussion

In the narrowest sense, our preliminary results indicate that our paradigm could detect emotion in applications where phrases from a small vocabulary are being uttered. These can be found for instance in computer games, or certain voice based speech recognition tasks such as booking or information retrieval using the telephone. Ways in which these results could be extended to suit applications with unconstrained speech should be examined further, through future versions of the game proposed here.

We were careful here not to attach psychological interpretation or a specific definition for the emotion detected in this game scenario, and arbitrarily used Damasio's original definition of "apprehension". However, when extracting prosodic features from a larger database from the same experiment, we hope to be able to detect several emotional classes. For example, disappointment/contentment/neutral are intuitively indicated as response classes to positive/negative/none stimuli in the game. In this pilot study, we could detect, however, an association between voice features and the *negative* game stimuli (a considerable drop in the game score), which were tagged as "emotion-eliciting". A "considerable loss" is a very subjective concept, which we determined experimentally by a rule based data mining algorithm. The results indicated that on average a decrease below -30 points the cumulative score in the game yielded a high probability of "emotion" in the voicing of the subsequent phrase across all subjects in our database.

The voice features in which the emotion was manifested were varied. As seen in Figure 1, a parabolic approximation can better follow the pitch contour in these phrases than a linear approximation, and the case of such short phrases as used here it can probably capture the important pitch movements. Nevertheless, duration, mean and linear slope also proved to be important cues. We intend to examine further whether other pitch features can also be extracted from these coefficients, based on perceptual criteria.

We plan to collect more data using the same paradigm to assess these preliminary findings. Subjects from both genders will be included as well as subjects from different cultures. A larger dataset will enable both to perform classification to 3 emotion types, and to add both voice features and features from other modalities collected in the experiment: facial expressions and physiological data.

4. References

[1] Banse, R. and Scherer, K., Acoustic profiles in emotion expression, *Journal of Personality and Social Psychology*, 70(3), 614-636, 1996

[2] Yang, L. and Yunxin, Z., Recognizing emotions in speech using short tem and long term features, *Proceedings of ICSLP 98*, Sydney

[3] Dellaert, F., Polzin, T, Waibel, A., Recognizing emotion in speech, *Proceedings of ICSLP 96*

[4] Douglas-Cowie, E., Cowie, R., Schroder, M., A new emotion database: consideration, sources and scope, *ISCA workshop on speech and emotion*, Belfast 2000

[5] Amir, N., and Ron, S. Towards an automatic classification of emotion in speech, *Proceedings of ICSLP 98*, Sydney

[6] Kehrein, R., Prosodie und Emotionen, Tuebingen Neimeyer 2002

[7] Cowie, R.,Describing the emotional states expressed in speech, *ISCA workshop on speech and emotion*, Belfast 2002

[8] R. Fernandez, R. W. Picard, (2000), Modeling drivers' speech under stress, *ISCA workshop on speech and emotion*, Belfast 2000

[9] A. R. Damasio (1994) Descartes' Error: Emotion, Reason and the Human Brain, New York: Gosset/Putnam Press.